Introduction to Number Theory

Franz Lemmermeyer
franzl@csusm.edu

November 28, 2000

Preface

This is the manuscript of the course "Introduction to number theory" (MATH 372) held in the fall semester 2000 at the CSU San Marcos.

I'd like to thank all the students for bearing with patience all the things that weren't perfect: Matthew Arvanitis, Charles Beck, Brent Colvin, Jessica Elder, Sosciety Hedge, Danielle Jones, Melinda Legg, Paula Melendrez, Aubi Mellin, Elizabeth Robbins, and Kim Strom.

Contents

Introduction

Number Theory deals with properties of all kinds of numbers; elementary number theory is the part of the theory concerned mostly (but not exclusively) with the natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$ and the integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$.

One of the oldest notions in number theory is that of a prime number, which is to the number theorist what the atom is to the chemicist (and just as physicists have learned that atoms aren't as indivisible as their name suggests, ordinary primes can be factored by algebraic number theorists).

Primes are natural numbers $p > 1$ such as 2, 3, 5, 7, 11, ... whose only divisors are 1 and p. Thus $6 = 2 \cdot 3$ is not prime but $2^{31} - 1$ is. Two questions that immediately suggest themselves are:

- How many primes are there?
- How can we check whether a given integer is prime or not?

One answer to the first question was already given by Euclid:

Theorem. (Euclid) *There are infinitely many primes.*

There are many proofs of this fact,[1] and the simplest and best known can already be found in Euclid's books (we will give his proof below).

Remark. For many results in these lectures, we will give two (or even more) proofs; the idea behind that is not that more proofs would make the result more "probable": in fact, different proofs may give new insights into the problem, and the more proofs you know, the better your chances are that one of them generalizes to give a deeper result.

[1] See e.g. P. Ribenboim's *The Book of Prime Number Records* or his *Little Book of Big Primes*; the books are better than their titles suggest.

There is a more precise answer to the first question that was first conjectured by Legendre and Gauss and first proved by Hadamard and de la Vallée-Poussin: let $\pi(x)$ denote the number of primes below x, that is

$$\pi(x) = \#\{p : p \text{ prime}, \ p \le x\}.$$

Here is a small table for the function $\pi(x)$:

x	5	10	100	1000
$\pi(x)$	3	4	25	168
$\frac{x}{\log x}$	3.1	4.3	21.7	144.7

The last line displays the values of the function $\frac{x}{\log x}$, where $\log x$ denotes the logarithm to the basis e. The precise version of the conjecture that $\pi(x) \approx \frac{x}{\log x}$ goes by the name of the

Prime Number Theorem. *We have* $\pi(x) \sim \frac{x}{\log x}$.

Here $f(x) \sim g(x)$ is short for $\lim_{x\to\infty} f(x)/g(x) = 1$. The proof of the Prime Number Theorem is quite hard without using complex analysis and will not be proved here.

Warning. $f \sim g$ does not imply that $|f(x) - g(x)|$ is bounded; it only implies that this difference is small when compared to $f(x)$ or $g(x)$. As far as I know it has been proved that $|\pi(x) - \frac{x}{\log x}|$ can become arbitrarily large.

The question on the infinity of primes can also be refined: there are primes 5, 13, 17, ... of the form $4n+1$ and 3, 7, 11, ... of the form $4n+3$. Are there infinitely many primes in each series? Again the answer is yes, and in fact the primes are more or less "evenly" distributed among these two classes. More generally we have

Dirichlet's Theorem on Primes in Arithmetic Progression. *Given any pair a, b of natural numbers such that a and b do not have a common prime divisor, there are infinitely many primes of the form $an + b$.*

As for the prime number theorem, the proof of Dirichlet's theorem uses complex analysis in an essential way.

We may also ask if there are infinitely many primes of the form n^2+1 such as 5, 17, 37, ... ; it is conjectured that the answer is yes, but a proof seems out of reach. A similar question concerns Mersenne primes: these are primes of the form $M_p = 2^p - 1$. In the pre-computer age, the following values of p

were known to yield primes: $p = 2, 3, 5, 7, 13, 17, 19, 31, 61, 89, 107$ and 127 (note that $2^{127} - 1 = 170141183460469231731687303715884105727$!!) At the time of writing, the largest known prime is the Mersenne prime M_p for $p = 6,972,593$.

The second question also has several answers: of course we can check that N is prime by trying to divide N by the numbers $2, 3, 4, \ldots, N-1$; if N is not divisible by any of these numbers, then N is prime.

This algorithm can be improved immediately; consider the following program:

```
0.  input $N$; if $N = 1$ print '$N$ is a unit' and terminate;
1.  if $2 \mid N$ print '$p = 2$' and terminate;
2.  put $q := 3$;
3.  if $q \mid N$ print '$p = q$' and terminate;
4.  put $q := q + 2$; if $q > \sqrt{N}$ print '$p = N$' and terminate;
   otherwise go to step 3.
```

First, this is an algorithm because it terminates: you cannot loop back to step 3 forever because eventually q will become larger than $\sqrt{N}$, and the program terminates with step 4.

Next, this algorithm determines the smallest prime factor of a given number N (in practice, the condition $q > \sqrt{N}$ has to replaced by something like $q > \sqrt{N} + 0.1$ in order to avoid rounding errors e.g. when $N = p^2$). It trial divides by 2 and all odd integers $\le \sqrt{N}$; if N is not divisible by any of these integers, then N must be prime. In fact, in this case the smallest prime divisor of N must be $> \sqrt{N}$. If N has a proper prime divisor $> \sqrt{N}$, then $N/p < \sqrt{N}$ shows that it also has a proper prime divisor $< \sqrt{N}$, so the only possible prime factor of N is N itself.

In the worst case (namely when N is prime) this method requires about $\sqrt{N}/2$ divisions, which is much better than the $N-1$ divisions needed when dividing by $2, 3, 4, \ldots, N-1$ (for a further (slight) improvement, see the Exercises). The big question is whether there are even better algorithms; as a matter of fact there are, but these are based on completely different methods. We will discuss some of them once we have developed the necessary tools.

It is usually quite surprising to beginners in number theory that it is possible to show that a number is composite without knowing any of its factors. The methods that allows us to do this are called primality test: any number that fails such a test must be composite. As the size of known Mersenne primes shows, there are some quite sophisticated primality test;

one that applies to Mersenne numbers is the Lucas-Lehmer test that we will discuss as an application of our theory of conics.

Most of the known factorization algorithms have been developed in the last 30 years, mainly because of the appearance of computers. Another reason behind the search for good factorization algorithms is the discovery of cryptographic applications of number theory: the difficulty of factoring large integers can be used to make the transmission of information extremely safe. But real world applications weren't the main motivation behind number theorists' effort during the last two millenia: I hope that after having attended this introduction to number theory you will be able to understand why it is called the Queen of Mathematics.

Chapter 1

Unique Factorization

I have no intention to start from scratch, that is, to develop the basic properties of natural numbers and integers from, say, the Peano Axioms: instead, we regard the structures $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$ as known, and we will also make free use of elementary properties such as the principle of induction (often in the form that every nonempty set of natural numbers has a smallest element). After having discussed divisibility and the Euclidean algorithm, we are going to prove the unique factorization theorem.

1.1 Divisibility

The basic notion on which everything that follows will be based is the notion of divisibility.

Let $a, b \in \mathbb{Z}$ be integers; we say that b *divides* a or that b *is a divisor of* a (and write $b \mid a$) if there is an integer $c \in \mathbb{Z}$ such that $a = bc$.

More generally, we can talk about divisibility in any ring R: we say that $b \mid a$ in R is there is a $c \in R$ such that $a = bc$. Divisibility is boring in rings F that are fields, because then every nonzero element b divides every other element for the simple reason that $c = a/b \in F$. Divisors of 1 are called units. Obviously 1 and -1 are units because $1 \cdot 1 = 1$ and $(-1) \cdot (-1) = 1$.

We claim that in $\mathbb{Z}$, there aren't any others. In fact, assume that $r \in \mathbb{Z}$ is a unit. Then there is an $s \in \mathbb{Z}$ such that $rs = 1$. Assume that $r \neq \pm 1$; then $|r| \geq 2$ since $rs = 1$ implies $r \neq 0$. But $|r| \geq 2$ and $rs = 1$ imply $|s| \leq \frac{1}{2}$, and the only integer s with this property is 0: but then $rs = 0$, contradiction.

Some of the basic properties of divisibility are collected in the following

Proposition 1.1. *For any integers $a, b, c \in \mathbb{Z}$, we have*

1. *$1 \mid a$, $a \mid a$, $a \mid 0$;*
2. *if $a \mid b$ and $b \mid c$, then $a \mid c$;*
3. *if $a \mid b$ and $a \mid c$, then $a \mid (b \pm c)$;*
4. *if $a \mid b$, then $(-a) \mid b$, $a \mid (-b)$ and $(-a) \mid (-b)$;*
5. *if $a \mid b$ and $b \neq 0$ then $|a| \le |b|$.*

Proof. These are formal consequences of the definition:

1. $a = a \cdot 1$; $0 = 0 \cdot a$.
2. By assumption there exist integers $r, s \in \mathbb{Z}$ such that $b = ar$ and $c = bs$. Then $c = bs = (ar)s = a(rs)$, hence $a \mid b$ since $rs \in \mathbb{Z}$.
3. We have $b = ar$ and $c = bs$ for some integers $r, s \in \mathbb{Z}$; but then $c = bs = a(rs)$, hence $a \mid c$.
4. We have $b = ar$ and $c = as$ for integers r, s; then $b \pm c = a(r \pm s)$ implies that $a \mid (b \pm c)$.
5. We have $b = ar$ for some $a \in \mathbb{Z}$; since $b \neq 0$, we deduce that $r \neq 0$, hence $|r| \ge 1$ and therefore $|b| = |ar| \ge |a|$.

Note that the very same proofs are valid for general rings! □

Irreducibles and Primes

The divisors ± 1 and $\pm n$ of an integer are called trivial divisors; nontrivial divisors are also called proper divisors. Nonunits without proper divisors are called – irreducible. This may surprise you, because you may have expected that we call them primes. Sticking with the tradition in modern algebra, however, we reserve the term prime for something else. Luckily, for our ring $\mathbb{Z}$, primes and irreducibles are the same.

Here comes our definition of primes: a nonunit $p \in \mathbb{Z}$ is called prime if the following conclusion holds for all integers $a, b \in \mathbb{Z}$: if $p \mid ab$, then $p \mid a$ or $p \mid b$.

While it is easy to see whether an integer is irreducible or not (simply factor the thing), showing that a number p is prime requires some effort. Take the simplest example $p = 2$: to show that 2 is prime we have to prove that if $2 \mid ab$, then $2 \mid a$ or $2 \mid b$. This is not too hard: if the claim were false, then there would be odd numbers $a = 2m + 1$ and $b = 2n + 1$ such that ab is even; but $ab = 4mn + 2m + 2n + 1$ is clearly odd.

We now state

Proposition 1.2. *Primes are irreducible.*

Proof. Assume not. Then $p = rs$ with $r, s \in \mathbb{Z}$ nonunits. In particular, $p \mid rs$. If we can show that $p \nmid r$ and $p \nmid s$, then p cannot be prime and we have won. But we have $r \mid p$ (and so $|r| \le |p|$), hence $p \mid r$ would imply $|p| \le |r|$, and together this gives $|p| = |r|$, that is, $p = \pm r$ and thus $s = \pm 1$: this is a contradiction since r and s were assumed to be nonunits. □

The proof we have given for Proposition 1.2 is not "nice" in the following sense: we have used absolute values, and this is a notion that does not exist in arbitrary rings. Since we are working with integers anyway, this my not seem to be much of a nuisance. Nevertheless, we can rewrite the proof in such a way that it becomes valid in arbitrary rings:

Proof version # 2. Assume that $p \mid r$ and $r \mid p$; then $r = pt$ and $p = rs$, so $p = pst$, hence $st = 1$, and this shows that r and s are units. □

Note that our effort to make the proof valid for arbitrary rings has made the proof even simpler!

The other half of the truth is

Proposition 1.3. *Irreducible elements are prime.*

If primes and irreducibles always coincide, then why did we bother to introduce two notions at all? The answer is that primes and irreducibles do not always coincide, although they do in $\mathbb{Z}$. As a matter of fact, this is one of the instances where generalization for generalizations sake turns out to be helpful: although we will not make use of divisibility in arbitrary rings, the examples these rings provide help us to understand that certain truths are not as obvious as they may seem if one only looks at the rational integers.

Here's a simple example: take the ring $2\mathbb{Z} = \{\ldots, -4, -2, 0, 2, 4, \ldots\}$ of even integers;[1] define $b \mid a$ if there is a $c \in R$ such that $a = bc$. Then $4 = 2 \cdot 2$

[1] Actually, $2\mathbb{Z}$ is not a ring but a rng, namely a ring without identity.

shows that $2 \mid 4$; on the other hand, 2 does not have a factor at all (it does not divide itself) because 1 is not an element of $\mathbb{Z}/2\mathbb{Z}$. In particular, 2 is an element without a factorization into irreducibles. Although 2 is irreducible (it has no factor at all, let alone proper divisors), it is not prime: in fact, 2 divides $4 = 2 \cdot 2$ without dividing one of the factors.

Note that the rng $2\mathbb{Z}$ does not have unique factorization: we have $36 = 2 \cdot 18 = 6 \cdot 6$, but 2, 6 and 18 are all irreducible.

Granted, this is a pathological example of a ring where primes and irreducibles are not the same; rest assured that there are completely natural rings (provided by algebraic number theory) that share the same defect.

The proof of Proposition 1.3 will be given in the next subsection (it will turn out to be a special case of Lemma 1.7; it is deeper than its converse for the following very simple reason: Proposition 1.2 is valid for any ring (we used only the rules for divisibility in the proof), while Proposition 1.3 is only true for rings with unique prime factorization.

The proof of Proposition 1.3 is based on the Euclidean algorithm; this is our next topic.

1.2 The Euclidean Algorithm

Let $m, n \in \mathbb{Z}$ be integers; a common divisor is any integer d such that $d \mid m$ and $d \mid n$. We could define the greatest common divisor as the maximal common divisor, but we prefer the following definition based only on divisibility properties (this allows us to transfer the definition to arbitrary rings): d is called a greatest common divisor of m and n if d is a common divisor of m and n with the following property: if e is any common divisor of m and n, then $e \mid d$. If d is a greatest common divisor of m and n, then so is $-d$ (Exercise!). Nevertheless we agree to write $d = \gcd(m, n)$ even though d is not uniquely determined (alternatively, we may normalize the gcd by demanding that $d > 0$; but this problem is irrelevant for most problems).

Theorem 1.4. *In $\mathbb{Z}$, any two integers m and n possess a greatest common divisor d, which is unique up to sign. Moreover, d has a "Bezout representation",*[2] *that is, there exist integers $x, y \in \mathbb{Z}$ such that $d = mx + ny$.*

The statement that the greatest common divisor d of two integers m and n is a $\mathbb{Z}$-linear combination of them is often referred to as Bezout's Lemma.

[2] Etienne Bezout: 1730 (Nemours, France) – 1783 (Basses-Loges, France)

There are two basically different proofs of this result. One is quite simple and proves the existence of the gcd and the Bezout representation, the other is a bit involved but provides us with an algorithm for computing gcd's. Both methods use the following result on division of integers:

Lemma 1.5. *Given $a, b \in \mathbb{Z}$ with $b \neq 0$, there exist unique integers $q, r \in \mathbb{Z}$ with $0 \le r < b$ such that $a = bq + r$.*

For the proof we introduce the floor function: Given $x \in \mathbb{R}$, we let $\lfloor x \rfloor$ denote the largest integer $\le x$. Examples: $\lfloor 2.3 \rfloor = 2$, $\lfloor -1.71 \rfloor = -2$, $\lfloor 3 \rfloor = 3$. For positive reals, the floor function basically cuts off the integer part from the decimal expansion. In more technical terms, $\lfloor x \rfloor$ is defined as the unique integer such that $0 \le x - \lfloor x \rfloor < 1$.

Proof. Put $q = \lfloor \frac{a}{b} \rfloor$; by definition, $0 \le \frac{a}{b} - q < 1$. Multiplying through by b gives $0 \le a - bq < q$. Now put $r = a - bq$.

This proves the existence. Now assume that $q', r' \in \mathbb{Z}$ are integers such that $0 \le r' < b$ and $a = bq' + r'$. Then $0 = a - a = bq + r - (bq' + r') = b(q - q') + r - r'$; if $q \neq q'$, then $|q - q'| \ge 1$, hence

$$\begin{aligned} |b| &> |r - r'| && \text{since } r, r' \in [0, b) \\ &= |b(q - q')| && \text{by definition of } r \text{ and } r' \\ &\ge |b| && \text{since } |q - q'| \ge 1. \end{aligned}$$

But $|b| > |b|$ is a contradiction. □

The first proof of the existence of the gcd goes like this: Consider the set $D = m\mathbb{Z} + n\mathbb{Z} = \{am + bn : a, b \in \mathbb{Z}\}$. Clearly D is a nonempty set, and if $c \in D$ then we also have $-c \in D$. In particular, D contains positive integers. Let d be the smallest positive integer in D; we claim that $d = \gcd(m, n)$. There are two things to show:

Claim 1: d is a common divisor of m and n. By symmetry, it is sufficient to show that $d \mid m$. Write $m = rd + s$ with $0 \le s < d$; we find $d = am + bn$, hence $s = rd - m = r(am + bn) - m = (ra - 1)m + bn \in D$. The minimality of d implies $s = 0$, hence $d \mid m$.

Claim 2: if e is a common divisor of m and n, then $e \mid d$. Assume that $e \mid m$ and $e \mid n$. Since $d = am + bn$, we conclude that $e \mid d$.

The existence of the Bezout representation is a simple consequence of the fact that $d \in D$.

The second proof is based on the Euclidean algorithm. Given integers m and n, there are uniquely determined integers q_1 and r_1 such that $m = q_1n + r_1$ and $0 \le r_1 < n$. Repeating this process with n and r_1, we get $n = r_1q_2 + r_2$ with $0 \le r_2 < r_1$, etc. Since $n > r_1 > r_2 > \ldots \ge 0$, one of the r_i, say r_{n+1}, must eventually be 0:

$$\begin{aligned} m &= q_1n + r_1 && (1.1)\\ n &= q_2r_1 + r_2 && (1.2)\\ r_1 &= q_3r_2 + r_3 && (1.3)\\ &\ldots \\ r_{n-2} &= q_nr_{n-1} + r_n && (1.4)\\ r_{n-1} &= q_{n+1}r_n && (1.5)\end{aligned}$$

Example: $m = 56$, $n = 35$

$$\begin{aligned} 56 &= 1 \cdot 35 + 21\\ 35 &= 1 \cdot 21 + 14\\ 21 &= 1 \cdot 14 + 7\\ 14 &= 2 \cdot 7\end{aligned}$$

Note that the last r_i that does not vanish (namely $r_3 = 7$) is the gcd of m and n. This is no accident: we claim that $r_n = \gcd(m,n)$ in general. For a proof, we have to verify two things:

Claim 1: r_n is a common divisor of m and n. Equation (1.5) shows $r_n \mid r_{n-1}$; plugging this into (1.4) we find $r_n \mid r_{n-2}$, and going back we eventually find $r_n \mid r_1$ from (1.3), $r_n \mid n$ from (1.2) and finally $r_n \mid m$ from (1.1). In particular, r_n is a common divisor of m and n.

Claim 2: if e is a common divisor of m and n, then $e \mid r_n$. This is proved by reversing the argument above: (1.1) shows that $e \mid r_1$, (1.2) then gives $e \mid r_2$, and finally we find $e \mid r_n$ from (1.5) as claimed.

The Euclidean algorithm does more than just compute the gcd: take our example $m = 56$ and $n = 35$; writing the third line as $\gcd(m,n) = 7 = 21 - 1 \cdot 14$ and replacing the 14 by $14 = 35 - 1 \cdot 21$ coming from the second line we get $\gcd(m,n) = 21 - 1 \cdot (35 - 1 \cdot 21) = 2 \cdot 21 - 1 \cdot 35$. Now $21 = 56 - 1 \cdot 35$ gives $\gcd(m,n) = 2 \cdot (56 - 1 \cdot 35) - 1 \cdot 35 = 2 \cdot 56 - 3 \cdot 35$, and we have represented the gcd of 56 and 35 as a $\mathbb{Z}$-linear combination of m and n.

This works in complete generality: (1.4) says $r_n = r_{n-2} - q_n r_{n-1}$; the line before, which $r_{n-1} = r_{n-3} - q_{n-1} r_{n-2}$, allows us to express r_n as a $\mathbb{Z}$-linear combination of r_{n-2} and r_{n-3}, and going back we eventually find an expression of r_n as a $\mathbb{Z}$-linear combination of a and b.

Finding a Bezout representation by working backwards after having run through the Euclidean algorithm (note that, on a computer, this means that you have to save all the intermediate results) is complicated. There is a much better method called the extended Euclidean algorithm or

Berlekamp's algorithm. Check that the following algorithm computes the gcd of two integers as well as the Bezout representation.

Given positive integers a and b, this algorithm computes integers $d \in \mathbb{N}$ and $u, v \in \mathbb{Z}$ such that $d = \gcd(a, b) = ax + by$:

```
1.  Set $a_1 \leftarrow a$, $a_2 \leftarrow b$; $x_1 \leftarrow 1$, $x_2 \leftarrow 0$; $y_1 \leftarrow 0$, $y_2 \leftarrow 1$.
2.  Let $q \leftarrow \lfloor a_1/a_2 \rfloor$.
3.  Set $a_3 \leftarrow a_1 - qa_2$; $x_3 \leftarrow x_1 + qx_2$; $y_3 \leftarrow y_1 + qy_2$.
4.  Set $a_1 \leftarrow a_2$, $a_2 \leftarrow a_3$; $x_1 \leftarrow x_2$, $x_2 \leftarrow x_3$; $y_1 \leftarrow y_2$, $y_2 \leftarrow y_3$.
5.  If $a_2 > 0$ goto 2.
6.  If $ax_1 - by_1 > 0$ return $(d, x, y) = (a_1, x_1, -y_1)$,
        else return $(d, x, y) = (a_1, -x_1, y_1)$.
```

While this algorithm is well suited for computers, for calculations by hand I suggest applying the usual Euclidean algorithm and then working backwards.

We note a few useful results concerning divisibility and coprime integers.

Lemma 1.6. *If a and b are coprime integers such that $a \mid n$ and $b \mid n$, then $ab \mid n$.*

Proof. Write $n = ar$ and $n = bs$; since $\gcd(a, b) = 1$, there exist $x, y \in \mathbb{Z}$ such that $ax + by = 1$. Now $bsx = nx = axr = r - byr$ implies $r = byr + bsx$, that is, $b \mid r$. But then $r = bt$ gives $n = ar = abt$, so $ab \mid n$. □

Lemma 1.7. *If $m \mid ab$ and $\gcd(m, a) = 1$, then $m \mid b$.*

Proof. By Bezout's Lemma, we have $1 = mx + ay$ (this is a number theorist's Pavlovian reflex upon seeing $\gcd(m, a) = 1$); since $ab = mn$ for some n, we have $mny = aby = (1 - mx)b = b - mxb$. Thus $b = mny + mxb = m(ny + xb)$. □

Actually, this last result contains Proposition 1.3 as a special case: we want to show that irreducibles and primes are the same. In view of our definitions, what we have to prove is the following:

If p is irreducible and $p \mid ab$, then $p \mid a$ or $p \mid b$.

We can formulate this in an equivalent way as follows:

If p is irreducible, $p \mid ab$ and $p \nmid a$, then $p \mid b$.

But since p is irreducible, $p \nmid a$ implies $\gcd(p, a) = 1$, because every common divisor would divide p, and the only divisors of p are ± 1 and $\pm p$. This completes the proof of Proposition 1.3. By induction, we can now show

Corollary 1.8. *If a prime p divides a product $a_1 \cdots a_r$, then p divides one of the a_j.*

Proof. Proposition 1.3 applied to the product $a_1 \cdot (a_2 \cdots a_r)$ shows that $p \mid a_1$ or $p \mid (a_2 \cdots a_r)$. In the first case we are done, in the second we apply Proposition 1.3 to $a_2 \cdot (a_3 \cdots a_r)$; after finitely many steps we have reached the conclusion that p divides one of the a_j. □

Note that, from now on, we may use the terms "prime" and "irreducible" interchangeably.

1.3 Unique Factorization

Our first result is quite innocent:

Proposition 1.9. *Every integer $n > 1$ has a prime factorization.*

Proof. We proceed by induction. We call an integer n "nice" if it has a prime factorization. Clearly $n = 2$ is nice because 2 is prime. Now assume that all integers $< n$ are nice; since $n > 1$, it is either prime (and thus nice) or it isn't; but if n is not prime, then n is not irreducible (since primes and irreducibles are the same), so n has proper divisors, say $n = ab$ with $a, b \in \mathbb{N}$. Since $a, b < n$, these factors are nice, hence they have prime factorizations, say $a = p_1 \cdots p_r$ and $b = q_1 \cdots q_s$. But then $n = p_1 \cdots p_r q_1 \cdots q_s$ is a prime factorization of n. □

Since any factor in the prime factorization of n is a prime factor of n, this takes care of our second remark on our proof of the infinitude of primes.

We also can attach a prime factorization to negative integers: if $n < 0$ and $-n = p_1 \cdots p_r$ is a prime factorization of $-n > 0$, then $n = -p_1 \cdots p_r$ is a prime factorization of n.

Note that we have talked about "a" prime factorization; as a matter of fact, the prime factorization of an integer n is essentially unique, but this needs to be proved.

Again you may think that this is obvious; after all, if, say, 11 divides an integer n, then there cannot be a prime factorization of n that does not contain 11 as a factor. Or can there?

Consider the set $S = \{1, 5, 9, 13, \ldots\}$ of positive integers of the form $4n + 1$. Let us call a number $p > 1$ in S irreducible if its only divisors in S are 1 and p. Thus 5 and 9 are irreducible, while 25 is not. Here every integer has a factorization into irreducibles, but it is not unique: for example, $21 \cdot 33 = 9 \cdot 77$, and 9, 21, 33 and 77 are all irreducible according to our definition. The reason why unique factorization fails is the existence of irreducibles that aren't prime: clearly $9 \mid 21 \cdot 33$ since $21 \cdot 33 = 9 \cdot 77$, but 9 does not divide 21 or 33.

The theorem of unique factorization asserts that every integer has a prime factorization, and that it is unique up to the order of the factors.

Theorem 1.10. *Every integer $n \geq 2$ has a prime factorization $n = p_1 \cdots p_r$ (with possibly repeated factors). This factorization is essentially unique, that is: if $n = p_1 \cdots p_r$ and $n = q_1 \cdots q_s$ are prime factorizations of an integer n, then $r = s$, and we can reorder the q_j in such a way that $p_j = q_j$ for $1 \leq j \leq r$.*

A partial result in the direction of Theorem 1.10 can already be found in Euclid's elements; the first explicit statement and proof was given by Gauss[3] in 1801.

We already know that prime factorizations exist, so we only have to deal with uniqueness. This will be proved by induction on $\min\{r, s\}$, i.e. on the minimal number of prime factors of n. We may assume without loss of generality that $r \leq s$.

If $r = 0$, then $n = 1$, and $n = 1 = q_1 \cdots q_s$ implies $s = 0$.

[3] Carl-Friedrich Gauss: 1777 (Braunschweig, Germany) – 1855 (Göttingen, Germany)

Now assume that every integer that is a product of at most $r-1$ prime factors has a unique prime factorization, and consider $n = p_1 \cdots p_r = q_1 \cdots q_s$. Since p_1 is a prime that divides $n = q_1 \cdots q_s$, it must divide one of the factors, say $p_1 \mid q_1$ (after rearranging the q_i if necessary). But q_1 is prime, so its only positive divisors are 1 and q_1; since p_1 is a prime, it is a nonunit, and we conclude that $p_1 = q_1$. Canceling p_1 shows that $p_2 \cdots p_r = q_2 \cdots q_s$, and by induction assumption we have $r = s$, and $p_j = q_j$ after rearranging the q_i if necessary.

Remark. There is a simple reason for doing induction on the minimal number of prime factors and not simply on the number of prime factors of n: the fact that the number of prime factors of an integer is well defined is a consequence of the result we wanted to prove!

1.4 The Infinitude of Primes

Let us now see how Euclid[4] proved that there are infinitely many primes:

Theorem 1.11 (Euclid). *There are infinitely many primes.*

Proof of Thm. 1.11. We prove this claim by deriving a contradiction from the assumption that it is false. Assume therefore that there are only finitely many primes $2 = p_1, \ldots, p_n$, and consider the integer $N = p_1 \cdots p_n + 1 \geq 3$. Since $N > p_j$ for $1 \leq j \leq n$, we deduce that N is not a prime. Since $N > 1$, it must have a prime factor (here we use Proposition 1.9), say p_k. But then p_k divides N as well as $N-1$; this implies that p_k divides the difference $1 = N - (N-1)$: contradiction. □

Remark. Euclid's formulation was different from ours: Book IX, Proposition 20 reads

> Prime numbers are more than any assigned multitude of prime numbers.

In other words: Given any finite list of primes, there exists a prime that is not on the list. Note that Euclid carefully avoids the notion of infinite sets; in fact, problems with infinities (recall Zenon's paradox, for example) have led

[4] Euclid of Alexandria, ca. 325 – 265 BC, if he lived at all. "Author" of the oldest textbook in mathematics, the *Elements*.

the Greeks to allow only finite quantities in mathematics. For example, lines in geometry were not infinite but 'could be prolonged as much as desired'. Infinite sets were given their proper place in mathematics by Cantor[5] at the end of the last century.

Euclid's idea can be used to give a proof of the following result:

Proposition 1.12. *There are infinitely many primes of the form* $4n-1$.

Proof. Assume that there are only finitely many such primes $p_1 = 3$, $p_2 = 7$, $\ldots$, p_n, and form the number $N = 4p_1 \cdots p_n - 1$. As in Euclid's proof, this number cannot be prime (since $N \neq p_j$, and by assumption the p_j exhaust the primes of the form $4n-1$). Thus it must have some prime divisors. If all of them are of the form $4n+1$, then N itself would have that form since $(4n+1)(4m+1) = 4(4nm+n+m)+1$. But N does not have that form, so at least one of the prime divisors of N, say p, must have the form $4n-1$. Since $p \mid 4p_1 \cdots p_n - 1$, this prime must be different from the p_j: this contradicts our assumption that there are only finitely many primes of the form $4n-1$. □

The same idea does not work for primes of the form $4n+1$, because numbers of the form $4n+1$ can be made up of primes of the form $4n-1$, as the example $21 = 3 \cdot 7$ shows. We will have to wait until Chapter 3 to see a proof of the infinitude of primes of the form $4n+1$.

There is another famous proof of the infinitude of primes due to Euler;[6] it runs as follows: assume that there are only finitely many primes $p_1, \ldots$, p_n. Consider the (finite) product

$$P = \prod_{j=1}^{n} \frac{1}{1-\frac{1}{p_j}}.$$

We can expand each of the factors into a series by plugging in $x = 1/p_j$ in the geometric series

$$\frac{1}{1-x} \; = \; 1 + x + x^2 + x^3 + \ldots\,;$$

[5] Georg Cantor, 1845 (St. Petersburg) – 1918 (Halle). We owe him, among others, the notion of countability: the rationals are a countable set, the reals aren't,

[6] Leonhard Euler: 1707 (Basel, Switzerland) – 1783 (St. Petersburg, Russia).

note that this series converges absolutely for any x with $|x| < 1$, in particular for $x = 1/p_j$. This allows us to manipulate the series by rearranging terms; we have

$$\begin{aligned} P &= \Big(\frac{1}{1-\frac{1}{p_1}}\Big)\cdots\Big(\frac{1}{1-\frac{1}{p_n}}\Big) \\ &= \Big(1+\frac{1}{p_1}+\frac{1}{p_1^2}+\ldots\Big)\cdots\Big(1+\frac{1}{p_n}+\frac{1}{p_n^2}+\ldots\Big). \end{aligned}$$

By multiplying out we find that the right hand side is

$$\begin{aligned} 1 &+ \frac{1}{p_1}+\ldots+\frac{1}{p_n}+\frac{1}{p_1^2}+\frac{1}{p_1p_2}+\ldots+\frac{1}{p_n^2} \\ &+\frac{1}{p_1^3}+\frac{1}{p_1^2p_2}+\ldots+\frac{1}{p_1p_2p_3}+\ldots+\frac{1}{p_n^3}+\ldots \end{aligned}$$

Clearly the sum is over all fractions $\frac{1}{n}$ where n runs through the integers of the form $n = p_1^{a_1}\cdots p_n^{a_n}$ exactly once. Using the fact that every integer has a unique representation as a product of primes, we find that every summand $\frac{1}{n}$ occurs in this last sum exactly once, in other words: we have

$$P = 1+\frac{1}{2}+\frac{1}{3}+\ldots = \sum_{n=1}^{\infty}\frac{1}{n}.$$

But this is the harmonic series, which is well known to be divergent; this contradiction completes Euler's proof of the infinitude of primes.

Euler's proof is clearly more complicated than Euclid's: it uses manipulation of series as well as unique factorization. Why would anyone find such a complicated proof attractive? One of the reasons is that Euler's proof can be generalized to give Dirichlet's theorem on primes in arithmetic progressions, whereas a generalization of Euclid's proof yields only special cases such as the infinitude of primes of the form $an+1$.

The "Euler factorization" in the proof above can be saved at a little cost: we know that the series $\zeta(s) = \sum n^{-s}$ converges for any real $s > 1$, and then the method above proves the following curious result:

$$\prod_p \frac{1}{1-\frac{1}{p^s}} \;=\; \zeta(s).$$

Riemann noticed that $\zeta(s)$ converges for complex numbers $s = x + yi \in \mathbb{C}$ with real part $\operatorname{Re} s = x > 1$, and showed that it is possible to extend

$\zeta(s)$ to an analytic function on $\mathbb{C} \setminus \{1\}$. His conjecture that every zero of $\zeta(s)$ in the vertical strip $0 \leq \text{Re } s \leq 1$ lies on the line $\text{Re } s = \frac{1}{2}$ is one of the outstanding conjectures in modern number theory. Note that Hadamard[7] and de la Vallée-Poussin[8] have shown that the prime number theorem follows from the fact that there are no zeros on the line $\text{Re } s = 1$.

1.5 Applications

In this section, we will apply our results to concrete problems in number theory.

Mersenne Numbers

Let $M_n = 2^n - 1$ denote the Mersenne numbers. They have a lot of nice properties:

Proposition 1.13. *If $d \mid n$, then $M_d \mid M_m$.*

Proof. Write $n = dr$; then the identity

$$X^{dr} - 1 = (X^d - 1)(X^{r(d-1)} + X^{r(d-2)} + \ldots + X^r + 1)$$

in the polynomial ring $\mathbb{Z}[X]$ is verified by multiplying out. Plugging in $X = 2$, we find that $M_d = (2^d - 1) \mid 2^{dr} - 1 = M_n$. □

Corollary 1.14. *If M_n is prime, then so is n.*

Proof. We prove that if n is composite, then so is M_n. Assume therefore that $n = ab$ with $1 < a, b, < n$. Then $M_a \mid M_n$ by Proposition 1.13, $M_a > 1$ since $a > 1$, and $M_a < M_n$ since $a < n$. □

This explains why p is prime for the Mersenne primes M_p listed in the Introduction. Note that $M_{11} = 2^{11} - 1 = 2047 = 23 \cdot 89$ is composite.

[7] Jacques Salomon Hadamard, 1865 (Versaille) – 1963 (Paris).

[8] Charles Jean Gustave Nicolas de la Vallée-Poussin, 1866 (Louvain, Belgium) – 1962 (Louvain, Belgium). It was once rumored that the first mathematicians to prove the prime number conjectured would be granted immortality. This legend lost some credibility when de la Vallée-Poussin died at the age of 95, and was finally laid to rest when Hadamard died at 97.

Proposition 1.15. *We have* $\gcd(M_m, M_n) = M_{\gcd(m,n)}$.

Proof. Put $d = \gcd(m, n)$; then $m = dr$ and $n = ds$. Proposition 1.13 gives $M_d \mid M_m$ and $M_d \mid M_n$, thus M_d is a common divisor of M_m and M_n. It is not so clear, however, that M_d is the *greatest* common divisor of M_m and M_n. One idea is to show that M_n/M_d and M_m/M_d are coprime. This boils down to showing that $2^{r(d-1)}+2^{r(d-2)}+\ldots+2^r+1$ and $2^{s(d-1)}+2^{s(d-2)}+\ldots+2^s+1$ are coprime. Playing around with these numbers (subtracting multiples of one from the other etc.) does not seem to lead anywhere. Let's face it: we're stuck.

Let us therefore try another method, namely applying the Euclidean Algorithm to M_m and M_n. Here's the basic trick:

> If $m = qn + r$ with $q \in \mathbb{N}$ and $0 \le r < n$, then $M_m = QM_n + M_r$ with $0 \le M_r < M_n$ and $Q = 2^r \frac{M_{qn}}{M_n} \in \mathbb{N}$.

That means: each line in the Euclidean algorithm applied to the pair (m, n) corresponds to a line in the Euclidean algorithm applied to (M_m, M_n), with m, n and the r_i replaced by M_m, M_n and M_{r_i}, and with the q_i replaced by suitable integers Q_i. Thus if the Euclidean algorithm applied to (m, n) yields $\gcd(m, n) = d$, it will give $\gcd(M_m, M_n) = M_d$ when applied to (M_m, M_n). □

GCD's via Unique Factorization

Now we want to see the Unique Factorization Theorem in action. First we will introduce a useful formalism for writing down prime factorizations. We numerate the primes as $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, ... , and write the prime factorization of an integer a in the form

$$a = \prod_{j=1}^{\infty} p_i^{a_i},$$

where the a_i are nonnegative integers, and where – of course – only finitely many exponents a_i are nonzero. This allows us to control the prime factorizations of products: if we have

$$b = \prod_{j=1}^{\infty} p_i^{b_i},$$

then clearly

$$ab = \prod_{j=1}^{\infty} p_i^{a_i+b_i}.$$

Lemma 1.16. *For integers $a, b \in \mathbb{N}$ let*

$$a = \prod_{j=1}^{\infty} p_i^{a_i}, \quad b = \prod_{j=1}^{\infty} p_i^{b_i}$$

be their prime factorizations, where of course only finitely many a_i and b_i are $\neq 0$. Then $b \mid a$ if and only if $a_i \geq b_i$ for all i.

Proof. Assume that $a_i \geq b_i$ for all i. Then

$$c = \prod_{j=1}^{\infty} p_i^{a_i-b_i}$$

is an integer, and we clearly have $a = bc$. The converse is just as clear. □

The fact that every integer has a unique prime factorization allows us to give a formula for the gcd of integers $m, n \in \mathbb{N}$: write

$$n = \prod_{j=1}^{\infty} p_i^{a_i}, \quad m = \prod_{j=1}^{\infty} p_i^{b_i}.$$

We claim that $d = \gcd(m, n)$, where

$$d = \prod_{j=1}^{\infty} p_i^{d_i}, \quad d_i = \min(a_i, b_i).$$

By now we know that there are two things to prove:
Claim 1: d is a common divisor of m and n. This is clear since, by Lemma 1.16, $d_i \leq a_i$ for each i.
Claim 2: if e is a common divisor of m and n, then $e \mid d$. To see this, write $e = \prod_{j=1}^{\infty} p_i^{e_i}$; by Lemma 1.16, $e \mid a$ implies that $e_i \leq a_i$, while $e \mid b$ implies that $e_i \leq b_i$. Thus $e_i \leq \min(a_i, b_i) = d_i$.

Note that this gives a method for computing the gcd of two integers; unfortunately, this method is not very useful for large integers because finding their prime factorization is extremely difficult (if we proceed by trial division, then we have to perform up to $\sqrt{N}/2$ divisions, while the Euclidean algorithm takes only about $c \cdot \log N$ divisions for some constant c. These crude estimates can be made much more precise; this activity is called "studying the complexity of an algorithm". The complexity also measures "how fast" algorithms run independent of any hardware.

1.6 Diophantine Equations

Our second application of the unique factorization theorem concerns Pythagorean[9] triples: these are integers $x, y, z \in \mathbb{N}$ such that $x^2 + y^2 = z^2$. The most famous of these triples is of course $(3, 4, 5)$. It is quite easy to give formulas for producing such triples: for example, take $x = 2mn$, $y = m^2 - n^2$ and $z = m^2 + n^2$. It is less straightforward to verify that there are no other solutions.

Assume that (x, y, z) is a Pythagorean triple. If d divides two of these, it divides the third, and then $(x/d, y/d, z/d)$ is another Pythagorean triple. We may therefore assume that x, y and z are pairwise coprime; such triples are called primitive. In particular, exactly one of them is even.

Claim 1. The even integer must be one of x or y. In fact, if z is even, then x and y are odd. Writing $x = 2X + 1$, $y = 2Y + 1$ and $z = 2Z$, we find $4X^2 + 4X + 4Y^2 + 4Y + 2 = 4Z^2$: but the left hand side is not divisible by 4: contradiction.

Exchanging x and y if necessary we may assume that x is even. Now we transfer the additive problem $x^2 + y^2 = z^2$ into a multiplicative one (if we are to use unique factorization, we need products, not sums) by writing $x^2 = z^2 - y^2 = (z - y)(z + y)$.

Claim 2. $\gcd(z - y, z + y) = 2$. In fact, put $d = \gcd(z - y, z + y)$. Then d divides $z - y$ and $z + y$, hence their sum $2z$ and their difference $2y$. Now $\gcd(2y, 2z) = 2\gcd(y, z) = 2$, so $d \mid 2$; on the other hand, $2 \mid d$ since $z - y$ and $z + y$ are even since z and y are odd. Thus $d = 2$ as claimed.

This is the point where Unique Factorization comes in:

Proposition 1.17. *Let $a, b \in \mathbb{N}$ be coprime integers such that ab is a square. Then a and b are squares.*

Proof. Write down the prime factorizations of a and b as

$$a = p_1^{a_1} \cdots p_r^{a_r}, \quad b = q_1^{b_1} \cdots q_s^{b_s}.$$

[9] Pythagoras of Samos (ca. 569 – 475 BC.).

Now a and b are coprime, so the set of p_i and the set of q_j are disjoint, and we conclude that the prime factorization of ab is given by

$$ab = p_1^{a_1} \cdots p_r^{a_r} q_1^{b_1} \cdots q_s^{b_s}.$$

Since ab is a square, all the exponents in the prime factorization of ab must be even. This implies that the a_i and the b_j are even, therefore a and b are squares. □

Corollary 1.18. *Let $a, b \in \mathbb{N}$ be integers with* $\gcd(a,b) = d$ *such that ab is a square. Then a/d and b/d are squares.*

Proof. Apply the proposition to the pair a/d and b/d. □

Applying the corollary to the case at hand (and observing that $z - y \in \mathbb{N}$, since $z + y > 0$ and $(z-y)(z+y) = x^2 > 0$) we find that there exist $m, n \in \mathbb{N}$ such that $z - y = 2n^2$ and $z + y = 2m^2$. Adding and subtracting these equations gives $z = m^2 + n^2$ and $y = m^2 - n^2$, and from $x^2 = (z-y)(z+y) = m^2n^2$ and $x \in \mathbb{N}$ we deduce that $x = 2mn$.

Note that we must have $\gcd(m,n) = 1$: in fact, any common divisor of m and n would divide x, y and z contradicting our assumption that our triple be primitive. We have shown:

Theorem 1.19. *If (x,y,z) is a primitive Pythagorean triple with x even, then there exist coprime integers $m, n \in \mathbb{N}$ such that $x = 2mn$, $y = m^2 - n^2$ and $z = m^2 + n^2$.*

Note that if y is even, then the general solution is given by $x = m^2 - n^2$, $y = 2mn$ and $z = m^2 + n^2$. Moreover, if we drop the condition that the triples be primitive then the theorem continues to hold if we also drop the condition that the integers m, n be relatively prime.

Fermat's Last Theorem for $n = 4$

The solution of $x^2 + y^2 = z^2$ is the godfather of the proof that the diophantine equation

$$X^4 + Y^4 = Z^4 \tag{1.6}$$

has only trivial solutions, namely those with $X = 0$ or $Y = 0$. As a matter of fact, it is a lot easier to prove more, namely that

$$X^4 + Y^4 = Z^2 \tag{1.7}$$

has only trivial solutions (this *is* more: if $X^4 + Y^4$ cannot be a square, it cannot be a fourth power). The proof is quite involved and uses a technique that Fermat[10] called infinite descent. It is related to the following proof of the irrationality of $\sqrt{2}$:

Proposition 1.20. *The number* $\sqrt{2}$ *is irrational.*

Proof. Assume not; then there are integers $a, b \in \mathbb{N}$ such that $\sqrt{2} = \frac{a}{b}$. Squaring and clearing denominators gives $2b^2 = a^2$. Thus $2 \mid a^2$, and by Proposition 1.3 $2 \mid a$ (since 2 is prime). This means that $a = 2c$, and canceling 2 gives $b^2 = 2c^2$. By the same argument, $b = 2d$ for some $d \in \mathbb{N}$, and we get $2d^2 = c^2$, that is, $\sqrt{2} = \frac{c}{d}$.

We have shown: if $\sqrt{2} = \frac{a}{b}$ for $a, b \in \mathbb{N}$, then we also have $\sqrt{2} = \frac{c}{d}$ with $c < a$ and $d < b$ (these inequalities follow from $a = 2c$ and $b = 2d$). But this cannot be, because nothing prevents us from repeating this argument and getting arbitrarily long series of integers $a > c > e > \ldots > 0$ and $b > d > f > \ldots > 0$ such that $\sqrt{2} = \frac{a}{b} = \frac{c}{d} = \frac{e}{f} = \ldots$: but natural numbers cannot become smaller and smaller, and this contradiction proves the theorem. □

We have rewritten the classical[11] proof of the irrationality of $\sqrt{2}$ in such a way that the idea of infinite descent becomes visible: if we want to prove that a certain diophantine equation is impossible in $\mathbb{N}$, it is sufficient to show that for every solution in natural numbers there is another solution that is "smaller", which eventually leads to a contradiction because there is no natural number smaller than 1.

Fermat used this idea to give a proof of

Theorem 1.21. *The Fermat equation* (1.7) *for the exponent* 4 *does not have any solution with* $X, Y, Z \in \mathbb{N}$.

[10] Pièrre de Fermat ca. 1607 (Beaumont-de-Lomagne, France) – 1665 (Castres, France).

[11] Yep: it was known to the Greeks and went like this: write $\frac{a}{b}$ in lowest terms; then proceed as above and show that $a = 2c$ and $b = 2d$: this contradicts our choice of a and b.

Proof. The following proof is due to Euler; there is no doubt, however, that Fermat must have possessed something similar, since he gave a detailed description of his method of infinite descent in his letters.

Assume that $X, Y, Z \in \mathbb{N}$ satisfy (1.7); we may (and will) assume that these integers are pairwise coprime (otherwise we can cancel common divisors). Now we vaguely follow our solution of the Pythagorean equation: Z must be odd (if Z were even, then X and Y would have to be odd, and we get a contradiction as in the proof of Theorem 1.19).

Thus we may assume that X is odd and Y is even, and write this equation in the form $Y^4 = (Z - X^2)(Z + X^2)$; since any common divisor d of $Z - X^2$ and $Z + X^2$ divides their sum and their difference, we easily get that $d = 2$. Thus $R = \frac{1}{2}(Z - X^2)$ and $S = \frac{1}{2}(Z + X^2)$ are coprime, and $RS = \frac{1}{4}y^4$. Since R and S are not both even, either R is odd (and then R and $4S$ are coprime), or S is odd (and then $4R$ and S are coprime). In the first case, $R \cdot 4S = y^4$ is a fourth power, hence $2R = Z - X^2 = 2a^4$ and $4S = 2(Z + X^2) = (2b)^4$, that is $Z + X^2 = 8b^4$ for integers $a, b \in \mathbb{N}$; in the second case, $4R \cdot S = y^4$, and then $Z - X^2 = 8a^4$ and $Z + X^2 = 2b^4$. The first possibility leads to $4b^4 - a^4 = X^2$, which is impossible modulo 4 (the equation gives $-a^4 \equiv X^2 \bmod 4$ with a and X odd; but squares of odd numbers are $\equiv 1 \bmod 4$, so the congruence is $-1 \equiv 1 \bmod 4$: contradiction). Thus we are in the second case and get $b^4 - 4a^4 = X^2$.

Now we repeat the trick and write $4a^4 = (b^2 - X)(b^2 + X)$. Since X and b are odd, we find $\gcd(b^2 - X, b^2 + X) = 2$ and $b^2 - X = 2r^4$, $b^2 + X = 2s^4$. Adding the equations yields $b^2 = r^4 + s^4$, that is, we have found a new solution (b, r, s) to our equation $Z^2 = X^4 + Y^4$; since $0 < b < X < Z$, this means that to every solution (X, Y, Z) in natural numbers there exists another solution with a smaller Z. This is impossible. □

The idea we used to solve the diophantine problems are above was turning additive into multiplicative problems and then using unique factorization. In the cases we have considered, this was quite easy. It also works e.g. for the diophantine equation $x^4 + 2y^4 = z^2$, because we can write $2y^4 = z^2 - x^4 = (z - x^2)(z + x^2)$. But what about e.g. $2x^4 - y^4 = z^2$? One possible approach is the introduction of algebraic numbers: we can factor the left hand side over the ring $\mathbb{Z}[\sqrt{2}\,]$ as $(\sqrt{2}x^2 - y^2)(\sqrt{2}z^2 + y^2)$. In this specific example, however, Lagrange found a very clever trick to avoid irrationalities: he observed that $4x^4 = 2y^4 + 2z^2 = (y^2 + z)^2 + (y^2 - z)^2$, which implies $(y^2 - z)^2 = (2x^2 - y^2 - z)(2x^2 + y^2 + z)$. Nice, huh?

Notes

Let us quote a few propositions from Euclid's elements. It is not easy to give a faithful translation (whatever that means): Euclid's language was geometric even when dealing with arithmetic notions such as divisibility and primes. In particular, "measure" means "divide" in our language.

Euclid's Elements, Book VII:

> **Proposition 32** *Any number is either prime or is measured by some prime number.*

This result tells us that any integer is prime or is divisible by a prime: in other words: it is the prototype of the existence of a prime factorization.

Euclid's Elements, Book IX:

> **Proposition 14** *If a number is the least that is measured by prime numbers, then it is not measured by any other prime number except those originally measuring it.*

Translated into our language, this would read

> If a number is the smallest common multiple of a set of primes, then it is not divisible by any other primes.

Or, more formally: if $n = p_1 \cdots p_n$ with the p_i distinct primes, then n is not divisible by any prime different from the p_i. Note that in his proof, Euclid deals only with three factors due to his lack of a proper language.

This is a weak form of the unique factorization theorem; it does not deal with repeated prime factors, and it does not exclude that, say, $p^2q = pq^2$ for primes $p \neq q$. It may be argued, however, that Euclid knew about unique factorization but could not express (or prove) it using the inappropriate geometric language he was using.

The first explicit formulation and proof of the theorem on unique factorization of integers was given by Gauss (1801) in his famous book "Disquisitiones Arithmeticae" (Arithmetical Investigations), of which there are translations into German, French, English, Spanish and Russian.

Finally, here's a proof for the irrationality of $\sqrt{n}$ distilled from one Paula gave in her homework: we claim that $\sqrt{n}$ is irrational for any integer n that is not the square of an integer.

In fact, if n is not a square of an integer, then we can find an integer a such that $a^2 < n < (a+1)^2$. Assume that $\sqrt{n} = \frac{p}{q}$ with $q > 0$ minimal. Then $p^2 = nq^2$, hence $p(p-aq) = p^2 - apq = nq^2 - apq = q(nq-ap)$, so

$$\frac{p}{q} = \frac{nq-ap}{p-aq}.$$

But $a < \frac{p}{q} < a+1$ implies $0 < p - aq < q$: this contradicts the minimality of the denominator q.

Summary

You should be able to
• prove the infinitude of primes with Euclid's proof, and give the main ideas of Euler's proof;
• compute gcd's and Bezout elements using the Euclidean algorithm.

You also should know that
• the fact that irreducibles are prime, which is the basis for the proof of the Unique Factorization Theorem, follows from the existence of the Euclidean Algorithm;
• if $\gcd(a,b) = 1$, $a \mid n$ and $b \mid n$, then $ab \mid n$;
• if $m \mid ab$ and $\gcd(m,a) = 1$, then $m \mid b$; and
• if $\gcd(a,b) = 1$ for $a, b \in \mathbb{N}$ and if ab is a square, then so are a and b.

Chapter 2

Congruences

Congruences are a very clever notation invented by Gauss (and published in 1801 in his "Disquisitiones Arithmeticae", or "Arithmetical Investigations")[1] to denote the residue of a number a upon division by another number m. More precisely, we write $a \equiv b \bmod m$ if $m \mid (a-b)$.

Examples. We have $13 \equiv 3 \bmod 5$, $14 \equiv 0 \bmod 7$, and $16 \equiv 13 \bmod 3$.

Our rules for divisibility can now be transferred painlessly to congruences: first we observe that congruence between integers is an equivalence relation, that is: For integers $a, b, c, d \in \mathbb{Z}$ and $m \in \mathbb{Z} \setminus \{0\}$ we have

- $a \equiv a \bmod m$;
- $a \equiv b \bmod m \Longrightarrow b \equiv a \bmod m$;
- $a \equiv b \bmod m$ and $b \equiv c \bmod m \Longrightarrow a \equiv c \bmod m$.

The proofs are straightforward. In fact, $a \equiv a \bmod m$ means $m \mid (a-a)$, and since every integer $m \neq 0$ divides 0, this is obvious. Similarly, $a \equiv b \bmod m$ is equivalent to $m \mid (a-b)$; but this implies $m \mid (b-a)$, hence $b \equiv a \bmod m$.

We also have

- $a \equiv b \bmod m \Longrightarrow a \equiv b \bmod n$ for every $n \mid m$;
- $a \equiv b \bmod m$ and $c \equiv d \bmod m \Longrightarrow a+c \equiv b+d \bmod m$ and $ac \equiv bd \bmod m$;
- $a \equiv b \bmod m \Longrightarrow ac \equiv bc \bmod m$ for any $c \in \mathbb{Z}$.

[1] The publication of this masterpiece was delayed for various reasons, one of them being the fact that Gauss wasn't very good in Latin. For modern readers, this is quite fortunate because it makes Gauss's Latin works easy to read.

Note, however, that canceling factors in congruences is dangerous: we have $2 \equiv 8 \bmod 6$, but not $1 \equiv 4 \bmod 6$. Here's what we're allowed to do:

Proposition 2.1. *If $ac \equiv bc \bmod m$, then $a \equiv b \bmod \frac{m}{\gcd(m,c)}$.*

Proof. We have $m \mid (ac - bc) = c(a-b)$. Write $d = \gcd(m,c)$, $m = dm'$, $c = dc'$, and note that $\gcd(m',c') = 1$. From $dm' \mid dc'(a-b)$ we deduce immediately that $m' \mid c'(a-b)$; since $\gcd(m',c') = 1$, we even have $m' \mid (a-b)$ by Lemma 1.7, i.e. $a \equiv b \bmod \frac{m}{\gcd(m,c)}$. □

Example. What are the last two digits of 9^{99}? In order to answer this question, we observe that the last two digits (in base 10) of an integer N are the same as the residue of N modulo 100. Our task is therefore to compute $9^{99} \bmod 100$. Here's how I would do that: $9^{99} = (9^3)^{33} = 729^{33} \equiv 29^{33} = 29 \cdot 29^{32} \bmod 100$. The second factor 29^{32} can be computed by repeated squaring:

$$\begin{aligned}
29^2 &= 841 & & \equiv 41 \bmod 100,\\
29^4 &\equiv 41^2 &= 1681 & \equiv 81 \bmod 100,\\
29^8 &\equiv 81^2 &= 6561 & \equiv 61 \bmod 100,\\
29^{16} &\equiv 61^2 &= 3721 & \equiv 21 \bmod 100,\\
29^{32} &\equiv 21^2 &= 441 & \equiv 41 \bmod 100,
\end{aligned}$$

so $9^{99} \equiv 29 \cdot 29^{32} \equiv 29 \cdot 41 = 1189 \equiv 89 \bmod 100$, hence the last two digits of 9^{99} are 89.

In general, $b^e \bmod m$ is of course not computed by repeatedly multiplying b to itself (which would take $e-1$ multiplications and reductions modulo m) but by repeated squaring and multiplication. Here's an example how to do it by hand: if you have to compute, say, a^{77}, write $77 = 64 + 8 + 4 + 1$; then $a^{77} = a^{64}a^8a^4a^1$, and these powers of a are computed by repeated squaring.

For computers, the following algorithm is well suited: it computes the binary expansion of the exponent e and the power of $b^e \bmod m$ as we go along:

```
0.  Set $t \leftarrow 1$; $a \leftarrow b$; $d \leftarrow e$;
1.  if d=0 return '$t \bmod m$' and terminate.
2.  If $d$ is odd, set $t \leftarrow ta - m\lfloor \frac{ta}{m} \rfloor$;
3.  Set $a \leftarrow a^2 - m\lfloor \frac{a^2}{m} \rfloor$, $d \leftarrow \lfloor \frac{d}{2} \rfloor$ and goto 1.
```

This algorithm terminates because $d \ge 0$ is reduced by a factor of at least 2 in each round. We have to check that the result t is indeed the correct answer. First observe that $t \leftarrow ta - m\lfloor \frac{ta}{m} \rfloor$ is nothing but $t \leftarrow ta \bmod m$, that is: t is multiplied by a and reduced modulo m. Similarly, the first part of step 3 replaces a by $a^2 \bmod m$.

We will now look at the values of t, a and d after the algorithm has completed i cycles. To this end we write e in the binary system as $e = e_0 + e_1 2 + e_2 4 + \ldots + e_n 2^n$. At the start of the algorithm ($i = 0$) we have $t = 1 \equiv b^0 \bmod m$, $a = b = b^1$, and $d = e$. After the first cycle ($i = 1$), we have $t = b$ if e is odd and $t = 1$ otherwise, that is: we have $t = b^{e_0}$. Moreover, we have $a = b^2$ and $d = \lfloor \frac{e}{2} \rfloor = e_1 + e_2 2 + \ldots + e_n 2^{n-1}$.

We now claim that, after i cycles, we have

$$t \equiv b^{e_0 + e_1 2 + \ldots + e_{i-1} 2^{i-1}}, \quad a = b^{2^i}, \quad d = e_i + e_{i+1} 2 + \ldots e_n 2^{n-i}.$$

Assume that this holds for some i; then cycle $i+1$ replaces t by $ta^{e_i} \equiv tb^{e_i 2^i} \equiv b^{e_0 + e_1 2 + \ldots + e_i 2^i} \bmod m$, a by $a^2 = b^{2^{i+1}}$, and d by $\lfloor \frac{d}{2} \rfloor = e_{i+1} + e_{i+2} 2 + \ldots + e_n 2^{n-i-1}$.

Thus after n cycles, we will have $e = 0$, and $t \equiv b^e \bmod m$ as desired.

The one thing that makes congruences *really* useful is the fact that we can define a ring structure on the set of residue classes. Since this is fundamental, let us do this in detail.

Fix an integer $m > 1$. We have already noted that the congruence relation modulo m is an equivalence relation. The set of integers congruent to a given integer a is called the residue class of a modulo m and will be denoted by $[a]$; for $m = 3$, for example, we have

$$\begin{aligned} [0] &= \{\ldots, -6, -3, 0, 3, 6, \ldots\}, \\ [1] &= \{\ldots, -5, -2, 1, 4, 7, \ldots\}, \\ [2] &= \{\ldots, -4, -1, 2, 5, 8, \ldots\}, \\ [3] &= \{\ldots, -3, 0, 3, 6, 9, \ldots\} = [0], \end{aligned}$$

etc. Note that $[0] = [3] = [6] = \ldots$ (in fact, $[0] = [a]$ for any $a \in [0]$), and similarly $[1] = [4] = \ldots$. In general, we have $[a] = [a']$ if and only if $a \equiv a' \bmod m$, that is, if and only if $m \mid (a - a')$.

In general, there are exactly m different residue classes modulo m, namely $[0]$, $[1]$, ..., $[m-1]$. First, they are pairwise distinct, since $[a] = [b]$ for $0 \le a, b < m$ implies $b \in [a]$, hence $a \equiv b \bmod m$ or $m \mid (b - a)$: but since $|b - a| < m$, this can only happen if $a = b$. Next, there are no other residue classes: given any class $[a]$, we write $a = mq + r$ with $0 \le r < m$, and then $[a] = [r]$ is one of the classes listed above. The set $\{0, 1, 2, \ldots, m-1\}$ is often called a complete set of representatives modulo m for this reason. Sometimes we write $r + m\mathbb{Z}$ instead of $[r]$.

These residue classes $[0], [1], \ldots, [m-1]$ modulo m will form the elements of our ring $\mathbb{Z}/m\mathbb{Z}$. We have to define an addition and a multiplication and then verify the ring axioms.

• Addition $\oplus$: Given two classes $[a]$ and $[b]$, we put $[a] \oplus [b] = [a+b]$. We have to check that this is well defined: after all, we have e.g. $[a] = [a+m]$, so if our addition makes sense then $[a] \oplus [b]$ and $[a+m] \oplus [b]$ should be equal.

More precisely we have to do the following: assume that $[a] = [a']$ and $[b] = [b']$; then we have to show that $[a+b] = [a'+b']$. But this is easy: we have $a - a' \in m\mathbb{Z}$, say $a - a' = mA$, and similarly $b - b' = mB$. But then $(a+b) - (a'+b') = m(A+B) \in m\mathbb{Z}$, hence $[a+b] = [a'+b']$.

The neutral element is the residue class $[0] = m\mathbb{Z}$, and the inverse element of $[a]$ is $[-a]$, or, if you prefer, $[m-a]$. In fact, we have $[a] \oplus [0] = [a+0] = [a]$ and $[a] \oplus [-a] = [a + (-a)] = [0]$. The law of associativity and the commutativity are transferred easily from the corresponding properties of integers: since e.g. $(a+b)+c = a+(b+c)$, we have a fortiori $([a] \oplus [b]) \oplus [c] = [a] \oplus ([b] \oplus [c])$.

• Multiplication $\odot$: of course we put $[a] \odot [b] = [ab]$. The verification that this is well defined is left as an exercise. The neutral element is the class $[1]$.

• Distributive Law: Again, $([a] \oplus [b]) \odot [c] = [a] \odot [c] \oplus [b] \odot [c]$ follows from the corresponding properties of integers.

Theorem 2.2. *The residue classes* $[0]$, $[1]$, … , $[m-1]$ *modulo* m *form a ring* $\mathbb{Z}/m\mathbb{Z}$ *with respect to addition* $\oplus$ *and multiplication* $\odot$*.*

Computing with residue classes is easy: here are the addition and multiplication table for the ring $\mathbb{Z}/3\mathbb{Z}$:

+	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

·	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

Now that we have introduced the rings that we will study for some time to come, we simplify the notation by writing $+$ and $\cdot$ instead of $\oplus$ and $\odot$. Moreover, we will drop our references to classes and deal only with the integers representing them; in order to make clear that we are dealing with residue classes, we write $\equiv$ instead of $=$ and add a "mod m" at the end. What this means in practice is that we identify $\mathbb{Z}/m\mathbb{Z}$ with the set of integers $\{0, 1, \ldots, m-1\}$.

Computing residue classes often occurs in everyday life: the main example is the "clock arithmetic": you add the hours on the clock by taking residues modulo 12 (or modulo 24, if you replace x pm by x+12).

A less trivial example are the ISBN (international standard book numbers) codes: the book "The Queen of Mathematics" by J. Goldman has the ISBN 1-56881-006-7; The first digit encodes the country in which the publishing company resides: 0 is for the USA, 1 for the UK, and 3 for Germany. The next string of digits give information about the publishing company; for example, 0-387 is for Springer Verlag New York, 3-540 for Springer Verlag Heidelberg. The third set of strings distinguishes the different books published by each company; the book "Measure Theory" by J.L. 3Doob has two ISBNs: 0-387-94055-3 and 3-540-94055-3.

Thus we can explain every digit in an ISBN except the last one. This last digit carries no information: its purpose is to check whether any errors have been made in copying the ISBN.

Here's how it works: assume that the digits of an ISBN are $n_1 n_2 \ldots n_9$ (if the code has only eight digits, put $n_9 = 0$); compute the sum $N = n_1 + 2n_2 + \ldots + 9x_9 = \sum_{j=1}^{9} in_i$; the residue class N mod 11 is one of $\{0, 1, \ldots, 10\}$, and if we write X for the residue class 10 mod 11, this i s the last digit of the ISBN number.

Example. Goldman's book has the ISBN 1-56881-006-7; in fact, we find $1 \cdot 1 + 2 \cdot 5 + 3 \cdot 6 + 4 \cdot 8 + 5 \cdot 8 + 6 \cdot 1 + 7 \cdot 0 + 8 \cdot 0 + 9 \cdot 6 \equiv 7$ mod 11.

The ISBN code is capable of detecting single errors: assume that the correct ISBN number (without the last digit) is $n_1 n_2 \ldots n_9$, and that someone copying that number made a single error. Let $m_1 \ldots m_9$ be the incorrect number. Since there is only one error, we must have $n_i = m_i$ for all indices $j = 0, 1, \ldots 9$ except one, say j. The correct check digit is $n_{10} \equiv \sum in_i$ mod 11; if we form $N \equiv im_i$ mod 11, then $N - n_{10} \equiv \sum i(m_i - n_i) = j(m_j - n_j)$ mod 11. Since $m_j \neq n_j$ and $11 \nmid j$, this difference does not vanish, indicating that some error must have occurred.

Now that we have the residue class rings $\mathbb{Z}/m\mathbb{Z}$, we arrest the usual suspects: can we determine the unit group $(\mathbb{Z}/m\mathbb{Z})^\times$? Recall that a unit is an element that divides 1; in our case this means that a residue class

$a \bmod m$ is a unit if there is a residue class $b \bmod m$ with $ab \equiv 1 \bmod m$ (in other words: the units are the elements having an inverse modulo m). Let us do an example: put $m = 15$ and compute the units by brute force.

a	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
b	–	1	8	–	4	–	–	13	2	–	–	11	–	7	14

The units modulo 15 are exactly those residue classes $[a]_{15}$ for which a and 15 are coprime: $\gcd(a, 15) = 1$. This holds in general:

Theorem 2.3. *We have* $(\mathbb{Z}/m\mathbb{Z})^\times = \{a \bmod m : \gcd(a, m) = 1\}$.

Proof. It is now that the Bezout representation begins to show its full power. If $\gcd(a, m) = 1$, then there exist integers $x, y \in \mathbb{Z}$ such that $ax + my = 1$. Reducing this equation modulo m gives $ax \equiv 1 \bmod m$, in other words: the residue class $a \bmod m$ is a unit! Not only that: the extended Euclidean algorithm gives us a method to compute the inverse elements.

To prove the converse, assume that $a \bmod m$ is a unit. Then $ac \equiv 1 \bmod m$, so $ac = km + 1$ for some $k \in \mathbb{Z}$. But then $ac - km = 1$ shows that $\gcd(a, m) = 1$. □

If $m = p$ is a prime, the unit groups are particularly simple: we have $\gcd(a, p) = 1$ if and only if $p \nmid a$, hence $(\mathbb{Z}/p\mathbb{Z})^\times = \{1, 2, \ldots, p - 1\} = \mathbb{Z}/p\mathbb{Z} \setminus \{0\}$. But if every element $\neq 0$ of a ring has an inverse, then that ring is a field, and we have proved

Corollary 2.4. *If p is a prime, then the residue class ring $\mathbb{Z}/p\mathbb{Z}$ is a field.*

Important Remark. Assume that we are given an integer m and an element a coprime to m; how do we compute the inverse of $a \bmod m$? We start by applying the Euclidean algorithm to m and a, and then compute a Bezout representation $ax + bm = 1$. If we read that equation modulo m, then bm vanishes, and we get $ax \equiv 1 \bmod m$, in other words: x is the inverse of a modulo m.

2.1 Fermat's Little Theorem

Let n be an odd integer and consider the factorization of the Mersenne[2] numbers $2^{n-1} - 1$:

n	$2^{n-1} - 1$
3	3
5	$3 \cdot 5$
7	$3^2 \cdot 7$
9	$3 \cdot 5 \cdot 17$
11	$3 \cdot 11 \cdot 31$

n	$2^{n-1} - 1$
13	$3 \cdot 5 \cdot 7 \cdot 13$
15	$3 \cdot 43 \cdot 127$
17	$3 \cdot 5 \cdot 17 \cdot 257$
19	$3 \cdot 7 \cdot 19 \cdot 73$
21	$3 \cdot 5 \cdot 11 \cdot 31 \cdot 41$

The fact that 3 divides these integers is not very surprising: If n is odd, then $n - 1$ is even, and 2^{n-1} is a power of 4. But $4 \equiv 1 \bmod 3$, so $2^{n-1} \equiv 1 \bmod 3$, hence $2^{n-1} - 1 \equiv 0 \bmod 3$. Note, however, that $n \mid 2^{n-1} - 1$ in this table if and only if n is prime. Could this be true? Unfortunately, it isn't, but you have to extend the table considerably before you can observe that $341 \mid 2^{340} - 1$ although $341 = 11 \cdot 31$: in fact we have

$$\begin{aligned} 2^{170} - 1 &= 3 \cdot 11 \cdot 31 \cdot 43691 \cdot 131071 \cdot 9520972806333758431 \\ &\quad \cdot 26831423036065352611 \\ 2^{170} + 1 &= 5^2 \cdot 41 \cdot 137 \cdot 953 \cdot 1021 \cdot 4421 \cdot 26317 \cdot 550801 \cdot 23650061 \\ &\quad \cdot 7226904352843746841 \end{aligned}$$

Well, at least the other direction is true in general; it is a special case of

Theorem 2.5 (Fermat's Little Theorem). *If p is a prime and a an integer not divisible by p, then $a^{p-1} \equiv 1 \bmod p$.*

Proof. The following proof is due to Leibniz[3] and probably the oldest proof known for Fermat's Little Theorem. It proves the equivalent (!) statement $a^p \equiv a \bmod p$ for all $a \in \mathbb{Z}$ via induction on a. The claim is clearly trivial for $a = 1$; assume it has been proved for some a; then

$$(a+1)^p = a^p + \binom{p}{1} a^{p-1} + \ldots + \binom{p}{p-1} a + 1.$$

[2] Marin Mersenne, 1588 (Oize, France) – 1648 (Paris).

[3] Gottfried Wilhelm von Leibniz, 1646 (Leipzig) – 1716 (Hannover).

By the induction assumption, $a^p \equiv a \bmod p$. Next we know that $\binom{p}{k} = \frac{p!}{k!(p-k)!}$; since the numerator is divisible by p while the denominator is not divisible by p unless $k = 0$ or $k = p$, we conclude that $p \mid \binom{p}{k}$ for $0 < k < p$. Thus

$$(a+1)^p \equiv a^p + 1 \equiv a + 1 \bmod p,$$

and the induction step is established. □

Assume that we are given an integer m and an integer a coprime to m. The smallest exponent $n > 0$ such that $a^n \equiv 1 \bmod m$ is called the order of $a \bmod m$; we write $n = \operatorname{ord}_m(a)$. Note that we always have $\operatorname{ord}_m(1) = 1$. Here's a table for the orders of elements in $(\mathbb{Z}/7\mathbb{Z})^\times$:

$a \bmod 7$	1	2	3	4	5	6
$\operatorname{ord}_7(a)$	1	3	6	3	6	2

If $m = p$ is prime, then Fermat's Little Theorem gives us $a^{p-1} \equiv 1 \bmod p$, i.e., the order of $a \bmod p$ is at most $p-1$. In general, the order of a is not $p-1$; it is, however, always a divisor of $p-1$ (as the table above suggested):

Proposition 2.6. *Given a prime p and an integer a coprime to p, let n denote the order of a modulo p. If m is any integer such that $a^m \equiv 1 \bmod p$, then $n \mid m$. In particular, n divides $p-1$.*

Proof. Write $d = \gcd(n,m)$ and $d = nx + my$; then $a^d = a^{nx+my} \equiv 1 \bmod p$ since $a^n \equiv a^m \equiv 1 \bmod p$. The minimality of n implies that $n \le d$, but then $d \mid n$ shows that we must have $d = n$, hence $n \mid m$. □

Here comes a pretty application to prime divisors of Mersenne and Fermat numbers.

Corollary 2.7. *If p is an odd prime and if $q \mid M_p$, then $q \equiv 1 \bmod 2p$.*

Proof. It suffices to prove this for prime values of q (why?). So assume that $q \mid 2^p - 1$; then $2^p \equiv 1 \bmod q$. By Proposition 2.6, the order of $2 \bmod p$ divides p, and since p is prime, we find that $p = \operatorname{ord}_p(a)$.

On the other hand, we also have $2^{q-1} \equiv 1 \bmod p$ by Fermat's little theorem, so Proposition 2.6 gives $p \mid (q-1)$, and this proves the claim because we clearly have $q \equiv 1 \bmod 2$. □

Fermat numbers are integers $F_n = 2^{2^n} + 1$ (thus $F_1 = 5$, $F_2 = 17$, $F_3 = 257$, $F_4 = 65537$, ...), and Fermat conjectured (and once even seemed to claim he had a proof) that these integers are all primes. These integers became much more interesting when Gauss succeeded in proving that a regular p-gon, p an odd prime, can be constructed with ruler and compass if p is a Fermat prime. Gauss also stated that he had proved the converse, namely that if a regular p-gon can be constructed by ruler and compass, then p is a Fermat prime, but the first (almost) complete proof was given by Pi èrre Wantzel.[4]

Corollary 2.8. *If q divides F_n, then $q \equiv 1 \bmod 2^{n+1}$.*

Proof. It is sufficient to prove this for prime divisors q. Assume that $q \mid F_n$; then $2^{2^n} + 1 \equiv 1 \bmod q$, hence $2^{2^n} \equiv -1 \bmod q$ and $2^{2^{n+1}} \equiv 1 \bmod q$. We claim that actually $2^{n+1} = \operatorname{ord}_q(2)$: in fact, Proposition 2.6 says that the order divides 2^{n+1}, hence is a power of 2. But 2^{n+1} is clearly the smallest power of 2 that does it.

On the other hand, $2^{q-1} \equiv 1 \bmod q$ by Fermat's Little Theorem, and Proposition 2.6 gives $2^{n+1} \mid (q-1)$, which proves the claim. □

In particular, the possible prime divisors of $F_5 = 4294967297$ are of the form $q = 64m + 1$. After a few trial divisions one finds $F_5 = 641 \cdot 6700417$. This is how Euler disproved Fermat's conjecture. Today we know the prime factorization of F_n for all $n \le 11$, we know that F_n is composite for $5 \le n \le 30$ (and several larger values up to $n = 382447$), and we don't know any factors for $n = 14, 20, 22$ and 24. See
`http://vamri.xray.ufl.edu/proths/fermat.html`
for more.

Euler-Fermat

The following proof of Fermat's Little Theorem is slightly more complicated than the one we have given before but has the advantage of being valid for any finite abelian group:

Second Proof of Theorem 2.5. We claim that if r runs though the residue classes $1, 2, \ldots, p-1 \bmod p$, then so does ra. Clearly, none of the ra is divisible by p, since $p \nmid a$ by assumption and $p \nmid r$ since $1 \le r < p$. Since

[4] Pièrre Wantzel, 1814 (Paris) – 1848 (Paris).

there are $p-1$ numbers ra and $p-1$ nonzero residue classes modulo p, all we need to show is that the residue classes $ra \bmod p$ are different. But assume that $ra \equiv sa \bmod p$ for $1 \le r < s < p$; a and p are coprime, we nay cancel a by Proposition 2.1, and we find $r \equiv s \bmod p$. Since $1 \le r < s < p$, this is impossible, hence the ra are different modulo p.

This means that $a, 2a, \ldots, (p-1)a$ are congruent to $1, 2, \ldots, p-1$ in some order. But then the products over all these elements must be congruent to each other:

$$a \cdot 2a \cdots (p-1)a \equiv 1 \cdot 2 \cdots (p-1) \bmod p.$$

The left hand side equals $a \cdot 2a \cdots (p-1)a = a^{p-1} \cdot 1 \cdot 2 \cdots (p-1)$, and since we may cancel the common factor $1 \cdot 2 \cdots (p-1)$, again by Proposition 2.1, our claim follows. □

As a matter of fact, Fermat's little theorem holds for any group G of finite order n:

Theorem 2.9 (Lagrange's Theorem). *If G is a finite abelian group of order n, then $a^n = 1$ for any $a \in G$.*

Proof. The proof is basically the same: write $G = \{1 = g_1, g_2, \ldots, g_n\}$; then $G = \{g_1 g, g_2 g, \ldots, g_n g\}$ because $g_i g = g_j g$ implies $g_i = g_j$ (just multiply by the inverse of g: in groups, every element has an inverse). Thus $\prod_{j=1}^n g_j = \prod_{j=1}^n g_j g = g^n \prod_{j=1}^n g_j$, (in the last equation we have used that G is abelian: observe that e.g. $g_1 g g_2 g = g^2 g_1 g_2$ since we may pull the g's up to the front) and canceling $\prod g_j$ (i.e. multiplying by its inverse) we get $g^n = 1$. □

Corollary 2.10. *We have $a^{p-1} \equiv 1 \bmod p$ for any prime p and any integer a not divisible by p.*

Proof. Apply Lagrange's Theorem to the group $G = (\mathbb{Z}/p\mathbb{Z})^\times$. Since $\#G = p-1$, we get $[a]^n = [1]$ for every element $[a] \in (\mathbb{Z}/p\mathbb{Z})^\times$. □

Before we apply Lagrange's Theorem to the group $(\mathbb{Z}/m\mathbb{Z})^\times$, we shall introduce a notation for the number of elements in $(\mathbb{Z}/m\mathbb{Z})^\times$: we write $\phi(m) = \#(\mathbb{Z}/m\mathbb{Z})^\times$; ϕ is called Euler's phi function (or totient, especially in older books).

Corollary 2.11 (Theorem of Euler-Fermat). *We have $a^{\phi(m)} \equiv 1 \bmod m$ for any $a \in (\mathbb{Z}/m\mathbb{Z})^\times$, that is whenever $\gcd(a, m) = 1$.*

Without a formula for computing $\phi(m)$ this result would be pretty useless.

We can compute $\phi(m)$ for small values of m by brute force. Since $\phi(m)$ counts the invertible residue classes modulo m (in other words: the number of a with $1 \le a < m$ and $\gcd(a,m) = 1$), this is no big deal for small values of m. For example, $\phi(6) = 2$ since exactly two integers satisfy these conditions, namely $a = 1$ and $a = 5$; in other words: $(\mathbb{Z}/6\mathbb{Z})^\times = \{[1],[5]\}$. Here are some more values:

m	1	2	3	4	5	6	7	8	9	10	11
$\phi(m)$	1	1	2	2	4	2	6	4	6	4	10

Observe that $\phi(m)$ is even for $m \ge 3$ and that $\phi(m) = m - 1$ if and only if m is prime.

We will derive a formula for $\phi(m)$ by combining two results: one gives $\phi(m)$ for prime powers, the other shows that ϕ is multiplicative for coprime values.

Proposition 2.12. *We have $\phi(p^k) = (p-1)p^{k-1}$ for all $k \ge 1$.*

As usual, we give two proofs. The first simply counts the number of elements in $\mathbb{Z}/p^k\mathbb{Z}$ that are nonunits. Since nonunits here are exactly the elements divisible by primes, we see that $\phi(p^k)$ is the number of all integers $0 \le n < p^k$ minus those that are divisible by p. But these are exactly the numbers pb for $0 \le b < p^{k-1}$, so $\phi(p^k) = p^k - p^{k-1} = (p-1)p^{k-1}$.

The other result we will need is

Proposition 2.13. *If m and n are coprime, then $\phi(mn) = \phi(m)\phi(n)$.*

Proof. We will show that there exists a bijection between the prime residue classes modulo mn and the pairs of prime residue classes modulo m and n. More precisely: the map

$$\psi : (\mathbb{Z}/mn\mathbb{Z})^\times \longrightarrow (\mathbb{Z}/m\mathbb{Z})^\times \oplus (\mathbb{Z}/n\mathbb{Z})^\times : a + mn\mathbb{Z} \longmapsto (a + m\mathbb{Z}, a + n\mathbb{Z})$$

is bijective. Since there are exactly $\phi(m)\phi(n)$ such pairs, the claim will follow.

Let us first prove that ϕ is onto: given integers r, s such that $\gcd(r,m) = \gcd(s,n) = 1$, we have to find an integer a such that $a \equiv r \bmod m$ and $a \equiv s \bmod n$. Of course we have to use $\gcd(m,n) = 1$ somehow, and in general this is best done by using a Bezout representation. Thus let x, y be integers with $1 = xm + yn$. We claim that $a = ryn + sxm$ does it: in fact,

$a = ryn + sxm \equiv ryn \equiv 1 \bmod m$ since $yn \equiv 1 \bmod m$ from the Bezout representation, and similarly $a = ryn + sxm \equiv sxm \equiv s \bmod n$. Moreover, $a \in (\mathbb{Z}/mn\mathbb{Z})^\times$ since $\gcd(a, mn) \mid \gcd(a, m)\gcd(a, n) = 1$.

It remains to show that ϕ is injective. Assume that there are residue classes $a \bmod mn$ and $b \bmod mn$ such that $a \equiv b \bmod m$ and $a \equiv b \bmod n$. By Lemma 1.6, this implies that $a \equiv b \bmod mn$ and proves the injectivity of ϕ. □

These formulas allow the fast computation of $\phi(m)$ if the prime factorization of m is known: for example, $\phi(200) = \phi(8 \cdot 25) = \phi(8)\phi(25)$ because $\gcd(8, 25) = 1$; moreover $\phi(8) = 4$ and $\phi(25) = 20$ because of Proposition 2.12.

In the general case, the last two results imply

Corollary 2.14. *For integers n with prime factorization $n = p_1^{a_1} \cdots p_r^{a_r}$, we have*

$$\phi(n) \;=\; \prod_{i=1}^{r-1}(p_i - 1)p_i^{a_i - 1}.$$

Proof. Proposition 2.13 implies that

$$\phi(n) = \phi(p_1^{a_1}) \cdots \phi(p_r^{a_r}),$$

while Proposition 2.12 then does the rest. □

2.2 Euler's ϕ revisited and primitive roots

Before we go on we have to prove that the group $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic, that is: there is an integer g such that each residue class $[a]$ in $(\mathbb{Z}/p\mathbb{Z})^\times$ can be written uniquely in the form $a \equiv g^k \bmod p$ for some $0 \le k < p - 1$. Any such integer g is called a primitive root modulo p. An integer g is a primitive root modulo p if and only if $\operatorname{ord}_p(g) = p - 1$, because then and only then do the powers of g generate $p - 1$ different residue classes.

Example: 2 and 3 are primitive roots modulo 5.

More generally, an integer g is called a primitive root modulo m if the powers of g generate all residue classes in $(\mathbb{Z}/m\mathbb{Z})^\times$. Since there are exactly $\phi(m)$ such residue classes, g is a primitive root modulo m if and only if g has order $\phi(m)$ modulo m.

Primitive roots do not always exist: there are no primitive roots modulo 8 or modulo 15. In this Chapter we shall prove that there are primitive roots modulo odd prime powers.

Example: 2 is a primitive root modulo 9, but 4 is not. In fact, the powers of 2 mod 9 are $2^0 \equiv 1$, $2^1 \equiv 2$, $2^2 \equiv 4$, $2^3 \equiv 8$, $2^4 \equiv 7$, and $2^5 \equiv 5 \bmod 9$ (of course, $2^6 \equiv 1 \bmod 9$ by Euler-Fermat). On the other hand, $4^0 \equiv 1$, $4^1 \equiv 4$, $4^2 \equiv 7$, and $4^3 \equiv 1 \bmod 9$, so the powers of 4 only generate 3 different classes modulo 9.

The fact that $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic is a special case of a far more general result on finite fields: these are fields with finitely many elements. So far we have only learned about finite fields with p elements, namely $\mathbb{Z}/p\mathbb{Z}$, but we shall see in Chapter 4 that there are others.

Theorem 2.15. *The multiplicative group $F^\times$ of any finite field is cyclic.*

Since $\mathbb{Z}/p\mathbb{Z}$ is a finite field, this proves the existence of primitive roots modulo p.

Let us collect a couple of results:

Lemma 2.16. *Let F be a field and $f \in F[X]$ a polynomial with coefficients in F. Then f has at most $\deg f$ roots.*

Proof. We prove this by induction. If $\deg f = 0$, then $f(X) = a$ is a nonzero constant function, hence f has no zero, and $0 = \deg f$.

Now assume that the result is true for all $g \in F[X]$ with degree $\deg g \le n$, and suppose $\deg f = n + 1$. If f has no zero, the claim is trivially true, so assume that $f(x) = 0$ for some $x \in F$. Then $f(X) = (X - x)g(X) + a$ for some $g \in F[X]$ and a constant $a \in F$ by the Euclidean division algorithm. Putting $X = x$ shows that $0 = f(x) = a$, hence $f(X) = (X - x)g(X)$, and clearly $\deg g = \deg f - 1 = n$. By induction assumption, g has at most n roots; since $f(X) = (X - x)g(X)$, we see that f has at most $n + 1$ roots. □

For an element a of a finite group G, its order is the smallest positive integer r such that $a^r = 1$.

Lemma 2.17. *If G is an abelian group, and if $a, b \in G$ are elements of order m and n respectively such that $\gcd(m, n) = 1$, then ab has order mn.*

Proof. Clearly $(ab)^{mn} = a^{mn}b^{mn} = (a^m)^n(b^m)^n = 1$, so ab has order dividing mn (note that we have used commutativity here). For the converse, we first show that a^n has order m: to this end, write $1 = mx + ny$; then $a^{ny} = a^{1-mx} = a$, and a has order m. Thus a^n has order $\ge m$, and on the other hand $(a^n)^m = 1$ shows it has order $\le m$. This proves the claim. Similarly, b^m has order n.

Now $(ab)^m = b^m$, hence ab has order divisible by n, and similarly $(ab)^n = a^n$ shows that ab has order divisible by m, so ab has order divisible by mn since m and n are coprime. □

Proof of Theorem 2.15. Let p be a prime divisor of the order n of $F^\times$. Then there is an element $a \in F$ such that $a^{n/p} \ne 1$. For if not, then every $a \in F$ is a root of the polynomial $f(X) = X^{n/p} - 1$; in particular, f has degree n/p and n roots. But polynomials f over fields can have at most $\deg f$ roots: contradiction.

Now let p^r be the exact power of a prime p that divides $n = \#F^\times$; then we claim that the element $x = a^{n/p^r}$ has order p^r. In fact, $x^{p^r} = a^n = 1$ by Lagrange's Theorem, so the order of x divides p^r. If the order were smaller, then we would have $x^{p^{r-1}} = 1$; but $x^{p^{r-1}} = a^{n/p} \ne 1$ by choice of a.

Now write $n = p_1^{r_1} \cdots p_t^{r_t}$. By the above, we can construct an element x_i of order $p_i^{r_i}$ for every $1 \le i \le t$. But then $x_1 \cdots x_t$ has order n by Lemma 2.17 (use induction). □

We now claim

Proposition 2.18. *If p is an odd prime, then $(\mathbb{Z}/p^k\mathbb{Z})^\times \simeq \mathbb{Z}/(p-1)p^{k-1}\mathbb{Z}$.*

In particular, there exist primitive roots modulo p^k (p odd) because the group $\mathbb{Z}/(p-1)p^{k-1}\mathbb{Z}$ is cyclic. For a proof we need the following congruence:

Lemma 2.19. *If p is an odd prime and a an integer not divisible by p, then*

$$(1 + ap)^{p^{k-1}} \equiv 1 + ap^k \bmod p^{k+1}. \tag{2.1}$$

In particular, $\operatorname{ord}_{p^k}(1 + ap) = p^{k-1}$ *for all $k \ge 1$.*

Proof. The claim is trivial for $k = 1$ because $(1 + ap)^{p^0} \equiv 1 + ap \bmod p^2$. Assume that (2.1) holds for some k; then $(1 + ap)^{p^{k-1}} = 1 + ap^k + p^{k+1}c$ for some integer c, hence $(1 + ap)^{p^k} = (1 + ap^k + p^{k+1}c)^p = (1 + ap^k)^p + \binom{p}{1}(1 + ap^k)^{p-1}p^kc + \ldots \equiv (1 + ap^k)^p \equiv 1 + \binom{p}{1}ap^k = 1 + ap^{k+1} \bmod p^{k+2}$. This proves (2.1).

The congruence (2.1) implies $(1+ap)^{p^{k-1}} \equiv 1 \bmod p^k$, hence $\operatorname{ord}_{p^k}(1+ap)$ divides p^{k-1}. In fact we must have equality: if not, then $(1+ap)^r \equiv 1 \bmod p^k$ for some proper divisor r of p^{k-1}. But $(1+ap)^{p^{k-2}} \equiv 1+ap^{k-1} \bmod p^k$ shows that $(1+ap)^{p^{k-2}} \not\equiv 1 \bmod p^k$, hence $\operatorname{ord}_{p^k}(1+ap) = p^{k-1}$. □

Another useful observation is

Lemma 2.20. *If g is a primitive root modulo some* odd *prime p, then g or $g+p$ is a primitive root modulo p^k for every $k \ge 1$.*

Proof. In fact, assume that g is a primitive root modulo p, i.e. $\operatorname{ord}_p(g) = p-1$. If $g^{p-1} \equiv 1 \bmod p^2$, then $(g+p)^{p-1} = g^{p-1} + \binom{p-1}{1} g^{p-2}p + \ldots \equiv 1 + (p-1)pg^{p-2} \equiv 1 - pg^{p-2} \bmod p^2$. If we replace g by $g+p$ in this case, we may (and will) assume that $g^{p-1} \not\equiv 1 \bmod p^2$.

Since $g^{p-1} \equiv 1 \bmod p$, we can write $g^{p-1} = 1+ap$ for some integer a, and our assumption implies that $p \nmid a$. By (2.1), $\operatorname{ord}_{p^k}(g^{p-1}) = \operatorname{ord}_{p^k}(1+ap) = p^{k-1}$. Thus the smallest $r > 0$ such that $(g^{p-1})^r \equiv 1 \bmod p^k$ is $r = p^{k-1}$. This in turn implies that the smallest $r > 0$ such that $g^r \equiv 1 \bmod p^k$ is $r = (p-1)p^{k-1}$. Since $(p-1)p^{k-1} = \phi(p^k)$, we see that g is a primitive root modulo p^k. □

Now we are ready for the

Proof of Prop. 2.18. Let g be a primitive root modulo p^k, and define a map $\phi : \mathbb{Z}/(p-1)p^{k-1}\mathbb{Z} \longrightarrow (\mathbb{Z}/p^k\mathbb{Z})^\times$ by $\phi(a+(p-1)p^{k-1}\mathbb{Z}) \equiv g^a \bmod p^k$. This is a homomorphism since $\phi((a+b)+(p-1)p^{k-1}\mathbb{Z})) \equiv g^{a+b} = g^a g^b \bmod p^k = \phi(a+(p-1)p^{k-1}\mathbb{Z})\phi(b+(p-1)p^{k-1}\mathbb{Z})$.

Next, $\ker\phi$ consists of all residue classes $a+(p-1)p^{k-1}\mathbb{Z}$ such that $g^a \equiv 1 \bmod p^k$. Since g is a primitive root, the integers a with $g^a \equiv 1 \bmod p^k$ are exactly those divisible by $\phi(p^k)$, hence $\ker\phi = 0 + (p-1)p^{k-1}\mathbb{Z}$, and ϕ is injective. Since injective maps between finite sets of the same cardinality are bijective, this proves the proposition. □

Observe that $\mathbb{Z}/(p-1)p^{k-1}\mathbb{Z} \simeq \mathbb{Z}/(p-1)\mathbb{Z} \oplus \mathbb{Z}/p^{k-1}\mathbb{Z}$ by Corollary A.4.

In particular, there exist primitive roots modulo p^k (this is any element of $(\mathbb{Z}/p^k\mathbb{Z})^\times$ that generates this group). Also note that the result does not hold for powers of 2:

Proposition 2.21. *We have $(\mathbb{Z}/2^k\mathbb{Z})^\times \simeq \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2^{k-2}\mathbb{Z}$ for every integer $k \ge 2$.*

Proof. We define a map $f : \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2^{k-2}\mathbb{Z} \longrightarrow (\mathbb{Z}/2^k\mathbb{Z})^\times$ by putting $f((a + 2\mathbb{Z}, b + 2^{k-2}\mathbb{Z}) = (-1)^a 5^b + 2^k\mathbb{Z}$. This is clearly a homomorphism (why?). If we can show that $\ker f = 0$, then f must be bijective since an injection be tween finite sets is always a bijection.

Now $\ker f = \{(a + 2\mathbb{Z}, b + 2^{k-2}\mathbb{Z}) : (-1)^a 5^b \equiv 1 \bmod 2^k\}$. Since $k \geq 2$, we must have in particular $1 \equiv (-1)^a 5^b \equiv (-1)^a \bmod 4$, and this implies $a \in 2\mathbb{Z}$. Thus

$$\ker f = \{(0 + 2\mathbb{Z}, b + 2^{k-2}\mathbb{Z}) : 5^b \equiv 1 \bmod 2^k\}.$$

We claim that $5^b \equiv 1 \bmod 2^k$ if and only if $b \in 2^{k-2}\mathbb{Z}$.

For a proof, we will use the congruence $5^{2^r} \equiv 1 + 2^{r+2} \bmod 2^{r+3}$. This congruence is true if $r = 0$. Assume that it holds for some $r \in \mathbb{Z}$. Then $5^{2^r} = 1 + 2^{r+2} + 2^{r+3}c$ for some $c \in \mathbb{Z}$, hence $5^{2^{r+1}} = 1 + 2\ cdot 2^{r+2} + 2^{2r+4} + 2(1 + 2^{r+2})2^{r+3}c \equiv 1 + 2^{r+3} \bmod 2^{r+4}$.

Back to our claim. Assume that $b \in 2^{k-2}\mathbb{Z}$. Then $b = 2^{k-2}c$, so $5^b = (5^{2^{k-2}})^c \equiv 1 \bmod 2^k$.

Conversely, assume that $5^b \equiv 1 \bmod 2^k$. There are integers r and c such that $b = 2^r c$ with c odd. Then $5^b = (5^{2^r})^c \equiv (1 + 2^{r+2})^c \equiv 1 + 2^{r+2} \bmod 2^{r+3}$, and this shows that the minimal b with $5^b \equiv 1 \bmod 2^k$ is $b = k - 2$. By Proposition 2.6, all b satisfying the last congruence are multiples of 2^{k-2}. This proves our proposition. □

Proposition 2.22. *If m and n are coprime integers, then $(\mathbb{Z}/mn\mathbb{Z})^\times \simeq (\mathbb{Z}/m\mathbb{Z})^\times \oplus (\mathbb{Z}/n\mathbb{Z})^\times$.*

Proof. We have to find a map sending a residue class modulo mn to two residue classes modulo m and n. As a matter of fact, a good memory is all we need because we have found one in our first proof:

$$\psi : (\mathbb{Z}/mn\mathbb{Z})^\times \longrightarrow (\mathbb{Z}/m\mathbb{Z})^\times \oplus (\mathbb{Z}/n\mathbb{Z})^\times : a + mn\mathbb{Z} \longmapsto (a + m\mathbb{Z}, a + n\mathbb{Z})$$

All that's left to do is check that it works.

First, ψ is a homomorphism. This is so obvious that it is hard to write down: $\psi(ab + mn\mathbb{Z}) = (ab + m\mathbb{Z}, ab + n\mathbb{Z}) = (a + m\mathbb{Z}, a + n\mathbb{Z}) \cdot (b + m\mathbb{Z}, b + n\mathbb{Z})$ by definition of the direct sum; the last product is $\psi(a + m\mathbb{Z}) \cdot \psi(b + n\mathbb{Z})$, and we are done.

Since we have already shown (in our first proof) that ψ is bijective, the proof is complete. □

Odds and Ends

Theorem 2.23 (Wilson's Theorem). *For any $p \geq 3$, we have $(p-1)! \equiv -1 \bmod p$ if and only if p is a prime.*

Proof. The idea is to look at pairs of the elements of $(\mathbb{Z}/p\mathbb{Z})^\times$. In fact, for every $a \in (\mathbb{Z}/p\mathbb{Z})^\times$ there is an element $a^{-1} \in (\mathbb{Z}/p\mathbb{Z})^\times$ such that $a \cdot a^{-1} \equiv 1 \bmod p$. If a and a^{-1} were different residue classes for every a, then $(\mathbb{Z}/p\mathbb{Z})^\times$ would be the union of $\frac{p-1}{2}$ such pairs, and we would have $(p-1)! \equiv 1 \bmod p$. There are two exceptions, however: the congruence $a \equiv a^{-1} \bmod p$ is equivalent to $a^2 \equiv 1 \bmod p$, and this congruence has two solutions, namely $a \equiv 1 \bmod p$ and $a \equiv -1 \bmod p$ (here we use that $\mathbb{Z}/p\mathbb{Z}$ is a field).

Thus $(\mathbb{Z}/p\mathbb{Z})^\times \setminus \{-1, +1\}$ is the union of pairs $\{a, a^{-1}\}$ with $a \not\equiv a^{-1} \bmod p$, hence the product over all elements of $(\mathbb{Z}/p\mathbb{Z})^\times \setminus \{-1, +1\}$ is congruent to $1 \bmod p$. We can get $(p-1)!$ by multiplying this product with the two missing classes $1 \bmod p$ and $-1 \bmod p$, and this gives the result.

We still have to prove the converse: assume that $(n-1)! \equiv -1 \bmod n$; if p is a prime divisor of n, this implies $(n-1)! \equiv -1 \bmod p$. But $p < n$ also implies that p occurs as a factor of $(p-1)!$ on the left hand side, hence we would have $0 \equiv (n-1)! \bmod p$. But then $0 \equiv -1 \bmod p$, a contradiction. □

Note that Wilson's theorem provides us with a primality test; unfortunately the only known way to compute $(n-1)!$ is via $n-2$ multiplications, so it takes even longer than trial division!

Proof # 2. Consider the polynomial $f(X) = X^{p-1} - [1] \in (\mathbb{Z}/p\mathbb{Z})[X]$; since the residue classes $[1], [2], \ldots, [p-1]$ modulo p are roots (by Fermat's Little Theorem) and since $\deg f = p-1$, these are all the roots of f, hence $f(X) = [c]_p(X-[1])(X-[2]) \cdots (X-[p-1])$ for some constant c. Since the leading coefficient of f is 1, we have $c = 1$. Multiplying out we find that the constant term of $(X-[1])(X-[2]) \cdots (X-[p-1])$ is $(-1)^{p-1}[(p-1)!]_p = (p-1)!$. On the other hand, the constant term of f is $[-1]$, which proves the theorem. □

Proof # 3. The classes $[1], [2], \ldots, [p-1]$ can be written (in some order) as $[g^0], [g^1], \ldots, [g^{p-2}]$, where g is a primitive root modulo p. Thus $(p-1)! = 1 \cdot 2 \cdots (p-1) \equiv g^0 g^1 \cdots g^{p-2} = g^{(p-1)(p-2)/2} \bmod p$. But $g^{(p-1)/2} \equiv -1 \bmod p$, so $(p-1)! \equiv (-1)^{p-2} = -1 \bmod p$.

The fact that $x := g^{(p-1)/2} \equiv -1 \bmod p$ can be proved as follows: squaring yields $x^2 = g^{p-1} \equiv 1 \bmod p$, and this shows $p \mid (x^2-1) = (x-1)(x+1)$.

Since p is prime, we have $p \mid (x-1)$ or $p \mid (x+1)$, i.e., $x \equiv 1 \bmod p$ or $x \equiv -1 \bmod p$. The first possibility $x = g^{(p-1)/2} \equiv 1 \bmod p$ cannot occur because g is a primitive root modulo p. Thus the second possibility must hold, and this concludes the third proof of Wilson's theorem. □

2.3 RSA

Cryptography deals with methods that allow us to transmit information safely, that is, in such a way that eavesdroppers have no chance of reading it. Simple methods for encrypting messages were known and widely used in military circles for several millenia; basically all of these codes are easy to break with computers.

An example of such a classical code is Caesar's cipher: permute the letters of the alphabet by sending X $\longmapsto$ A, Y $\longmapsto$ B, Z $\longmapsto$ C, A $\longmapsto$ D etc; the text "ET TU, BRUTE" would be encrypted as "BQ QR, YORQB". For longer texts, analyzing the frequency of letters (for given languages) makes breaking this and similar codes a breeze if you are equipped with a computer.

Another common feature of these ancient methods of encrypting messages is the following: anyone who knows the key, that is, the method with which messages are encrypted, can easily break the code by inverting the encryption. In 1976, Diffie and Hellman suggested the existence of public key cryptography: these are methods for encrypting messages that do not allow you to read encrypted messages even if you know the key. The most famous of all public key cryptosystems is called RSA after its discoverers Ramir, Shamir and Adleman (1978).

Here's the simple idea: assume that Bob wants to receive secure messages; he selects two (large) primes p and q and forms their product $n = pq$. Bob also chooses an integer $E < n$ coprime to $(p-1)(q-1)$. The integers n and E are made public and constitute the key, so everybody can encrypt messages. For decrypting messages, however, one needs to know the prime factors p and q, and if p and q are large enough (say about 150 digits each) then known factorization methods cannot factor n in any reasonable amount of time (say 100 years).

How does the encryption work? It is a simple matter to transform any text into a sequence of numbers, for example by using $a \longmapsto 01$, $b \to 02, \ldots$, with a couple of extra numbers for blanks, commas, etc. We may therefore assume that our message is a sequence of integers $T < n$ (if the text is longer, break it up into smaller pieces). Alice encrypts each integer T as $C \equiv T^E \bmod n$ and sends the sequence of C's to Bob (by email, say). Now Bob can decrypt the message as follows: since he knows p and q, he can form the product $m = (p-1)(q-1)$ and run the Euclidean algorithm on the pair (E, m) to find an integer D such that $DE \equiv 1 \bmod m$. Now he takes the message C and computes $C^D \bmod n$. The result is $C^D \equiv (T^E)^D = T^{DE} \bmod n$, but since $DE \equiv 1 \bmod m = \phi(n)$, the theorem of Euler-Fermat shows that $C^D \equiv T \bmod n$, and Bob has got the original text that Alice sent him.

Now assume that Celia is eavesdropping. Of course she knows the pair (n, E) (which is public anyway), and she also knows the message C that Alice sent to Bob. That does not suffice for decrypting the message, however, since one seems to need an in verse D of $E \bmod (p-1)(q-1)$ to do that; it is likely that one needs to know the factors of n in order to compute D.

Baby Example. The following choice of $n = 1073$ with $p = 29$ and $q = 37$ is not realistic because this number can be factored easily; its only purpose is to illustrate the method.

So assume that Bob picks the key $(n, E) = (1073, 25)$. Alice wants to send the message "miss piggy" to Bob. She starts by transforming the message into a string of integers as follows:

	m	i	s	s		p	i	g	g	y
T	13	9	19	19	27	16	9	7	7	25

Next she encrypts this sequence by computing $C \equiv T^{25} \bmod n$ for each of these T: starting with $13^{25} \equiv 671 \bmod 1073$, she finds

T	13	9	19	19	27	16	9	7	7	25
C	671	312	901	901	656	1011	312	922	922	546

Alice sends this string of C's to Bob. Knowing the prime factorization of n, Bob is able to compute the inverse of 25 mod $(p-1)(q-1)$ as follows: he multiplies $p - 1 = 28$ and $q - 1 = 36$ to get $(p-1)(q-1) = 28 \cdot 36 = 1008$. Then he applies the extended Euclidean algorithm to $(25, 1008)$ and finds $1 = 25 \cdot 121 - 1008 \cdot 3$, and this shows that $D = 121$.

Now Bob takes the string of C's he got from Alice and decrypts them: starting with $671^{121} \equiv 13 \bmod n$ he can get back the string of T's, and hence the original message.

Remark. There is a big problem with this baby example: if we encrypt the message letter for letter, then equal letters will have equal code, and the cryptosystem can be broken (if the message is long enough) by analyzing the frequency with which each letter occurs (say in English). This problem vanishes into thin air when we use (realistic) key sizes of about 10^{200} digits: there we encrypt the message in blocks of about 100 letters, and since the chance that any two blocks of 100 letters inside a message coincide is practically 0, an attack based on the frequency of letters will not be successful for keys of this size.

RSA can also be applied to the signature problem. Assume that Alice receives an email from someone claiming to be Bob. How can Alice verify that this is true? Here's the simple trick in a nutshell: both Bob and Alice choose public keys, say (n_A, E_A) for Alice and (n_B, E_B) for Bob. Moreover, Alice knows D_A with $D_A E_A \equiv 1 \bmod \phi(n_A)$, while Bob knows D_B with $D_B E_B \equiv 1 \bmod \phi(n_B)$. Now Bob encrypts his message as above, but instead of sending the T's to Alice, he computes $U = T^{D_B} \bmod n_B$ and sends the U's. In order to decrypt the message, Alice computes first $T \equiv U^{E_D} \bmod n_B$ and then decrypts the T's as in the original version of RSA using her D_A. If this works, then Alice can be sure that the message came from Bob because in order to encrypt the message this way, the sender has to know D_B.

2.4 Primality Tests

Fermat's Little Theorem says that if p is a prime and a an integer not divisible by p, then $a^{p-1} \equiv 1 \bmod p$. This can be turned into a primality test:

```
1.  Pick some random integer $0 < a < n$;
2.  Check whether $d := \gcd(a, n) = 1$;
     if not, print ``d is a factor of n'' and terminate;
3.  Check whether $a^{p-1} \equiv 1 \bmod p$;
     if not, print ``n is composite''.
```

Any integer n surviving this test is called a pseudoprime to basis a; as the example $a = 2$, $n = 341$ shows, there exist composite pseudoprimes.

The primality test given above can be turned into an algorithm that *proves* n to be a prime if it is one; here's the idea: we know that if n is prime, then $(\mathbb{Z}/n\mathbb{Z})^\times$ is cyclic, generated by a primitive root g. We know that $p - 1$ is the smallest positive exponent k of g such that $g^k \equiv 1 \bmod p$. In particular, $g^{(p-1)/q} \not\equiv 1 \bmod p$ for every prime divisor q of $p - 1$. Now we claim

Theorem 2.24. *If n and a are integers such that*

1. $a^{n-1} \equiv 1 \bmod n$ *and*

2. $a^{(n-1)/q} \not\equiv 1 \bmod n$ *for every prime divisor* q *of* $p-1$*,*

then n *is a prime, and* a *is a primitive root modulo* n*.*

Proof. Let r be the order of $a \bmod n$. Then r divides $n-1$ by Proposition 2.6. We claim that $r = n-1$. If not, then $n-1 = rs$ with $s > 1$, hence there is a prime factor $q \mid s$, i.e., $s = qt$ and $n-1 = rqt$. Then $a^{(n-1)/q} = a^{rt} = (a^r)^t \equiv 1^t = 1 \bmod n$ contradicting 2, so we conclude that $r = n-1$.

Since $a^{n-1} \equiv 1 \bmod n$, we must have $\gcd(a,n) = 1$ (if we had $q \mid a$ and $q \mid n$, then $q \mid a \Longrightarrow q \mid a^{n-1}$ and $q \mid n \Longrightarrow q \mid a^{n-1} - 1$ (from 1.), and this implies $q \mid 1$: contradiction). Thus the powers of $a \bmod n$ generate $n-1$ different residue classes modulo n, all of them coprime to n. Thus every nonzero residue class mod n has an inverse, hence n is prime (and a is a primitive root mod p): this is because if $n = de$ is a nontrivial factorization, then the residue class $d \bmod n$ is nonzero but does not have an inverse. □

Here's a baby-example: take $n = 127$; then $n-1 = 126 = 2 \cdot 3^2 \cdot 7$. Let us start with $a = 2$ (why not?). We first check that $2^{126} \equiv 1 \bmod 127$. Next we have to make sure that $2^{126/q} \not\equiv 1 \bmod 127$ for $q = 2, 3$ and 7. But already for $q = 2$ we find $2^{63} \equiv 1 \bmod 127$, and our algorithm fails.

Let's see if we are more successful with $a = 3$; again we find $3^{126} \equiv 1 \bmod 127$; now

$$\begin{aligned} 3^{63} &\equiv -1 \bmod 127, \\ 3^{42} &\equiv -20 \bmod 127, \\ 3^{18} &\equiv 18 \bmod 127, \end{aligned}$$

so 127 is indeed prime.

Thus it seems that with a few additional computations we can turn Fermat's little theorem into an algorithm that allows us to prove that a given integer is prime (or not). The problem, however, is this: in step 2, we need the complete factorization of $n-1$. Sometimes this is not a big problem, especially for numbers of the form $n = 2^k m + 1$ with small m, but for general integers this is indeed the bottleneck.

There are, however, improvements to this simple test: first, it can be shown that it suffices to know the factorization of a large part of $n-1$: the part of $n-1$ that we can factor has to be $> \sqrt{n}$.

The primality test given above works well if the prime factorization of $N-1$ is known. This is the case e.g. for Fermat numbers $N = F_n = 2^{2^n}+1$, where $N-1$ is a power of 2. We find:

Proposition 2.25. *A Fermat number F_n is prime if and only if $3^{(F_n-1)/2} \equiv -1 \bmod F_n$.*

Proof. The proof of the "only if" part will be deferred until we know about quadratic residues. Assume therefore that $3^{(F_n-1)/2} \equiv -1 \bmod F_n$; then Theorem 2.24 is satisfied with $a = 3$. □

2.5 Pollard's $p-1$-factorization Method

Pollard is definitely the world champion in inventing new methods for factoring integers. One of his earliest contributions were the $p-1$-method (ca. 1974), his ρ-method followed shortly after, and his latest invention is the number field sieve (which is based on ideas from algebraic number theory).

The idea behind Pollard's $p-1$-method is incredibly simple. Assume that we are given an integer N that we want to factor. Fix an integer $a > 1$ and check that $\gcd(a,N) = 1$ (should $d = \gcd(a,N)$ be not trivial, then we have already found a factor d and continue with N replaced by N/d).

Let p be a factor of N; by Fermat's Little Theorem we know that $a^{p-1} \equiv 1 \bmod p$, hence $D := \gcd(a^{p-1}-1, N)$ has the properties $p \mid D$ and $D \mid N$. Thus D is a nontrivial factor of N unless $D = N$ (which should not happen too often).

The procedure above is not much of a factorization algorithm as long as we have to know the prime factor p beforehand. The prime p occurs at two places in the method above: first, as the modulus when computing $a^{p-1} \bmod p$. But this problem is easily taken care of because we may simply compute $a^{p-1} \bmod N$. It is more difficult to get rid of the p in the exponent: the fundamental observation is that we can replace the exponent $p-1$ above by any multiple, and D still will be divisible by p (note though that the chance that $D = N$ has become slightly larger). Does this help us? Not always; assume, however, that $p-1$ is the product of *small* primes (say of primes below a bound B that in practice can be taken to be $B = 10^5$ or $B = 10^6$, depending on the computing power of your hardware). Then it is not too hard to come up with good candidates for multiples of $p-1$: we might simply pick $k = B!$, or, in a similar vein,

$$k = \prod_i p_i^{a_i}, \quad \text{where } p_i^{a_i} \le B < p_i^{a_i+1}. \tag{2.2}$$

If we $(p-1) \mid k$, then $a^k \equiv 1 \bmod p$, hence $p \mid D = \gcd(a^k - 1, N)$.

Thus the following algorithm has a good chance of finding those factors p of N for which $p-1$ has only small prime factors:

```
1.  Pick $a > 1$ and check that $\gcd(a, N) = 1$
2.  Choose a bound $B$, say $B = 10^4$, $10^5$, $10^6$, ...
3.  Pick $k$ as in (2.2) and compute $D = \gcd(a^k - 1, N)$.
```

Note that the computation of a^k can be done modulo N; if $p \mid N$ and $(p-1) \mid k$, then $a^k \equiv 1 \bmod p$, hence $p \mid D$.

If $D = 1$, we may increase k; if $D = N$, we can reduce k and repeat the computation.

Among the record factors found by the $p-1$-method is the 37-digit factor $p = 6902861817667290192729108442204980121$ of $71^{77} - 1$ with $p - 1 = 2^3 \cdot 3^3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 \cdot 401 \cdot 409 \cdot 3167 \cdot 83243 \cdot 83983 \cdot 800221 \cdot 2197387$ discovered by Dubner. A list of record factors can be found at
`http://www.users.globalnet.co.uk/~aads/Pminus1.html`

Here's a baby example: take $N = 1769$, $a = 2$ and $B = 6$. Then we compute $k = 2^2 \cdot 3 \cdot 5$ and we find $2^{60} \equiv 306 \bmod 1769$, $\gcd(305, 1769) = 61$ and $N = 29 \cdot 61$. Note that $61 - 1 = 2^2 \cdot 3 \cdot 5$, so the factor 61 was found, while $29 - 1 = 2^2 \cdot 7$ explains why 29 wasn't (although $29 < 61$).

Pollard's $p-1$-algorithm was the father of Lenstra's ECM (elliptic curve method) algorithm. While Pollard's method is based on the group $(\mathbb{Z}/p\mathbb{Z})^\times$ of order $p-1$, Lenstra's ECM is based on the group $E(\mathbb{Z}/p\mathbb{Z})$ defined by an elliptic curve, an object defined by equations of the form $y^2 = x^3 + ax + b$. The order $E(\mathbb{Z}/p\mathbb{Z})$ satisfies $p + 1 - 2\sqrt{p} \le \#\, E(\mathbb{Z}/p\mathbb{Z}) \le p + 1 + 2\sqrt{p}$, and by varying the elliptic curve we may eventually find a curve for which the group order is ”smooth”, that is, divisible only by small primes. Lenstra's ECM can find factors of up to 50 digits whereas Pollards method has not yet found any factors with more than 40 digits.

Another large class of factorization algorithms is based on an algorithm invented by Fermat: the idea is to write an integer n as a difference of squares. If $n = x^2 - y^2$, then $n = (x-y)(x+y)$, and unless this is the trivial factorization $n = 1 \cdot n$, we have found a factor.

Another baby example: take $n = 1073$; then $\sqrt{n} = 32.756\ldots$, so we start by trying to write $n = 33^2 - y^2$. Since $33^2 - 1073 = 16$, we find $n = 33^2 - 4^2 = (33-4)(33+4) = 29 \cdot 37$. If the first attempt would have been unsuccessful, we would have tried $n = 34^2 - y^2$, etc.

In modern algorithms (continued fractions, quadratic sieve, number field sieve) the equation $N = x^2 - y^2$ is replaced by a congruence $x^2 \equiv y^2 \bmod N$: if we have such a thing, then $\gcd(x-y, N)$ has a good chance of being a nontrivial factor of N. The first algorithm above constructed such pairs (x,y) by computing the continued fraction expansion of $\sqrt{n}$ (which we have not discussed), the number field sieve produces such pairs by factoring certain elements in algebraic number fields.

Notes

In this Chapter, we have introduced congruences. Even the concept of congruence alone has applications, for example to ISBN's and similar (simple) "error detecting codes".

Apart from the fundamental notion of a congruence, we introduced

- the order of an element g in a (multiplicative) group G as the smallest $n \geq 1$ such that $g^n = 1$.
- the order of a finite group G: the numbers of elements of G.

In the special case $G = (\mathbb{Z}/p\mathbb{Z})^\times$, we proved in Proposition 2.6 that the order of any $a \in G$ divides the order $p-1$ of the group. This is true in general (the proof goes through as well): in finite abelian groups, the order of any element divides the order of the group.

Among the main results proved in this Chapter are

- Fermat's Little theorem: $a^{p-1} \equiv 1 \bmod p$ for primes p and integers a not divisible by p. Fermat's little theorem was used to prove that prime divisors of Fermat and Mersenne numbers have a special form, for devising a primality algorithm in Section 2.4, as well as Pollard's $p-1$-method for factorization in Section 2.5.
- The Theorem of Euler-Fermat saying that $a^{\phi(m)} \equiv 1 \bmod m$ if $\gcd(a,m) = 1$. This contains Fermat's Little Theorem as a special case (take $m = p$)

and is the basis for RSA-cryptography. Note that the Theorem of Euler-Fermat is just a special case of Lagrange's Theorem that $a^n = 1$ in any (multiplicatively written) finite abelian group with n elements.

- Multiplicative groups of finite fields are cyclic. This was used to prove that primitive roots exist modulo p, and from there we used induction to prove the same thing for odd prime powers.

As for the methods of proof, we used two different approaches: a combinatorial one based on counting, and the more abstract approach using algebra.

elementary	abstract
$a^{p-1} \equiv 1 \bmod p$ if $p \nmid a$	Lagrange's Theorem
$\phi(p) = p - 1$	$(\mathbb{Z}/p\mathbb{Z})^\times \simeq \mathbb{Z}/(p-1)\mathbb{Z}$
$\phi(p^k) = (p-1)p^{k-1}$	p odd: $(\mathbb{Z}/p^k\mathbb{Z})^\times \simeq \mathbb{Z}/(p-1)p^{k-1}\mathbb{Z}$
$\gcd(a,b) = 1$: $\phi(ab) = \phi(a)\phi(b)$	$(\mathbb{Z}/ab\mathbb{Z})^\times \simeq (\mathbb{Z}/a\mathbb{Z})^\times \times (\mathbb{Z}/b\mathbb{Z})^\times$

The results on the right hand side are much stronger than those on the left hand side; for example, $(\mathbb{Z}/p^k\mathbb{Z})^\times \simeq \mathbb{Z}/(p-1)p^{k-1}\mathbb{Z}$ for odd primes p also implies that there exist primitive roots modulo p^k.

The isomorphisms on the right hand side also allow us to improve on the theorem of Euler-Fermat:

- For $G = (\mathbb{Z}/8\mathbb{Z})^\times$, Euler-Fermat says $a^4 \equiv 1 \bmod 8$ for every $a \in (\mathbb{Z}/8\mathbb{Z})^\times$ because $\phi(8) = 4$. On the other hand, $(\mathbb{Z}/8\mathbb{Z})^\times \equiv \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ implies the stronger claim that $a^2 \equiv 1 \bmod 8$.

- For $G = (\mathbb{Z}/24\mathbb{Z})^\times$, we have $\phi(24) = 8$, hence $a^8 \equiv 1 \bmod 24$ by uler-Fermat. On the other hand, $(\mathbb{Z}/24\mathbb{Z})^\times \simeq (\mathbb{Z}/8\mathbb{Z})^\times \times (\mathbb{Z}/3\mathbb{Z})^\times \simeq \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ implies that $a^2 \equiv 1 \bmod 24$ whenever $\gcd(a, 24) = 1$.

Thus the abstract approach is much more powerful than the elementary one. Of course, abstract algebra is perceived as being difficult by beginners; but difficult mathematics is not mastered by avoiding it.

Chapter 3

Quadratic Reciprocity

Quadratic Reciprocity belongs to the highlights of every introduction to number theory. Conjectured by Euler and partially proved by Legendre in the late 18th century, the first complete proof was published 1801 in Gauss's Disquisitiones Arithmeticae (actually he gave two proofs there, followed later by six others).

3.1 Quadratic Residues

Let F be a field; it is an apparently simple question to ask for a characterization of the squares in F, that is, the set of elements $a \in F$ such that $a = b^2$ for some $b \in F$. This question is trivial for $F = \mathbb{C}$ because every complex number is a square (you may deduce this from the fundamental theorem of algebra by looking at the polynomial $x^2 - a$, but it is instructive to give a direct proof, that is: given a complex number $r + si$, find its square roots). The answer is also easy for $F = \mathbb{R}$: a real number x is a square if and only if $x \geq 0$.

Knowledge about squares is important for solving quadratic equations: $x^2 + ax + b = 0$ has solutions in the reals if and only if the discriminant $a^2 - 4b$ of the polynomial is a square. The same thing is true for finite fields $\mathbb{Z}/p\mathbb{Z}$ for odd p (the case $p = 2$ is different because the formula for solving quadratic equations has a 2 in the denominator, and $2 = 0$ in $\mathbb{Z}/2\mathbb{Z}$): consider e.g. $x^2+2x-1 = 0$ over $\mathbb{Z}/p\mathbb{Z}$. The well known fomula gives the two solutions $\frac{1}{2}(-2 \pm \sqrt{8}\,) = -1 \pm \sqrt{2}$, so there are exactly two solutions if 2 is a square in $\mathbb{Q}(\sqrt{p}\,)$, and none otherwise. Example: for $p = 7$, $2 \equiv 3^2 \bmod 7$, so the formula gives the two solutions $-1 \pm 3 \equiv 2, -4 \bmod 7$, and in fact $2^2 - 2 \cdot 2 - 1 \equiv (-4)^2 + 2 \cdot (-4) - 1 \equiv 0 \bmod 7$.

Quadratic reciprocity helps us deciding whether certain elements are squares in $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ or not. We will call the squares in $\mathbb{F}_p$ (or, more exactly, the integers whose residue classes in $\mathbb{F}_p$ are squares) *quadratic residues* modulo p, the nonsquares *quadratic nonresidues*. Let us make some experiments; since 0 is always a square, we restrict ourselves to $\mathbb{F}_p^\times$.

prime	squares	nonsquares
3	1	2
5	$1, 4$	$2, 3$
7	$1, 2, 4$	$3, 5, 6$
11	$1, 3, 4, 5, 9$	$2, 6, 7, 8, 10$
13	$1, 3, 4, 9, 10, 12$	$2, 5, 6, 7, 8, 11$

There are hardly any regularities to discover. One may notice that the sums of the squares in $\mathbb{F}_p$ are divisible by p for $p > 3$ (can you prove that?), but we want to get a grip on the elements, not on sums (what about products?) of them.

Clearly 1 is always a square; but even the question when 2 is a quadratic residue seems hard to answer. An easier case to figure out is the residue class of -1, which is a square exactly for $p = 5$ and 13, and a nonsquare for $p = 3, 7$ and 11. Based on this observation, we might conjecture that

> The integer -1 is a quadratic residue modulo an odd prime p if and only if $p \equiv 1 \bmod 4$.

As a matter of fact, this is quite easy to prove; we will do that using a characterization of quadratic residues due to Euler:

Proposition 3.1 (Euler's Criterion). *If $a \in \mathbb{Z}$ is not divisible by p, then a is a quadratic residue or nonresidue modulo p according as $a^{(p-1)/2} \equiv +1$ or $a^{(p-1)/2} \equiv -1 \bmod p$.*

Proof. This is easy: assume that $a \equiv x^2 \bmod p$; then $a^{(p-1)/2} \equiv x^{p-1} \equiv 1 \bmod p$ by Fermat's Little Theorem.

Conversely, assume that $a^{(p-1)/2} \equiv +1 \bmod p$ and let g be a primitive root modulo p. Then $a \equiv g^r \bmod p$ for some $0 \le r < p - 1$; if r were odd, then $a^{(p-1)/2} \equiv (g^{(p-1)/2})^r \equiv (-1)^r = -1 \bmod p$, hence r must be even, say $r = 2s$. But then $a \equiv (g^s)^2 \bmod p$ is a quadratic residue. □

At this point it is appropriate to introduce the Legendre symbol. Given a prime p and an integer $a \in \mathbb{Z}$ with $p \nmid a$, we put

$$\Big(\frac{a}{p}\Big) = \begin{cases} +1, & \text{if } a^{(p-1)/2} \equiv +1 \bmod p, \\ -1, & \text{if } a^{(p-1)/2} \equiv -1 \bmod p. \end{cases}$$

By Euler's criterion, we have $(\frac{a}{p}) = +1$ if a is a quadratic residue modulo p, and $(\frac{a}{p}) = -1$ if a is a quadratic nonresidue. Observe that we have $a^{(p-1)/2} \equiv (\frac{a}{p}) \bmod p$ whenever a is not divisible by p. If we put $(\frac{a}{p}) = 0$ whenever $p \mid a$, the congruence $a^{(p-1)/2} \equiv (\frac{a}{p}) \bmod p$ holds for all integers a.

Euler's criterion has a couple of applications. While it can be proved directly from the definition that the product of two quadratic residues is again a quadratic residue, it is not as easy to show that the product of two nonresidues mod p is a quadratic residue:

Corollary 3.2. *The Legendre symbol* $(\frac{\cdot}{p})$ *induces a homomorphism* $(\mathbb{Z}/p\mathbb{Z})^\times \longrightarrow (\mathbb{Z}/4\mathbb{Z})^\times \simeq \mathbb{Z}/2\mathbb{Z}$*; in other words: we have* $(\frac{a}{p})(\frac{b}{p}) = (\frac{ab}{p})$ *for all* $a, b \in (\mathbb{Z}/p\mathbb{Z})^\times$.

Proof. Homework. □

Corollary 3.3. *The integer* -1 *is a quadratic residue modulo an odd prime* p *if and only if* $p \equiv 1 \bmod 4$. *In other words:* $(\frac{-1}{p}) = (-1)^{(p-1)/2}$.

Proof. By Euler's criterion, -1 is a quadratic residue modulo p if and only if $(-1)^{(p-1)/2} = +1$ (modulo p, but since p is odd, this implies equality). This in turn holds if and only if the exponent $\frac{p-1}{2}$ is even, that is, if and only if $p \equiv 1 \bmod 4$. □

That's not much, but better than nothing. As a matter of fact, this simple result allows us to prove that there are infinitely many primes of the form $4n - 1$. We first formulate a little

Lemma 3.4. *If* $p > 0$ *is an odd prime divisor of an integer of the form* $n^2 + 1$*, then* $p \equiv 1 \bmod 4$.

Proof. From $p \mid n^2 + 1$ we deduce that $n^2 \equiv -1 \bmod p$. Thus -1 is a quadratic residue modulo p, hence $p \equiv 1 \bmod 4$. □

Corollary 3.5. *There are infinitely many primes of the form* $4n + 1$.

Proof. Assume there are only finitely many primes of the form $4n+1$, say $p_1 = 5, p_2, \ldots, p_n$. Then $N = 4p_1^2 \cdots p_n^2 + 1$ is of the form $4n+1$ and greater than all the primes p_k of this form, hence N must be composite. Now N is odd, hence so is any prime divisor p of N, and since any such p is of the form $4n+1$ by Corollary 3.3, we conclude that $p = p_k$ for some index k. But then $p_k \mid N$ and $p_k \mid N - 1 = 4p_1^2 \cdots p_n^2$, and we get the contradiction that $p_k \mid (N - (N-1)) = 1$. □

Now let us study the behaviour of the prime 2:

p	3	5	7	11	13	17	19	23	29	31
$(2/p)$	-1	-1	$+1$	-1	-1	$+1$	-1	$+1$	-1	$+1$
$\sqrt{2}$	$-$	$-$	± 3	$-$	$-$	± 6	$-$	± 5	$-$	± 8

Thus 2 is a quadratic residue modulo 7, 17, 23, and 31; among the primes in this table, these are exactly the primes of the form $p \equiv \pm 1 \bmod 8$. Thus we conjecture:

Proposition 3.6. *The prime 2 is a quadratic residue modulo an odd prime p if and only if $p \equiv \pm 1 \bmod 8$. In other words: we have $(\frac{2}{p}) = (-1)^{(p^2-1)/8}$.*

The fact that the second claim is equivalent to the first is easy to check: Basically, the proof boils down to the following table:

$a \bmod 8$	1	3	5	7
$\frac{1}{8}(a^2-1) \bmod 2$	0	1	1	0

Great. Now how would one prove such a conjecture? Euler's criterion does not really seem to help, because we have no idea how to evaluate $2^{\frac{p-1}{2}} \bmod p$.

There is a simple proof that 2 is a quadratic residue modulo primes $p = 8k+1$: let g be a primitive root modulo p and put $s = g^k + g^{-k}$. Then $s^2 \equiv g^{2k} + g^{-2k} + 2 \bmod p$; but $g^{2k} + g^{-2k} = g^{-2k}(g^{4k}+1) \equiv 0 \bmod 4$ since $g^{4k} \equiv -1 \bmod p$. Thus $s^2 \equiv 2 \bmod p$.

There's actually a classical idea behind this trick: look at the eighth roots of unity in the complex numbers, say $\zeta = e^{2\pi i/8}$. Then $\sqrt{2} = \zeta + \zeta^{-1}$: in fact, $(\zeta + \zeta^{-1})^2 = i + i^{-1} + 2 = 2$ since $1/i = -i$, and moreover $\zeta + \zeta^{-1} > 1$ on the real line (sketch!). Our proof above was merely a translation of this computation from $\mathbb{C}$ to $\mathbb{Z}/p\mathbb{Z}$.

It is possible to do something similar (even for the general reciprocity law) by constructing $\sqrt{p}$ out of roots of unity; this requires some algebra, however, and we will choose a different proof with an elementary flavor.

For computing $(-3/p)$, the algebra involved is simple enough. We claim

Proposition 3.7. *For primes $p > 3$, we have*

$$\Big(\frac{-3}{p}\Big) = \begin{cases} +1 & \text{if } p \equiv 1 \bmod 3, \\ -1 & \text{if } p \equiv 2 \bmod 3. \end{cases}.$$

Proof. Before we go into details, here's the idea: if $p = 3n + 1$, let g be a primitive root mod p, put $\rho = g^n$, and show that $(\rho^2 - \rho)^2 \equiv -3 \bmod p$.

Conversely, if $x^2 \equiv -3 \bmod p$, put $\rho \equiv (-1 + x)/2 \bmod 3$ and show that ρ has order 3; since the order of ρ divides $p - 1$ by Proposition 2.6, we must have $p \equiv 1 \bmod 3$.

Now let's do it properly. Assume first that $p = 3n + 1$. We want to construct a square root of $-3 \bmod p$. To this end, pick a primitive root $g \bmod p$ and put $\rho = g^n \bmod p$. Then $\rho^3 \equiv 1 \bmod p$ by Fermat's Little Theorem, and $\rho \not\equiv 1 \bmod p$ since g is a primitive root. Thus $0 \equiv \rho^3 - 1 = (\rho - 1)(\rho^2 + \rho + 1) \bmod p$, and since $p \nmid (\rho - 1)$, we conclude that $\rho^2 + \rho + 1 \equiv 0 \bmod p$. But then $(\rho^2 - \rho)^2 = \rho^4 - 2\rho^3 + \rho^2 \equiv \rho - 2 + \rho^2 \equiv 1 + \rho + \rho^2 - 3 \equiv -3 \bmod p$, and we have shown that -3 is a square modulo p.

Now assume conversely that $x^2 \equiv -3 \bmod p$. We put $\rho \equiv \frac{1}{2}(-1 + x) \bmod p$, where $\frac{1}{2} \bmod p$ denotes the inverse of $2 \bmod p$, and find that $\rho^2 \equiv \frac{1}{4}(1 - 2x + x^2) \equiv \frac{1}{2}(-1 - x) \bmod p$ since $x^2 \equiv -3 \bmod p$. But then $\rho^3 \equiv \frac{1}{4}(1 - x^2) \equiv 1 \bmod p$, so the order of $\rho \bmod p$ divides 3. We claim that the order is 3; if not, the order would have to be 1, and this implies $\rho \equiv 1 \bmod p$; but $p \mid (\rho - 1) = \frac{1}{2}(-3 + x)$ implies $x \equiv 3 \bmod p$, hence $x^2 \equiv 9 \bmod p$ contradicting $x^2 \equiv -3 \bmod p$ whenever $p \neq 3$. □

3.2 Gauss's Lemma

The main ingredient of the elementary proofs of the quadratic reciprocity law is a lemma that Gauss invented for his third proof. Recall how we proved Fermat's Little Theorem: we took a complete set of prime residue classes $\{1, 2, \ldots, p - 1\}$, multiplied everything by a, and pulled out the factor a^{p-1}. For quadratic reciprocity, Euler's criterion suggests that we would like to pull out a factor $a^{(p-1)/2}$. That's what made Gauss introduce a halfsystem

modulo p: this is any set $A = \{a_1, \ldots, a_m\}$ of representatives for residue classes modulo $p = 2m+1$ with the following properties:
a) the a_j are distinct modulo p, that is: if $a_i \equiv a_j \bmod p$, then $i = j$;
b) every integer is either congruent modulo p to a_i or to $-a_i$ for some $1 \le i \le \frac{p-1}{2}$.

In other words: a halfsystem A is any set of integers such $A \cup -A$ is a complete set of prime residue classes modulo p. A typical halfsystem modulo p is the set $A = \{1, 2, \ldots, \frac{p-1}{2}\}$.

Now consider the prime $p = 13$, choose $A = \{1,2,3,4,5,6\}$, and look at $a = 2$. Proceeding as in the proof of Fermat's Little Theorem, we multiply everything in sight by 2 and find

$$\begin{aligned} 2\cdot 1 &\equiv +2 \bmod 13, \\ 2\cdot 2 &\equiv +4 \bmod 13, \\ 2\cdot 3 &\equiv +6 \bmod 13, \\ 2\cdot 4 &\equiv -5 \bmod 13, \\ 2\cdot 5 &\equiv -3 \bmod 13, \\ 2\cdot 6 &\equiv -1 \bmod 13. \end{aligned} \tag{3.1}$$

Thus three products still lie in A, while three others lie in $-A$. Thus there is an odd number of sign changes, and 2 is a quadratic nonresidue.

What about $a = 3$? Here we find

$$\begin{aligned} 3\cdot 1 &\equiv +3 \bmod 13, \\ 3\cdot 2 &\equiv +6 \bmod 13, \\ 3\cdot 3 &\equiv -4 \bmod 13, \\ 3\cdot 4 &\equiv -1 \bmod 13, \\ 3\cdot 5 &\equiv +2 \bmod 13, \\ 3\cdot 6 &\equiv +5 \bmod 13. \end{aligned} \tag{3.2}$$

Here the number of sign changes is even (there are two), and 3 is a quadratic residue modulo 13.

Gauss realized that this is not an accident: in fact, if you multiply the congruences (3.1), you get

$$2^6 \cdot 6! \equiv (-1)^3 \cdot 6! \bmod 13,$$

and since $\gcd(6!, 13) = 1$, canceling gives $2^6 \equiv -1 \bmod 13$, so by Euler's criterion we see that 2 must be a quadratic nonresidue modulo 13. Similarly, (3.2) gives $3^6 \equiv (-1)^2 = 1 \bmod 13$, so 3 is a quadratic residue modulo 13. We can do this in complete generality:

Lemma 3.8 (Gauss's Lemma). *Let $p = 2n+1$ be an odd prime, pick a half system $A = \{a_1, \ldots, a_n\}$, and let a be an integer not divisible by p. Write*

$$a_i a \equiv (-1)^{s(i)} a_{t(i)} \bmod p \tag{3.3}$$

for every $a_i \in A$, where $s(i) \in \{0,1\}$ and $t(i) \in \{1, 2, \ldots, n\}$. Then

$$a^n \equiv \prod_{i=1}^{n} (-1)^{s(i)} \bmod p.$$

Thus a is a quadratic residue or nonresidue modulo p according as the number of sign changes is even or odd. The proof is quite simple:

Proof. Observe that the $a_{t(i)}$ in (3.3) run through A if the a_i do, that is: the $a_{t(i)}$ are just the a_i in a different order. In fact, if we had $a_i a \equiv (-1)^{s(i)} a_{t(i)} \bmod p$ for $i \neq k$ and $a_k a \equiv (-1)^{s(k)} a_{t(k)} \bmod p$ with $a_{t(i)} = a_{t(k)}$, then dividing the first congruence by the second gives $a_{t(i)}/a_{t(k)} \equiv (-1)^{s(i)-s(k)} \bmod p$, that is, we have $a_{t(i)} \equiv \pm a_{t(k)} \bmod p$ for some choice of sign. But this is impossible since $1 \leq a_{t(i)}, a_{t(k)} \leq \frac{p-1}{2}$.

Now we apply the usual trick: if two sets of integers coincide, then the product over all elements must be the same. In our case, this means that $\prod_{i=1}^n a_i a \equiv \prod_{i=1}^n (-1)^{s(i)} a_{t(i)} \bmod p$. The left hand side equals $(a_1 a) \cdot (a_2 a) \cdots (a_n a) = a^n \prod_{i=1}^n a_i$, whereas the right hand side is $\prod_{i=1}^n (-1)^{s(i)} \cdot \prod_{i=1}^n a_{t(i)}$. But $\prod_{i=1}^n a_{t(i)} = \prod_{i=1}^n a_i$ by the preceding paragraph. Thus we have $a^n \prod_{i=1}^n a_i \equiv \prod_{i=1}^n (-1)^{s(i)} \prod_{i=1}^n a_i$, and since the product over the a_i is coprime to p, it may be canceled; this proves the claim. □

Let's apply this to give a proof for our conjecture that $(\frac{2}{p}) = (-1)^{(p^2-1)/8}$. We have to count the number of sign changes when we multiply the "half system" $A = \{1, 2, \ldots, \frac{p-1}{2}\}$ by 2. Assume first that $p = 4k+1$, i.e. $\frac{p-1}{2} = 2k$.

Then

$$\begin{aligned} 2\cdot 1 &\equiv 2 \bmod p,\\ 2\cdot 2 &\equiv 4 \bmod p,\\ &\dots\\ 2\cdot k &\equiv 2k \bmod p,\\ 2\cdot(k+1) &\equiv 2k+2 \equiv -2k+1 \bmod p,\\ 2\cdot(k+2) &\equiv -2k-1 \bmod p,\\ &\dots\\ 2\cdot 2k &\equiv -1 \bmod p. \end{aligned}$$

Thus there are no sign changes for the first k congruences, and there are sign changes for the last k congruences. This implies by Gauss's Lemma that $(\frac{2}{p}) = (-1)^k$.

Now it remains to check that k is even if and only if $\frac{p^2-1}{8}$ is, i.e., that $k \equiv \frac{p^2-1}{8} \bmod 2$. But this follows directly from $\frac{p^2-1}{8} = \frac18(p-1)(p+1) = \frac18\cdot 4k(4k+2) = k(2k+1)$.

Now assume that $p = 4k-1$; then there are no sign changes whenever $1 \le a \le k-1$, and there are exactly k sign changes for $k \le a < 2k$, so again we have $(\frac{2}{p}) = (-1)^k$. But now $\frac{p^2-1}{8} = \frac18(p-1)(p+1) = (2k-1)k$ shows that $k \equiv \frac{p^2-1}{2} \bmod 2$, and we have proved

Proposition 3.9. *The prime* 2 *is a quadratic residue of the odd prime p if and only if $p = 8k \pm 1$; in other words:* $(\frac{2}{p}) = (-1)^{(p^2-1)/8}$.

As a corollary, consider the Mersenne numbers M_q, where q is odd and $p = 2q+1$ is prime. If $q \equiv 3 \bmod 4$, then $p \equiv 7 \bmod 8$, hence $(2/p) = 1$. By Euler's criterion, this means that $2^q = 2^{(p-1)/2} \equiv 1 \bmod p$, and this in turn shows that $p \mid M_q$.

Corollary 3.10. *If* $p = 2q+1 \equiv 7 \bmod 8$ *is prime, then* $p \mid M_q$*, the q-th Mersenne number.*

In particular, $23 \mid M_{11}$ and $83 \mid M_{41}$. Thus some Mersenne numbers can be seen to be composite without applying the Lucas-Lehmer test. There are similar (but more complicated) rules for $p \mid M_q$ when $p = 4q+1$; in this case, we have to study $2^{(p-1)/4} \bmod p$, which leads us to quartic reciprocity. There is a quartic reciprocity law, but this cannot be formulated in $\mathbb{Z}$: Gauss realized in 1832[1] one has to enlarge $\mathbb{Z}$ to the ring $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}, i^2 = -1\}$ to do that.

3.3 The Quadratic Reciprocity Law

Here it comes:

Theorem 3.11. *For distinct odd primes p and q, we have*

$$\Big(\frac{p}{q}\Big)\Big(\frac{q}{p}\Big) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}.$$

Moreover, we have the first and the second supplementary law:

$$\Big(\frac{-1}{p}\Big)(-1)^{\frac{p-1}{2}}, \quad \Big(\frac{2}{p}\Big)(-1)^{\frac{p^2-1}{8}}.$$

What Theorem 3.11 says is that p is a square modulo q if and only if q is a square modulo p, except in the case where both p and q are $\equiv 3 \bmod 4$, when p is a square modulo q if and only if q is a nonsquare modulo p. This is a very surprising result, because at first sight the worlds $Z/p\mathbb{Z}$ and $\mathbb{Z}/q\mathbb{Z}$ seem totally different, and there is no apparent reason why they should be related at all. A preliminary version of the reciprocity law was discovered already around 1742 by Euler in his research on prime divisors of numbers of the form $a^n \pm b^n$ (like Mersenne or Fermat numbers), and Euler's final version was published 1785 (two years after his death). It was rediscovered by Legendre in 1788, who gave an incomplete proof. When Gauss rediscovered it at the age of 18, it took even him a whole year to find a proof (April 8, 1796); he found a simpler proof ten weeks later, but this proof used the theory of binary quadratic forms. The proof using Gauss's Lemma was his third published proof, and he gave eight different proofs altogether.

The following proof is particularly elegant and due to Eisenstein. It uses a variant of Gauss's Lemma. The idea is to extract the essential information form congruences like (3.1), namely the number of sign changes. Eisenstein noticed that the sine function is well suited for this job: given a congruence $a \equiv b \bmod m$, we see that $a = b + mr$ for some integer r, and applying $\sin \frac{2\pi}{m}$ to this equation we get $\sin 2\pi \frac{a}{m} = \sin 2\pi(\frac{b}{m} + r) = \sin 2\pi \frac{b}{m}$. In other

[1] Actually, around 1816; he was a bit slow in publishing results, if he published them at all.

words, $\mathbb{Z}$-periodic functions like the sine function are able to turn congruences $a \equiv b \bmod m$ into equalities $\sin 2\pi\frac{a}{m} = \sin 2\pi\frac{b}{m}$.

In the special case (3.1), we get

$$\begin{aligned}
\sin 2\pi\frac{2\cdot 1}{13} &= +\sin 2\pi\frac{2}{13},\\
\sin 2\pi\frac{2\cdot 2}{13} &= +\sin 2\pi\frac{4}{13},\\
\sin 2\pi\frac{2\cdot 3}{13} &= +\sin 2\pi\frac{6}{13},\\
\sin 2\pi\frac{2\cdot 4}{13} &= -\sin 2\pi\frac{5}{13},\\
\sin 2\pi\frac{2\cdot 5}{13} &= -\sin 2\pi\frac{3}{13},\\
\sin 2\pi\frac{2\cdot 6}{13} &= -\sin 2\pi\frac{1}{13}.
\end{aligned}$$

where we have used the fact that the sine is an odd function, i.e., that $\sin(-x) = -\sin x$. Multiplying gives

$$(-1)^3 = \prod_{a=1}^{6} \frac{\sin 2\pi\frac{2\cdot a}{13}}{\sin 2\pi\frac{a}{13}}.$$

In this way, we can express $(-1)^r$, where r is the number of sign changes, as a product of values of the sine function.

Here's the general version:

Lemma 3.12. *Let $p = 2n+1$ be an odd prime, and let $f : \mathbb{Q} \longrightarrow \mathbb{C}$ be a function with the following properties:*

i) $f(-z) = -f(z)$ for all $z \in \mathbb{Q} \setminus \mathbb{Z}$;

ii) $f(r) = f(r+z)$ for any $z \in \mathbb{Z}$;

iii) $f(\frac{a}{p}) \neq 0$ for all integers a not divisible by p.

Then

$$\Big(\frac{q}{p}\Big) = \prod_{a\in A} \frac{f(\frac{qa}{p})}{f(\frac{a}{p})},$$

where $A = \{1, 2, \ldots, \frac{p-1}{2}\}$.

Proof. Assume that we have $a_i q \equiv (-1)^{s(i)} a_{t(i)} \bmod p$ for $a_i, a_{t(i)} \in A$. Then $f(a_i q/p) = f((-1)^{s(i)} a_{t(i)}/p)$ by ii), and $f((-1)^{s(i)} a_{t(i)}/p) = (-1)^{s(i)} f(a_{t(i)}/p)$ by i), so $f(a_i q/p) = (-1)^{s(i)} f(a_{t(i)}/p)$. As a_i runs through A, so does $a_{t(i)}$ by the proof of Gauss's Lemma, so forming the product over all $a_i \in A$ gives

$$\prod_{i=1}^{n} f\Big(\frac{a_i q}{p}\Big) = \prod_{i=1}^{n} (-1)^{s(i)} f\Big(\frac{a_{t(i)}}{p}\Big) = \Big(\frac{q}{p}\Big) \prod_{i=1}^{n} f\Big(\frac{a_{t(i)}}{p}\Big) = \Big(\frac{q}{p}\Big) \prod_{i=1}^{n} f\Big(\frac{a_i}{p}\Big)$$

If you observe that $(\frac{q}{p}) = (\frac{q}{p})^{-1}$ since $(\frac{q}{p}) = \pm 1$, then this is exactly what was claimed. □

Two functions satisfying these properties are

- $f(x) = (-1)^{\lfloor 2x \rfloor}$;
- $f(x) = \sin 2\pi x$.

The proof we shall give uses the sine function.

Proposition 3.13. *Let $A = \{\alpha \in \mathbb{Z} \mid 1 \le \alpha \le \frac{p-1}{2}\}$ be a half-system modulo an odd prime p; then*

$$\Big(\frac{q}{p}\Big) = \prod_{\alpha \in A} \frac{\sin(\frac{2\pi}{p} q\alpha)}{\sin(\frac{2\pi}{p}\alpha)}. \tag{3.4}$$

We now claim that, for any odd integer q, we have

$$\frac{\sin qz}{\sin z} = P(\sin z), \tag{3.5}$$

where P is a polynomial with integral coefficients and leading coefficient $(-4)^{(q-1)/2}$.

This is done by induction; as a matter of fact, in order to be able to do induction we must prove a similar formula for the cosine simultaneously. For the induction step, we will need the addition formulas for trigonometric functions, so let us derive these first.

By Euler's formula, we know that

$$e^{it} = \cos t + i \sin t$$

for real numbers $t \in \mathbb{R}$. This implies that

$$e^{i(\alpha+\beta)} = \cos(\alpha+\beta) + i \sin(\alpha+\beta).$$

On the other hand, the functional equation of the exponential function gives

$$\begin{aligned} e^{i(\alpha+\beta)} &= e^{i\alpha}e^{i\beta} = (\cos\alpha + i\sin\alpha)(\cos\beta + i\sin\beta) \\ &= [\cos\alpha\cos\beta - \sin\alpha\sin\beta] + i[\cos\alpha\sin\beta + \cos\beta\sin\alpha] \end{aligned}$$

Comparing the real and the imaginary parts of the two expressions for $e^{i(\alpha+\beta)}$ immediately gives

$$\begin{aligned} \sin(\alpha+\beta) &= \sin\alpha\cos\beta + \sin\beta\cos\alpha \\ \cos(\alpha+\beta) &= \cos\alpha\cos\beta - \sin\alpha\sin\beta \end{aligned}$$

Back to our claims on $\sin qz$. By the addition formulas, we know that

$$\sin 2z = 2\sin z\cos z, \quad \cos 2z = \cos^2 z - \sin^2 z = 1 - 2\sin^2 z.$$

This in turn implies

$$\begin{aligned} \sin 3z &= \sin(z+2z) = \sin z\cos 2z + \sin 2z\cos z \\ &= \sin z(1-2\sin^2 z) + 2\sin z\cos^2 z \\ &= \sin z(1-2\sin^2 z) + 2\sin z(1-\sin^2 z) \\ &= \sin z(3-4\sin^2 z). \end{aligned}$$

A similar calculation for $\cos 3z$ gives

$$\begin{aligned} \cos 3z &= \cos(z+2z) = \cos z\cos 2z - \sin z\sin 2z \\ &= \cos z(1-2\sin^2 z) - 2\sin^2 z\cos z \\ &= \cos z(1-4\sin^2 z). \end{aligned}$$

Now we claim

Lemma 3.14. *For all odd integers $q \ge 1$, there exist polynomials $P, Q \in Z[X]$ of degree $q-1$ and with leading coefficient $(-4)^{(q-1)/2}$ such that*

$$\frac{\sin qz}{\sin z} = P(\sin z), \quad \frac{\cos qz}{\cos z} = Q(\sin z).$$

Proof. For $q = 1$, the claims are trivial, for $q = 3$ we have just proved them. So assume the assertions are correct for some odd integer q. Then

$$\begin{aligned} \sin(q+2)z &= \sin qz \cos 2z + \sin 2z \cos qz \\ &= \sin z P(\sin z)[1 - 2\sin^2 z] + 2 \sin z \cos^2 z Q(\sin z) \\ &= \sin z \Big[P(\sin z)(1 - 2\sin^2 z) + 2(1 - \sin^2 z)Q(\sin z)\Big] \end{aligned}$$

This implies that $\frac{\sin(q+2)z}{\sin z}$ is a polynomial in $\sin z$ of degree 2 greater than $\deg P = \deg Q$, and with leading coefficient $(-4)^{(q-1)/2}[-2-2] = (-4)^{(q+1)/2}$. This proves our claims. □

Now the zeros of $f(z) = \frac{\sin 2\pi qz}{\sin 2\pi z}$ are given by $\pm\frac{\beta}{2q}$, where β runs through the integers not divisible by q. If we substitute $X = \sin 2\pi z$, then $\frac{\sin 2\pi qz}{\sin 2\pi z} = P(X)$ is a polynomial of degree $q-1$ in X; clearly P has zeros $x = \sin 2\pi(\pm\frac{\beta}{q})$ with β as above, but these will not all be different. In fact, since $\sin 2\pi z$ is $\mathbb{Z}$-periodic, we see that only the values $\pm\frac{\beta}{q}$ with $1 \le \beta \le \frac{q-1}{2}$ give rise to different zeros; but since we have just found $q - 1 = \deg P$ zeros, there can't be any others.

Now we use

Proposition 3.15. *Let f and g be monic (leading coefficient $= 1$) polynomials of degree n with coefficients in some field F. If f and g have n different zeros in common, then they are equal.*

Proof. Clearly $f - g$ is a polynomial of degree $< n$ (because the term x^n cancels), and the common zeros of f and g are zeros of $f - g$. But over fields, nonzero polynomials can have at most as many roots as the degree indicates, so $f - g$ can have no n distinct roots unless it is the zero polynomial, i.e., unless $f = g$. □

Now $P(X)$ is a polynomial with leading coefficient $(-4)^{(q-1)/2}$, of degree $q - 1$, and with roots $\pm \sin 2\pi\frac{\beta}{q}$, where $1 \le \beta \le \frac{q-1}{2}$, and so is

$$f(X) = (-4)^{(q-1)/2} \prod_{\beta=1}^{(q-1)/2} \Big(X^2 - \sin^2 2\pi\frac{\beta}{q}\Big).$$

Thus $P(X) = f(X)$ by Proposition 3.15. Plugging $X = \sin 2\pi z$ back in we get

$$\frac{\sin 2\pi qz}{\sin 2\pi z} = (-4)^{\frac{q-1}{2}} \prod_{\beta\in B} \Big(\sin^2 2\pi z - \sin^2 2\pi \frac{\beta}{q}\Big). \tag{3.6}$$

Now we can reap the harvest of our work: let A and B denote half-systems mod p and q, respectively. We put $z = \frac{\alpha}{p}$ and find

$$\begin{aligned} \Big(\frac{q}{p}\Big) &= \prod_{\alpha\in A} \frac{\sin 2\pi(\frac{q\alpha}{p})}{\sin 2\pi(\frac{\alpha}{p})} = \prod_{\alpha\in A}(-4)^{\frac{q-1}{2}} \prod_{\beta\in B} \Big(\sin^2 2\pi\tfrac{\alpha}{p} - \sin^2 2\pi\tfrac{\beta}{q}\Big) \\ &= (-4)^{\frac{p-1}{2}\frac{q-1}{2}} \prod_{\alpha\in A} \prod_{\beta\in B} \Big(\sin^2 2\pi\tfrac{\alpha}{p} - \sin^2 2\pi\tfrac{\beta}{q}\Big). \end{aligned} \tag{3.7}$$

Exchanging p and q on the right hand side of (3.7) gives rise to a factor $(-1)^{\frac{p-1}{2}\frac{q-1}{2}}$; hence $\big(\frac{q}{p}\big) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}\big(\frac{p}{q}\big)$, which is the quadratic reciprocity law.

Let me explain parts of the proof by going through the example $q = 3$. We have seen that $\frac{\sin 3z}{\sin z} = P(X)$, where $X = \sin z$ and $P(X) = 3 - 4X^2$. In order to find the zeros of $P(X)$, which is hard if the degree $q - 1$ of P is larger than 2, we look at the zeros of the left hand side. Here it is clear that $\pm\frac{\pi}{3}$ are zeros. On the other hand, $\frac{\sin 3z}{\sin z} = P(\sin z)$ for $P(X)$ with $X = \sin z$, where $P(X)$ is a polynomial of degree 2 with leading coefficient -4. Since the left hand side vanishes for $z = \pm\frac{\pi}{3}$, so does the right hand side. Therefore, $P(X)$ has zeros $x = \sin(\pm\frac{\pi}{3}) = \pm\sin\frac{\pi}{3}$. Since these zeros are different, since $P(X)$ can have at most two zeros, and since it has leading coefficient -4, we conclude that $P(X) = -4(X - \sin\frac{\pi}{3})(X + \sin\frac{\pi}{3})$.

In fact, $P(X) = 3 - 4X^2 = -4(X - \frac{1}{2}\sqrt{3}\,)(X + \frac{1}{2}\sqrt{3}\,)$. Comparing both sides shows that $\sin\frac{\pi}{3} = \frac{1}{2}\sqrt{3}$.

Some Applications

We have already seen that 2 is a quadratic residue modulo an odd prime p if and only if $p \equiv \pm 1 \bmod 8$. Can we find a similar rule for 3? As a matter of fact, we can; since the corresponding result for -3 is even simpler, let's start with that one:

Corollary 3.16. *The integer -3 is a quadratic residue modulo primes $p \neq 3$ if and only if $p \equiv 1 \bmod 3$.*

For example, $-3 \equiv 4 \equiv 2^2 \bmod 7$ and $-3 \equiv 6^2 \bmod 13$, but $-3 \equiv 2 \bmod 5$ is a quadratic nonresidue modulo 5.

Proof. By the quadratic reciprocity law, we have

$$\begin{aligned} \Big(\frac{-3}{p}\Big) &= \Big(\frac{-1}{p}\Big)\Big(\frac{3}{p}\Big) = (-1)^{(p-1)/2}\Big(\frac{p}{3}\Big)(-1)^{(3-1)(p-1)/4} \\ &= \Big(\frac{p}{3}\Big) = \begin{cases} +1 & \text{if } p \equiv 1 \bmod 3, \\ -1 & \text{if } p \equiv 2 \bmod 3. \end{cases} \end{aligned}$$

This proves our claim. □

Now consider the integers $S_q = 2^{2q} + 1$ for odd numbers q. Since $2q \equiv 2 \bmod 4$ and $2^4 \equiv 1 \bmod 5$, we find $S_q \equiv 2^2 + 1 \equiv 0 \bmod 5$.

Now $S_q = A_qB_q$, where $A_q = 2^q - 2^{(q+1)/2} + 1$ and $B_q = 2^q + 2^{(q+1)/2} + 1$. If $q = 8k+1$, then $B_q = 2^{8k+1} + 2^{4k+1} + 1 \equiv 2 + 2 + 1 \equiv 0 \bmod 5$, hence $5 \mid B_q$ in that case, while $A_q \equiv 2 - 2 + 1 \equiv 1 \bmod 5$. If $q = 8k+3$, then $A_q = 2^{8k+3} - 2^{4k+2} + 1 \equiv 8 - 4 + 1 \equiv 0 \bmod 5$, and proceeding this way we find

Corollary 3.17. *With* $S_q = A_qB_q$ *as above, we have* $5 \mid B_q$ *if and only if* $q \equiv \pm 1 \bmod 8$, *and* $5 \mid A_q$ *otherwise.*

Finally, consider Fibonacci numbers. Binet's formula says that

$$F_n = \frac{1}{\sqrt{5}}(\alpha^n - \beta^n),$$

where $\alpha = \frac{1}{2}(1+\sqrt{5}\,)$ and $\beta = \frac{1}{2}(1-\sqrt{5}\,)$.

Let $p \neq 5$ be a prime; we want to compute $F_p \bmod p$. Using the binomial expansion we find

$$(1+\sqrt{5}\,)^p = 1 + \binom{p}{1}\sqrt{5} + \binom{p}{2}\cdot 5 + \ldots + \binom{p}{p-1}\sqrt{5}^{p-1} + \sqrt{5}^p,$$

hence

$$(1+\sqrt{5}\,)^p - (1-\sqrt{5}\,)^p = 2\sqrt{5}\left\{\binom{p}{1} + 5\binom{p}{3} + \ldots + 5^{\frac{p-1}{2}}\binom{p}{p}\right\}.$$

Dividing through by 2^p, we see that

$$F_p = 2^{1-p}\left\{ \binom{p}{1} + 5\binom{p}{3} + \ldots + 5^{(p-2)/2}\binom{p}{p} \right\}.$$

Now $2^{p-1} \equiv 1 \bmod p$ by Fermat's Little Theorem, moreover $\binom{p}{1} \equiv \ldots \equiv \binom{p}{p-2} \equiv 0 \bmod p$, hence

$$F_p \equiv 5^{(p-1)/2} \equiv \Big(\frac{5}{p}\Big) \bmod p.$$

Now we do a similar computation for F_{p+1}. As above, we find

$$(1+\sqrt{5}\,)^{p+1}-(1-\sqrt{5}\,)^{p+1} = 2\sqrt{5}\left\{\binom{p+1}{1}+5\binom{p+1}{3}+\ldots+5^{\frac{p-1}{2}}\binom{p+1}{p}\right\}.$$

It is immediate from Pascal's triangle that $p \mid \binom{p+1}{k}$ for all $2 \le k \le p-1$; moreover, $\binom{p+1}{1} = \binom{p+1}{p} = p+1 \equiv 1 \bmod p$. Dividing through by 2^{p+1} gives

$$F_{p+1} \equiv \tfrac{1}{2}\big(1+5^{(p-1)/2}\big) \;\equiv\; \tfrac{1}{2}\big(1+(\tfrac{5}{p})\big) \equiv \begin{cases} 0 & \text{if } p \equiv \pm 1 \bmod 5, \\ 1 & \text{if } p \equiv \pm 2 \bmod 5. \end{cases}$$

Since $F_{p-1} = F_{p+1} - F_p$, we have proved

Corollary 3.18. *For primes $p \ne 5$, we have $p \mid F_{p-1}$ if $p \equiv \pm 1 \bmod 5$, and $p \mid F_{p+1}$ if $p \equiv \pm 2 \bmod 5$. More exactly, the following congruences hold:*

p	F_{p-1}	F_p	F_{p+1}
$\pm 1 \bmod 5$	$0 \bmod p$	$1 \bmod p$	$1 \bmod p$
$\pm 2 \bmod 5$	$1 \bmod p$	$-1 \bmod p$	$0 \bmod p$

Second Proof of Quadratic Reciprocity

The original version of Gauss's Lemma is different from the version we gave above. Consider an odd prime $p = 2m+1$ and take the half-system $A = \{1, 2, \ldots, m\}$. For finding $(\frac{a}{p})$, we multiplied every element of A by a and reduced the product modulo p such that the remainder has minimal absolute value, i.e., to an element in A up to sign. Now we choose the residues from the elements $1, 2, \ldots, p-1$. If there are exactly r negative remainders in the original versions, then there are exactly r remainders in the interval $[m+1, 2m]$. In other words:

Lemma 3.19 (Gauss's Lemma). *Let $p = 2m+1$ be an odd prime, a an integer not divisible by p, and $A = \{1, 2, \ldots, m\}$ a half-system modulo p. Write*

$$a \cdot i = pq_i + r_i, \quad 1 \le r_i \le p-1, \tag{3.8}$$

for $1 \le i \le m$. Then $(\frac{a}{p}) = (-1)^r$, where r is the number of residues r_i that are $> \frac{p}{2}$.

Now let's look at $a \cdot i = pq_i + r_i$; we clearly have $pq_i = ai - r_i$ with $0 < a_i < p$, hence $q_i = \lfloor \frac{ai}{p} \rfloor$. If we sum up all the n equations in (3.8), we therefore get

$$a \sum_{i=1}^{m} i = p \sum_{i=1}^{m} \left\lfloor \frac{ai}{p} \right\rfloor + \sum_{i=1}^{m} r_i.$$

What can we say about the r_i? We know that exactly r of them are from the interval $[m+1, 2m]$, hence are equal to $p - a_j$ for some a_j, while the other $m - r$ residues are elements from the half system A. Thus $r_i \equiv 1 + a_j \bmod 2$ for r of the equations, and $r_i \equiv a_j \bmod 2$ for the other $n - r$ equations. This implies $r_1 + \ldots + r_m \equiv r + a_1 + \ldots a_m \bmod 2$, and we get

$$\sum_{i=1}^{m} \left\lfloor \frac{ai}{p} \right\rfloor \equiv p \sum_{i=1}^{m} \left\lfloor \frac{ai}{p} \right\rfloor = a \sum_{i=1}^{m} ia_i - \sum_{i=1}^{m} r_i \equiv r \bmod 2$$

assuming that a is odd. Using Gauss's Lemma, we deduce

Proposition 3.20. *For odd integers a and odd primes $p = 2m+1$ with $p \nmid a$ we have*

$$\left(\frac{a}{p}\right) = (-1)^r, \quad \textit{where } r = \sum_{i=1}^{m} \left\lfloor \frac{ai}{p} \right\rfloor.$$

In particular, if $q = 2n+1$ is a prime different from p, then we have

$$\begin{aligned} \left(\frac{q}{p}\right) &= (-1)^r, \quad \text{where } r = \sum_{i=1}^{m} \left\lfloor \frac{qi}{p} \right\rfloor, \\ \left(\frac{p}{q}\right) &= (-1)^s, \quad \text{where } s = \sum_{i=1}^{n} \left\lfloor \frac{pi}{q} \right\rfloor \end{aligned}$$

The quadratic reciprocity theorem therefore boils down to the statement that

$$\sum_{i=1}^{m} \left\lfloor \frac{qi}{p} \right\rfloor + \sum_{i=1}^{n} \left\lfloor \frac{pi}{q} \right\rfloor \equiv \frac{p-1}{2} \cdot \frac{q-1}{2} \bmod 2.$$

But this follows immediately from Eisenstein's observation that

$$\sum_{i=1}^{m} \left\lfloor \frac{qi}{p} \right\rfloor + \sum_{i=1}^{n} \left\lfloor \frac{pi}{q} \right\rfloor$$

is the number of lattice points inside the rectangle R with corners $(1,1)$ and $(\frac{p-1}{2}, \frac{q-1}{2})$.

In fact, consider the line L through the origin and (p,q), that is, with the equation $y = \frac{q}{p}x$. There is no lattice point (a point with integral coordinates) on L between $x = 0$ and $x = p$: in fact, if (r,s) were such a point, then $s = \frac{q}{p}r$, that is, $\frac{s}{r} = \frac{q}{p}$ with $0 < r < p$. But the fraction $\frac{q}{p}$ is in its lowest terms since p and q are different primes. The number of lattice points inside the rectangle R with x-coordinate $x = i$ are $(i,1)$, $(i,2)$, ... , $(i, \lfloor \frac{qi}{p} \rfloor)$. This means that

$$\sum_{i=1}^{m} \left\lfloor \frac{qi}{p} \right\rfloor$$

is the number of lattice points inside R below L. By the same reasoning (as can be seen by switching the x- and y-axis),

$$\sum_{i=1}^{n} \left\lfloor \frac{pi}{q} \right\rfloor$$

is the number of lattice points inside R and above the line L.

3.4 The Jacobi Symbol

The reciprocity law for the Legendre symbol is an amazing piece of insight; for computing Legendre symbols, it is less suited. The reason is simple: before we can invert a symbol (n/p), we have to find the prime factorization of n. Here's an example: suppose you want to compute $(39/59)$; then $39 = 3 \cdot 13$, so $(39/59) = (3/59)(13/59) = -(59/3)(59/13)$ bye the quadratic reciprocity law, hence $(39/59) = -(2/3)(7/13) = (7/13)$ by the second supplementary law, so $(39/59) = (13/7) = (-1/7) = -1$.

Now we know that finding the prime factorization of an integer n isn't much fun if n is big. Fortunately, there's a better way: the reciprocity law for the Jacobi symbol. The trick is simple: invert the Legendre symbols as if the composites that occur were primes. In our example, $(39/59) = -(59/39) = -(20/39) = -(5/39) = -(39/5) = -1$.

Why does this work? Well, for a start we have do define what a symbol like $(59/39)$ should mean. This is easy: assume that n is an odd positive integer with prime factorization $= p_1 \cdots p_r$; then we put $(m/n) := (m/p_1) \cdots (m/p_r)$, where the symbols on the right hand side are Legendre symbols; (m/n) is called the Jacobi symbol. Now we claim

Theorem 3.21 (Reciprocity Law for Jacobi Symbols). *If m and n are coprime positive odd integers, then*

$$\Big(\frac{m}{n}\Big)\Big(\frac{n}{m}\Big) = (-1)^{\frac{m-1}{2}\frac{n-1}{2}}.$$

Moreover, we have the supplementary laws

$$\Big(\frac{-1}{n}\Big) = (-1)^{\frac{n-1}{2}}, \quad \Big(\frac{2}{n}\Big) = (-1)^{\frac{n^2-1}{8}}.$$

Thus the quadratic reciprocity law holds for Jacobi symbols! There are two possible approaches to a proof: either we redo our proof of the reciprocity law for the Legendre symbols (the only problem is generalizing Gauss's Lemma to composite values of m), or we reduce the reciprocity law for Jacobi symbols to the reciprocity law for Legendre symbols. We will do the latter here.

Proof. Let us start with the first supplementary law. Write $n = p_1 \cdots p_r$; then

$$\Big(\frac{-1}{n}\Big) = \Big(\frac{-1}{p_1}\Big) \cdots \Big(\frac{-1}{p_r}\Big) = (-1)^{\frac{p_1-1}{2}+\ldots+\frac{p_r-1}{2}}.$$

Thus it remains to show that

$$\frac{n-1}{2} \equiv \frac{p_1-1}{2} + \ldots + \frac{p_r-1}{2} \bmod 2. \tag{3.9}$$

This is done by induction. We start with the observation that $(a-1)(b-1) \equiv 0 \bmod 4$ for odd integers a, b, hence $ab - 1 \equiv (a-1) + (b-1) \bmod 4$, and dividing by 2 gives

$$\frac{ab-1}{2} \equiv \frac{a-1}{2} + \frac{b-1}{2} \bmod 2.$$

Now use induction.

Now let us treat the reciprocity law similarly. Write $m = p_1 \cdots p_r$ and $n = q_1 \cdots q_s$; then

$$\Big(\frac{m}{n}\Big)\Big(\frac{n}{m}\Big) = \prod_{i=1}^{r}\prod_{j=1}^{s}\Big(\frac{p_i}{q_j}\Big)\Big(\frac{q_j}{p_i}\Big) = \prod_{i=1}^{r}\prod_{j=1}^{s}(-1)^{(p_i-1)(q_j-1)/4},$$

and our claim will follow if we can prove that

$$\frac{m-1}{2}\frac{n-1}{2} \equiv \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{p_i-1}{2}\frac{q_j-1}{2} \bmod 4.$$

But this follows by multiplying the two congruences you get by applying (3.9) to m and n.

Finally, consider the second supplementary law. Similar to the above, everything boils down to showing

$$\frac{n^2-1}{8} \equiv \frac{p_1^2-1}{8} + \ldots + \frac{p_r^2-1}{8} \bmod 2.$$

Now clearly $16 \mid (a^2-1)(b^2-1)$ (as a matter of fact, even this product is even divisible by 64), hence

$$(ab)^2 - 1 \equiv a^2 - 1 + b^2 - 1 \bmod 16.$$

Now induction does the rest. □

Chapter 4

Conics

4.1 Rational Points on Conics

Conic sections (or conics for short) have been studied since antiquity:

Appolonius[1] even wrote what we could call a textbook on conics consisting of eight volumes. Much later, Copernicus[2] discovered that the orbits of planets are ellipses, and later Newton realized that the possible paths of an object in our solar system are exactly the conics: ellipses, parabolas and hyperbolas.

These objects can be described by equations: $ax^2+by^2=r^2$ with $a,b>0$ describes an ellipse (the special case $a=b=1$ gives a circle with radius r), while the case $a>0,b<0$ leads to hyperbolas. Finally, equations like $y=ax^2+bx+c$ describe parabolas as long as $a\neq 0$.

In this chapter we study conics not over the reals (this would be geometry) but over residue class rings. The unit circle over $\mathbb{Z}/5\mathbb{Z}$, for example, is the set of points (x,y) with $x,y\in\mathbb{Z}/5\mathbb{Z}$ such that $x^2+y^2\equiv 1 \bmod 5$, namely $\{(0,\pm 1),(\pm 1,0)\}$. If we plot this set, the result doesn't look at all like the circles we know, so we have to be careful when using geometric intuition for studying objects like these.

[1] Appolonius, ca. 262 BC in Perga (Greek Ionia, now Turkey) –ca. 190 BC (Alexandria, Egypt).

[2] Nicolaus Copernicus, 1473 (Torun, Poland) – 1543 (Frombork, Poland).

Pythagorean Triples Revisited

Recall that a Pythagorean Triple consists of three nonzero integers (a,b,c) with $a^2+b^2=c^2$; a triple is called primitive if these integers are coprime. Dividing through by c^2 and putting $x=a/c$, $y=b/c$ we see that primitive Pythagorean triples correspond to rational points on the unit circle $x^2+y^2=1$, that is, to elements of the set

$$\mathcal{C}(\mathbb{Q})=\{(x,y): x,y\in\mathbb{Q},\ x^2+y^2=1\}.$$

There are numerous methods for finding all rational points on the unit circle: in his lectures on number theory,[3] Kronecker gave two methods[4] for deriving the formulas for Pythagorean triples: we have already seen the arithmetic proof using the Unique Factorization Theorem of the integers, and now we will present an analytic proof based on the parametrization of $\mathcal{C}$ by trigonometric function.

It uses the fact that $x=\cos\alpha$, $y=\sin\alpha$ is a parametrization of

$$\mathcal{C}(\mathbb{R})=\{(x,y): x,y\in\mathbb{R}, x^2+y^2=1\}$$

by trigonometric functions. Using the identities $\cos^2\alpha-\sin^2\alpha=\cos 2\alpha$ and $\cos^2\alpha+\sin^2\alpha=1$ we get

$$\begin{aligned} x &= \cos\alpha = \frac{\cos^2\frac{\alpha}{2}-\sin^2\frac{\alpha}{2}}{\cos^2\frac{\alpha}{2}+\sin^2\frac{\alpha}{2}} = \frac{1-m^2}{1+m^2}, \\ y &= \sin\alpha = \frac{2\sin\frac{\alpha}{2}\cos\frac{\alpha}{2}}{\cos^2\frac{\alpha}{2}+\sin^2\frac{\alpha}{2}} = \frac{2m}{1+m^2}, \end{aligned}$$

where we have put $m=\tan\frac{\alpha}{2}$. If we choose m to be a rational number, we get rational points on $\mathcal{C}$. Conversely, if x and y are rational and $y\neq 0$, then $m=\frac{1-x}{y}$ is rational, too. Thus this parametrization gives us all rational points $\neq(-1,0)$ on $\mathcal{C}$.

One of the best known derivations of the formulas for Pythagorean triples is the parametrization of rational points on $\mathcal{C}$ via the technique of sweeping lines: pick any rational point on $\mathcal{C}$, say $P=(-1,0)$, and consider the lines l

[3] L. Kronecker, *Vorlesungen über Zahlentheorie*, (K. Hensel, ed.), reprint Springer-Verlag 1978

[4] T. Ono, *Variations on a theme of Euler*, Plenum Press, New York, 1994 even gives five different methods.

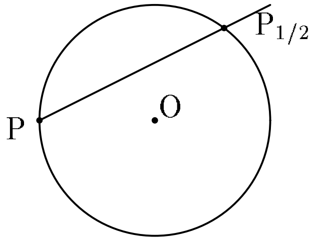

Figure 4.1: Unit Circle and Line with slope 1/2

through P with rational slope m; the line l is given by $y = m(x+1)$, and it intersects $\mathcal{C}$ in the second point $P_m = (\frac{1-m^2}{1+m^2}, \frac{2m}{1+m^2})$. Conversely, any rational point $Q \neq (-1,0)$ defines a line PQ with rational slope, hence $m \longmapsto P_m$ is a bijection between the set $\mathbb{Q}$ of rational numbers and the set $\mathcal{C}(\mathbb{Q}) \setminus \{P\}$ of rational points on $\mathcal{C}$ different from P.

As far as the algebra behind this derivation is concerned, one might say that this technique was known already to Diophantus. Since analytic geometry had not been invented then, he could not possibly have realized the geometric interpretation of his technique: this was left for Newton to do, who, however, chose not to publish this valuable piece of insight, leaving it to be rediscovered by several mathematicians during the second half of the 19th century.

The Unit Circle over Arbitrary Rings

The definition of $\mathcal{C}(\mathbb{Q})$ can be generalized to arbitrary rings R (all our rings are commutative and have a unit element 1): we put

$$\mathcal{C}(R) = \{(x,y) : x, y \in R,\ x^2 + y^2 = 1\}.$$

If R is a finite ring, then $\mathcal{C}(R)$ is also finite, and it is a natural problem to determine its cardinality. Let us consider this problem for the rings $R = \mathbb{Z}/n\mathbb{Z}$. One possible line of attack consists in adapting the method of finding all points on $\mathcal{C}(\mathbb{Q})$ that we discussed in Section 4.1. When working over $\mathbb{Z}/n\mathbb{Z}$, however, there are a number of problems to overcome: the line through $P = (-1,0)$ with slope $m \in \mathbb{Z}/n\mathbb{Z}$ intersects $\mathcal{C}(\mathbb{Z}/n\mathbb{Z})$ in the second point $P_m = (\frac{1-m^2}{1+m^2}, \frac{2m}{1+m^2})$ unless $\gcd(m^2+1, n) \neq 1$. Conversely, if $Q = (x,y)$ is any point $\neq P$ on $\mathcal{C}(\mathbb{Z}/n\mathbb{Z})$, then $Q = P_m$ for some $m = \frac{y}{x+1} \in \mathbb{Z}/n\mathbb{Z}$ unless $\gcd(x+1, n) \neq 1$.

Thus let us assume for the sake of simplicity that p is prime; since $\mathcal{C}(\mathbb{Z}/2\mathbb{Z}) = \{(1,0),(0,1)\}$ we may assume that p is odd. If $p \equiv 3 \bmod 4$, then $\gcd(m^2+1,p) \neq 1$ for all m, hence each $m \in \mathbb{Z}/p\mathbb{Z}$ gives a point on $\mathcal{C}(\mathbb{Z}/n\mathbb{Z})$. We now claim that two different m give two different points. In fact, assume that $P_m = P_n$; equating the x-coordinates, multiplying through by $(1-n^2)(1-m^2)$ and simplifying we find $n^2 \equiv m^2 \bmod p$, i.e. $m \equiv \pm n \bmod p$; plugging this into the equation $\frac{2m}{1+m^2} \equiv \frac{2n}{1+n^2} \bmod p$ gives $m \equiv n \bmod p$. The points $(x,y) \in \mathcal{C}(\mathbb{F}_p)$ that we don't get are exactly those with $(x+1,p) \neq 1$, that is, with $x \equiv -1 \bmod p$; since this implies $(x,y) = (-1,0)$, we have a bijection between $\mathbb{Z}/p\mathbb{Z}$ and $\mathcal{C}(\mathbb{F}_p) \setminus \{(-1,0)\}$: in particular, $\#\mathcal{C}(\mathbb{F}_p) = p+1$.

If $p \equiv 1 \bmod 4$, on the other hand, then $m^2+1 \equiv 0 \bmod p$ has exactly two solutions, which implies as before that $\#\mathcal{C}(\mathbb{F}_p) = p-1$. We have shown

Proposition 4.1. *For odd primes p, there are $p - (\frac{-1}{p})$ points on $\mathcal{C}(\mathbb{Z}/p\mathbb{Z})$.*

A similar analysis for composite m is more complicated: you might try your hands at the simplest case where $m = pq$ is the product of two different primes. If you count correctly, your answer will be $\#\mathcal{C}(\mathbb{Z}/pq\mathbb{Z}) = \#\mathcal{C}(\mathbb{Z}/p\mathbb{Z}) \cdot \#\mathcal{C}(\mathbb{Z}/q\mathbb{Z})$. This in turn suggests that there may be an algebraic explanation, and in fact there is: see Corollary 4.5.

Proposition 4.1 implies the second supplementary law of quadratic reciprocity: in fact, the solutions of the congruence $x^2+y^2 \equiv 1 \bmod p$ come in octuples $(\pm x, \pm y)$, $(\pm y, \pm x)$, except for the quadruple $(0,\pm 1)$, $(\pm 1, 0)$ and the quadruple $(\pm r, \pm r)$ with $2r^2 \equiv 1 \bmod p$ which exists if and only if 2 is a quadratic residue modulo p, i.e. iff $(2/p) = 1$. This shows that

$$\#\mathcal{C}(\mathbb{Z}/p\mathbb{Z}) \equiv \begin{cases} 4 \bmod 8 & \text{if } (\frac{2}{p}) = -1, \\ 0 \bmod 8 & \text{if } (\frac{2}{p}) = +1. \end{cases}$$

Comparing this with the explicit formula in Proposition 4.1 implies immediately that $(\frac{2}{p}) = +1$ if $p \equiv \pm 1 \bmod 8$ and $(\frac{2}{p}) = -1$ if $p \equiv \pm 3 \bmod 8$. Actually, the whole quadratic reciprocity law can be proved in a similar way.

4.2 The Group Law on Circles

The unit circle $\mathcal{C}(\mathbb{R})$ in the Euclidean plane is described as the set of points $(x,y) \in \mathbb{R} \times \mathbb{R}$ satisfying the equation $x^2+y^2 = 1$. We have seen in Section 4.1 that $\mathcal{C}(\mathbb{R})$ can be parametrized by trigonometric functions: the map

$$\lambda : \mathbb{R} \longrightarrow \mathcal{C}(\mathbb{R}) : \alpha \longmapsto (\cos 2\pi\alpha, \sin 2\pi\alpha)$$

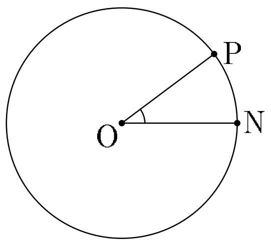

Figure 4.2: Unit Circle with Angle determined by two points

"covers" the unit circle, and it does so infinitely often since $\lambda(\alpha+n)=\lambda(\alpha)$ for any integer $n\in\mathbb{Z}$. In fact, λ induces a bijection $\mu:\mathbb{R}/\mathbb{Z}\longrightarrow\mathcal{C}(\mathbb{R})$.

Now observe that the object $\mathbb{R}/\mathbb{Z}$ on the left hand side carries the structure of an additive abelian group: the sum of $\alpha_1+\mathbb{Z}$ and $\alpha_2+\mathbb{Z}$ is $(\alpha_1+\alpha_2)+\mathbb{Z}$. Using the bijection μ, we can make $\mathcal{C}(\mathbb{R})$ into an abelian group as follows: to add two points P_1 and P_2 on the circle, find their inverse images $\alpha_1=\mu^{-1}(P_1)$ and $\alpha_2=\mu^{-1}(P_2)$, and put $P_1+P_2=\mu(\alpha_1+\alpha_2)$. This is of course not very exciting: each point P defines an angle $\measuredangle NOP$, where $O=(0,0)$ and $N=(1,0)$, and adding points on the circle amounts to adding their angles.

Nevertheless, let us compute the addition law explicitly: for $j=1,2$ write $P_j=(x_j,y_j)$ and find $\alpha_j\in\mathbb{R}$ (defined modulo $\mathbb{Z}$) with $x_j=\cos 2\pi\alpha_j$ and $y_j=\sin 2\pi\alpha_j$. The addition formulas give

$$\begin{aligned}\cos 2\pi(\alpha_1+\alpha_2) &= \cos 2\pi\alpha_1\,\cos 2\pi\alpha_2-\sin 2\pi\alpha_1\,\sin 2\pi\alpha_2,\\ \sin 2\pi(\alpha_1+\alpha_2) &= \cos 2\pi\alpha_1\,\sin 2\pi\alpha_2+\cos 2\pi\alpha_2\,\sin 2\pi\alpha_1,\end{aligned}$$

hence

$$(x_1,y_1)+(x_2,y_2)=(x_1x_2-y_1y_2,x_1y_2+x_2y_1). \tag{4.1}$$

The miracle here is that these formulas are polynomials in the coordinates x_i,y_j. What this means is that, given any commutative ring R with 1, we can define

$$\mathcal{C}(R)=\{(x,y)\in R\times R: x^2+y^2=1\}$$

as the set of "R-rational points" on $\mathcal{C}$ and make $\mathcal{C}(R)$ into a group via (4.1). It can be easily checked that the neutral element of $\mathcal{C}(R)$ is $(1,0)$ (as expected, since $(1,0)=\mu(0+\mathbb{Z})$ is the neutral element in the case $R=\mathbb{R}$). Similarly, $-(x,y)=(x,-y)$ since $(x,y)+(x,-y)=(1,0)$. Finally, the associativity follows by a brute force calculation: one has to show that $[(x_1,y_1)+(x_2,y_2)]+(x_3,y_3) = (x_1,y_1)+[(x_2,y_2)+(x_3,y_3)]$, and in fact both sides are equal to (x,y) with $x = x_1x_2x_3 - x_1y_2y_3 - x_2y_1y_3 - x_3y_1y_2$ and $y = x_1x_2y_3 + x_1x_3y_2 + x_2x_3y_1 - y_1y_2y_3$. Another method is to notice that associativity for $R = \mathbb{R}$ implies a polynomial identity in $\mathbb{Z}[x_1,\ldots,y_3]$ which then is a fortiori correct in any commutative ring with 1.

Theorem 4.2. *Let R be a ring with identity 1. Then the formula (4.1) defines a group law on $\mathcal{C}(R) = \{(x,y) \in R^2 : x^2+y^2 = 1\}$. The neutral element is $(1,0)$, and $-(x,y) = (x,-y)$.*

Thus the unit circle $\mathcal{C}$ is a machine that eats rings with 1 and spits out groups; machines such as $\mathcal{C}$ are quite common in mathematics: another well known example is GL_n, which turns a ring with 1 into the group of invertible $n \times n$-matrices with entries in R.

4.3 Factoring Integers with the Unit Circle

Suppose that we are given an integer N that we want to factor into primes. Assume moreover that we know a nontrivial point P on $\mathcal{C}(\mathbb{Z}/N\mathbb{Z})$ (nontrivial in the sense that its order isn't too small; in particular, the points $P = (\pm 1, 0), (0, \pm 1)$ won't do); for integers $N = n^2+3$, for example, we have the point $P = (n,2)$.

Assume for a moment that we already know a prime factor $p \mid N$; let us see what happens when we compute $kP = (x,y)$ for $k = p - (\frac{-1}{p})$. Since the group order k is a multiple of the order of P on $\mathcal{C}(\mathbb{Z}/p\mathbb{Z})$, we see that $x \equiv 1 \bmod p$ and $y \equiv 0 \bmod p$. Thus, unless accidentally $N \mid y$, we can recover a nontrivial factor of N from the coordinates of kP by computing $\gcd(y,N)$ or $\gcd(x-1,N)$

Now we get a factorization algorithm if we only notice that we can replace k by some multiple. If $k = p - (\frac{-1}{p})$ is composed of small prime factors, then it is easy to write down such a multiple M of k without knowing p: choose a bound B (say $B = 10^4, 10^5, 10^6 \ldots$) and form the product $M = \prod p^{a(p)}$, where $p^{a(p)}$ is the largest power of p smaller than B.

Here's a description of the algorithm for factoring integers $N = n^2+3$:

1. Pick a bound B; put $m = 0$, $p_m = 1$, $P_m = (n,2)$.

2. Let p_{m+1} be the smallest prime $> p_m$; if $p_{m+1} > B$, terminate. Otherwise choose $e \in \mathbb{N}$ maximal with $p_m^e < B$.

3. Compute $P_{m+1} = (x,y) := p_m^e P_m$; if $(y,N) = 1$, replace m by $m+1$ and goto step 2; if $(y,N) = N$, repeat the algorithm with a smaller bound B (or redo the computation of $p_m^e P_m$ but check whether $(y,N) \neq 1$ after each step); otherwise put out (y,N) as a factor.

The computation of $p^n P$ is of course not done by adding P sufficiently often to itself but by the method of duplication and addition (squaring and multiplying in the multiplicative language). Here's a simple example: take $N = 56^2 + 3 = 3139$, $P_0 = (56,2)$ and $B = 10$. Our first prime is $p = 2$, and $2^3 = 8$ is the smallest power < 10; we get $2P = (-7,224)$, $4P = (97,3)$ and $P_1 = 8P = (-17,582)$, and since $(582,N) = 1$ we continue with $p = 3$. Here we have to compute $P_2 = 9P_1$, and this is done by doubling P_1 three times and adding P_1: $2P_1 = (577,-954)$, $4P_1 = (389,873)$, $8P_1 = (1297,1170)$, $9P_1 = (1520,438)$, and now $\gcd(438,N) = 73$, hence $N = 73 \cdot 43$. Note that we cannot expect to find the second factor 43 with this method and $B = 10$ since $43+1 = 4 \cdot 11$ has a prime factor larger than B. We did find 73 on the other hand since $73 - 1 = 2^3 3^2$ is a product of prime powers $< B$. In fact, P has order 9 on $\mathcal{C}(\mathbb{Z}/73\mathbb{Z})$, so we would have found it by simply computing $9P$. Check this!

The main problem when working with the group law on the unit circle is finding a nontrivial $\mathbb{Z}/N\mathbb{Z}$-rational point on it; this problem is overcome by replacing the unit circle with a conic of the form $ax^2 + y^2 = 1$, choosing integers x, y at random and then taking $a \equiv (1-y^2)/x^2 \bmod N$. But in order to do so we need a group law on conics: this will be discussed below. The only implementation of this method that I am aware of is due to Zhang.[5]

The Structure of $\mathcal{C}(\mathbb{Z}/n\mathbb{Z})$

What do the groups $\mathcal{C}(R)$ look like? Let us start with rings $R = \mathbb{Z}/n\mathbb{Z}$; computing $\mathcal{C}(\mathbb{Z}/n\mathbb{Z})$ for odd $n \leq 15$ yields the results presented in the following table:

For every $m \geq 3$, the point $(0,1)$ generates a cyclic subgroup of order 4 (this is "obvious": add the angles; observe that $2(0,1) = (-1,0) = (1,0)$ in

[5] M. Zhang, *Factoring integers with conics*, J. Sichuan Univ., Nat. Sci. Ed. **33**, No.4 (1996), 356–359.

n	$\mathcal{C}(\mathbb{Z}/n\mathbb{Z})$	structure
2	$(0,1),(1,0)$	$\mathbb{Z}/2\mathbb{Z}$
3	$(0,\pm1),(\pm1,0)$	$\mathbb{Z}/4\mathbb{Z}$
5	$(0,\pm1),(\pm1,0)$	$\mathbb{Z}/4\mathbb{Z}$
7	$(0,\pm1),(\pm1,0),(\pm2,\pm2)$	$\mathbb{Z}/8\mathbb{Z}$
9	$(0,\pm1),(\pm1,0),(\pm1,\pm3),(\pm3,\pm1)$	$\mathbb{Z}/12\mathbb{Z}$
11	$(0,\pm1),(\pm1,0),(\pm3,\pm5),(\pm5,\pm3)$	$\mathbb{Z}/12\mathbb{Z}$
13	$(0,\pm1),(\pm1,0),(\pm2,\pm6),(\pm6,\pm2)$	$\mathbb{Z}/12\mathbb{Z}$
15	$(0,\pm1),(0,\pm4),(\pm1,0),(\pm4,0),(\pm5,\pm6),(\pm6,\pm5)$	$\mathbb{Z}/4\mathbb{Z}\oplus\mathbb{Z}/4\mathbb{Z}$

$\mathbb{Z}/2\mathbb{Z}$), hence $\mathcal{C}(\mathbb{Z}/9\mathbb{Z}) \simeq \mathbb{Z}/12\mathbb{Z}$ follows from the fact that $\#\mathcal{C}(\mathbb{Z}/9\mathbb{Z}) = 12$. The fact that $\mathcal{C}(\mathbb{Z}/15\mathbb{Z}) \simeq \mathcal{C}(\mathbb{Z}/3\mathbb{Z}) \oplus \mathcal{C}(\mathbb{Z}/5\mathbb{Z})$ suggests the following result which, once conjectured, is immediately verified:

Proposition 4.3. *Let R and S be rings with 1, and let $\phi : R \longrightarrow S$ be a ring homomorphism. Then ϕ induces a group homomorphism $\phi_C : \mathcal{C}(R) \longrightarrow \mathcal{C}(S)$ defined by $(x,y) \longmapsto (\phi(x),\phi(y))$, and if ϕ is an isomorphism then so is ϕ_C.*

Recall that a ring homomorphism $\phi : R \longrightarrow S$ is called injective if $\phi^{-1}(0) = \{0\}$. Since ϕ is a ring homomorphism, we have $1 = \phi(1)$; if $\phi(a) = 1$, then $0 = \phi(a) - \phi(1) = \phi(a-1)$, so if ϕ is injective, then $\phi^{-1}(1) = \{1\}$.

Now we claim

Lemma 4.4. *If $\phi : R \longrightarrow S$ is an injective ring homomorphism, then so is $\phi_C : \mathcal{C}(R) \longrightarrow \mathcal{C}(S)$.*

Proof. Assume that $\phi(P) = (1,0)$ for $P = (x,y)$; then $\phi(x) = 1$ and $\phi(y) = 0$, hence $x = 1$ and $y = 0$ by what we said above. □

Warning. Observe that ϕ_C need not be surjective even if ϕ is: the homomorphism $\mathcal{C}(\mathbb{Z}) \longrightarrow \mathcal{C}(\mathbb{Z}/n\mathbb{Z})$ induced by the ring homomorphism $\mathbb{Z} \longrightarrow \mathbb{Z}/n\mathbb{Z}$ is not surjective in general since $\mathcal{C}(\mathbb{Z})$ only has four elements!

Since $\mathbb{Z}/mn\mathbb{Z} \simeq \mathbb{Z}/m\mathbb{Z} \oplus \mathbb{Z}/n\mathbb{Z}$ for coprime integers $m,n \in \mathbb{N}$, we get

Corollary 4.5. *If m and n are coprime, then $\mathcal{C}(\mathbb{Z}/mn\mathbb{Z}) \simeq \mathcal{C}(\mathbb{Z}/m\mathbb{Z}) \oplus \mathcal{C}(\mathbb{Z}/n\mathbb{Z})$.*

This reduces the problem of determining the structure of $\mathcal{C}(\mathbb{Z}/m\mathbb{Z})$ to the case of prime powers $m = p^n$. For odd p and $n \geq 1$, one finds without too much difficulties that $\mathcal{C}(\mathbb{Z}/p^n\mathbb{Z})$ has cardinality $p^{n-1}\#\mathcal{C}(\mathbb{Z}/p\mathbb{Z})$; showing that $\mathcal{C}(\mathbb{Z}/p^n\mathbb{Z})$ is cyclic demands more care. The case $p = 2$ is particularly interesting (or nasty, depending on your point of view).

4.4 Finite Fields

Recall that a field is a commutative ring in which every nonzero element has an inverse. A field with only finitely many elements is called a finite field. The only finite fields we have seen so far are the fields $\mathbb{Z}/p\mathbb{Z}$ for primes p. In this section, we will meet a few more.

A basic method for constructing fields is the following:

Proposition 4.6. *If K is a field of characteristic $\neq 2$ and if $x \in K$ is a nonsquare, then $L = \{a + b\sqrt{x} : a, b \in K\}$ is a field with respect to addition $(a + b\sqrt{x}) + (c + d\sqrt{x}) = (a + c) + (b + d)\sqrt{x}$ and multiplication $(a + b\sqrt{x}) \cdot (c + d\sqrt{x}) = (ac + bdx) + (ac + bd)\sqrt{x}$. This field is denoted by $L = K(\sqrt{x}\,)$.*

Proof. $(L, +)$ is clearly an additive group with neutral element $0 = 0 + 0\sqrt{x}$. Moreover, $L \setminus \{0\}$ is a multiplicative group: given any $a + b\sqrt{x} \neq 0$, we claim that its inverse is given by $c + d\sqrt{x}$ with $c = \frac{a}{a^2-xb^2}$ and $d = -\frac{b}{a^2-xb^2}$. In order for these expressions to make sense, we have to show that $a^2 - xb^2 \neq 0$. So assume that $a^2 - xb^2 = 0$; if $b = 0$, then $a^2 = 0$ and hence $a = 0$ (since K is a field): contradiction, since then $a + b\sqrt{x} = 0$. Thus $b \neq 0$, hence $x = (a/b)^2$ is a square in K contradicting our assumption. Finally, we have to check that the given element really is the desired inverse: $(a + b\sqrt{x})\,\frac{a-b\sqrt{x}}{a^2-xb^2} = 1$.

Proving that the other field axioms hold is easy. □

Now assume that $K = \mathbb{Z}/p\mathbb{Z}$; if $x \in K$ is a quadratic nonresidue, then the proposition above tells us that $L = K(\sqrt{x})$ is a field. Since L has exactly p^2 elements, L is a finite field with p^2 elements; in the mathematical literature, finite fields with q elements are denoted by $\mathbb{F}_q$.

How many fields with p^2 elements are there? At first sight, we have constructed $\frac{p-1}{2}$ such fields above, since this is the number of quadratic nonresidues modulo p. It turns out, however, that these fields are isomorphic: if x and y are quadratic nonresidues, then $y = xz^2$ for some nonzero $z \in \mathbb{Z}/p\mathbb{Z}$;

but then the element $a+b\sqrt{y}\in K(\sqrt{y}\,)$ is nothing but $a+bz\sqrt{x}\in K(\sqrt{x}\,)$. As a matter of fact, it can be shown that for every prime power $q=p^n$ there exists exactly one finite field with q elements.

The Structure of $\mathcal{C}(\mathbb{F}_q)$

Determining the structure of $\mathcal{C}(\mathbb{F}_q)$, where $\mathbb{F}_q$ is a finite field with $q=p^n$ elements, is much less difficult than the corresponding problem for rings $\mathbb{Z}/p^n\mathbb{Z}$. The table for $\mathcal{C}(\mathbb{Z}/p\mathbb{Z})$ suggests $\mathcal{C}(\mathbb{F}_p)\simeq\mathbb{Z}/(p\mp1)\mathbb{Z}$ for primes $p\equiv\pm1 \bmod 4$. The proof is not difficult and works for arbitrary fields of characteristic $\ne 2$.

Suppose first that K contains a square root i of -1. Then we consider the map

$$\psi:\mathcal{C}(K)\longrightarrow K^\times:(x,y)\longmapsto x+iy.$$

It is easy to check that ψ is a group homomorphism: $\psi(P_1)\cdot\psi(P_2)=(x_1+iy_1)(x_2+iy_2)=x_1x_2-y_1y_2+i(x_1y_2+x_2y_1)=\psi(P_1+P_2)$. The kernel of ψ consists of points $(x,y)\in\mathcal{C}(K)$ with $x+iy=1$. Since $1=x^2+y^2=(x+iy)(x-iy)$, we have $x-iy=1$. But then $x=1$ and $y=0$, hence $(x,y)=(1,0)$ is the neutral element of the circle, and ψ is injective. In order to show that ψ is surjective, we have to show that every $r\in K^\times$ can be written in the form $r=x+iy$ with $x^2+y^2=1$. But this is easy: since 2 has an inverse in K, we only need to put $x=\frac12(r+\frac1r)$ and $y=\frac1{2i}(r-\frac1r)$.

Next we consider the case where $L=K(i)$ is a quadratic extension of K. Then $\psi:(x,y)\longmapsto x+iy$ defines a homomorphism $\mathcal{C}(K)\longrightarrow L^\times$. The proof of injectivity given above continues to hold, and the image of ψ is clearly equal to the subgroup

$$\mathbb{G}_m[-1]:=\{x+iy\in L^\times: x^2+y^2=1\}$$

of $L^\times$. We have proved:

Proposition 4.7. *If K is a field of characteristic $\ne 2$, then*

$$\mathcal{C}(K)\simeq\begin{cases}\mathbb{G}_m & \textit{if } i\in K;\\ \mathbb{G}_m[-1] & \textit{if } i\notin K.\end{cases}$$

Here we have put $\mathbb{G}_m=K^\times$. In the special case $K=\mathbb{R}$ we have $\mathbb{G}_m[-1]=S^1$, the group of complex numbers with absolute value 1. In fact, this is how one would discover the isomorphism of Proposition 4.7: one works over the complex numbers, where $\mathcal{C}(\mathbb{R}) \longrightarrow S^1$ is an isomorphism, and observes that this map makes sense over arbitrary fields if only $i = \sqrt{-1}$ can be given a meaning.

Now consider the special case $K = \mathbb{F}_q$ of a finite field with $i \notin K$: since $L^\times \longrightarrow K^\times : x + iy \longmapsto x^2 + y^2$ is the norm map, and since norm maps between finite fields are surjective, we see that $\#\mathbb{G}_m[-1] = \#L^\times/\#K^\times = \frac{q^2-1}{q-1} = q + 1$. This proves

Corollary 4.8. *If $K = \mathbb{F}_q$ is a finite field of characteristic $\neq 2$, then*

$$\mathcal{C}(K) \simeq \begin{cases} \mathbb{Z}/(q-1)\mathbb{Z} & \text{if } q \equiv 1 \bmod 4; \\ \mathbb{Z}/(q+1)\mathbb{Z} & \text{if } q \equiv 3 \bmod 4. \end{cases}$$

Indeed, if $i \in K$, then $q = \#K^\times \equiv 0 \bmod 4$; conversely, if $4 \mid \#K^\times$, then there exists an element of order 4, since the multiplicative group of finite fields is cyclic. Therefore $i \in K$ if and only if $q = \#K \equiv 1 \bmod 4$.

4.5 The Group Law on Conics

The unit circle is a special case of a conic, and here we will show that it is possible to define a group law on any irreducible[6] conic $\mathcal{C}$. The most naive way to achieve this is to pick a point $P \in \mathcal{C}(K)$ and use the parametrization of $\mathcal{C}(K)$, that is, the bijection between K and $\mathcal{C}(K) \setminus \{P\}$, to turn $\mathcal{C}(K) \setminus \{P\}$ into a group (or even a field).

Such a "group law" was first studied by von Staudt[7] in his influential books *Geometrie der Lage* (Nuremberg 1847) and *Beiträge zur Geometrie der Lage* (Nuremberg 1856–60). His work is almost illegible for most of today's readers (it is for me; even Klein,[8] who knew infinitely more about geometry than I do, confessed that "For me, Staudt's presentation has always been completely inaccessible").

[6] A conic $f(X, Y) = 0$ is said to be defined over K if $f(X, Y) \in K[X, Y]$; it is called irreducible if it cannot be written as a product of two lines over the algebraic closure of K. For example, the conic $X^2 - Y^2 = 0$ defined over $\mathbb{Q}$ is obviously reducible, but so is $X^2 + Y^2 = 0$ since $X^2 + Y^2 = (X + Yi)(X - Yi)$.

[7] Karl Georg Christian von Staudt (1798–1867), professor at the universities of Nuremberg and Erlangen.

[8] p. 133 in F. Klein, *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, Chelsea 1967; (orig. Berlin 1926; new ed. Springer-Verlag 1979).

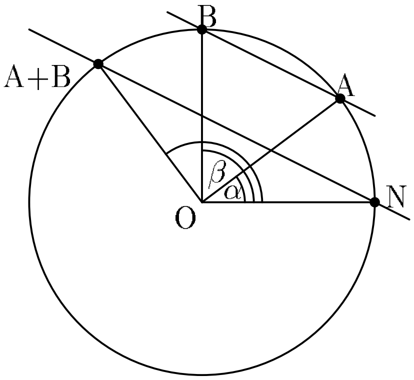

Figure 4.3: Group Law on Unit Circle

Juel [9] arrived at the correct definition of a group law on conics (he stated it only for circles and hyperbolas, though; Prasolov & Solovyev[10] give the general case) defined over a field K: the idea is to choose an arbitrary point N (the addition law will be defined over the field you get by adjoining to K the coordinates of N; note that there are conics defined over $\mathbb{Q}$ without rational points such as the one described by $x^2 + y^2 = 3$) as your neutral element; in order to add two points P and Q on the conic $\mathcal{C}$, draw a parallel to PQ through N. This line will intersect $\mathcal{C}$ in a second point R; now put $P + Q = R$. If $\mathcal{C}$ is the unit circle, this coincides with the addition law given above if we pick $N = (1, 0)$.

For subfields K of $\mathbb{R}$, we can prove geometrically that the above construction defines an abelian group law on the unit circle which coincides with the one discussed before. In fact, put $\mathcal{C} = A + B$ and let Q denote the point of intersection of the lines BA and ON (if these lines are parallel, the proof is clear). Then we have

$$\begin{aligned} \measuredangle ONC &= \measuredangle OQB = \pi - \beta - \measuredangle OBA, && (4.2)\\ \beta - \alpha + 2\measuredangle BA &= \pi, && (4.3) \end{aligned}$$

where the last equation follows from $\measuredangle OBA = \measuredangle OAB$. Multiply (4.2) by 2

[9] *Ueber die Parameterbestimmung von Punkten auf Curven zweiter und dritter Ordnung. Eine geometrische Einleitung in die Theorie der logarithmischen und elliptischen Funktionen*, Math. Ann. **47**, 72–104; in particular, p. 101.

[10] Viktor Prasolov & Yuri Solovyev, *Elliptic Functions and Elliptic Integrals*, AMS 1997

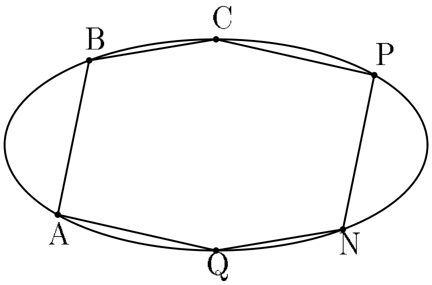

Figure 4.4: Pascal's Theorem

and subtract (4.3): this gives $2\measuredangle ONC = \pi - \alpha - \beta$, hence $\measuredangle CON = \alpha + \beta$ as desired.

If the field K is not a subfield of $\mathbb{R}$, the identity of the two definitions has to be proved computationally; another possibility is to derive it using some algebraic geometry: see e.g. Fulton[11] or Shafarevich.[12] If K is not a field but only a ring, it may seem that our definition of the group law may not even make sense: intersecting a line through a point $P \in K \times K$ with a conic defined over K leads to a quadratic equation, and in rings quadratic equations may have more than two roots (consider $x^2 = 1$ in $\mathbb{Z}/8\mathbb{Z}$, for example). Nevertheless, since our quadratic equation comes equipped with a solution in K (corresponding to the point P), it is an easy exercise to show that there is exactly one other root (counted with multiplicity), hence the addition law works even for rings.

Before we can talk about a group law on general conics, we have to check that our addition is associative, that is, that $A+(B+C) = (A+B)+C$. Put $P = A + B$ and $Q = B + C$; then associativity is equivalent to the following geometric statement: if A, B, $\mathcal{C}$, P, Q and N are points on the conic such that $AB \parallel NP$ and $BC \parallel QN$, then $AQ \parallel CP$. This is not obvious, at least to those of us who are not familiar anymore with classical geometry beyond the theorems of Pythagoras and Thales. In fact, the statement above is a special case of the formerly famous Pascal's Theorem:

> Opposite sides of a hexagon inscribed in a conic intersect on a straight line.

[11] W. Fulton, *Algebraic curves. An introduction to algebraic geometry,* New York-Amsterdam, 1969; reprint Addison-Wesley 1989

[12] I.R. Shafarevich, *Basic algebraic geometry,* vol. I, Springer-Verlag 1994

Let us now derive the addition formulas for the group on $\mathcal{C} : y^2 - ax^2 = 1$ (observe that the associativity follows from the addition formulas, although by a tedious calculation):

Proposition 4.9. *Consider the conic $\mathcal{C} : y^2 - ax^2 = 1$ over a ring R, where $a \in R$ is nonzero, and the point $N = (0,1)$ on $\mathcal{C}(R)$. Then the group law on $\mathcal{C}$ with neutral element N is given by $(r,s) + (t,u) = (ru + st, rt + asu,)$; the inverse of (r,s) is $(r,-s)$.*

Proof. For adding the points $P = (r,s)$ and $Q = (t,u)$, we have to draw a parallel to the line PQ through N and compute its second point of intersection with $\mathcal{C}$. The line through PQ has slope $m = \frac{s-u}{r-t}$, hence the parallel through N i s given by the equation $y - 1 = mx$. Intersecting this line with $\mathcal{C}$ leads to $m^2x^2 + 2mx - ax^2 = 0$; since $x = 0$ gives the point N, we may divide by x to find $x = \frac{2m}{a-m^2}$ and $y = mx + 1 = \frac{a+m^2}{a-m^2}$.

We now claim that $(x,y) = (ru + st, su + art)$. To this end, we first observe that $x = \frac{2(r-t)(s-t)}{a(r-t)^2-(s-u)^2}$; the denominator can be transformed into $(ar^2 - s^2) + (at^2 - u^2) - 2art + 2su = -2(1 + art - su)$, hence

$$x = -\frac{(r-t)(s-t)}{1 + art - su} = -\frac{(r-t)(s-t)(ru+st)}{(1+art-su)(ru+st)}.$$

Now $(r-t)(s-u) = (rs+tu) - (ru+st)$ and $(1+art-su)(ru+st) = (ru+st)+rs(at^2-u^2)+tu(ar^2-t^2) = (ru+st)-(rs+tu)$, hence $x = ru+st$ as claimed.

The formula for y now follows easily from

$$\begin{aligned} y &= mx + 1 = \frac{s-u}{r-t}(ru+st) + 1 \\ &= \frac{(r-t)su + t(s^2-1) - r(u^2-1)}{r-t} \\ &= \frac{(r-t)su + art(r-t)}{r-t} = su + art, \end{aligned}$$

and this concludes the proof. □

Let us now determine the structure of $\mathcal{C}(\mathbb{Z}/p\mathbb{Z})$, the group of $\mathbb{Z}/p\mathbb{Z}$-rational points on $\mathcal{C}$. First we leave it as an exercise to prove

Proposition 4.10. *Consider $\mathcal{C} : y^2 - ax^2 = 1$, and let p be an odd prime not dividing a; then $\#\mathcal{C}(\mathbb{Z}/p\mathbb{Z}) = p - (\frac{a}{p})$.*

The proof is the same as the one we gave for the unit circle. Now we claim

Theorem 4.11. *Consider $\mathcal{C} : y^2 - ax^2 = 1$, and let p be an odd prime not dividing a; then*

$$\mathcal{C}(\mathbb{Z}/p\mathbb{Z}) \simeq \begin{cases} \mathbb{Z}/(p-1)\mathbb{Z} & \textit{if } (\frac{a}{p}) = +1, \\ \mathbb{Z}/(p+1)\mathbb{Z} & \textit{if } (\frac{a}{p}) = -1. \end{cases}$$

Proof. We already know that the $\mathcal{C}(\mathbb{Z}/p\mathbb{Z})$ have the correct order: all that remains is to show that these groups are cyclic. Since the multiplicative group of a finite field is cyclic, and since subgroups of cyclic groups are again cyclic, it is sufficient to show that $\mathcal{C}(Z/p\mathbb{Z})$ is a subgroup of the multiplicative group of a finite field.

To this end, we distinguish two cases:

i) $(\frac{a}{p}) = +1$: then $a \equiv b^2 \bmod p$ for some $b \in \mathbb{Z}$, and $\psi(r,s) = s + rb$ defines a map $\psi : \mathcal{C}(\mathbb{Z}/p\mathbb{Z}) \longrightarrow (\mathbb{Z}/p\mathbb{Z})^\times$. We can easily check that ψ is a homomorphism: on one hand, we have $(r,s) + (t,u) = (ru + st, su + art)$, on the other hand $\psi(r,s) \cdot \psi(t,u) = (s + rb)(u + tb) = su + art + (ru + st)b$.

   Now we claim that ψ is injective; in fact, $\ker \psi$ consists of all points $(r,s) \in \mathcal{C}(\mathbb{Z}/p\mathbb{Z})$ with $\psi(r,s) = s + rb \equiv 1 \bmod p$. Since (r,s) has to satisfy $s^2 - ar^2 \equiv 1 \bmod p$, we find $s - rb \equiv (s^2 - ar^2)/(s + rb) = s - rb \equiv 1 \bmod p$, hence $2s \equiv (s + rb) + (s - rb) \equiv 2 \bmod p$. This implies $s \equiv 1 \bmod p$, thus $r \equiv 0 \bmod p$, and finally $\ker \psi = (1,0)$.

ii) $(\frac{a}{p}) = -1$: then $L = \mathbb{Z}/p\mathbb{Z}(b)$, where $b = \sqrt{a}$, is a finite field with p^2 elements. If we define the homomorphism $\psi : \mathcal{C}(\mathbb{Z}/p\mathbb{Z}) \longrightarrow L^\times$ exactly as in i), then the proof above shows that ψ is injective, a nd again we have realized $\mathcal{C}(\mathbb{Z}/p\mathbb{Z})$ as a subgroup of the multiplicative group of a finite field.

□

The Lucas-Lehmer Test

The largest primes known today are Mersenne primes, that is, primes of the form $M_p = 2^p - 1$ for prime values p. The reason for this is the existence of a very fast primality test for Mersenne numbers, namely the Lucas-Lehmer test. Given a number M_p, it states that M_p is prime if and only if $S_{p-2} \equiv 0 \bmod p$, where the sequence S_n is defined recursively by $S_0 = 4$ and $S_{n+1} = S_n^2 - 2$.

Example. Take $p = 5$; then $M_5 = 2^5 - 1 = 31$, and we find $S_0 = 4$, $S_1 = 14$, $S_2 = 194 \equiv 8 \bmod M_p$, and $S_3 \equiv 62 \equiv 0 \bmod M_p$. Therefore, M_5 is prime by the Lucas-Lehmer test.

Recall that the reason why a simple primality test for Fermat numbers N exists is that $N - 1$ is a power of 2. In the case of Mersenne primes N, we know that $N + 1$ is a power of 2, and the idea behind a primality test for M_p is replacing the group $(\mathbb{Z}/p\mathbb{Z})^\times \simeq \mathbb{Z}/(p-1)\mathbb{Z}$ used for Fermat numbers by a group of order $p + 1$. From what we know about conics, we might try $\mathcal{C} : y^2 - ax^2 = 1$ for some integer a such that $(\frac{a}{q}) = -1$ for Mersenne numbers $q = M_p = 2^p - 1$. A possible choice is $a = 3$, since $q = M_p \equiv 3 \bmod 4$ for $p \ge 3$ and $M_p \equiv 1 \bmod 3$ for odd p imply that $q = M_p \equiv 7 \bmod 12$, and we have $(\frac{3}{q}) = -(\frac{q}{3}) = -1$ for such q.

Thus we consider the curve $\mathcal{C} : x^2 - 3y^2 = 1$; a nontrivial rational (even integral) point on $\mathcal{C}$ is $P = (2, 1)$. The primality test analogous to the one for Fermat numbers is the following:

Proposition 4.12. *Let $\mathcal{C} : y^2 - ax^2 = 1$ be a conic, and assume that $q \equiv 7 \bmod 8$ is an integer such that $(\frac{a}{q}) = -1$. Then q is prime if and only if there exists a point $P \in \mathcal{C}(\mathbb{Z}/q\mathbb{Z})$ such that*

i) $(q + 1)P = (0, 1)$;

ii) $\frac{q+1}{r}P \neq (0, 1)$ *for any prime r dividing $(q + 1)$.*

Proof. Assume that $\frac{q+1}{2}P = (0, -1)$ in $\mathcal{C}(\mathbb{Z}/q\mathbb{Z})$; then we claim that q is prime. In fact, let p be any prime divisor of q. Then $\frac{q+1}{2}P = (0, -1)$ in $\mathcal{C}(\mathbb{Z}/p\mathbb{Z})$; but p is prime, so either $(p + 1)P = (0, 1)$ or $(p - 1)P = (0, 1)$.

On the other hand, condition i) implies that the order of P divides $q + 1$, while ii) implies that the order of P cannot be smaller. Thus $q + 1$ divides any integer m with $mP = (0, 1)$, in particular it divides $p - 1$ or $p + 1$. Since $p \mid q$, this is only possible if $p = q$, hence q is prime.

For the converse, assume that q is prime and let P be a point generating $\mathcal{C}(\mathbb{Z}/q\mathbb{Z})$. Then $(q+1)P = (0, 1)$ since $\mathcal{C}(\mathbb{Z}/q\mathbb{Z})$ has $q+1$ elements If $(\frac{a}{q}) = -1$, and moreover $\frac{q+1}{r}P \neq (0, 1)$ for any prime r dividing $(q + 1)$ since P is a generator. □

In the special case of Mersenne numbers $q = 2^p - 1$ (note that $q \equiv 7 \bmod 12$ for $p \geq 3$), we have $\frac{q+1}{2} = 2^{p-1}$. Picking $cC : y^2 - 3x^2 = 1$ and $P = (1,2)$ we have to show that P generates $\mathcal{C}(\mathbb{Z}/q\mathbb{Z})$ if q is prime. This is done as follows: since $q+1$ is a power of 2, so is the order of P. Thus if P does not generate $\mathcal{C}(\mathbb{Z}/q\mathbb{Z})$, then $P = 2Q$ for some point $Q = (r,s) \in \mathcal{C}(\mathbb{Z}/q\mathbb{Z})$. We find $s^2 - 3r^2 = 1$ as well as $s^2 + 3r^2 = 2$ from $P = 2Q$ and the addition law; thus $2s^2 = 3$, contradicting the fact that $(2/q) = +1$ and $(3/q) = -1$ for primes $q \equiv 7 \bmod 12$.

Thus in order to compute $\frac{q+1}{2} P = (0,-1)$, we only have to double the point P repeatedly. Here are the first few terms: $P = (1,2)$, $2P = (4,7)$, and $4P = (56,97)$. If we double the second coordinates of these points, we get the sequence 4, 14, 194 that we know from the Lucas-Lehmer test described above. This is no coincidence:

Lemma 4.13. *Consider $\mathcal{C} : y^2 - 3x^2 = 1$ and the point $P = (1,2)$ on $\mathcal{C}$. Define a sequence S_n by $S_0 = 4$ and $S_{n+1} = S_n^2 - 2$ for $n \geq 0$. Then $\frac{1}{2}S_r$ is the second coordinate of $2^r P$.*

Proof. This is of course done by induction: we have $2^0 P = P = (1,2)$ and $2 = \frac{1}{2}S_0$. Assume that $2^r P = (T, \frac{1}{2}S_r)$ for some integer T. Then $2^{r+1}P = (S_r T, \frac{1}{4}S_r^2 + 3T^2)$. But since $2^r P$ is on $\mathcal{C}$, we have $\frac{1}{4}S_r^2 - 3T^2 = 1$, hence $3T^2 = \frac{1}{4}S_r^2 - 1$, and the second coordinate of $2^{r+1}P$ is $\frac{1}{4}S_r^2 + \frac{1}{4}S_r^2 - 1 = \frac{1}{2}(S_r^2 - 2) = \frac{1}{2}S_{r+1}$. This proves the lemma. □

Now the condition $2^{p-1}P = (0,-1)$ in $\mathbb{Z}/q\mathbb{Z}$ implies that $\frac{1}{2}S_{p-1} \equiv -1 \bmod q$, thus $S_{p-1} \equiv -2 \bmod q$ and hence $q \mid S_{p-2}$. Conversely, if $q \mid S_{p-2}$, then the first coordinate of $2^{p-1}P$ must be -1, hence $2^{p-1}P = (-1,0)$. Thus the primality test of Proposition 4.12 is nothing but the Lucas-Lehmer test, expressed in a slightly different language.

Appendix A

Algebraic Preliminaries

In this appendix we introduce some algebra.

A.1 Groups and Rings

Groups

A group consists of two things: a set A of elements and a composition $*$ that maps two elements $a, b \in A$ to a third element $a * b \in A$ such that the following properties are satisfied:

(G1) Existence of neutral element: there is an $e \in A$ such that $e*a = a*e = a$ for all $a \in A$;

(G2) Existence of inverse: for every $a \in A$, there is an element $b \in A$ such that $a * b = b * a = e$;

(G3) Associativity: $(a * b) * c = a * (b * c)$.

All our groups will be abelian (commutative): they have to satisfy the additional property that $a * b = b * a$ for all $a, b \in A$.

Examples:

- $\mathbb{Z}$ is a group with respect to addition (we will say that $(\mathbb{Z}, +)$ is a group);

- $(\mathbb{Z}, \cdot)$ is not a group (2 has no inverse);

- $(\mathbb{N}, +)$ and $(\mathbb{N}_0, +)$ are not groups (1 has no inverse);
- $(\mathbb{Z}/m\mathbb{Z}, +)$ is a group with m elements;
- $(\mathbb{Z}/m\mathbb{Z})^\times, \cdot)$ is a group with $\phi(m)$ elements.

There is a simple way to make new groups out of old ones: given two groups $(G, *)$ and $(H, \circ)$, we can define the direct sum $(G \oplus H, \cdot)$ as the set of elements $(g, h) \in G \times H$ with composition $(g, h) \cdot (g', h') = (g * g', h \circ h')$. The group axioms are readily checked. The direct sum of two groups is often called the direct product, and then we write $G \times H$.

Rings

A ring $(R, +, \cdot)$ consists of a set R on which there are two (different) ways of composing elements (more exactly: two maps $+ : R \times R \longrightarrow R$ and $\cdot : R \times R \longrightarrow R$). We say that R is a ring if

(R1) $(R, +)$ is an abelian group;

(R2) multiplication is associative;

(R3) distributivity: $a(b + c) = ab + ac$ for all $a, b, c \in R$.

In addition, all our rings will have a unit element that we will denote by 1 (so $1a = a$ for all $a \in R$), and they will be commutative (so $ab = ba$ for all $a, b \in R$.

- $(\mathbb{Z}, +, \cdot)$ is a ring, $(\mathbb{N}, +, \cdot)$ is not;
- the polynomials $\mathbb{Z}[x]$ with coefficients in $\mathbb{Z}$ form a ring (more generally: if R is a ring, then so is $R[X]$, the ring of polynomials with coefficients in R).
- $(\mathbb{Z}/m\mathbb{Z}, +, \cdot)$ is a ring;
- the set of $n \times n$-matrices form a ring with respect to coordinate-wise addition and usual multiplication;
- the set of continuous functions $\mathbb{R} \longrightarrow \mathbb{R}$ form a ring;
- the set of converging series in $\mathbb{R}$ form a ring.

Elements a, b in a ring R are called zero divisors if $a \neq 0$, $b \neq 0$ and $ab = 0$. In the ring $\mathbb{Z}/pq\mathbb{Z}$, where p and q are primes, the residue classes $[p]$ and $[q]$ are nonzero, but their product $[pq] = [0]$ is: therefore, $[p]$ and $[q]$ are zero divisors.

A ring having no zero divisors (such as $\mathbb{Z}$) is called a domain.

Given any ring R, we can define its *unit group* $R^\times$ by

$$R^\times = \{r \in R : rs = 1 \text{ for some } s \in R\}.$$

The unit group is in fact a group: if $u, v \in R^\times$ are units, then $ur = vs = 1$ for suitable $r, s \in R$, and then $(uv)(rs) = 1$: but this proves that uv is a unit. Therefore multiplication is a composition on $R^\times$.

Next $R^\times$ contains an identity element because $1 \in R^\times$ (well, $1 \cdot 1 = 1$, so 1 is a unit), and if $u \in R^\times$, then there exists a unit v such that $uv = 1$ (this follows directly from the definition of a unit).

The unit group of $\mathbb{Z}$ is rather boring: $\mathbb{Z}^\times = \{-1, +1\}$. More interesting examples are provided by the unit groups $(\mathbb{Z}/m\mathbb{Z})^\times$ of the rings of residue classes modulo m.

Actually, $\mathbb{Z}$ is not only a ring but a domain: it does not contain zero divisors, i.e., $ab = 0$ implies $a = 0$ or $b = 0$. The residue class rings that we will discuss in the next chapters in general do have zero divisors.

Fields

A commutative ring is called a field if every nonzero element has a multiplicative inverse. For example, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{Z}/p\mathbb{Z}$ are fields, but $\mathbb{Z}/pq\mathbb{Z}$ (say, for primes p, q) is not since the class $[p]$ does not have an inverse.

If R is any ring, then so is the set $R[X]$ of polynomials with coefficients in R. If F is a field, then the ring $F[X]$ has a "Euclidean Algorithm". In fact, given any nonzero polynomials $A, B \in F[X]$, there are polynomials $Q, R \in F[X]$ such that $A = BQ + R$ and $\deg R < \deg B$.

Using the Euclidean Algorithm, we can define and compute greatest common divisors in the ring $F[X]$, define primes and irreducibles, show that they're the same, and eventually prove unique factorization in $F[X]$. In the special case where F is the finite field $\mathbb{Z}/p\mathbb{Z}$, we could even prove a reciprocity law ...

A.2 Homomorphisms

Here's the definition: a map $f : G \longrightarrow H$ from a group $(G, *)$ to another group $(H, \circ)$ is called a homomorphism if it respects the group structure, that is, if $f(g * g') = f(g) \circ f(g')$ for all $g, g' \in G$ (in other words: if it doesn't matter whether you compose first and then map, or whether you first map and then compose).

As for the motivation, consider the group of mappings of the Euclidean plane to itself consisting of the identity map I and the reflection R at the y-axis, and where composition is given by composition of maps. The multiplication table for this group is

$*$	I	R
I	I	R
R	R	I

Now we observe that this resembles the multiplication table of $\mathbb{Z}/2\mathbb{Z}$: in fact, if we write $[0]$ for I and $[1]$ for R, then the multiplication table above becomes the addition table for $\mathbb{Z}/2\mathbb{Z}$:

$+$	$[0]$	$[1]$
$[0]$	$[0]$	$[1]$
$[1]$	$[1]$	$[0]$

The map f sending $I \longmapsto [0]$ and $R \longmapsto [1]$ is an isomorphism: it is bijective, and it respects the group law: for example $f(I) = f(R * R) = f(R) + f(R) = [1] + [1] = [0]$.

Consider the three groups $\mathbb{Z}/4\mathbb{Z}$, $(\mathbb{Z}/5\mathbb{Z})^\times$ and $(\mathbb{Z}/8\mathbb{Z})^\times$ (you have determined their multiplication table in the Exercises). It is easy to see that the third group differs considerably from the first two: the equation $x^2 = 1$ has exactly two solutions in the first two groups (namely $x = [1]_4, [3]_4$ and $x = [1]_5, [4]_5$ respectively). There are, however, four solutions in the third: in fact, every residue class $[a]_8 \in (\mathbb{Z}/8\mathbb{Z})^\times$ satisfies $[a]_8^2 = [1]_8$.

The first two groups, on the other hand, do have the "same structure" in the following sense: if we rename the elements in the multiplication table of $\mathbb{Z}/4\mathbb{Z}$, then we get the multiplication table of $(\mathbb{Z}/5\mathbb{Z})^\times$. In other words:

Table A.1:

for	0	1	2	3
write	$\overline{1}$	$\overline{2}$	$\overline{4}$	$\overline{3}$

the only difference between the two groups is the name of the elements; their structure is the same.

Let's do this explicitly. Here are the two multiplication tables for $\mathbb{Z}/4\mathbb{Z}$ and $(\mathbb{Z}/5\mathbb{Z})^\times$:

$+$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

$\cdot$	1	2	3	4
1	1	2	3	4
2	2	4	1	3
3	3	1	4	2
4	4	3	2	1

We now relabel the elements of $\mathbb{Z}/4\mathbb{Z}$ in such a way that the multiplication table of $\mathbb{Z}/4\mathbb{Z}$ becomes the same as the one for $(\mathbb{Z}/5\mathbb{Z})^\times$. Here's how it's done: We do not have much freedom in choosing where to send the elements of $\mathbb{Z}/4\mathbb{Z}$: If we send 1 to $\overline{2}$, and if the multiplication tables should coincide, then we must send $2 = 1 + 1$ to $\overline{2} \cdot \overline{2} = \overline{3}$. We easily check that sending 1 to $\overline{2}$ determines the whole table.

The new multiplication table for $\mathbb{Z}/4\mathbb{Z}$ now looks like this:

$+$	$\overline{1}$	$\overline{2}$	$\overline{4}$	$\overline{3}$
$\overline{1}$	$\overline{1}$	$\overline{2}$	$\overline{4}$	$\overline{3}$
$\overline{2}$	$\overline{2}$	$\overline{4}$	$\overline{3}$	$\overline{1}$
$\overline{4}$	$\overline{4}$	$\overline{3}$	$\overline{1}$	$\overline{2}$
$\overline{3}$	$\overline{3}$	$\overline{1}$	$\overline{2}$	$\overline{4}$

Comparing it with the multiplication table for $(\mathbb{Z}/5\mathbb{Z})^\times$ we can now check that the two are indeed the same.

This whole process of showing that $\mathbb{Z}/4\mathbb{Z}$ and $(\mathbb{Z}/5\mathbb{Z})^\times$ have the same "structure" is quite complicated: luckily, most of the complications vanish if we proceed more abstractly. Table A.1 defines a map $\phi : \mathbb{Z}/4\mathbb{Z} \longrightarrow (\mathbb{Z}/5\mathbb{Z})^\times$; we have constructed ϕ in such a way that $\phi(a+b) = \phi(a)\phi(b)$ for any $a, b \in \mathbb{Z}/4\mathbb{Z}$, and it is exactly this property that is responsible for the equality between the multiplication tables of $\mathbb{Z}/4\mathbb{Z}$ (with the elements renamed as $\overline{1}, \overline{2}, \overline{4}, \overline{3}$) and $(\mathbb{Z}/5\mathbb{Z})^\times$.

In fact, assume we have two groups $(G, *)$ and $(H, \circ)$. Let a and b be two elements of G, and put $\overline{a} = f(a)$ and $\overline{b} = f(b)$. The product $a * b \in G$ gets renamed to $f(a * b)$; the multiplication tables will coincide after renaming if and only if $f(a * b)$ is the product of $f(a)$ and $f(b)$ in H, that is, if and only if $f(a * b) = f(a) \circ f(b)$:

$*$		a	
b		$a * b$	

$\circ$		$\overline{a}$	
$\overline{b}$		$\overline{a} \circ \overline{b}$	

Now consider any two groups $(G, *)$ and $(H, \circ)$. A map $\phi : G \longrightarrow H$ with $\phi(g * g') = \phi(g) \circ \phi(g')$ is called a group homomorphism. It is called injective (one-to-one) if $\phi(g) = \phi(g')$ implies $g = g'$, and it is called surjective (onto) if for every $h \in H$ there is a $g \in G$ such that $h = \phi(g)$. Homomorphisms $G \longrightarrow H$ that are injective and bijective are called isomorpisms, and the groups G and H isomorphic (we write $G \simeq H$). Isomorphic groups have the same multiplication table (possibly after relabeling the elements).

Isomorphisms $f : (G, *) \longrightarrow (H, \circ)$ have some nice properties:

- Isomorphisms map neutral elements to neutral elements. In fact, let 0_G and 0_H denote the neutral elements of G and H, respectively. Then $f(0_G) = f(0_G + 0_G) = f(0_G) + f(0_G)$, and subtracting $f(0_G)$ gives $0_H = f(0_G)$.

- Isomorphisms map elements of order n to elements of order n. In fact. let n denote the order of an element $g \in G$ and put $h = f(g)$. Then $nh = nf(g) = f(ng) = f(0_G) = 0_H$, so the order of h divides n. Conversely, if $mh = 0_H$, then $0_H = mf(g) = f(mg)$, and since f is an isomorphism, this implies that $mg = 0_G$. But then $m \geq n$, and this proves that $m = n$.

The map $G \longrightarrow H$ sending every $g \in G$ to e_H is always a homomorphism; it is called the trivial homomorphism. Nontrivial homomorphisms do not always exist: for example, the only homomorphism $f : \mathbb{Z}/m\mathbb{Z} \longrightarrow \mathbb{Z}$ is the trivial homomorphism. For let $f([1]) = a$ for some $a \in \mathbb{Z}$; then $ma =$

$mf([1]) = f([m]) = f([0]) = 0$, hence $a = 0$. But then $f([r]) = rf([1]) = r \cdot 0 = 0$.

An important result that we will use repeatedly is the following

Lemma A.1. *If $f : A \longrightarrow B$ is an injective map between finite sets of the same cardinality, then f is bijective.*

Proof. Since f is injective, $f(a) = f(a')$ implies that $a = a'$, in other words: f maps different elements of A to different elements of B. Let $f(A)$ be the set of all $f(a)$ with $a \in A$. Then $\#f(A) = \#A$ by injectivity, and $\#A = \#B$ by assumption, so $\#B = \#f(A)$. But this plainly shows that every element of B occurs in the image of f. □

In the case at hand, we can define ϕ simply by putting $\phi([a]_4) = [2^a]_5$, where $[a]_n$ denotes the residue class of a modulo n (since I need a map going from an additive to a multiplicative group, I am "forced" to map $[a]_4$ to $[g^a]_5$ for some g; choosing $g = 1$ or $g = -1$ would give maps that are not bijective, so we end up with the choices $g = 2$ or $g = 3$). This map is well defined because Fermat's little theorem shows that $[2^a]_5 = [2^{a+4}]_5$. Next it is a homomorphism since $\phi([a+b]_4) = [2^{a+b}]_5 = [2^a]_5[2^b]_5 = \phi(a)\phi(b)$. Finally, table A.1 implies that ϕ is bijective.

For homomorphisms, checking injectivity is simplified by the following

Lemma A.2. *A homomorphism $f : G \longrightarrow H$ between two additively written groups is injective if and only if $f(g) = 0$ implies $a = 0$.*

Proof. If f is injective, then $f(g) = 0$ implies $a = 0$. Conversely, assume that $f(g) = 0$ implies $a = 0$, and assume that there are elements $g, g' \in G$ such that $f(g) = f(g')$. Then $0 = f(g) - f(g') = f(g - g')$, hence $g - g' = 0$, and this implies $g = g'$. This f is injective. □

The set of all $g \in G$ such that $f(g) = 0$ is called the kernel of f And will be denoted by $\ker f$. It is easy to see that $\ker f$ is a subgroup of G: if $g, g' \in \ker f$, then $f(g) = f(g') = 0$, hence $f(g+g') = f(g)+f(g') = 0+0 = 0$, hence $g + g' \in \ker f$.

As another exercise, let us now prove that $(\mathbb{Z}/9\mathbb{Z})^\times \simeq \mathbb{Z}/6\mathbb{Z}$. To this end we define a map $f : \mathbb{Z}/6\mathbb{Z} \longrightarrow (\mathbb{Z}/9\mathbb{Z})^\times$ by $f([a]_6) = [2^a]_9$. As above, f is clearly a homomorphism. Its kernel consists of all $[a]_6$ such that $2^a \equiv 1 \bmod 9$; the smallest exponent a with this property is $a = 6$, hence the set of all a with this property are the multiples of 6 (by Proposition 2.6), so

$\ker f = [0]_6$ and f is inject ive. Moreover, f is surjective because $\phi(9) = 6$ and 2^6 is the smallest power of 2 congruent to 1 mod 9.

Let us now prove that $(\mathbb{Z}/8\mathbb{Z})^\times \simeq \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$. To this end we define a map $f : \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z} \longrightarrow (\mathbb{Z}/8\mathbb{Z})^\times$ by putting $f([a]_2, [b]_2) = [(-1)^a 5^b]_8$ (one of many possible choices; $g([a]_2, [b]_2) = [(-1)^a 3^b]_8$ would do as well). This is well defined because changing a or b by even numbers does not change the residue class $(-1)^a 5^b$ mod 8: in fact, $(-1)^2 \equiv 5^2 \equiv 1$ mod 8. Moreover, f is a homomorphism: $f([a+a']_2, [b+b']_2) = [(-1)^{a+a'} 5^{b+b'}]_8 = \ldots = f([a]_2, [b]_2) \cdot f([a']_2, [b']_2)$. Finally, f is bijective: this is most easily seen by going through all elements of $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$.

One of the fundamental results in the theory of finite abelian groups is the classification theorem that we formulate as follows:

Theorem A.3. *If G is an abelian group, then there exist prime numbers p_i and integers $a_i > 0$ such that $G \simeq \mathbb{Z}/p_1^{a_1}\mathbb{Z} \oplus \ldots \oplus \mathbb{Z}/p_r^{a_r}\mathbb{Z}$. Moreover, $G \simeq \mathbb{Z}/q_1^{b_1}\mathbb{Z} \oplus \ldots \oplus \mathbb{Z}/q_s^{b_r}\mathbb{Z}$ for primes q_i and integers $b_i > 0$ if and only if $r = s$ and $(p_i, a_i) = (q_i, b_i)$, possibly after rearranging the q_i.*

Observe the similarity to the unique factorization theorem. We will neither prove nor use this result. Note that it implies e.g. that $\mathbb{Z}/4\mathbb{Z}$ is not isomorphic to $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ since $p_1 = 2$, $a_1 = 2$ for $\mathbb{Z}/4\mathbb{Z}$ and $p_1 = p_2 = 2$, $a_1 = a_2 = 1$ for $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$.

Corollary A.4. *If G and H are cyclic groups of coprime order, then $G \oplus H$ is cyclic.*

Proof. Let g and h be generators of G and H, respectively. Then the elements $(g, 1)$ and $(1, h)$ of $G \oplus H$ have coprime orders, so by Lemma 2.17 their product (g, h) generates $G \oplus H$. □

Ring Homomorphisms

In a ring, there are two structures: addition and multiplication. A map $f : R \longrightarrow S$ between rings is called a ring homomorphism if it respects both of them, that is: if $f(a + b) = f(a) + f(b)$ and $f(ab) = f(a)f(b)$ for all $a, b \in R$. Bijective ring homomorphisms are of course called isomorphisms. Our proof that showed $(\mathbb{Z}/ab\mathbb{Z})^\times \simeq (\mathbb{Z}/a\mathbb{Z})^\times \times (\mathbb{Z}/b\mathbb{Z})^\times$ for coprime integers a, b actually shows that there is a ring isomorphism $\mathbb{Z}/ab\mathbb{Z} \simeq \mathbb{Z}/a\mathbb{Z} \times \mathbb{Z}/b\mathbb{Z}$.

Index