M14: NUMBER THEORY

Prof. David Loeffler

Assistant: **Dr. Francesco Zerman**

Autumn 2025

Last updated 9th December 2025

FernUni.ch
UniDistance.ch

Contents

Preamble

Acknowledgements

This course is loosely based on a lecture course taught by my former colleague Prof. John Cremona at the University of Warwick. It also incorporates a number of suggestions from Sarah Zerbes of ETH Zürich. I am grateful to Daniele Bolla for pointing out a number of errors and inconsistencies in the 2024 version of these notes, which should now be fixed.

HTML Version

These lecture notes are also available in an HTML version and in app form.

https://apptest.fernuni.ch

The HTML version contains the lecture notes, and additional resources such as model solutions to exercises.

Updates during the semester

[None so far]

CHAPTER 1

Divisibility and GCD

Preliminaries

Remark 1.1 (Recommended textbooks) All the material you need to know for the examination is in this script; but you might find some of the books below useful for an alternative viewpoint, or if you are curious to learn more beyond what is in this course.

- For elementary number theory (i.e. everything in chapters 1–7 of these notes), I recommend Davenport's classic text *The Higher Arithmetic*. Since Davenport died in 1969, the copyright on this book has now expired and you can download it for free – entirely legally – from `https://archive.org/details/h.-davenport-the-higher-arithmetic/`.
- For algebraic number fields (chapters 8–12), I highly recommend Stewart and Tall's *Algebraic Number Theory and Fermat's Last Theorem*, although this goes a long way beyond what we can cover here. Cox's lovely book *Primes of the form* $x^2 + ny^2$ gives a very interesting and original perspective on some of these ideas.

Remark 1.2 (General notations) In this module we use the following symbols:

- $\mathbb{N}$ denotes the natural numbers $\{0, 1, 2, 3, \dots\}$;
- $\mathbb{Z}$ the integers (positive, negative or zero);
- $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ the fields of rational, real, and complex numbers respectively.
- $\mathbb{N}_+$ denotes the *positive*[a] integers $\{1, 2, 3, \dots\}$.
- If a is in $\mathbb{R}$ (and in particular if $a \in \mathbb{Z}$), the symbol $|a|$ means the absolute value of a, i.e. $|a| = a$ if $a \geqslant 0$, and $|a| = -a$ if $a \leqslant 0$.
- For $n \in \mathbb{N}$, we write $n!$ (read as "n factorial") for the product $1 \times 2 \times \cdots \times n$, with $0!$ defined to be 1.
- The logical symbols $\Rightarrow$, $\iff$, $\exists$, $\forall$ have their usual meanings.
- The symbol $\square$ denotes the end of a proof.
- The symbol $\circledast$ is used to mark unsolved problems and conjectures.

[a] Beware that some other texts use $\mathbb{N}$ for positive integers!

Remark 1.3 (Reminders on induction) We're going to use **induction** quite a lot in this module, so it might be a good idea to revise it if your memory has got rusty.

As a reminder: the *principle of mathematical induction*, which you saw way back in M01 Algorithmics, is a very powerful tool for proving statements about $\mathbb{N}$. It goes as follows. Suppose $P(n)$ is some statement about the natural number n, and:

- $P(0)$ is true,

- for any $n \in \mathbb{N}$ the implication $P(n) \implies P(n+1)$ is true.

Then $P(n)$ is true for all n.

There are a few variants of induction which are useful:

Different starting points: Let $t \in \mathbb{N}$ be given. If $P(t)$ is true, and for any $n \geqslant t$ we have $P(n) \implies P(n+1)$, then $P(n)$ is true for all $n \geqslant t$. (The usual induction is $t = 0$, but the $t = 1$ case also occurs frequently.)
This can, of course, easily be derived from "usual" induction applied to the new statement $Q(n)$ defined by "if $n \geqslant t$ then $P(n)$", which is vacuously true for $n < t$.

Strong induction: Suppose P is a statement such that:

- for any $n \in \mathbb{N}$, if $P(r)$ is true for all $r < n$, then $P(n)$ is true.

Then $P(n)$ holds for all n.
This looks far more powerful than usual induction (because we have to prove only one thing, and we're allowed to assume something that looks a lot stronger); but in fact it easily follows from usual induction.

Exercise 1.4 Deduce Strong Induction from usual Induction. (Hint: consider the statement $Q(n)$ defined as "$P(r)$ holds for all r with $r < n$".)

Minimal elements: Our final induction variant is known as the *well-ordering principle* for $\mathbb{N}$.

- Let $S \subset \mathbb{N}$ be a non-empty set. Then S has a minimal element; that is, there exists $n \in S$ such that every $m \in S$ satisfies $m \geqslant n$.

It's not immediately obvious that this has anything to do with induction at all! But it's clearly something quite special about $\mathbb{N}$: it's obviously false for $\mathbb{Z}$, or for the non-negative reals[a].
To see this, suppose S doesn't have a minimal element, and let $P(n)$ be the statement "$m \geqslant n$ for all $m \in S$". Clearly $P(0)$ holds, since every natural number is $\geqslant 0$. Now, if $P(n)$ holds, then we must have $n \notin S$, since otherwise n would be the minimal element of S. So for $m \in S$, we have $m \geqslant n$ and $m \neq n$. So $m \geqslant n+1$, and thus $P(n+1)$ holds. By induction, $P(n)$ holds $\forall n$; so S is empty, a contradiction.

Exercise 1.5 Give an example of a subset of the non-negative reals $\mathbb{R}_{\geqslant 0}$ which does not have a minimal element.

[a] Of course, it's hugely important in real analysis that any bounded-below subset of the real numbers has a *greatest lower bound*, but this is not the same thing as a *minimal element* (why?)

1.1 Divisibility

Recall the following familiar definition:

Definition 1.6 Let $a, b \in \mathbb{Z}$. We say "a divides b", or "b is a multiple of a", if there exists $n \in \mathbb{Z}$ such that $na = b$. If so, we write "$a \mid b$", and we say a is a *divisor* or *factor* of b. Otherwise we write "$a \nmid b$".

Example 1.7 We have $3 \mid -15$, since $3 \times (-5) = -15$.

Notice that this still makes sense if a or b is zero; and since $n \cdot 0 = 0$ for all n, it follows that everything divides 0, but 0 does not divide anything except itself. On the other hand, 1 and -1 both divide everything, and nothing except ± 1 can divide them. (So 0 is the "most divisible" element of $\mathbb{Z}$, while 1 and -1 are the "least divisible".)

Remark 1.8 The "divides" symbol is a *relation*: for any given values of a and b, "$a \mid b$" is a self-contained statement which is either true or false. Don't confuse it with division a/b, which is a number (if it is defined at all, which it might not be if $b = 0$).

Exercise 1.9 Check that if $a, b \in \mathbb{N}$, then $a \mid b$ if and only if there exists $\underline{n \in \mathbb{N}}$ such that $na = b$. (Take care with the case $a = 0$!)

Proposition 1.10 (Elementary properties of divisibility) *Let $a, b, c, \dots \in \mathbb{Z}$. Then:*

(i) If $a \mid b$, then $a \mid kb$ for all $k \in \mathbb{Z}$.
(ii) If $a \mid b$ and $a \mid c$, then $a \mid b \pm c$.
(iii) If $a \mid b$ and $b \mid c$ then $a \mid c$.
(iv) If $a \mid b$ and $b \mid a$, then $a = \pm b$.
(v) If $a \mid b$ and $b \neq 0$, then $|a| \leqslant |b|$; so nonzero integers have only finitely many divisors.
(vi) We have $a \mid |a|$ (the notation is awkward; read it as "a divides the absolute value of a").
(vii) If $k \neq 0$, then $a \mid b \iff ka \mid kb$.

Proof Exercise. □

The following innocent-looking proposition will turn out to be crucial in understanding divisibility and factorisation of integers:

Proposition 1.11 (Division with remainder) *Let $a, b \in \mathbb{Z}$ with $a \neq 0$. Then there exists a unique pair of integers (q, r) such that $0 \leqslant r < |a|$ and $b = qa + r$.*

Example 1.12

(i) For $a = 5$ and $b = 21$, we have $(q, r) = (4, 1)$.
(ii) For $a = 5$ and $b = -21$, we have $(q, r) = (-5, 4)$ [*not* $(-4, -1)$!]

Proof Let S be the set of of integers which are of the form $b - qa$, for some $q \in \mathbb{Z}$; and let $S' = S \cap \mathbb{N}$ be the non-negative elements of S.

The set S' is always non-empty (if $b \geqslant 0$, then $b \in S'$, and if $b < 0$, then one checks that $(|a| - 1) \cdot |b| \in S'$).

We know that a non-empty subset of $\mathbb{N}$ always has a smallest element. So let r be the smallest element of S'. If $r \geqslant |a|$, then $r - |a|$ is a strictly smaller element of S', contradiction; so $0 \leqslant r < |a|$. By definition of S' there exists q with $r = b - qa$ so we are done. □

Remark 1.13 More concretely, if $a > 0$, then q is given by $\lfloor b/a \rfloor$, where $\lfloor x \rfloor$ is the *floor* function: the function which converts a real (or rational) number to an integer by rounding towards $-\infty$ (meaning that $\lfloor 1.5 \rfloor = 1$ and $\lfloor -1.5 \rfloor = -2$). Thus we can easily compute q and r from the decimal expansion of b/a.

1.2 The greatest common divisor

Proposition 1.14 *Let $a, b \in \mathbb{Z}$. Then there exists $c \in \mathbb{Z}$ such that the following holds:*

$$\forall x \in \mathbb{Z}, \qquad x \mid c \iff x \mid a \text{ and } x \mid b.$$

This c is uniquely determined by a and b up to sign; and we write $\gcd(a, b)$ for the unique non-negative c with this property, which we call the greatest common divisor *of a and b.*

Example 1.15 If we let $a = 20$ and $b = 30$, the integers dividing a are $\{\pm 1, \pm 2, \pm 4, \pm 5, \pm 10, \pm 20\}$, and the integers dividing b are $\{\pm 1, \pm 2, \pm 3, \pm 5, \pm 6, \pm 10, \pm 15, \pm 30\}$. The intersection of these sets, i.e. the set $\{x : x \mid a \text{ and } x \mid b\}$, is $\{\pm 1, \pm 2, \pm 5, \pm 10\}$, which are precisely the divisors of 10. So $\gcd(20, 30) = 10$.

Proof It is clear that if c and c' both satisfy the condition, then $c \mid c'$ and $c' \mid c$, so $c' = \pm c$. Conversely, if c works then $-c$ does too. So it suffices to prove existence.

If a, b are both zero, then the result is trivial; so assume not. Let T denote the set of integers of the form $ma + nb$ for $m, n \in \mathbb{Z}$, and T' its intersection with the *strictly* positive integers. We check easily that T' is non-empty (since at least one of $|a|$ and $|b|$ is in T'); so it contains a smallest element. Let c be this element. Clearly c has the form $ma + nb$, so anything which divides a and b also divides c.

We claim c itself divides both a and b. By symmetry it suffices to show $c \mid a$. By division-with-remainder, we can write $a = qc + r$, for some r with $0 \leqslant r < c$. But $r = a - qc = a - q(ma + nb)$ is also in T, and it is non-negative and strictly smaller than c. If $r > 0$, then $r \in T'$, contradicting the minimality of c. So we must have $r = 0$, i.e. c divides a. □

Remark 1.16 Note that (except in the trivial case $a = b = 0$), the greatest common divisor $\gcd(a, b)$ is, as its name suggests, the largest element of the set of common divisors of a and b (the set $\{x \in \mathbb{Z} : x \mid a \text{ and } x \mid b\}$). This follows from the much stronger fact proved above that this set consists precisely of the divisors of c (and since $c > 0$, the largest divisor of c is clearly c itself).

However, if we just *defined* $\gcd(a, b)$ to be the largest element of this set, it wouldn't be clear that all other elements of this set divided it.

Corollary 1.17 (Bézout's identity) *We can always write* $\gcd(a, b)$ *in the form* $ma + nb$*, for some* $m, n \in \mathbb{Z}$*.*

Proof Clear from the proof of existence above. □

Example 1.18 Since 11 and 13 are distinct primes, their GCD must be 1; and indeed we have $6 \cdot 11 + (-5) \cdot 13 = 66 - 65 = 1$.

Exercise 1.19 (Basic Properties of GCD) For all $a, b, k \in \mathbb{Z}$:

(i) $\gcd(a, b) = gcd(b, a) = gcd(|a|, |b|)$;
(ii) $\gcd(ka, kb) = |k| \gcd(a, b)$;
(iii) $\gcd(a, 0) = |a|$ and $\gcd(a, 1) = 1$;
(iv) $\gcd(a, b) = \gcd(a, b + ka)$.

Pairs of numbers whose greatest common divisor is 1 are quite special. We call such pairs *coprime*.

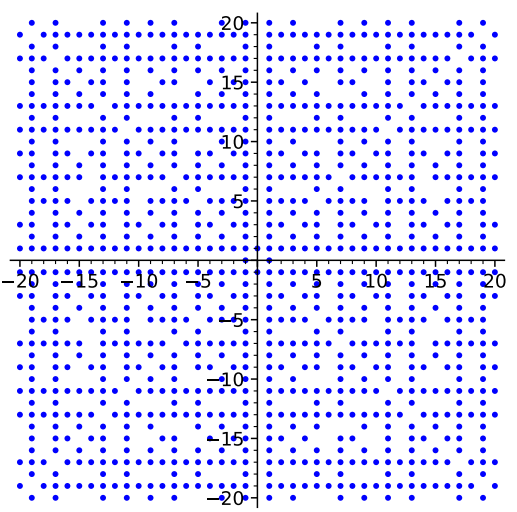

FIGURE 1.1. Pairs of coprime integers (m, n) with $\max(|m|, |n|) \leqslant 20$

Lemma 1.20 (Euler's Lemma) *If* $a \mid bc$*, and* a *and* b *are coprime, then* $a \mid c$*.*

Proof Write $1 = am + bn$. Then $c = c(am + bn) = (mc)a + n(bc)$. Since $a \mid bc$, it divides both terms in the sum, so it divides c. □

Exercise 1.21 Show that if x has the form $ma+nb$, for some $m, n \in \mathbb{Z}$, and x divides both a and b, then $x = \pm \gcd(a, b)$.

1.3 Euclid's algorithm

From our existence proof of the GCD, it's very difficult to see how one could compute it explicitly: we're asking for the smallest element of an infinite set. We can do slightly better using Theorem 1.16 – in principle we can make a list of the (finitely many) divisors of both a and b, and find the greatest element appearing in both lists. But there is a way to do much better.

Proposition 1.22 *Let $a, b \in \mathbb{Z}$ with $a \neq 0$, and suppose $b = aq + r$ for some q, r. Then*

$$\gcd(a, b) = \gcd(a, r).$$

Proof Clear from Exercise 1.19 (iv). □

If $r \neq 0$ then we can now repeat the process, replacing the larger number with its remainder on division by the smaller. Since the quantity $\max(|a|, |b|)$ gets strictly smaller each time, we must eventually reach a remainder of zero; and since $\gcd(a, 0) = |a|$ for all a, we are done.

It's convenient to arrange this in a table. Suppose we want to calculate $\gcd(113, 251)$. Then we write

$$\begin{aligned} 251 &= 2\times 113 + 25 \\ 113 &= 4\times 25 + 13 \\ 25 &= 1\times 13 + 12 \\ 13 &= 1\times 12 + 1 \\ 12 &= 12\times 1 + 0. \end{aligned}$$

Note how the numbers move diagonally to the left each time. The grey numbers (the q's in the division-with-remainder steps) aren't important for calculating the GCD (although we'll find a different use for them in a moment); the key things are the remainders.

We claim that *the last non-zero remainder* in the table is always equal to the GCD of the original two numbers. In the above example, this is 1, so 251 and 113 are coprime. To see this, apply the last proposition repeatedly, once for each division step:

$$\begin{aligned} \gcd(251, 113) &= \gcd(113, 25) \\ &= \gcd(25, 13) \\ &= \gcd(13, 12) \\ &= \gcd(12, 1) \\ &= \gcd(1, 0) = 1. \end{aligned}$$

So we have a method for computing GCD's: *Euclid's algorithm*. It's actually a very effective algorithm in practice.

Exercise 1.23 Recall the *Fibonacci numbers*, defined by $F_0 = 0$, $F_1 = 1$, and $F_n = F_{n-1} + F_{n-2}$ for $n \geqslant 2$. Show that $\gcd(F_n, F_{n-1}) = 1$ for all $n \geqslant 1$.

Now let's see how to use the grey numbers. Working up the table from the last-but-one row, we have

$$\begin{aligned} 1 &= 13 - 1 \times 12 \\ &= 13 - 1 \times (25 - 1 \times 13) && = -1 \times 25 + 2 \times 13 \\ &= -1 \times 25 + 2 \times (113 - 4 \times 25) && = \ 2 \times 113 - 9 \times 25 \\ &= 2 \times 113 - 9 \times (251 - 2 \times 113) && = -9 \times 251 + 20 \times 113 \end{aligned}$$

So we've written 1 as a sum of integer multiples of 251 and 113. This is a "free bonus" that Euclid's algorithm gives us: for any a, b, we can compute an expression for $\gcd(a, b)$ in the form $ma + nb$.

Remark 1.24 Finding these m, n (as well as just the GCD itself) is so useful that it has its own name: computing the triple $(\gcd(a, b), m, n)$ is called the *extended GCD problem* (XGCD). Lots of computer algebra systems have a command called `xgcd`, or something similar[a], which computes this in one step.

[a] Not to be confused with `xkcd`, an online comic strip popular with mathematicians.

Exercise 1.25 Show that 351 and 451 are coprime, and find integers m, n such that $351m + 451n = 1$.

CHAPTER 2

Prime numbers and unique factorisation

2.1 Prime numbers

I'm sure you all know this definition:

Definition 2.1 An integer $p \in \mathbb{N}$ is said to be *prime* if $p > 1$, and the only divisors of p in $\mathbb{N}$ are 1 and p itself. We write $\mathbb{P}$ for the set of primes.

The first few elements of $\mathbb{P}$ are $\{2, 3, 5, 7, 11, \dots\}$. A number which is not prime is said to be *composite*.

Exercise 2.2 Show that if $p > 1$ and p is not divisible by any integer a with $1 < a \leqslant \sqrt{p}$, then p is prime. Use this to show that 127 is prime. (Hint: $127 < 12^2 = 144$.)

Remark 2.3 It is conjectured, but not known, that there are infinitely many *twin primes* – that is, pairs (p, q) of primes with $q = p + 2$, such as $(59, 61)$. ⊛

We're going to show that any $n \in \mathbb{N}_+$ can be written *uniquely* in terms of the primes. First, we need a lemma:

Lemma 2.4 *Suppose p is prime, and $a, b \in \mathbb{Z}$. If $p \mid ab$, then $p \mid a$ or $p \mid b$.*

Proof Clearly $\gcd(p, a)$ is a divisor of p, so it must be 1 or p. If $\gcd(p, a) = p$, then $p \mid a$ and we're done. If $\gcd(p, a) = 1$, then Euler's lemma (Lemma 1.20) applies and shows that $p \mid b$. □

This extends in the obvious way to products of three or more factors: if $p \mid a_1 \dots a_r$, then $p \mid a_i$ for some i.

Exercise 2.5 Prove the converse: if $p > 1$ and p is *not* prime, there exist integers a, b with $p \mid ab$ but $p \nmid a$ and $p \nmid b$.

Remark 2.6 Lemma 2.4, and its converse, show that a positive integer $n \in \mathbb{N}_+$ is a prime number iff it is a *prime element* of the ring $\mathbb{Z}$ in the sense of the *M11 Algebra* course.

FIGURE 2.1. The insect *Magicicada septendecim* lives most of its life underground, emerging in huge swarms every 17 years to mate and die. A related species *M. tredecim* has a 13-year cycle. There are various theories why these insects have evolved to use prime numbers of years.

2.2 Unique factorisation

Theorem 2.7 (Fundamental Theorem of Arithmetic) *Every positive integer n can be written as a product of prime numbers, and its factorisation into primes is unique up to the order of the factors.*

Note that this includes $n = 1$, which is an empty product (the product of no primes); and the primes themselves, with only one factor in the product.

Proof *Existence*: Let $n \in \mathbb{N}_+$. By Strong Induction, we may suppose the theorem is true for all m with $m < n$.

If $n = 1$, then the statement is trivial (product of no primes). So let's suppose $n > 1$. If n is prime, we're again fine (product of one prime). So n must be of the form ab with $1 < a, b < n$. By the induction hypothesis, both a and b are products of primes, hence so is n.

Uniqueness: Suppose $n = p_1 p_2 \dots p_r = q_1 q_2 \dots q_s$ are two prime factorisations of n. We want to deduce that $s = r$ and the q's can be re-ordered such that $q_i = p_i$. We shall argue by induction on r.

If $r = 0$, then $n = 1$; thus $s = 0$ as well (since any nontrivial product of primes is > 1) so we're done.

Now suppose $r \geqslant 1$ and the theorem is true for $r - 1$. Then $p_r \mid q_1 \dots q_s$. Hence $p_r \mid q_i$ for some i, and after reordering we may suppose $p_r \mid q_s$. Since q_s is prime (and $p_r > 1$), this implies $p_r = q_s$. Since p_r is not zero, we deduce that $p_1 \dots p_{r-1} = q_1 \dots q_{s-1}$. By the induction hypothesis, $r - 1 = s - 1$, so $r = s$; and $q_1, \dots, q_{s-1}$ are $p_1, \dots, p_{r-1}$ in some order. So we are done. □

Collecting together any powers of primes which occur in a prime factorization, we obtain two alternative formulations:

Corollary 2.8 *Every positive integer n may be expressed uniquely in the form*

$$n = p_1^{r_1} p_2^{r_2} \dots p_k^{r_k}$$

where $k \geqslant 0$, $p_1, \ldots, p_k$ are primes with $p_1 < p_2 < \cdots < p_k$, and r_i are integers with $r_i \geqslant 1$.
Alternatively, every positive integer n may be expressed uniquely in the form

$$n = \prod_{p \in \mathbb{P}} p^{e_p}$$

where $e_p \in \mathbb{N}$ for all p, and all but finitely many e_p are zero. □

The exponent e_p which appears in this standard factorization of n is denoted $\mathrm{ord}_p(n)$; it is characterized by the following property:

$$e = \mathrm{ord}_p(n) \iff p^e | n \text{ and } p^{e+1} \nmid n.$$

For example, $700 = 2^2 \cdot 5^2 \cdot 7$, so $\mathrm{ord}_2(700) = \mathrm{ord}_5(700) = 2$, $\mathrm{ord}_7(700) = 1$, and $\mathrm{ord}_p(700) = 0$ for primes $p \neq 2, 5, 7$. Every positive integer n is uniquely determined by the sequence of exponents $\mathrm{ord}_p(n)$. From the uniqueness of the factorisation, it follows that

$$\mathrm{ord}_p(mn) = \mathrm{ord}_p(m) + \mathrm{ord}_p(n) \qquad \forall m, n \in \mathbb{N}_+, p \in \mathbb{P}. \tag{2.1}$$

Proposition 2.9 *Let $m, n \in \mathbb{N}_+$. Then $m \mid n$ iff $\mathrm{ord}_p(m) \leqslant \mathrm{ord}_p(n)$ for all $p \in \mathbb{P}$.*

Proof If $m \mid n$, then $n = km$ for some $k \in \mathbb{N}_+$. From (2.1) it follows that $\mathrm{ord}_p(n) = \mathrm{ord}_p(k) + \mathrm{ord}_p(m) \geqslant \mathrm{ord}_p(m)\ \forall p$.

Conversely, if $\mathrm{ord}_p(m) \leqslant \mathrm{ord}_p(n)$ for every p, let $k = \prod_p p^{\mathrm{ord}_p(n) - \mathrm{ord}_p(m)}$, which is in $\mathbb{N}_+$ since all the exponents are non-negative (and all but finitely many of them are zero). Then we have $n = km$. □

Corollary 2.10 *We have $\gcd(m, n) = 1$ iff there is no prime which divides both m and n.*

Proof The primes which divide $\gcd(m, n)$ are precisely the primes dividing both m and n, by the characterising property of the gcd. It follows from the existence of prime factorisations that for any $k \in \mathbb{N}_+$, we have $k > 1$ iff some prime divides k; applying this to $k = \gcd(m, n)$ we are done. □

Exercise 2.11 Show that for any $m, n \in \mathbb{N}_+$, we have

$$\gcd(m, n) = \prod_{p \in \mathbb{P}} p^{\min(\mathrm{ord}_p(m), \mathrm{ord}_p(n))}.$$

2.3 Infinitude of primes

No introductory course on number theory could possibly omit the following theorem:

Theorem 2.12 (Euclid) *There are infinitely many primes.*

Proof Suppose there are only finitely many primes $p_1, \dots, p_k$. Consider the integer $N = (p_1 p_2 \dots p_k) + 1$. Then all the p_i divide $N - 1$; so none of them can divide N (since otherwise they'd have to divide 1). But $N > 1$, so N must have some prime factors. This contradicts our assumption that $\{p_1, \dots, p_k\}$ are all the primes. □

There are lots of variants of this argument which can be used to construct primes with some special shape; we'll see a few in the next chapter.

Remark 2.13 Although there are infinitely many primes, they get "thinner and thinner" as you go further out. Gauss and Legendre conjectured around 1800 that the ratio

$$\frac{\#\{p \in \mathbb{P} : p \leqslant X\}}{X/\log X}$$

tends to 1 as $X \to \infty$. So for large X, the fraction of integers up to X which are prime is roughly $1/\log(X)$, which tends very slowly to 0.
This conjecture was open for over 100 years, until it was finally proved by Hadamard and de la Vallée Poussin in 1896. A measure of the importance of this theorem is that, among all of the thousands of theorems about prime numbers, theirs is universally known as "*the* prime number theorem".

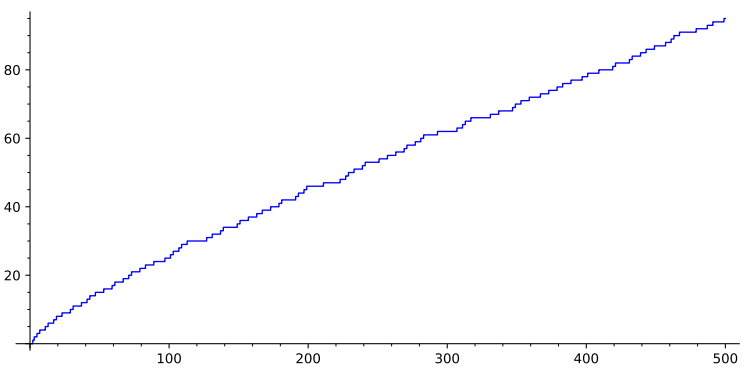

FIGURE 2.2. Graph of the number of primes $\leqslant x$, as a function of x, for $x \leqslant 500$.

Exercise 2.14 Does there exist $n \in \mathbb{N}$ such that all of the numbers $n + 1, n + 2, \dots, n + 20$ are composite?

CHAPTER 3

Congruences and modular arithmetic

3.1 Congruences

The following definition (originally due to Gauss) is a wonderful way of simplifying and organising lots of number-theoretic arguments:

Definition 3.1 Let $a, b, m \in \mathbb{Z}$, with $m \geqslant 1$. We say "a is congruent to b modulo m" if m divides $a - b$ (i.e. there exists $k \in \mathbb{Z}$ such that $a - b = km$).

Example 3.2 For example, a is congruent to 0 modulo 2 iff it's even, and to 1 modulo 2 iff it's odd.

It's easy to see that, for a fixed m, this is an *equivalence relation* in a and b. So the equivalence classes (the *congruence classes* modulo m) form a partition of $\mathbb{Z}$ into disjoint sets. There's exactly m of these congruence classes, represented by the integers $\{0, 1, \dots, m-1\}$, corresponding to the different remainders of a on division by m.

3.2 Modular arithmetic

Definition 3.3 We write $\mathbb{Z}/m\mathbb{Z}$ for the set of congruence classes modulo m.

You saw in *M11 Algebra* that this is a ring: the set $m\mathbb{Z}$ of multiples of m is an ideal of $\mathbb{Z}$, and $\mathbb{Z}/m\mathbb{Z}$ is the corresponding quotient ring. Moreover, the map $\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z}$, sending a to its congruence class, is a ring homomorphism.

Remark 3.4 Take a moment to reflect on what this is really saying: it's saying that, for $a, b \in \mathbb{Z}$, the congruence classes of $a \pm b$ and ab are uniquely determined by the congruence classes of a and b.

That might sound like a lot of abstract nonsense; but it's actually immensely useful for solving concrete questions about $\mathbb{Z}$.

Example 3.5 *"Do there exist integer solutions to the equation $x^2 - 3y^2 = 2$?"*

Suppose (x, y) was a solution. Then, reducing modulo 3, we would have a solution to the equation $(x \bmod 3)^2 = 2$ in $\mathbb{Z}/3\mathbb{Z}$. But $x \bmod 3$ must be one of $\{0, 1, 2\}$, and we have $0^2 = 0$, $1^2 = 2^2 = 1$ in $\mathbb{Z}/3\mathbb{Z}$. So there are no solutions.

Exercise 3.6 Show that if $m, n \in \mathbb{N}_+$ with $n \mid m$, then there is a unique ring homomorphism $\mathbb{Z}/m\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ sending the congruence class $a + m\mathbb{Z}$ to $a + n\mathbb{Z}$ for every $a \in \mathbb{Z}$.

3.3 Primes in congruence classes

Notice that if p is a prime, and $p \neq 2$, then p is odd, so p must be either $1 \bmod 4$ (like 5) or $3 \bmod 4$ (like 7).

Theorem 3.7 *There are infinitely many primes p with $p = 3 \bmod 4$.*

Proof Suppose there are finitely many such primes, namely $p_1, \ldots, p_k$ (with $p_1 = 3$, $p_2 = 7$, etc). Consider the product $N = 4p_2 \ldots p_k + 3$ (note p_1 is not included!)

Clearly N can't be divisible by any of the primes $p_2, \ldots, p_k$, since these all divide $4p_2 \ldots p_k$ but don't divide 3. Moreover, it's also not divisible by $p_1 = 3$ either (since 3 doesn't divide $4p_2 \ldots p_k$, but does divide 3). Finally, it is clearly odd and thus not divisible by 2 either. Hence all of its prime factors must be $1 \bmod 4$.

However, a product of numbers that are all $1 \bmod 4$ must itself be $1 \bmod 4$, while N is obviously $3 \bmod 4$. So we have a contradiction. □

Exercise 3.8 Why doesn't this argument adapt to show that there are infinitely many primes which are $1 \bmod 4$?

Remark 3.9 This is a special case of a much more general theorem: for any $a, b \in \mathbb{N}_+$ with $\gcd(a, b) = 1$, there are infinitely many primes p with $p = a \bmod b$ (Dirichlet's theorem on primes in arithmetic progressions.) However, this is a rather deep theorem and we won't prove it in this module.

3.4 The Chinese remainder theorem

The next theorem will tell us that if m and n are coprime, then congruences mod m and congruences mod n are in some sense "independent of each other": they give totally complementary information.

Theorem 3.10 (Chinese remainder theorem, or CRT) *Let $m, n \in \mathbb{N}_+$ be coprime, and let $x, y \in \mathbb{Z}$. Then there exist integers a such that $a = x \bmod m$ and $a = y \bmod n$; and the set of integers a with this property forms a congruence class modulo mn.*

Remark 3.11 The theorem has this name because it was discovered by ancient Chinese mathematicians (long before it was known in Europe); there is a complete proof in Qin Jiushao's *Mathematical Treatise in Nine Sections* from 1247.

Exercise 3.12 Find an integer a satisfying $a = 5 \bmod 7$ and $a = 6 \bmod 9$.

Proof *Existence*: We first show that there exist integers r, s with the following property:

- $r = 1 \bmod m$ and $r = 0 \bmod n$;
- $s = 0 \bmod m$ and $s = 1 \bmod n$.

To see this, use Euclid's algorithm to write $1 = um + vn$. Then we can take $r = vn$, since $vn = 1 - um = 1 \bmod m$ and clearly $vn = 0 \bmod n$. Similarly, we can take $s = um$. This proves the claim.

Having proved the claim, for any x, y we can take $a = rx + sy$.

Uniqueness: If a is one solution, then for any integer a', it follows that a' is a solution iff $a - a'$ is divisible by both m and n. Since m and n are coprime, the set of integers that are divisible by both m and n is precisely the set of integers divisible by mn. So the set of solutions is precisely the congruence class of $a \bmod mn$, as claimed. □

Remark 3.13 (i) In more abstract language, we've shown that the natural map from $\mathbb{Z}/mn\mathbb{Z}$ to the *direct product* $(\mathbb{Z}/m\mathbb{Z}) \times (\mathbb{Z}/n\mathbb{Z})$ is a bijection. Since it's also a ring homomorphism, these two rings are isomorphic.

(ii) Note that we can compute everything here explicitly, using Euclid's algorithm applied to (m, n) as the starting point.

(iii) By induction on k, one can prove the following more general theorem: if $m_1, \dots, m_k \in \mathbb{N}_+$ are *pairwise coprime*[a], and $x_1, \dots, x_k$ are arbitrary integers, then we can find an $a \in \mathbb{Z}$ with $a_i = x_i \bmod m_i$ for all i, and this a is uniquely determined modulo $m_1 m_2 \dots m_k$.

In particular, if $m = \prod_{i=1}^{k} p_i^{e_i}$ is the prime factorisation of m, then

$$\mathbb{Z}/m\mathbb{Z} \cong (\mathbb{Z}/p_1^{e_1}\mathbb{Z}) \times \cdots \times (\mathbb{Z}/p_k^{e_k}\mathbb{Z}).$$

[a] This means that there is no integer > 1 which divides *more than one* of the m_i. This is strictly stronger than requiring that there is no integer > 1 dividing *all* of the m_i; e.g. if $(m_1, m_2, m_3) = (6, 10, 15)$, then any two of the m_i have a prime in common, but there is no prime dividing all three.

CHAPTER 4

The group of units mod m

4.1 Inverses modulo m

Recall that if R is a (commutative) ring, an element $r \in R$ is said to be *invertible*, or a *unit* in R, if there exists $s \in R$ such that $rs = 1$.

Proposition 4.1 *Let $m \in \mathbb{N}_+$, and $a \in \mathbb{Z}$. Then $a \bmod m$ is invertible in $\mathbb{Z}/m\mathbb{Z}$ if and only if $\gcd(a, m) = 1$.*

Proof We have

$$\begin{aligned} \gcd(a, m) = 1 &\iff \exists u, v \in \mathbb{Z} \text{ such that } ua + vm = 1 \\ &\iff \exists u, v \in \mathbb{Z} \text{ such that } ua = 1 - vm \\ &\iff \exists u \in \mathbb{Z} \text{ such that } ua = 1 \bmod m \\ &\iff a \bmod m \text{ is invertible.} \end{aligned}$$

□

In particular, if p is prime, then any non-zero element in $\mathbb{Z}/p\mathbb{Z}$ is invertible, so $\mathbb{Z}/p\mathbb{Z}$ is not just a ring but a *field* (and conversely, if n is non-prime, then $\mathbb{Z}/n\mathbb{Z}$ is not a field.)

Definition 4.2 For p prime we frequently use the alternative notation $\mathbb{F}_p$ for $\mathbb{Z}/p\mathbb{Z}$ (to emphasise that it is a field).

Remark 4.3 One can show that for any $p \in \mathbb{P}$ and $k \geqslant 1$, there exists a finite field of size p^k, unique up to isomorphism (and any finite field must be one of these). It's conventional to denote this field by $\mathbb{F}_{p^k}$; but be warned that if $k > 1$, then $\mathbb{F}_{p^k}$ and $\mathbb{Z}/p^k\mathbb{Z}$ aren't the same thing (one is a field and the other is not). Some of you may have seen the field $\mathbb{F}_4$ before, as an example in the *Linear Algebra* module.
We won't use the fields $\mathbb{F}_{p^k}$ for $k > 1$ in this course (except very briefly in section 7.3); but you might encounter them in textbooks, so it's worth being aware that if $k > 1$ the notations $\mathbb{F}_{p^k}$ and $\mathbb{Z}/p^k\mathbb{Z}$ both mean something, but they aren't the same object.

4.2 The group of units and the φ function

Recall that for R a ring, the symbol $R^\times$ denotes the set of invertible elements in R, and this is naturally a group under multiplication (an abelian group if R is commutative).

Definition 4.4 We write $U_m = (\mathbb{Z}/m\mathbb{Z})^\times$ for the units in $\mathbb{Z}/m\mathbb{Z}$ (as a group under multiplication); and we define a function $\varphi : \mathbb{N}_+ \to \mathbb{N}_+$ by $\varphi(m) = \#U_m$.

Concretely, $\varphi(m)$ is the number of integers in the range $\{0, \dots, m-1\}$ which are coprime to m. (By convention $\varphi(1) = 1$.)

Example 4.5

- We have $\varphi(12) = 4$, since the only integers in the range $\{0, \dots, 11\}$ that are coprime to 12 are $\{1, 5, 7, 11\}$.
- If p is prime, then $\varphi(p) = p - 1$, since every non-zero integer $< p$ is coprime to p.

Exercise 4.6 Notice that $\varphi(12)/12 = \frac{1}{3}$ is quite small. Can you find an integer with $\varphi(n)/n < \frac{1}{4}$?

Proposition 4.7 *If m, n are coprime, then we have an isomorphism $U_{mn} \cong U_m \times U_n$ (direct product of groups). In particular, $\varphi(mn) = \varphi(m)\varphi(n)$.*

Proof Thanks to the Chinese remainder theorem, we know that the rings $\mathbb{Z}/mn\mathbb{Z}$ and $(\mathbb{Z}/m\mathbb{Z}) \times (\mathbb{Z}/n\mathbb{Z})$ are isomorphic. It follows that their unit groups are isomorphic; but the unit group of $(\mathbb{Z}/m\mathbb{Z}) \times (\mathbb{Z}/n\mathbb{Z})$ is obviously just $U_m \times U_n$. □

This means $\varphi(n)$ is determined for all n by its values when n is a prime power, which are computed as follows:

Proposition 4.8 *If $n = p^k$ is a prime power, then $\varphi(p^k) = p^{k-1}(p-1)$.*

Proof An integer is coprime to p^k iff it is not a multiple of p. Out of the p^k integers $\{0, 1, \dots, p^{k-1} - 1\}$, exactly p^{k-1} of them are multiples of p. So $\varphi(p^k) = p^k - p^{k-1} = p^{k-1}(p-1)$. □

Exercise 4.9

(a) Show that for any k there are only finitely many n with $\varphi(n) = k$.
(b) Does there exist an $n \in \mathbb{N}_+$ with $\varphi(n) = 14$?

Remark 4.10 *Carmichael's conjecture* is that for any k, if the equation $\varphi(n) = k$ has any solutions, then it has at least two solutions. (This has been an open problem for over 100 years.) ⊛

One of the main reasons for introducing φ is the following:

Theorem 4.11

(i) (**Euler's theorem**)*: Let $m \in \mathbb{N}_+$. Then for all $a \in \mathbb{Z}$ coprime to m, we have $a^{\varphi(m)} = 1 \bmod m$.*

(ii) (**Fermat's little theorem**)*: Let $p \in \mathbb{P}$. Then for all $a \in \mathbb{Z}$ with $p \nmid a$, we have $a^{(p-1)} = 1 \bmod p$. Moreover, for any $a \in \mathbb{Z}$ we have $a^p = a \bmod p$.*

Proof Euler's result is just Lagrange's theorem from group theory ("the order of any element of a group divides the size of the group") applied to the group U_m.

For Fermat's little theorem, specialising Euler's theorem shows that $a^{p-1} = 1 \bmod p$ for all a coprime to p, and it follows that $a^p = a \bmod p$. On the other hand, if a is not coprime to p, then $p \mid a$, so $a^p = a = 0 \bmod p$ and the result holds in this case too. □

Exercise 4.12 If $n \in \mathbb{N}$ satisfies $n > 1$ and $a^n = a \bmod n$ for all a, but n is not prime, then n is said to be a *Carmichael number*.

- Show that 561 is a Carmichael number. (Note $561 = 3 \times 11 \times 17$).
- Prove that the product of two distinct primes cannot be Carmichael.

4.3 Primitive roots

We'll now prove an important result about the structure of U_p for p prime. First we need a preparatory lemma:

Lemma 4.13 *For any $n \in \mathbb{N}_+$, we have*

$$\sum_{\substack{d \in \mathbb{N}_+ \\ d \mid n}} \varphi(d) = n.$$

Proof For each d dividing n, the map $r \mapsto \frac{n}{d} \cdot r$ gives a bijection between the sets

$$S_d = \{r \in \{0, \dots, d-1\} : \gcd(r, d) = 1\}$$

and

$$T_d = \left\{s \in \{0, \dots, n-1\} : \gcd(s, n) = \tfrac{n}{d}\right\}.$$

So we have $\#T_d = \#S_d = \varphi(d)$. However, each $s \in T = \{0, \dots, n-1\}$ lies in exactly one of the sets T_d; so the sum of their sizes must be $\#T = n$. □

Theorem 4.14 *If p is prime, then U_p is a* cyclic *group. That is, there exists an element $g \in U_p$ such that every $x \in U_p$ is equal to some power of g.*

Such a g is called a *primitive root* mod p.

Proof Note that a is a primitive root iff the order of a in U_p is exactly $p - 1$ (so Euler's theorem is the "best possible" bound).

Let $n = p - 1$; and for $d \mid n$, let $\psi(d)$ denote the number of elements of U_p whose order is precisely d. We claim that for any $d \mid n$, either $\psi(d) = 0$, or $\psi(d) = \varphi(d)$.

To see this, suppose $\psi(d) > 0$. Then there exists *some* element a of order exactly d. Hence the set $\{1, a, \dots, a^{d-1}\}$ has d distinct elements, and all of them have order dividing d; that is, they are roots of $X^d - 1$. Since this polynomial has degree d (and $\mathbb{F}_p$ is a field), it can't have more than d roots in $\mathbb{F}_p$. So our set is actually *all* of the elements of U_p of order dividing d. In particular, $\psi(d)$ is the number of h in $\{0, \dots, d-1\}$ such that a^h has order exactly d. However, a^h has order exactly d iff h is coprime to d; so we conclude that $\psi(d) = \varphi(d)$.

So it certainly follows that $\psi(d) \leqslant \varphi(d)$ for every d. But every element of U_n must have some order, so

$$\sum_{d|n} \psi(d) = n = \sum_{d|n} \varphi(d).$$

It follows that in fact $\psi(d) = \varphi(d)$ for all d, and in particular $\psi(n) = \varphi(n)$. As $\varphi(n) > 0$, this shows that elements of order exactly n exist. □

Example 4.15 The integer 2 is a primitive root mod 11: we have

$$\{1, 2, 2^2, \dots, 2^{10}\} = \{1, 2, 4, 8, 5, 10, 9, 7, 3, 6\} = U_{11}.$$

However, 2 isn't a primitive root modulo 7.
(*Artin's primitive root conjecture* predicts that there are infinitely many primes p such that 2 is a primitive root mod p. This is an open problem. ⊛)

Exercise 4.16 (hard!) The converse of Theorem 4.14 is false: for instance, $(\mathbb{Z}/18\mathbb{Z})^\times$ is cyclic (but 18 is clearly not prime). Can you classify, in terms of their prime factorisations, which integers n have the property that U_n is cyclic?

CHAPTER 5

Computing in U_n and RSA cryptography

As well as being interesting just from a pure theoretical standpoint, the group of units U_n is highly important in a major real-world application of number theory: *cryptography* – devising codes for securely transmitting secret information.

5.1 Powers mod n

Suppose we want to compute $3^{123456789} \bmod 7$ (more precisely: to compute the unique representative in $\{0, \dots, 6\}$ of its congruence class). How might we do this? One obvious idea would be to compute $3^{1234\cdots}$ as an integer, and then reduce it modulo 7.

This would work, eventually, but it would be a horrendous mess, because $3^{123456789}$ is *huge*, with millions of digits. So we need a better approach.

Using the φ function: Since 7 is prime, we know that $\varphi(7) = 6$; and $123456789 = 3 \bmod 6$, so it is $6q + 3$ for some q. Since 3 is coprime to 7, we conclude that

$$3^{123456789} = 3^{6q+3} = (3^6)^q \cdot 3^3 = 1^q \cdot 27 = 6 \bmod 7.$$

This algorithm works very well if the modulus n is small (but the exponent is large), as in the previous example. But if n is a bit bigger, there are two problems.

Example 5.1 Compute $3^{123456789} \bmod 21311$.

Here we hit two snags. Firstly, to compute $\varphi(21311)$, we have to factor 21311 into primes (which is doable on a computer, but takes a while, and would rapidly become impractical for larger moduli). Secondly, even once we've computed $\varphi(21311) = 21000$ and $123456789 \bmod 21000 = 18797$, we still have to compute 3^{18797}, which has about 9000 digits! So this is clearly not a sensible method.

Instead, we'll use a method called **repeated squaring**. The idea is to write the exponent as a *sum of powers of 2*, which we can always do; this is just the binary expansion of n. (This is easy to compute from the base-10 expansion, and if you're working on a computer, the computer probably converted your input to binary as soon as you entered it.) Now, we can easily make a table of values of $3^{(2^i)} \bmod 21311$ for small i by repeated

squaring:

$$\begin{aligned} 3^2 &= 3^2 = 9 \\ 3^4 &= 9^2 = 81 \\ 3^8 &= 81^2 = 6561 \\ 3^{16} &= 6561^2 = 19812 \\ 3^{32} &= 19812^2 = 9346 \\ 3^{64} &= 9346^2 = 15238 \\ &\vdots \end{aligned}$$

Because we reduce modulo $n = 21311$ *after every squaring step*, we never have to deal with integers bigger than n^2, so the computations are manageable. Once we have computed a table of $3^{(2^i)}$ for all i up to 26, we can use the formula

$$123456789 = 2^{26} + 2^{25} + 2^{24} + 2^{22} + \cdots + 2^2 + 1,$$

to compute $3^{123456789} = 20878 \bmod 21311$.

Remark 5.2 There's nothing very special about U_n here: if G is a finite group, and you have a practical way of representing elements of G on a computer and calculating the group operation, then you can use repeated squaring to efficiently compute g^n for any $g \in G$ and $n \in \mathbb{N}$.

5.2 Polynomial vs. exponential time

To formalise the ideas of "hard to compute" versus "easy to compute", we use the notion of *polynomial-time* and *exponential-time* algorithms. These compare the number of steps needed for some computational method, as a function of the *length of the input* (the amount of space required to write it down) – e.g. the number of decimal (or binary) digits needed to write down an integer. We say some algorithm is *polynomial-time* if the number of steps required, for input of length N, is bounded above by a constant multiple of N^k for some constant k. Similarly, if it's bounded above by a constant multiple of C^N for some C, we say it's *exponential-time*. Since exponentials grow much faster than polynomials, any polynomial-time algorithm will eventually beat any exponential-time one.

Remark 5.3 Note that since we ignore constant factors, it doesn't matter exactly how we measure the input length, as long as we stay within a constant factor of the original measure. E.g. if the input is a number, we could count its decimal digits, or its binary digits (bits); since these differ by a factor $\log_2 10$, this does not change whether an algorithm is polynomial or exponential time.

For example, computing the product $a \cdot b$ of two integers via the standard school-book "long multiplication" method requires approximately $N_a N_b$ steps, where N_r is the number of binary digits of r. Since $N_a N_b \leqslant \frac{1}{4}(N_a + N_b)^2$, and $N_a + N_b$ is the total length of the input, this is clearly a polynomial-time algorithm. The "repeated squaring" algorithm above, for computing $a^b \bmod N$, is also polynomial-time.

On the other hand, testing whether a number r is prime by trying all potential factors up to $\sqrt{r}$ ("trial division") involves at least $\sqrt{r}$ steps, which is clearly exponential in N_r.

There's a big difference here between *primality testing* – answering the yes/no question "is N prime?" – and *factorisation* – computing the prime factors of N. These might seem like the same problem, but they aren't: there are situations where you know N cannot be prime without being able to produce a specific factor of N.

Example 5.4 For instance, suppose you compute $2^{N-1} \bmod N$, and it's not 1. Then N cannot be prime, since otherwise it would contradict Fermat's little theorem). So you know that N has a nontrivial factor; but there's no obvious way to work out what that factor *is* using the information you have about $2^{N-1} \bmod N$.

- Primality testing can be done in polynomial time. This was proved by Miller in 1976 assuming an open conjecture in analytic number theory, the *generalised Riemann hypothesis*. In 2004, Agrawal, Kayal and Saxena gave a different algorithm, for which they could prove unconditionally (without assuming any conjectures) that it gave the correct answer in polynomial time.
- It's widely believed that factorisation *cannot* be done in polynomial time on a conventional computer[1]. There are algorithms (such as the *Number Field Sieve*) which are much better than trial division, but they are still much slower than any polynomial time algorithm.

It is this "gap" – that the complexity of *factoring* integers grows much faster than the complexity of *testing* whether integers are prime – that is vitally important in many applications of number theory.

5.3 Public key cryptography

We'll now learn about applications to secure communication – the science of cryptography. This could be used by a spy sending intelligence reports back to his home base; or it could be something much more mundane, like you logging into your bank account from a smartphone. This has two steps: *encryption* – the process the sender uses to transform a message into a coded form – and *decryption*, the opposite process that recovers the readable text from the coded message.

Traditional cryptographic techniques (prior to the 1970's) relied on the existence of a *shared secret*: both *sender* and *recipient* needed to know some piece of information which, if revealed to an outsider, would allow them to read the secret message themselves. This can be difficult to achieve: it requires coordination in advance between the sender and recipient.

Remark 5.5 Sometimes the entire system *is* the shared secret; but then any security lapse means redesigning the whole system from scratch. So most practical cryptographic systems rely on choosing a "secret key" which is an arbitrary number, string of letters, etc; and then scrambling up the input message in a way that depends on this secret key. It doesn't matter if an attacker knows how the system

[1] "Conventional" as opposed to "quantum". Quantum computers could, theoretically, factorise large numbers much faster than any conventional computer could; but building quantum computers on a realistic scale has proved to be somewhat difficult. This would be an interesting project topic.

works, as long as they don't know the secret key that was used for a particular message. That way, if one of your agents is captured, you just need to choose a new key, not a whole new algorithm!

Public key cryptography refers to a class of systems where the information needed to *encrypt* a message is different from the information needed to *decrypt* it. In such systems, each participant has a *public key* and a *private key*. If Alice wants to send a message to Bob, she can encrypt it knowing only Bob's public key, but only someone knowing his private key can encrypt it again. So Bob doesn't need to tell Alice – or anybody else – what his private key is; and as long as he keeps his private key secret, he can announce his public key openly to the world, without compromising the security of the system.

Of course, such a system can only work if it is impossible to determine the private key from the public one without an impractically lengthy computation. This is where number theory comes in: primes and prime factorisations are a rich source of difficult calculations!

Remark 5.6 There are some obvious security holes, of course. If Bob asks Alice a question that has only a few possible answers (e.g. just "yes" or "no") then an attacker can try encrypting both "yes" and "no" with the public key. This will give two different gibberish messages, but if one of those exactly matches the gibberish message Alice has just radioed to Bob, then the attacker knows the message.

(This is typically solved by padding messages with randomly chosen nonsense phrases. However, this is not without its dangers too, as is shown by a famous cryptographic mix-up during a World War II naval battle, involving a nonsense padding phrase being accidentally interpreted as part of the message, changing its meaning completely.)

5.4 The RSA cryptosystem

The first practical public-key cryptosystem is the *RSA* algorithm, announced by Rivest, Shamir and Aldeman in 1977.[2]

RSA relies on the following observation: *factorising large numbers into primes is difficult*. If I give you two 20-digit numbers p, q, then you can compute $N = pq$ in a few minutes. But if I give you a 40-digit number, and tell you that it's the product of two 20-digit primes, then it would take a very long time indeed to compute those prime factors.

In RSA, each participant chooses the following data:

- two large prime numbers p, q;
- an *encryption exponent* e, with $1 < e < \varphi(pq) = (p-1)(q-1)$ and e coprime to $\varphi(pq)$.

They announce to the world the product $N = pq$ and the encryption exponent e, but keep the factors p and q secret. Using this secret information, they can compute the

[2] They were not in fact the first to discover it; 20 years later it was revealed that Brtitish security services had already discovered the algorithm in 1973, but kept the discovery secret, and Rivest et al. rediscovered it independently.

decryption exponent

$$d = e^{-1} \bmod \varphi(N).$$

Suppose one participant (Bob) wants to send information to another (Alice). Bob finds out Alice's modulus N and encryption exponent e. He converts his message into a series of chunks, each of which is represented by an integer m in the range $1 < m < N$, and for each chunk he computes

$$c = m^e \bmod N.$$

These c's are the encrypted message he transmits to Alice.

Alice then takes each chunk c and computes

$$c^d \bmod N = (m^e)^d = m^{de} \bmod N.$$

Since $de = 1 \bmod \varphi(N)$, this is just $m \bmod N$, recovering the original message.

The security of this system relies on the fact that it's impossible to compute $\varphi(N)$ from N without factorising N, and factorising large integers is hard – much harder than any of the other steps in the algorithm.

Example 5.7 Suppose Alice's public key is

$$N = 21311, \qquad e = 11$$

Bob wants to send the message "TINKER".
Bob converts this into 3-letter blocks 'TIN | KER' and converts each one into a number in base 26,

$$TIN = 13065, \qquad KER = 6881.$$

For the first block, he computes $13065^{11} = 2460 \bmod 21311$, and the second $6881^{11} = 14867 \bmod 21311$. So he sends the message 02460 14867.
Alice knows that $21311 = 101 \times 211$, so $\varphi(N) = 21000$, and hence the decryption exponent is 19091, since $11 \times 19091 = 21001$. So she just computes $2460^{19091} = 13065 \bmod N$, etc, and recovers the original message.

Remark 5.8 In real-world applications, p, q would be chosen so that N has roughly 600 digits (corresponding to 2048 binary bits). With keys this size, the encryption and decryption steps are still reasonably practical[a] (each encryption taking fractions of a second). However, to crack the code – computing the private key from the public one, by factorising N – would take longer than the age of the universe, even using all the computing power of Google's datacentres put together.

[a] That said, RSA is becoming less popular nowadays because other algorithms – typically based on *elliptic curves* – can offer similar levels of security while using smaller keys and quicker encryption/decryption times. The widely used elliptic-curve algorithm ECDSA, used with a 256-bit key, is estimated to be roughly as secure as RSA with a 3000-bit key.

CHAPTER 6

Quadratic residues

We're now going to investigate what the image of the *squaring* map $x \mapsto x^2$ on $\mathbb{Z}/m\mathbb{Z}$ looks like. The elements which are in the image have a special name:

Definition 6.1 We say $a \in \mathbb{Z}/m\mathbb{Z}$ is a *quadratic residue* (QR) modulo m if there exists $x \in \mathbb{Z}/m\mathbb{Z}$ with $x^2 = a$.

For example, in $\mathbb{Z}/6\mathbb{Z}$, we have

x	0	1	2	3	4	5
x^2	0	1	4	3	4	1

so $\{0, 1, 3, 4\}$ are quadratic residues mod 6, and $\{2, 5\}$ are not.

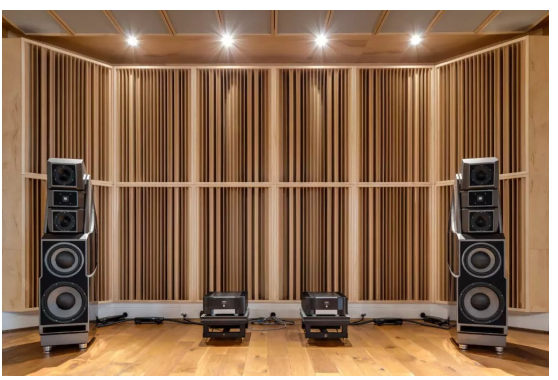

FIGURE 6.1. Quadratic residues are used in the design of echo-reducing wall panels for recording studios and concert halls. (Image: Dennis Foley, acousticfields.com)

6.1 Reducing to the prime case

From the Chinese remainder theorem, it's clear that if $n = \prod_i p_i^{e_i}$, then a is a QR mod n iff it's a QR modulo $p_i^{e_i}$ for all i. Rather less obvious is the following:

Proposition 6.2 *Let $p \in \mathbb{P}$ with $p > 2$, and let $a \in \mathbb{Z}$, with $p \nmid a$. If a is a QR mod p, then a is a QR mod p^k, for every $k \geqslant 1$.*

Proof Let's prove this by induction on k, the case $k = 1$ being true by assumption. So suppose $b \in \mathbb{Z}$ is such that $b^2 = a \bmod p^k$, for some $k \geqslant 1$, and let's try to cook up a solution modulo p^{k+1}.

By assumption, we have $b^2 = a + p^k r$, for some r. Let's consider integers of the form $b' = b + p^k s$. Then we have

$$(b')^2 = (b + p^k s)^2 = (a + p^k r) + 2bp^k s + p^{2k} s^2$$

and modulo p^{k+1} this is just $a + p^k(r + 2bs)$. So it suffices to show that we can choose s such that $r + 2bs = 0 \bmod p$.

Since $b^2 = a \neq 0 \bmod p$, and $p \neq 2$, it follows that $2b$ is a unit mod p, and we are done. □

Remark 6.3

- The argument breaks down for $p = 2$: if $a = 5$, then a is a QR modulo 2 and modulo 4, but not modulo 8. However, one can adapt the proof to show that an odd integer is a QR modulo every power of 2 iff it is 1 mod 8.
- The argument above is a preview of a much more general theorem called *Hensel's Lemma* which we'll see in the last chapter of the course.

6.2 QRs modulo primes

We can now concentrate on quadratic residues when the modulus is a **prime p with $p \neq 2$**. We first note that any nonzero quadratic residue mod p always has exactly 2 square roots mod p (if x is one, then $-x$ is the other, and $x \neq -x$). Since each unit mod p has to square to something, it follows that there are exactly $(p-1)/2$ nonzero quadratic residues; in other words, *exactly half* of the elements of U_p are squares.

Definition 6.4 Let p be an odd prime, and $a \in \mathbb{Z}$. Then the *Legendre symbol* is defined by

$$\left(\frac{a}{p}\right) = \begin{cases} 0 & \text{if } a = 0 \bmod p, \\ 1 & \text{if } a \text{ is a nonzero quadratic residue mod } p, \\ -1 & \text{if } a \text{ is a non-residue mod } p. \end{cases}$$

Then we have the following:

Theorem 6.5 (Euler, 1748) *We have*

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

Proof If $a = 0 \bmod p$ the result is obvious, so assume $p \nmid a$. Then $\left(a^{(p-1)/2}\right)^2 = 1 \bmod p$ by Fermat's little theorem, so $a^{(p-1)/2}$ must be either 1 or -1 modulo p.

If $a = b^2 \bmod p$ for some b, then $a^{(p-1)/2} = b^{(p-1)} = 1$, again by Fermat's little theorem. So every nonzero QR is a root of $X^{(p-1)/2} - 1 = 0$. However, since this polynomial has degree $(p-1)/2$, and we've just exhibited $(p-1)/2$ roots of it, there can't be any more. So all quadratic non-residues a must satisfy $a^{(p-1)/2} = -1 \bmod p$. □

Here's an easy consequence:

Proposition 6.6 *-1 is a quadratic residue modulo the odd prime p if $p = 1 \bmod 4$, and a non-residue if $p = 3 \bmod 4$.* □

Another important consequence is the *multiplicativity* of the Legendre symbol:

Corollary 6.7 *For any integers a, b we have*

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

Proof Since both sides are equal to 0 or ± 1, and $p > 2$, it suffices to show that $\left(\frac{ab}{p}\right)$ and $\left(\frac{a}{p}\right)\left(\frac{b}{p}\right)$ are congruent mod p. But this follows from Euler's criterion and the formula

$$(ab)^{(p-1)/2} = a^{(p-1)/2} b^{(p-1)/2}.$$ □

Remark 6.8 It's quite a strange and surprising thing that the product of two *non*-squares is always a square. This can be seen in an elementary way as follows. Take $a \in U_p$ which isn't a square, and consider the map $U_p \to U_p$ sending b to ab. This is a bijection; and it sends squares to non-squares, because if $b = x^2$ and $ab = y^2$ are both (nonzero) squares, then $a = (y/x)^2$ would have to be a square itself. Since there are equally many squares and non-squares, that "uses up" all the possible non-square images. Hence the non-squares have to go to squares, i.e. if b is non-square then ab is square.

Exercise 6.9 How many quadratic residues are there mod 15? How many of the *units* mod 15 are quadratic residues?

Give an example of integers a, b such that a, b and ab are all units and all quadratic non-residues mod 15.

Exercise 6.10 Suppose that there are finitely many primes p with $p = 1 \bmod 4$. By considering the integer $4(p_1 \dots p_k)^2 + 1$ where $\{p_1, \dots, p_k\}$ is the set of all such primes, deduce a contradiction.

CHAPTER 7

The reciprocity law

7.1 The statement

In the previous section the prime p was fixed, and we are asking "which a are quadratic residues mod p"? But we can also do something else: we can fix an integer a, and ask "for which (odd) primes $p \nmid a$ is a a quadratic residue mod p?" For instance, with $a = 5$, we see the following:

Residue: $\{11, 19, 29, 31, 41, 59, 61, 71, \dots\}$

Non-residue: $\{3, 7, 13, 17, 23, 37, 43, 47, \dots\}$

Notice the last digits! Amazingly, the answer seems to depend only on p mod 5 – which is strange, since the question is about $5 \bmod p$, not $p \bmod 5$, and these are totally different things.

If you try other values of a, the answer doesn't always depend on $p \bmod a$, but it's not far off – it suffices to know $p \bmod 4a$. This is the first hint at the following beautiful and important theorem:

Theorem 7.1 (Gauss' law of quadratic reciprocity) *If p, q are two distinct odd primes, then*

$$\left(\frac{p}{q}\right) \cdot \left(\frac{q}{p}\right) = (-1)^{\frac{(p-1)}{2} \cdot \frac{(q-1)}{2}} = \begin{cases} 1 & \text{if at least one of } p, q \text{ is } 1 \bmod 4 \\ -1 & \text{if both are } 3 \bmod 4. \end{cases}$$

Along with Gauss' law there are two related theorems (the "supplements to quadratic reciprocity"): one for $\left(\frac{-1}{p}\right)$ (which we have already proved as Proposition 6.6 above), and another for $\left(\frac{2}{p}\right)$ (which we will prove below). These say that for any odd prime p we have

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

and

$$\left(\frac{2}{p}\right) = (-1)^{\frac{(p^2-1)}{8}} = \begin{cases} 1 & \text{if } p = \pm 1 \bmod 8 \\ -1 & \text{if } p = \pm 3 \bmod 8. \end{cases}$$

The quadratic reciprocity law has many different proofs; Gauss himself published six different proofs in his lifetime, and hundreds more have been found since. However, none of them are particularly easy – whichever way you approach it, you have to do some genuine work. We'll give a proof shortly, which is quite close to one of Gauss' original arguments. First, we note that this does explain the observations above:

Corollary 7.2 *Let $a \in \mathbb{Z}$ be non-zero, and p, q odd primes, not dividing a, such that $p = q \bmod 4|a|$. Then $\left(\frac{a}{p}\right) = \left(\frac{a}{q}\right)$.*

Proof Considering the prime factorisation of $|a|$ and using the multiplicativity of the Legendre symbol, we may suppose that we are in one of three cases: $a = -1$, $a = 2$, or a is an odd prime. The first two cases are OK by the two supplementary laws, so we suppose we are in the third case.

Since $p = q \bmod 4|a|$, either $p = q = 1 \bmod 4$ or $p = q = 3 \bmod 4$. If $p = q = 1 \bmod 4$, or if $a = 1 \bmod 4$, then we have

$$\left(\frac{a}{p}\right) = \left(\frac{p}{a}\right) = \left(\frac{q}{a}\right) = \left(\frac{a}{q}\right).$$

If a, p, q are all $3 \bmod 4$, then we have similarly

$$\left(\frac{a}{p}\right) = -\left(\frac{p}{a}\right) = -\left(\frac{q}{a}\right) = \left(\frac{a}{q}\right). \qquad \square$$

7.2 Gauss sums

We'll first give a fairly "hands-on" proof of quadratic reciprocity in a special case: let p, q be odd primes, and assume that $p = 1 \bmod q$.

Then there exists an element $\zeta \in \mathbb{F}_p^\times$ of exact order q. So we may consider the "Gauss sum"

$$G(\zeta) = \sum_{a \in \mathbb{F}_q^\times} \left(\frac{a}{q}\right) \zeta^a \qquad \in \mathbb{F}_p.$$

(Since $\zeta^q = 1$, the power ζ^n only depends on $a \bmod q$, so it makes sense to consider ζ^a for $a \in \mathbb{F}_q$.)

Exercise 7.3 Does $G(\zeta)$ depend on *which* order q element we choose?

Proposition 7.4 *We have $G(\zeta)^2 = \left(\frac{-1}{q}\right) q$.*

Proof This follows by an explicit manipulation of sums. By definition we have

$$G(\zeta)^2 = \sum_{a,b \in \mathbb{F}_q^\times} \left(\frac{a}{q}\right)\left(\frac{b}{q}\right) \zeta^{a+b}.$$

Clearly, for any a, $b \in \mathbb{F}_q^\times$ there's a unique $k \in \mathbb{F}_q^\times$ such that $b = ka$, so we can write

$$G(\zeta)^2 = \sum_{a,k \in \mathbb{F}_q^\times} \left(\frac{a}{q}\right)\left(\frac{ka}{q}\right) \zeta^{a(1+k)} = \sum_{a,k \in \mathbb{F}_q^\times} \left(\frac{k}{p}\right) \zeta^{a(1+k)}.$$

We're going to group this into a sum over a on the inside, and a sum over k on the outside:

$$G(\zeta)^2 = \sum_{k \in \mathbb{F}_q^\times} \left(\frac{k}{q}\right) \left(\sum_{a \in \mathbb{F}_q^\times} \zeta^{a(1+k)} \right).$$

So, what is $\left(\sum_{a\in\mathbb{F}_q^\times} \zeta^{a(1+k)}\right)$? If $k=-1$, then it is $1+\cdots+1$ (with $q-1$ terms), so just $q-1$. On the other hand, if $k\neq -1$, then $\xi=\zeta^{k+1}$ also has exact order q, so using the formula for the sum of a geometric progression,

$$\begin{aligned}\xi+\cdots+\xi^{q-1} &= \frac{\xi^q-\xi}{\xi-1}\\ &= \frac{1-\xi}{\xi-1} = -1.\end{aligned}$$

Hence

$$G(\zeta)^2 = (q-1)\left(\frac{-1}{q}\right) - \sum_{\substack{k\in\mathbb{F}_q^\times\\ k\neq -1}}\left(\frac{k}{q}\right) = q\left(\frac{-1}{q}\right) - \sum_{k\in\mathbb{F}_q^\times}\left(\frac{k}{q}\right).$$

To complete the proof we have to show that the sum $S=\sum_{k\in\mathbb{F}_q^\times}\left(\frac{k}{q}\right)$ is zero. But if we take any $u\in\mathbb{F}_q^\times$ which isn't a square, and substitute $j=uk$ in the sum, then the sum gets multiplied by $\left(\frac{u}{q}\right)=-1$. This shows that $S=-S$, so $S=0$ as required. □

Corollary 7.5 *If $p=1 \bmod q$, then q is a square mod p, unless $p=q=3 \bmod 4$, in which case q is not a square mod p.*

Proof If $q=1 \bmod 4$, then $\left(\frac{-1}{q}\right)=1$ and $G(\zeta)$ is a square root of $q \bmod p$. If $q=3 \bmod 4$, then $G(\zeta)$ is a square root of $-q \bmod p$, so q is a square mod p iff -1 is. □

7.3 Enlarging the field

The above Gauss-sums proof for $p=1 \bmod q$ is very tidy, but it leaves many cases unsolved: if p is a square mod q, but $p\neq 1 \bmod q$, we can't find an element of exact order q in $\mathbb{F}_p^\times$ with which to form a Gauss sum. The solution is to work in *field extensions* of $\mathbb{Z}/p$: fields $\mathbb{K}$ containing $\mathbb{Z}/p$ as a subfield.

Let's first look at the case of p, q odd primes (with $p\neq q$). We'll use the following theorem, which we'll prove in a moment:

Theorem 7.6 *There exists a field $\mathbb{K}$, and an element $\zeta\in\mathbb{K}^\times$, such that*
- *$\mathbb{K}$ contains $\mathbb{F}_p$ as a subfield;*
- *ζ has exact order q in $\mathbb{K}^\times$ (that is, $\zeta^q=1$ but $\zeta\neq 1$).*

Then we can run the same machine as before, defining

$$G(\zeta) = \sum_{a\in\mathbb{F}_q^\times}\left(\frac{a}{q}\right)\zeta^a \qquad \in \quad \mathbb{K},$$

and the same proof as before shows that $G(\zeta)^2=\left(\frac{-1}{q}\right)q$. But we want to know if q (or $\left(\frac{-1}{q}\right)q$) is the square of something *in* $\mathbb{F}_p$, not just in the larger field $\mathbb{K}$. So, we need to ask: when does $G(\zeta)$ lie in the subfield $\mathbb{F}_p$ of $\mathbb{K}$?

The trick is to compute $G(\zeta)^p$. Since $p = 0$ in $\mathbb{K}$, we have $(x+y)^p = x^p + y^p$ for any $x, y \in \mathbb{K}$; hence

$$G(\zeta)^p = \sum_{a \bmod q} \left(\frac{a}{q}\right) \zeta^{pa} = \sum_{b \bmod q} \left(\frac{p^{-1}b}{q}\right) \zeta^b,$$

since multiplying by p is a bijection on $\mathbb{F}_q^\times$. Using the multiplicativity of the Legendre symbol this is just

$$\left(\frac{p^{-1}}{q}\right) G(\zeta) = \left(\frac{p}{q}\right) G(\zeta).$$

If $\left(\frac{p}{q}\right) = 1$, then we've shown that $G(\zeta)$ is a root of $X^p - X$ in $\mathbb{K}$. However, all the elements of $\mathbb{F}_p$ are roots of $X^p - X$ (Fermat's little theorem); and since $\mathbb{K}$ is a field, a polynomial of degree p can't have more than p roots in $\mathbb{K}$. Hence $G(\zeta)$ must actually be in $\mathbb{F}_p$, and we've shown that $\left(\frac{-1}{q}\right) q$ is a square mod p.

Conversely, if $\left(\frac{p}{q}\right) \neq 1$, then $G(\zeta)$ is *not* a root of $X^p - X$, so $G(\zeta)$ *cannot* be in $\mathbb{F}_p$. So $G(\zeta)$, and likewise $-G(\zeta)$, are square roots of $\left(\frac{-1}{q}\right) q$ in $\mathbb{K}$ that aren't in $\mathbb{F}_p$. Since $\left(\frac{-1}{q}\right) q$ can't have more than two square roots in $\mathbb{K}$, it follows that $\left(\frac{-1}{q}\right) q$ has no square root in $\mathbb{F}_p$.

Thus $\left(\frac{-1}{q}\right) q$ is a square mod p iff p is a square mod q, which is equivalent to Quadratic Reciprocity.

7.4 The supplementary law for 2

We can handle the supplementary law for $\left(\frac{2}{p}\right)$ similarly, with a little more notation. For $a \in (\mathbb{Z}/8)^\times$, define

$$\psi(a) = \begin{cases} 1 & \text{if } a = \pm 1 \bmod 8 \\ -1 & \text{if } a = \pm 3 \bmod 8. \end{cases}$$

Then one has $\psi(ab) = \psi(a)\psi(b)$ (an easy check).

If p is an odd prime, then we can find a field extension $\mathbb{K}$ of $\mathbb{F}_p$ containing an element $\zeta \in \mathbb{K}$ of order exactly 8 (so $\zeta^4 = -1$). We define

$$G(\zeta) = \sum_{a \in (\mathbb{Z}/8)^\times} \psi(a)\zeta^a = \zeta - \zeta^3 - \zeta^5 + \zeta^7 \qquad \in \quad \mathbb{K}.$$

A calculation now shows that $G(\zeta)^2 = 8$, and $G(\zeta)^p = \psi(p)G(\zeta)$. So if $\psi(p) = 1$, then $\frac{1}{2}G(\zeta) \in \mathbb{Z}/p$ is a square root of 2; conversely, if $\psi(p) = -1$, then the square roots of 2 in $\mathbb{K}$ do not lie in $\mathbb{Z}/p$, so 2 cannot be a square in $\mathbb{Z}/p$. So we deduce

$$\left(\frac{2}{p}\right) = \psi(p),$$

which is the supplementary law.

7.5 Finding the field

Now we finish the argument by explaining why the field $\mathbb{K}$ exists. For q an odd prime, consider the polynomial (a *cyclotomic polynomial*)

$$\Phi_q(X) = \frac{X^q - 1}{X - 1} = 1 + X + X^2 + \cdots + X^{q-1} \in \mathbb{Z}[X].$$

If we reduce Φ_q mod p, for $p \neq q$ another odd prime, we get a polynomial $\bar{\Phi}_q(X) \in \mathbb{F}_p[X]$. This will factor into a product of powers[1] of irreducibles (since polynomials over any field are a Euclidean domain). Let F be any such irreducible factor.

Then $\mathbb{K} = \mathbb{F}_p[X]/F(X)$ is a field (because F is irreducible, see *Algebra* chapter 10); it clearly contains $\mathbb{F}_p$; and the element $\zeta = X \bmod F(X) \in \mathbb{K}$ satisfies $\Phi_q(X) = 0$. Hence $\zeta^q = 1$, but $\zeta \neq 1$ (since $\Phi_q(1) = q \bmod p \neq 0$). Thus ζ has exact order q in $\mathbb{K}^\times$. as required.

Exercise 7.7 By considering the mod p reduction of $\frac{X^8-1}{X^4-1} = X^4 + 1$, show that for any odd prime p there exists a field extension $\mathbb{K}$ of $\mathbb{F}_p$, and an element $\zeta \in \mathbb{K}^\times$, such that ζ has order exactly 8.

[1] In fact one can show that the reduction of Φ_q is square-free, i.e. all the irreducible factors appear to the power 1, but we don't need this here.

CHAPTER 8

Gaussian integers

Having investigated the arithmetic of $\mathbb{Z}$ quite thoroughly, we're now going to look at how factorisation, primes, etc work out in some other algebraic structures – in particular, some subrings of the complex numbers which behave a bit like $\mathbb{Z}$.

8.1 Definitions

Definition 8.1 For any $\alpha \in \mathbb{C}$, we let $\mathbb{Z}[\alpha]$ denote the subgroup of $(\mathbb{C}, +)$ generated by the powers $\{1, \alpha, \alpha^2, \dots\}$.

This is clearly a sub*ring* of $\mathbb{C}$, not just an additive subgroup, and in fact it's the smallest subring containing α. It is always an integral domain (since it's a subring of $\mathbb{C}$, which is a field and hence an integral domain, and any subring of an integral domain is an integral domain.)

Definition 8.2 The ring of *Gaussian integers* is the ring $\mathbb{Z}[i]$, where $i = \sqrt{-1}$ as usual.

Since $i^2 = -1$, any element of $\mathbb{Z}[i]$ can be written uniquely as $a + bi$ for some $a, b \in \mathbb{Z}$; so $\mathbb{Z}[i]$ is isomorphic to $\mathbb{Z}^2$ as an additive group. We can visualise it as a "square lattice" inside the complex plane:

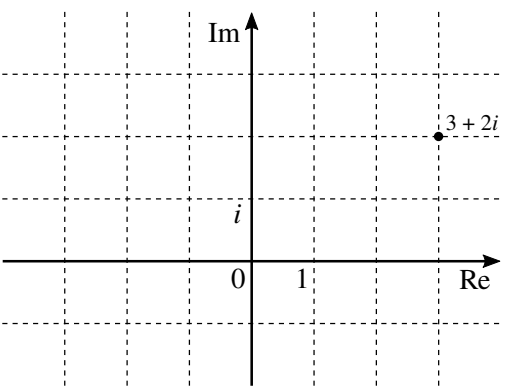

FIGURE 8.1. Gaussian integer grid (image: Wikipedia)

Definition 8.3 If $x = a + bi \in \mathbb{Z}[i]$, we define $N(x) = |x|^2 = a^2 + b^2$.

Note that $N(x) \in \mathbb{N}$, and $N(xy) = N(x)N(y)$. Moreover, we have $N(x) = x\bar{x}$, where $\bar{x}$ is the complex conjugate of x (which is in $\mathbb{Z}[i]$ if x is).

Let's use this to compute the *units* in $\mathbb{Z}[i]$. If x is invertible in $\mathbb{Z}[i]$, then $N(x)$ is invertible in $\mathbb{N}$; so it must be 1. Conversely, if $N(x) = 1$, then x is invertible, since its inverse is $\bar{x}$. So the units are exactly the x with $N(x) = 1$.

However, for integers a, b we have $a^2 + b^2 > 1$ unless $(a, b) = (\pm 1, 0)$ or $(0, \pm 1)$. So we've shown that:

Proposition 8.4 *The set $\mathbb{Z}[i]^\times$ of invertible Gaussian integers consists precisely of $\{1, -1, i, -i\}$.*

So we have more invertible elements than we do in $\mathbb{Z}$ (where the only units are ± 1). This means we need to take care of them when making divisibility statements. So we'll introduce the following notation:

Definition 8.5 We say $\alpha, \beta \in \mathbb{Z}[i]$ are *associates* if $\alpha = u\beta$ for a unit u.

Clearly this is an equivalence relation; moreover, α and β are associates iff $\alpha \mid \beta$ and $\beta \mid \alpha$.

8.2 Euclidean division

Proposition 8.6 *Let $\alpha, \beta \in \mathbb{Z}[i]$ with $\alpha \neq 0$. Then there exist $\kappa, \rho \in \mathbb{Z}[i]$ such that*

- $\beta = \kappa\alpha + \rho$,
- $0 \leqslant N(\rho) < N(\alpha)$.

Proof Let $q = \beta/\alpha \in \mathbb{C}$. Clearly $q = u + vi$ with $u, v \in \mathbb{Q}$; but u and v won't necessarily be in $\mathbb{Z}$.

We shall define $\kappa = x + yi \in \mathbb{Z}[i]$ by rounding u, v to the nearest integer, so that $|x - u| \leqslant \frac{1}{2}$ and $|y - v| \leqslant \frac{1}{2}$. Then we compute that

$$\rho = \beta - \kappa\alpha = \alpha\left((u + vi) - (x + yi)\right).$$

Since the norm on $\mathbb{C}$ is multiplicative, we have

$$N(\rho) = N(\alpha) \cdot \left((u - x)^2 + (v - y)^2\right).$$

But both $|u - x|$ and $|v - y|$ are $\leqslant \frac{1}{2}$, so $(u - x)^2 + (v - y)^2 \leqslant (\frac{1}{2})^2 + (\frac{1}{2})^2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. So we've shown that

$$N(\rho) \leqslant \frac{1}{2} N(\alpha) < N(\alpha).$$

□

Example 8.7 For $\beta = 11 + 8i$ and $\alpha = 2 + 3i$, we compute

$$\frac{\beta}{\alpha} = \frac{(11 + 8i)(2 - 3i)}{(2 + 3i)(2 - 3i)} = \frac{46}{13} + \frac{-17}{13} i.$$

Rounding $\frac{46}{13}$ and $\frac{-17}{13}$ to the *nearest* integers we obtain $\kappa = u + vi = 4 - i$, and hence

$$\rho = \beta - \kappa\alpha = (11 + 8i) - (2 + 3i) \cdot (4 - i) = -2i.$$

Remark 8.8 Note that, unlike in the case of $\mathbb{Z}$, we haven't claimed any uniqueness for κ and ρ. Can you find a different pair (κ, ρ) which also works, for the same (α, β) as above?

Proposition 8.6 is precisely the statement that $\mathbb{Z}[i]$ is a *Euclidean ring* (Algebra, Chapter 9). This is exactly what we need to make the Euclidean algorithm work in $\mathbb{Z}[i]$: for any two elements α, β there exists an (explicitly computable) element $\gcd(\alpha, \beta)$, well-defined up to multiplication by units, such that we have

$$\forall x \in \mathbb{Z}[i], \qquad x \mid \gcd(\alpha, \beta) \iff x \mid \alpha \text{ and } x \mid \beta.$$

Moreover, $\gcd(\alpha, \beta)$ can always be written as $r\alpha + s\beta$ for $r, s \in \mathbb{Z}[i]$.

Remark 8.9 Note that in general there are four equally valid possibilities for the GCD – it is only well-defined up multiplication by $\{\pm 1, \pm i\}$ and there's no obvious "best" choice among these four options.

Example 8.10 From the calculation above, $\gcd(11+8i, 2+3i) = \gcd(2+3i, -2i)$. We also have

$$(2 + 3i) = (-1 + i) \cdot (-2i) + i,$$

so

$$\gcd(2 + 3i, -2i) = \gcd(-2i, i).$$

Since i is a unit, this shows that $11 + 8i$ and $2 + 3i$ are coprime in $\mathbb{Z}[i]$.

Corollary 8.11 *Let $\alpha \in \mathbb{Z}[i]$. Then the following are equivalent:*

- *α is an* indecomposable element*: that is, if $\beta \mid \alpha$, then either β is a unit or it is an associate of α.*
- *α is a* prime element*: that is, if $\rho, \sigma \in \mathbb{Z}[i]$ and $\alpha \mid \rho\sigma$, then $\alpha \mid \rho$ or $\alpha \mid \sigma$.*

Proof Cf. *Algebra*, Prop 9.21. Since $\mathbb{Z}[i]$ is Euclidean, it is a PID; and in a PID, prime elements and indecomposable elements coincide. Alternatively, we can repeat exactly the same argument as for $\mathbb{Z}$, using Euler's Lemma 1.20. □

Remark 8.12 Recall that you saw in *Algebra* that in the similar-looking ring $\mathbb{Z}[\sqrt{-5}]$, the element 3 is indecomposable but not prime, since it divides $(1 - \sqrt{-5})(1+\sqrt{-5})$ but doesn't divide either factor. So this is something rather special about $\mathbb{Z}[\sqrt{-1}]$.

Corollary 8.13 (Fundamental Theorem of Arithmetic for $\mathbb{Z}[i]$) *Any non-zero $\alpha \in \mathbb{Z}[i]$ can be written as a product of prime elements. Moreover, if*

$$\alpha = \pi_1\pi_2 \dots \pi_r = \mu_1\mu_2 \dots \mu_s$$

are two factorisations of α as products of prime elements, then $r = s$, and we can re-order the factors so that μ_i is an associate of π_i for $i = 1, \dots, r$.

Exactly as before, we can also gather together the factors and write

$$\alpha = u \cdot \prod_i \pi_i^{e_i},$$

with u a unit, $e_i \in \mathbb{N}_+$, and π_i primes which are pairwise non-associate.

8.3 Gaussian primes

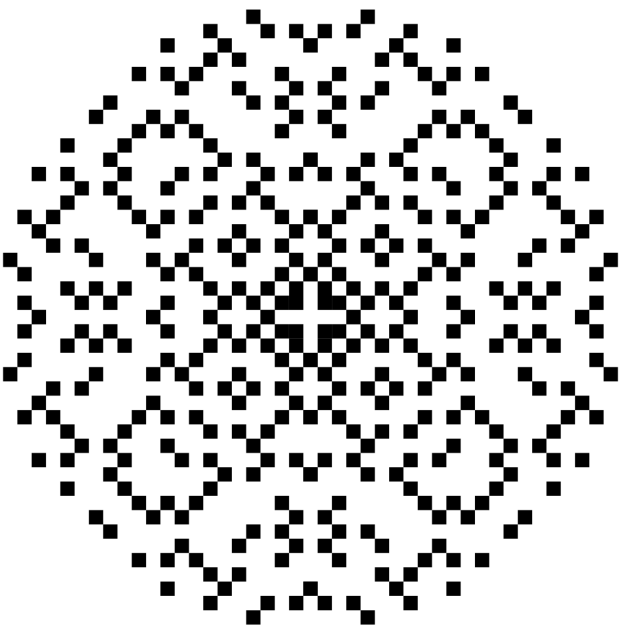

FIGURE 8.2. Gaussian primes α with $N(\alpha) \leqslant 500$ (image: Wikipedia)

We'll now classify all the primes in $\mathbb{Z}[i]$. We start with the following easy remark:

Proposition 8.14 *Suppose $\alpha \in \mathbb{Z}[i]$ is a prime element. Then there is a unique prime integer $p \in \mathbb{P}$ such that α divides p. (We say α* lies above *the prime integer p.)*

Proof Consider the norm $N(\alpha)$, which is a non-zero integer. Since $\alpha\bar{\alpha} = N(\alpha)$, we have $\alpha \mid N(\alpha)$. From the factorisation theory of $\mathbb{Z}$, we can write $N(\alpha)$ as a product of prime integers; but since α is prime, it must divide one of these factors. This shows that α must divide some $p \in \mathbb{P}$. But if α divides two distinct elements $p, q \in \mathbb{P}$, then it must divide $mp + nq$ for all $m, n \in \mathbb{Z}$; so it must divide 1, which is a contradiction since α is not a unit. □

So we can study all Gaussian primes by asking, for each $p \in \mathbb{P}$, which Gaussian primes lie above it.

Proposition 8.15 *Given $p \in \mathbb{P}$, exactly one of the following two possibilities occurs:*

- *p factors as $\alpha\bar{\alpha}$, for some prime $\alpha \in \mathbb{Z}[i]$ with $N(\alpha) = p$. Then the primes above p are the associates of α and $\bar{\alpha}$.*
- *p is itself a prime element of $\mathbb{Z}[i]$ (so the only primes above p are $\pm p$ and $\pm ip$).*

Moreover, the first case occurs if and only if there exist integers (x, y) with $p = x^2 + y^2$.

Proof First let us assume that (x, y) exists with $p = x^2 + y^2$. Then $\alpha = x + yi \in \mathbb{Z}[i]$ satisfies $N(\alpha) = p$. Since $N(\alpha) = \alpha\bar{\alpha}$, we have $\alpha \mid p$; and α must be indecomposable, and hence prime, since if α factors as a product $\beta\gamma$ then we must have $N(\beta)N(\gamma) = N(\alpha) = p$, so one of β and γ has norm 1 and is thus a unit. Since $N(\bar{\alpha}) = N(\alpha)$ we see that $\bar{\alpha}$ is also prime. Moreover, any prime above p must divide $\alpha\bar{\alpha}$; so it divides one of α and $\bar{\alpha}$, and must therefore be an associate of it, since they are both prime.

Conversely, if no such (x, y) exists, then p is indecomposable, since any nontrivial factor β of p would have to satisfy $N(\beta) = p$. □

We'd like to know which $p \in \mathbb{P}$ remain prime, and which do not. Clearly $p = 2$ factors as $(1 + i)(1 - i)$, so we can restrict to odd p. It turns out that the answer depends only on p mod 4. One direction is easy:

Proposition 8.16 *Let $p \in \mathbb{P}$. If $p = 3 \bmod 4$, then p is a Gaussian prime.*

Proof The only squares mod 4 are 0 and 1, so if $p = 3 \bmod 4$, the equation $x^2 + y^2 = p$ has no solutions mod 4 and hence no solutions in $\mathbb{Z}$. □

It turns out that the converse is also true, but this is a much deeper theorem:

Theorem 8.17 (Fermat) *Let $p \in \mathbb{P}$ with $p = 1 \bmod 4$. Then p is not a Gaussian prime. Equivalently, p is the sum of two integer squares.*

Proof Consider the equation $X^2 + 1 = 0 \bmod p$. This has a solution, since $p \neq 3 \bmod 4$. Choose $t \in \mathbb{Z}$ such that $t^2 + 1 = 0 \bmod p$; and let $\alpha = \gcd(t - i, p)$.

Clearly $\alpha \mid p$, but α is not an associate of p, since $p \nmid t - i$. Thus $N(\alpha) = 1$ or p; and it suffices to prove that $N(\alpha) \neq 1$, i.e. that $t - i$ and p aren't coprime in $\mathbb{Z}[i]$.

Consider the map

$$\lambda : \mathbb{Z}[i] \to \mathbb{F}_p, \qquad a + bi \mapsto a + tb \bmod p.$$

This is obviously compatible with addition; we claim it's also compatible with multiplication. This can be checked explicitly: if $u = a + bi$, $v = c + di$, then

$$\lambda(uv) = \lambda((ac - bd) + (ad + bc)i) = (ac - bd) + t(ad + bc) \bmod p,$$

while

$$\lambda(u)\lambda(v) = (a + tb)(c + td) = (ac + t^2 bd) + t(ad + bc) \bmod p,$$

and since $t^2 = -1 \bmod p$ these are the same.

Clearly λ kills both p and $t - i$. So if these elements were coprime, there would be u, v with $up + v(t - i) = 1$, and we'd have $1 = \lambda(up + v(t - i)) = 0 + 0 = 0 \bmod p$, which is a contradiction. So p and $t - i$ can't be coprime. □

Remark 8.18 Fermat announced that he had proved the theorem in a letter dated Christmas Day 1640, but he never revealed his method of proof (sound familiar?). Just like Quadratic Reciprocity, this theorem now has many different proofs; these include a famous "one-sentence proof" due to Don Zagier (link).

Exercise 8.19 Show that if $p = 1 \bmod 4$ and $\alpha \in \mathbb{Z}[i]$ satisfies $N(\alpha) = p$, then $\bar{\alpha}$ is not an associate of α. (Hint: if α divides $t - i$, for some $t \in \mathbb{Z}$ as above, then $\bar{\alpha}$ divides $t + i$.)

8.4 Euclidean rings

Recall the following construction from Algebra:

Definition 8.20 Let R be an integral domain. A *Euclidean function* on R is a map

$$\delta : R \setminus \{0\} \to \mathbb{N}$$

such that for every $a, b \in R$ with $b \neq 0$, we can find $q, r \in R$ with $a = bq + r$ such that either $r = 0$, or $\delta(r) < \delta(b)$.

We know that $\mathbb{Z}$, $k[X]$ for k a field, and $\mathbb{Z}[i]$ are examples of Euclidean domains. One can check similarly that $\mathbb{Z}[\sqrt{-2}]$ (i.e. the subring of $\mathbb{C}$ consisting of numbers of the form $a + b\sqrt{-2}$, with $a, b \in \mathbb{Z}$), is a Euclidean domain, with the Euclidean function again given by $N(x) = x\bar{x}$, so $N(a + b\sqrt{-2}) = a^2 + 2b^2$.

Exercise 8.21 Prove that N is a Euclidean function on $\mathbb{Z}[\sqrt{-2}]$. (You will need the fact that if $|p|, |q| \leqslant \frac{1}{2}$ then $p^2 + 2q^2 \leqslant \frac{3}{4} < 1$.)

It follows that factorisation in $\mathbb{Z}[\sqrt{-2}]$ works in just the same elegant way as before; the ring is a PID and a UFD, and and we can characterise exactly which primes remain prime in $\mathbb{Z}[\sqrt{-2}]$ in terms of congruences mod 8.

Exercise 8.22 (hard!) Show that an odd prime p has the form $x^2 + 2y^2$ iff $p = 1, 3 \bmod 8$, and not if $p = 5, 7 \bmod 8$.
[Hint: First show that -2 is a square mod p iff $p = 1, 3 \bmod 8$, using the supplementary laws of quadratic reciprocity.]

8.5 The Eisenstein integers

On the other hand, the ring $\mathbb{Z}[\sqrt{-3}]$ is *not* Euclidean. It can't be, because

$$(1 - \sqrt{-3})(1 + \sqrt{-3}) = 4 = 2 \cdot 2$$

and 2 does not divide either factor on the right. So 2 is not a prime element; but it is obviously indecomposable since $a^2 + 3b^2 = 2$ has no solutions. So $\mathbb{Z}[\sqrt{-3}]$ is not a PID, and hence not Euclidean either. However, we can fix this by embedding $\mathbb{Z}[\sqrt{-3}]$ inside a slightly larger ring:

Definition 8.23 The *Eisenstein integers* is the subring $\mathbb{Z}[\omega] \subset \mathbb{C}$, where $\omega = \frac{-1+\sqrt{-3}}{2}$.

This clearly contains $\mathbb{Z}[\sqrt{-3}]$ (since $\sqrt{-3} = 2\omega + 1$), but it is slightly larger, since $\omega \notin \mathbb{Z}[\sqrt{-3}]$.

One can check that $\mathbb{Z}[\omega]$ consists precisely of the linear combinations $a + b\omega$ with $a, b \in \mathbb{Z}$. This is because ω satisfies the equation $\omega^2 = -1 - \omega$, and we can use this (and induction on n) to show that ω^n is a $\mathbb{Z}$-linear combination of 1 and ω for all $n \in \mathbb{N}$.

Exercise 8.24 Show that the abelian-group quotient $\mathbb{Z}[\omega]/\mathbb{Z}[\sqrt{-3}]$ has order 2.

One can picture $\mathbb{Z}[\omega]$ as a triangular lattice inside the complex plane, as in the following figure[1]:

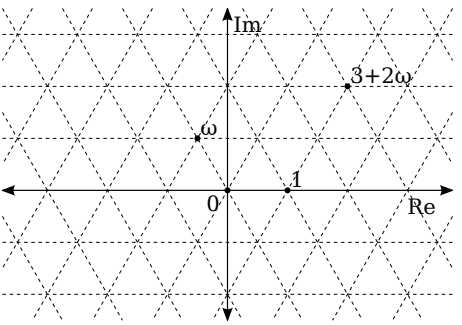

FIGURE 8.3. Eisenstein integer grid (image: Wikipedia)

Exercise 8.25 Find all the units of $\mathbb{Z}[\omega]$. (Hint: There are 6 of them.)

From Figure 8.3, it's easy to convince yourself that for every $x \in \mathbb{C}$, there exists $y \in \mathbb{Z}[\omega]$ with $|x - y| < 1$. (In fact we can do a little better: we're never more than $\frac{1}{\sqrt{3}} \cong 0.58$ away from an element of $\mathbb{Z}[\omega]$.) This suffices to show that $\mathbb{Z}[\omega]$ is Euclidean, with $N(x) = x\bar{x}$ as the Euclidean function, just as before. So $\mathbb{Z}[\omega]$ is a PID and a UFD; and we can deduce the following:

Proposition 8.26 *Let $p \in \mathbb{P}$. Then $p = N(\alpha)$ for some $\alpha \in \mathbb{Z}[\omega]$ if and only if $p = 1 \bmod 3$.*

Exercise 8.27 Can you show that, despite $\mathbb{Z}[\sqrt{-3}]$ not being a PID, nonetheless every prime that is 1 mod 3 has the form $x^2 + 3y^2$? (Hint: Show that if $\alpha \in \mathbb{Z}[\omega]$ then at least one of its associates lies in the subring $\mathbb{Z}[\sqrt{-3}]$.)

8.6 Another non-Euclidean ring

Now consider $\mathbb{Z}[\sqrt{-5}]$. Again, this is not a UFD, because

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$$

are two distinct factorisations of the same element.

[1] From Wikipedia, with thanks to Wikipedia contributor `gunther`.

It turns out that this failure is in some sense *much worse* than with $\mathbb{Z}[\sqrt{-3}]$. This is for two reasons:

- One can show that the non-uniqueness of factorisation in $\mathbb{Z}[\sqrt{-3}]$ is "local at the prime 2". Let's say $a + b\sqrt{-3}$ is *good* if $a + b$ is odd; one checks that the product of good elements is good (exercise). It turns out that good indecomposable elements are prime, and good elements have unique prime factorisations. So we can restore uniqueness of factorisation by imposing congruences modulo 2. In contrast, the failure in $\mathbb{Z}[\sqrt{-5}]$ is "global"; we can't get rid of it by imposing congruences to any fixed modulus.

- In $\mathbb{Z}[\sqrt{-3}]$ we were able to restore unique factorisation by going up to the "finitely larger" ring $\mathbb{Z}[\omega]$. In $\mathbb{Z}[\sqrt{-5}]$ this doesn't work: there aren't any rings finitely larger than $\mathbb{Z}[\sqrt{-5}]$. More precisely, if R is a ring with $\mathbb{Z}[\sqrt{-5}] \subseteq R \subseteq \mathbb{C}$, then R is either equal to $\mathbb{Z}[\sqrt{-5}]$, or *much* bigger (the index $[R : \mathbb{Z}[\sqrt{-5}]]$ is infinite).

(These two explanations are related by the fact that $2 = [\mathbb{Z}[\omega] : \mathbb{Z}[\sqrt{-3}]]$, so $\mathbb{Z}[\sqrt{-3}]$ is "a factor of 2 away from a UFD".)

CHAPTER 9

Real quadratic fields and Pell's equation

9.1 Setup

Having studied which integers can be written as $X^2 + Y^2$, $X^2 + 2Y^2$ etc, we're now going to change tack a bit, and study representations of integers in the form $X^2 - nY^2$ for $n > 0$.

This change of sign might seem insigificant, but it actually changes the whole flavour of the theory. For $n > 0$, it's obvious that there are at most finitely many solutions to $X^2 + nY^2 = r$ for any given r, and all solutions will have $|X| \leqslant \sqrt{|r|}$ and $|Y| \leqslant \sqrt{\frac{|r|}{n}}$, so we can find all of them with a finite search. However, for $X^2 - nY^2 = r$, it's perfectly possible that there could be infinitely many solutions; and if solutions do exist, it's not at all obvious how big they might be, so we can't rely on finding solutions with a computer search.

Definition 9.1 An equation of the form $X^2 - nY^2 = 1$, for a given $n \in \mathbb{N}_+$, is called a *Pell equation*. More generally, an equation of the form $X^2 - nY^2 = r$, for given $n \in \mathbb{N}_+$ and $r \in \mathbb{Z}$, is called a *generalized Pell equation*.

We want to know if Pell and generalized Pell equations have solutions with $(X, Y) \in \mathbb{Z}^2$, and find a way of describing all such solutions.

Remark 9.2 Note the crucial importance of the condition $n > 0$. If we were to allow $n < 0$, then it is obvious that there can be at most finitely many solutions for a given n and r, and we can find all of them by a routine search. In contrast, when $n > 0$ (and n is not a square), we'll see that the solution set is always either empty or infinite.

The case of n a perfect square is easy: if $n = m^2$, then our equation becomes $(X - mY)(X + mY) = r$, and since $X \pm mY$ are integers, they must be in the finite list of divisors of r. So we can easily find all solutions by considering the prime factorization of r.

Example 9.3 Let's find all integer solutions of $X^2 - 4Y^2 = 9$.

If (X, Y) is a solution, then $X - 2Y$ must divide 9, so it is one of $\{\pm 1, \pm 3, \pm 9\}$. Trying each possibility leads to the six solutions $(\pm 5, \pm 2)$ and $(\pm 3, 0)$.

If n is not a perfect square, then Pell's equation is closely related to the arithmetic of the ring $\mathbb{Z}[\sqrt{n}] = \{a + b\sqrt{n} : a, b \in \mathbb{Z}\}$. Note that since $n > 0$, this is a subring of $\mathbb{R}$ (unlike

the rings $\mathbb{Z}[i]$, $\mathbb{Z}[\sqrt{-3}]$, $\mathbb{Z}[\omega]$ from the previous chapters, which are contained in $\mathbb{C}$ but not in $\mathbb{R}$). We'll mostly focus on $n = 2$.

Definition 9.4 If $x = a + b\sqrt{n} \in \mathbb{Z}[\sqrt{n}]$, we write $x^* = a - b\sqrt{n}$, and we define

$$N(x) = xx^* = a^2 - nb^2.$$

We call this the *norm* of x (although it's not a norm in the sense of analysis). We already used this for $n = -1, -2, -3$ in the previous chapter (in which case $x^* = \bar{x}$ is the complex conjugate of x, and $N(x)$ is the square of the complex absolute value); but now we want to take $n > 0$. Observe that x^* is still well-defined, since $\sqrt{n}$ is irrational and so there is a unique way of writing x as $a + b\sqrt{n}$. The map N still respects multiplication, i.e. $N(xy) = N(x)N(y)$. However, what's new is that the map N can take negative values. For example, $N(1 + \sqrt{2}) = -1$.

9.2 Pell's equation and units

Proposition 9.5 *Assume n is not a perfect square. Then there is a bijection between pairs of integers (X, Y) with $X^2 - nY^2 = r$, and the set $\{\alpha \in \mathbb{Z}[\sqrt{n}] : N(\alpha) = r\}$, given by $(X, Y) \mapsto X + Y\sqrt{n}$.* □

Now we have much more structure to work with, because the norm is *multiplicative*. In particular, if $r = 1$, the solution set is actually a *group*: it is clearly closed under multiplication, and since $N(x) = 1$ implies $x^{-1} = x^*$, it is also closed under inverses. So it is a subgroup of the unit group $\mathbb{Z}[\sqrt{n}]^\times$. This allows us to get new solutions from old:

Example 9.6 Consider the equation $X^2 - 2Y^2 = 1$.
One obvious non-trivial solution is $(X, Y) = (3, 2)$; that is, $N(3 + 2\sqrt{2}) = 1$. So $N((3 + 2\sqrt{2})^k) = 1$ for all $k \geqslant 1$, giving us the solutions

$$(3 + 2\sqrt{2})^2 = 17 + 12\sqrt{2} \Rightarrow 17^2 - 2 \cdot 12^2 = 1$$
$$(3 + 2\sqrt{2})^3 = 99 + 70\sqrt{2} \Rightarrow 99^2 - 2 \cdot 70^2 = 1$$
$$\vdots$$

Clearly this cannot repeat (since $3 + 2\sqrt{2}$ is a real number and not ± 1, so no power of it can be equal to 1). So we obtain infinitely many non-trivial solutions.

Notice that if (X, Y) is a non-trivial solution of the Pell equation, then so is $(\pm X, \pm Y)$. This corresponds to replacing $u = X + Y\sqrt{n}$ with $\pm u$ or $\pm u^{-1}$. So we can assume $X, Y > 0$, which corresponds to assuming $u > 1$.

Definition 9.7 A positive-integer solution (A, B) of Pell's equation is said to be *fundamental* if for all other positive-integer solutions (X, Y), we have $A + B\sqrt{n} < X + Y\sqrt{n}$.
If (A, B) is the fundamental solution, then $u = A + B\sqrt{n}$ is called the *fundamental positive unit* of $\mathbb{Z}[\sqrt{n}]$.

Proposition 9.8 *If Pell's equation has non-trivial solutions (for a given n), then it has a fundamental solution.*

Proof If (X, Y) is a positive solution, then there are clearly only finitely many pairs of positive integers (X', Y') with $X' + Y'\sqrt{n} \leqslant X + Y\sqrt{n}$ (since both X' and Y' must lie in a bounded range). So there are certainly only finitely many such pairs that are solutions of the equation; and a finite subset of $\mathbb{R}$ must have a least element. □

Example 9.9 Returning to the previous example, if X, Y are positive with $X + Y\sqrt{2} < 3 + 2\sqrt{2}$, then $X < 3 + 2\sqrt{2} < 5$, and $Y < (3 + 2\sqrt{2})/\sqrt{2} < 4.5$. So $1 \leqslant X, Y \leqslant 4$, and none of these leads to a smaller solution. Thus $(3, 2)$ is the fundamental solution of Pell's equation for $n = 2$.

Proposition 9.10 *If Pell's equation has non-trivial solutions (for a given n), and u is the fundamental positive unit, then every solution with $X, Y > 0$ is of the form*

$$X + Y\sqrt{n} = u^k$$

for a uniquely determined $k \in \mathbb{N}_+$.
Equivalently, the full set of solutions (with no assumption on signs) is given by

$$X + Y\sqrt{n} = \{\pm u^k : k \in \mathbb{Z}\}.$$

Proof Let (X, Y) be any solution with $X, Y > 0$, and consider $v = X + Y\sqrt{n}$. Then $v > 1$, but the sequence v/u^k obviously tends to 0, so there must be a largest k such that $v/u^k \geqslant 1$.

For this k, the ratio $v' = v/u^k$ satisfies $1 \leqslant v' < u$. If $1 < v'$, then this would contradict the minimality of u; so we must have $v' = 1$, i.e. $v = u^k$ is a power of u. □

Example 9.11 To complete our $n = 2$ example, the complete set of solutions to $X^2 - 2Y^2 = 1$ is given by

$$X + Y\sqrt{2} = \pm(3 + 2\sqrt{2})^k$$

for any choice of sign and $k \in \mathbb{Z}$.

More algebraically, we're saying that the group $\{v \in \mathbb{Z}[\sqrt{n}]^\times : \mathrm{N}(v) = 1\}$ must be either ± 1, or the product of ± 1 and an infinite cyclic group generated by the fundamental positive unit. In fact the former case never occurs:

Theorem 9.12 *For any non-square n, the Pell equation $X^2 - nY^2 = 1$ does have non-trivial solutions.*

Remark 9.13 We won't prove this here. It can be deduced from a much more general theorem of Dirichlet about units in a general number field (see Appendix B of Stewart and Tall). Alternatively, there is a more elementary, but still rather difficult,

proof in §IV.11 of Davenport's book, based on studying rational approximations to $\sqrt{n}$.

Exercise 9.14 Show that $\mathbb{Z}[\sqrt{7}]$ does have a fundamental positive unit (and find it explicitly).

9.3 The negative Pell equation

We'll now consider the case $r = -1$. These don't form a group, obviously; but we can fix this by considering the set of solutions for $r = -1$ and $r = +1$ together:

Proposition 9.15 *The set of solutions of $X^2 - nY^2 = \pm 1$ bijects with the unit group $\mathbb{Z}[\sqrt{n}]^\times$.* □

Proof If α is invertible in $\mathbb{Z}[\sqrt{n}]$, then $\mathrm{N}(\alpha)$ is invertible in $\mathbb{Z}$, so it must be ± 1. Conversely, if $\alpha \in \mathbb{Z}[\sqrt{n}]$ has $\mathrm{N}(\alpha) = \pm 1$, then $\alpha^{-1} = \pm\alpha^* \in \mathbb{Z}[\sqrt{n}]$. □

Arguing exactly as in the previous section, if the set of numbers of the form $X + Y\sqrt{n}$ with $X, Y > 0$ and $X^2 - nY^2 = \pm 1$ is non-empty, then it has a smallest element, which we call the *fundamental unit*. For instance, the fundamental unit of $\mathbb{Z}[\sqrt{2}]$ is $1 + \sqrt{2}$, which is the *square root* of the fundamental positive unit $3 + 2\sqrt{2}$.

In general, there are two cases which can arise:

- The fundamental unit has $N(u') = +1$. In this case, the fundamental unit and fundamental positive unit coincide, we have $N(\alpha) = 1$ for all $\alpha \in \mathbb{Z}[\sqrt{n}]^\times$, and $X^2 - nY^2 = -1$ has no solutions.
- The fundamental unit has $N(u') = -1$. In this case, the fundamental positive unit is the *square* of the fundamental unit (as we saw for $n = 2$), and multiplying by the fundamental unit gives a bijection between solutions of $X^2 - nY^2 = -1$ and solutions of $X^2 - nY^2 = +1$. Algebraically, $\{\alpha \in \mathbb{Z}[\sqrt{n}]^\times : N(\alpha) = 1\}$ is an index 2 subgroup of $\mathbb{Z}[\sqrt{n}]^\times$, and both groups are isomorphic to $\{\pm 1\} \times \mathbb{Z}$.

Remark 9.16 If n is prime, then the fundamental unit always has norm 1 if $n = 3 \bmod 4$, and has norm -1 otherwise. For composite n the situation is more complicated and not fully understood. ⊛

9.4 Generalized Pell equations

For the generalized Pell equation with $r \neq \pm 1$, we need to combine the information we've just gathered about units, with information about factorization. Using basically the same argument as before, we can show:

Theorem 9.17 *The ring $\mathbb{Z}[\sqrt{2}]$ is Euclidean, with Euclidean function $\delta(x) = |N(x)|$.*

Hence it is a PID, and thus a UFD; but we need to remember that – as with any UFD – factorizations are only unique up to units, and we've seen that the unit group of $\mathbb{Z}[\sqrt{2}]$ is quite large.

Theorem 9.18 *If $p \in \mathbb{P}$, then $X^2 - 2Y^2 = p$ has integer solutions if and only if $p = 2$ or $p \equiv \pm 1 \bmod 8$.*

Proof The non-trivial point is to show that if p is odd with $\left(\frac{2}{p}\right) = 1$, then $X^2 - 2Y^2 = p$ has solutions. This is essentially the same argument as we have already seen for $X^2 + Y^2$ and $X^2 + 2Y^2$, with a little more care about signs: if t is a square root of $2 \bmod p$, and $\alpha = \gcd(t - \sqrt{2}, p)$, then $N(\alpha)$ must be ± 1 or $\pm p$. But α is not a unit, since there is a ring homomorphism $\mathbb{Z}[\sqrt{2}] \to \mathbb{F}_p$ which sends both $t - \sqrt{2}$ and p to 0. Hence α must have norm $\pm p$; and since α is only defined up to units, we can multiply by $1 + \sqrt{2}$ if needed so that $N(\alpha) = p$. □

Exercise 9.19 Show that if $p \in \mathbb{P}$ with $p > 3$, then $X^2 - 3Y^2 = p$ has solutions iff $p = 1 \bmod 12$, and $X^2 - 3Y^2 = -p$ has solutions iff $p = -1 \bmod 12$. (Hint: $\mathbb{Z}[\sqrt{3}]$ is Euclidean, but its fundamental unit has norm $+1$).

To classify *all* solutions, we need to keep track of which units can occur:

Example 9.20 Consider the equation $X^2 - 2Y^2 = 7$.
I claim that every solution has the form $X + Y\sqrt{2} = \pm(5 + 3\sqrt{2}) \cdot (3 + 2\sqrt{2})^k$ *or* $\pm(5 - 3\sqrt{2})(3 + 2\sqrt{2})^k$ for some $k \in \mathbb{Z}$; and these possibilities are mutually exclusive.

To see this, note that $7 = \alpha\beta$ where $\alpha = 5 + 3\sqrt{2}$ and $\beta = 5 - 3\sqrt{2}$. Since $N(\alpha) = N(\beta) = 7$ is prime, they are both indecomposable, and hence prime, elements of $\mathbb{Z}[\sqrt{2}]$. Moreover, they are not associates (they do not divide each other), since $\alpha/\beta = \frac{43+30\sqrt{2}}{7} \notin \mathbb{Z}[\sqrt{2}]$.
If (X, Y) is any solution of $X^2 - 2Y^2 = 7$, then $\gamma = X + Y\sqrt{2}$ satisfies $\gamma\gamma^* = 7$, so $\gamma \mid \alpha\beta$. Moreover, γ is also indecomposable and hence prime (because $N(\gamma) = 7$). Hence either γ is an associate of α, or γ is an associate of β.

The general picture is as follows. As we've seen, the solutions of $X^2 - 2Y^2 = 1$ are a group. For any r, multiplication in $\mathbb{Z}[\sqrt{2}]$ gives a *group action* of the group of solutions of $X^2 - 2Y^2 = 1$ on the set of solutions of $X^2 - 2Y^2 = r$. Using uniqueness of prime factorisations, we can show that this group action has finitely many orbits (and we can determine the orbits explicitly from the prime factorisation of r).

Exercise 9.21 Find a similar description of the solutions of $X^2 - 2Y^2 = 14$ (you should find that there are again two orbits). How many orbits are there for $X^2 - 2Y^2 = 119$?

Remark 9.22 If we replace $X^2 - 2Y^2$ with $X^2 - nY^2$, for general non-square $n > 0$, then it's still true that the solutions of $X^2 - nY^2 = r$ for any r fall into finitely many orbits up to the action of the units. However, since $\mathbb{Z}[\sqrt{n}]$ is not always a UFD, there may not be a simple criterion describing the r for which the equation is solvable.

(It is an open problem whether there are infinitely many square-free $n > 0$ such that $\mathbb{Z}[\sqrt{n}]$ is a UFD. $\circledast$)

CHAPTER 10

Arithmetic in number fields

10.1 Number fields

Remember the following definition from *Algebra*:

Definition 10.1 For $\alpha \in \mathbb{C}$, we say α is an *algebraic number* if there is a non-constant polynomial $f(X) \in \mathbb{Q}[X]$ with $f(\alpha) = 0$; and we write $\overline{\mathbb{Q}}$ for the set of all algebraic numbers.

Moreover, you saw that:

- If α is algebraic, then there is a unique "simplest" polynomial that it satisfies – the *minimal polynomial* of α, which is the smallest-degree monic f with $f(\alpha) = 0$.
- For any $\alpha \in \mathbb{C}$, there is a unique smallest subfield[1] $\mathbb{Q}(\alpha) \subset \mathbb{C}$ containing α, and α is algebraic if and only if $\mathbb{Q}(\alpha)$ is finite-dimensional over $\mathbb{Q}$.
- If α is algebraic, then $\mathbb{Q}(\alpha)$ has basis $\{1, \alpha, \dots, \alpha^{d-1}\}$ where d is the degree of its minimal polynomial.
- $\overline{\mathbb{Q}}$ is a field.

We're going to study "little pieces" of $\overline{\mathbb{Q}}$, rather than all of $\overline{\mathbb{Q}}$ at once:

Definition 10.2 A *number field* is a subfield of $\mathbb{C}$ which is finite-dimensional as a $\mathbb{Q}$-vector space.

Example 10.3 The field $\mathbb{Q}(i) = \{a + bi : a, b \in \mathbb{Q}\}$ is a number field, with $[\mathbb{Q}(i) : \mathbb{Q}] = 2$.

Note that any number field must be contained in $\overline{\mathbb{Q}}$: if K is a number field, and $\alpha \in K$, then $\mathbb{Q}(\alpha) \subseteq K$. Since K has finite dimension, so does $\mathbb{Q}(\alpha)$, hence α is algebraic. Conversely, for any $\alpha \in \overline{\mathbb{Q}}$, the field $\mathbb{Q}(\alpha)$ is a number field.

What's less obvious, but true, is that *every* number field can be written in this form: for any number field $K \subset \mathbb{C}$, we can find some $\alpha \in K$ such that $K = \mathbb{Q}(\alpha)$ (a "primitive element" for K).

[1] Pedantic notational remark: by definition $\mathbb{Q}(\alpha)$ is the smallest *subfield* of $\mathbb{C}$ containing $\mathbb{Q}$ and α, while $\mathbb{Q}[\alpha]$ is the smallest *subring* of $\mathbb{C}$ containing $\mathbb{Q}$ and α; we have $\mathbb{Q}[\alpha] \subseteq \mathbb{Q}(\alpha)$, because every subfield is a subring, and equality holds iff α is algebraic (e.g. $\frac{1}{\pi} \notin \mathbb{Q}[\pi]$.) In this module we only care about the algebraic case, so it doesn't matter if we write $\mathbb{Q}(\alpha)$ or $\mathbb{Q}[\alpha]$; we're going to standardize on $\mathbb{Q}(\alpha)$.

Example 10.4 Let's start with $\mathbb{Q}$, and let $K = \mathbb{Q}(i)$ be the extension of $\mathbb{Q}$ generated by i; and then let L be the extension of K generated by $\sqrt{2}$ (which is not in K).
Then L is an extension of $\mathbb{Q}$ of degree 4: a $\mathbb{Q}$-vector-space basis is given by $\{1, i, \sqrt{2}, \sqrt{-2}\}$.
Clearly none of these basis elements is a primitive element, but one can check that $\alpha = i + \sqrt{2}$ is a primitive element: the powers of $i + \sqrt{2}$ are a basis of L.

Exercise 10.5 For α as in the example, write each of $1, \alpha, \alpha^2, \alpha^3, \alpha^4$ in terms of the basis $\{1, i, \sqrt{2}, \sqrt{-2}\}$. Hence verify that $\{1, \alpha, \alpha^2, \alpha^3\}$ span L as a $\mathbb{Q}$-vector space, and calculate the minimal polynomial of α.

Remark 10.6 (Non-examinable) Here is a sketch of why every number field has a primitive element. It follows from Galois theory (cf. *Algebra* script) that for any number field K, there are only finitely many possible subfields K' with $\mathbb{Q} \subseteq K' \subsetneq K$. So the union of these subfields can't be the whole of K, and we can choose an $\alpha \in K$ which isn't contained in any smaller field. This must be a primitive element for K.
(This also makes it clear that primitive elements are very non-unique; in some sense "most" elements of K are primitive elements.)

10.2 Algebraic integers

We'd like to find more examples of rings like $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ above, which have interesting factorisation theories attached to them. Number fields themselves are not interesting in this way (in a field, every non-zero element is a unit). We want to pick out those algebraic numbers which "don't have any denominators" in some sense, just like $\mathbb{Z}[i]$ inside $\mathbb{Q}(i)$.

It turns out the good definition is the following:

Definition 10.7 We say $\alpha \in \mathbb{C}$ is an *algebraic integer* if there exists a *monic* polynomial $f(X) \in \mathbb{Z}[X]$ with $f(\alpha) = 0$. We write $\bar{\mathbb{Z}}$ for the set of algebraic integers.

Note the similarity to the definition of "algebraic number"; but here it really matters that f be monic. (Exercise: show that for *any* algebraic number α, we can find an $f \in \mathbb{Z}[X]$, usually not monic, with $f(\alpha) = 0$.)

Example 10.8 Clearly we have $\mathbb{Z} \subseteq \bar{\mathbb{Z}}$, since for any $n \in \mathbb{Z}$, $f(X) = X - n$ is a monic polynomial that it satisfies. Moreover, if $n \in \mathbb{Z}$ then $\sqrt{n} \in \bar{\mathbb{Z}}$.
Less obviously, $\omega = \frac{-1+\sqrt{-3}}{2} \in \bar{\mathbb{Z}}$, since it satisfies $X^2 + X + 1 = 0$.

Exercise 10.9 Show that if $\alpha \in \overline{\mathbb{Z}}$, then $\sqrt{\alpha} \in \overline{\mathbb{Z}}$.

Proposition 10.10 *For any algebraic number α, there exists some $N \in \mathbb{N}_+$ such that $N\alpha \in \bar{\mathbb{Z}}$.*

Proof Exercise. (Hint: if $f(X) = X^n + a_{n-1}X^{n-1} + \cdots + a_0 \in \mathbb{Q}[X]$ is the minimal polynomial of α, and $\beta = N\alpha$ for some N, then what is the minimal polynomial of β?) □

What's less obvious is how one would show that anything is *not* an algebraic integer! Fortunately, we have the following criterion:

Proposition 10.11 *An algebraic number $\alpha \in \mathbb{C}$ is an algebraic integer if and only if its* minimal *polynomial has integer coefficients.*

Proof Let $f \in \mathbb{Q}[X]$ be the minimal polynomial of α. If $f \in \mathbb{Z}[X]$, then clearly f is an algebraic integer.

Conversely, suppose f does not have integer coefficients, but there is some (larger-degree) monic integral polynomial h with $h(\alpha) = 0$. Then we must have $h(X) = f(X)g(X)$ for some $g \in \mathbb{Q}[X]$.

Let C be the least common multiple of the denominators of the coefficients of f, so that $Cf \in \mathbb{Z}[X]$, and similarly D for g. Then we clearly have $(Cf)(Dg) = (CD)h$. Now let p be a prime dividing CD. Clearly at least one coefficient of Cf is not divisible by p (since otherwise C/p would be the LCM of the denominators). Similarly at least one of the coefficients of Dg is not divisible by p. So $Cf \bmod p$ and $Dg \bmod p$ are non-zero in $\mathbb{F}_p[X]$. But their product CDh is zero, since $p \mid CD$ and h has integral coefficients. This contradicts the fact that $\mathbb{F}_p[X]$ is an integral domain. So CD must in fact be 1, i.e. both f and g are integral. □

Example 10.12

- If $x \in \mathbb{Q} - \mathbb{Z}$, then x is not an algebraic integer. (That is, we have $\bar{\mathbb{Z}} \cap \mathbb{Q} = \mathbb{Z}$).
- The number $\frac{1+\sqrt{3}}{2}$ is not an algebraic integer: it is a root of the polynomial $x^2 - x - \frac{1}{2}$, and since it clearly isn't in $\mathbb{Q}$, this must be the minimal polynomial.

It follows that a rational number is an algebraic integer iff it's an integer in the usual sense.

Exercise 10.13 (Warning!) Give a counterexample to show that is *not* true that if α is an algebraic integer, then every monic polynomial that f satisfies has to have integral coefficients.

10.3 Arithmetic with algebraic integers

For doing arithmetic with algebraic integers, the following characterisation is useful:

Proposition 10.14 *$\alpha \in \mathbb{C}$ is an algebraic integer if and only if $\mathbb{Z}[\alpha]$ is finitely generated as an abelian group.*

Proof If α satisfies a polynomial $f(X) = X^n + a_{n-1}X^{n-1} + \dots$, then α^n is in the $\mathbb{Z}$-span of $1, \dots, \alpha^{n-1}$, and by induction one can show that α^{n+1}, α^{n+2} etc are also in this span.

Conversely, if this group is finitely generated, then each generator can only mention finitely many powers of α, so there is some N such that $\{1, \dots, \alpha^N\}$ is a generating set. Hence α^{N+1} is in the $\mathbb{Z}$-span of $\{1, \dots, \alpha^N\}$, giving a monic integral polynomial that α satisfies. □

Corollary 10.15 *If α, β are algebraic integers then so are $\alpha \pm \beta$ and $\alpha\beta$.*

Proof Suppose α, β satisfy polynomials of degree M, N respectively. Consider the subgroup of $\mathbb{C}$ generated by $\{\alpha^i\beta^j : 0 \leqslant i < N, 0 \leqslant j < M\}$. This is finitely generated and contains $\alpha^r\beta^s$ for all $r, s \in \mathbb{N}$, so in particular it contains $(\alpha\beta)^j$ and $(\alpha \pm \beta)^k$ for all j, k. Since a subgroup of a finitely generated abelian group is finitely generated, the result follows. □

Thus the set $\overline{\mathbb{Z}}$ of all algebraic integers is a subring of of $\mathbb{C}$.

Remark 10.16 Note that the above proofs are *not constructive*: we've proved that $\alpha \pm \beta$ and $\alpha\beta$ satisfy monic polynomials in $\mathbb{Z}[X]$, but we haven't shown how to explicitly write down those polynomials.

Exercise 10.17 Find a monic polynomial $f(X) \in \mathbb{Z}[X]$ with $f(\sqrt{2} + \sqrt{3}) = 0$.

10.4 Rings of integers

Definition 10.18 If K is a number field, then we define $\mathcal{O}_K$, the *ring of integers of K*, as $K \cap \bar{\mathbb{Z}}$.

Note that if α is an algebraic integer, $\mathbb{Z}[\alpha]$ is contained in the ring of integers of $\mathbb{Q}(\alpha)$, but it might be smaller. For instance, $\mathbb{Z}[\sqrt{-3}]$ is not the ring of integers of $\mathbb{Q}(\sqrt{-3})$, because it doesn't contain ω.

Proposition 10.19 (Rings of integers of quadratic fields) *Let $d \in \mathbb{Z}$ with $d \neq 1$, and suppose d is not divisible by n^2 for any $n > 1$ (d is "square-free"). Then the ring of integers of $\mathbb{Q}(\sqrt{d})$ is given by*

$$\mathcal{O}_{\mathbb{Q}(\sqrt{d})} = \begin{cases} \mathbb{Z}\left[\frac{1+\sqrt{d}}{2}\right] & \text{if } d = 1 \bmod 4, \\ \mathbb{Z}[\sqrt{d}] & \text{otherwise.} \end{cases}$$

Proof First, note that $\frac{1+\sqrt{d}}{2}$ is a root of $X^2 - X + \frac{1-d}{4}$, so it is an algebraic integer iff $d = 1 \bmod 4$.

Conversely, let $\alpha = u + v\sqrt{d}$ with $u, v \in \mathbb{Q}$, and suppose $\alpha \in \bar{\mathbb{Z}}$. Then $\alpha' = u - v\sqrt{d}$ is also in $\bar{\mathbb{Z}}$, since it satisfies the same polynomial that α does; and hence $\alpha + \alpha' = 2u \in \bar{\mathbb{Z}} \cap \mathbb{Q} = \mathbb{Z}$. Similarly, $\alpha - \alpha' = 2v\sqrt{d} \in \bar{\mathbb{Z}}$; thus $(2v)^2 d \in \mathbb{Z}$, but since d is squarefree, this implies that $2v \in \mathbb{Z}$.

So, if α is an algebraic integer but doesn't lie in $\mathbb{Z}[\sqrt{d}]$, then we can subtract a $\mathbb{Z}$-linear combination of 1 and $\sqrt{d}$ to deduce that one of $\{\frac{1}{2}, \frac{\sqrt{d}}{2}, \frac{1+\sqrt{d}}{2}\}$ is an algebraic integer. Clearly $\frac{1}{2}$ and $\frac{\sqrt{d}}{2}$ are never algebraic integers (since $4 \nmid d$); and $\frac{1+\sqrt{d}}{2}$ is an algebraic integer iff $d = 1 \bmod 4$. □

Remark 10.20 Note that $\mathcal{O}_{\mathbb{Q}(\sqrt{d})}$ is isomorphic to $\mathbb{Z}^2$ as an abelian group: every element can be written uniquely in the form $a + b\lambda$ for $a, b \in \mathbb{Z}$, where $\lambda = \frac{1+\sqrt{d}}{2}$ or $\sqrt{d}$ respectively.
We will prove in the next chapter that for any number field K, $\mathcal{O}_K$ is isomorphic to $\mathbb{Z}^d$ as an abelian group, where $d = [K : \mathbb{Q}]$; but this requires a little more work.

Exercise 10.21 Use Propositions 10.14 and 10.19 to justify the claim we made in Chapter 8 that there are no rings "finitely larger than" $\mathbb{Z}[\sqrt{-5}]$.

We finish this section with a useful little result which will be helpful later on:

Proposition 10.22 *For any number field K and any non-zero $\alpha \in \mathcal{O}_K$, there exists a non-zero $\beta \in \mathcal{O}_K$ such that $\alpha\beta \in \mathbb{Z}$. That is, α divides some non-zero integer.*

Proof This is a disguised version of Proposition 10.10. Let $\gamma = 1/\alpha$. Then $\gamma \in \overline{\mathbb{Q}}$, so there is some $N \in \mathbb{N}_+$ such that $N\gamma$ is an algebraic integer. Let $\beta = N\gamma$ for any such N. Then $\beta = N/\alpha$ is in K, and it's an algebraic integer, so it's in $\mathcal{O}_K$; and we have $\alpha\beta = N$. □

CHAPTER 11

Determining the integer ring

We'll now study the ring $\mathcal{O}_K$, for K a number field, a bit more closely.

11.1 Norm and trace

If K is a number field, and $x \in K$, then we can consider the "multiplication by x" map $M_x : K \to K$, defined by $M_x(y) = xy$. This is clearly $\mathbb{Q}$-linear.

Definition 11.1 The norm $\mathrm{Nm}_{K/\mathbb{Q}}(x)$ and trace $\mathrm{Tr}_{K/\mathbb{Q}}(x)$ are the determinant and trace (in the sense of linear algebra) of M_x, viewed as a $\mathbb{Q}$-linear map $K \to K$.

One checks easily that norm is compatible with multiplication, and trace compatible with addition:

$$\mathrm{Nm}_{K/\mathbb{Q}}(xy) = \mathrm{Nm}_{K/\mathbb{Q}}(x)\,\mathrm{Nm}_{K/\mathbb{Q}}(y),$$
$$\mathrm{Tr}_{K/\mathbb{Q}}(x \pm y) = \mathrm{Tr}_{K/\mathbb{Q}}(x) \pm \mathrm{Tr}_{K/\mathbb{Q}}(y).$$

Moreover, if $x \neq 0$, then taking $y = x^{-1}$ in the first equation we deduce that $\mathrm{Nm}_{K/\mathbb{Q}}(x) \neq 0$, so $\mathrm{Nm}_{K/\mathbb{Q}}$ is a group homomorphism $K^\times \to \mathbb{Q}^\times$.

Example 11.2 Let $K = \mathbb{Q}(\sqrt{d})$ for a square-free integer d, and $x = a + b\sqrt{d}$. We claim that

$$\mathrm{Tr}_{K/\mathbb{Q}}(x) = 2a, \qquad \mathrm{Nm}_{K/\mathbb{Q}}(x) = a^2 - db^2.$$

To prove this, consider the basis $\{1, \sqrt{d}\}$ of K. In this basis, the matrix of M_x is

$$M_x = \begin{pmatrix} a & db \\ b & a \end{pmatrix},$$

and the result is now clear.

Remark 11.3 Notice that this depends on K: if we have two number fields K, L, and $x \in K \cap L$, then $\mathrm{Tr}_{K/\mathbb{Q}}(x)$ and $\mathrm{Tr}_{L/\mathbb{Q}}(x)$ are both well-defined, but they aren't the same in general. So it is a little dangerous to write "$\mathrm{Tr}(x)$" without specifying K, although we'll allow ourselves to do this sometimes when K is clear from context.

(Thus, for quadratic number fields, $\mathrm{Nm}_{K/\mathbb{Q}}(x)$ is what we were calling $N(x)$ before.)

Proposition 11.4 *If $\alpha \in \mathcal{O}_K$, then $\mathrm{Nm}_{K/\mathbb{Q}}(\alpha)$ and $\mathrm{Tr}_{K/\mathbb{Q}}(\alpha)$ are in $\mathbb{Z}$.*

Proof Let's suppose first that $K = \mathbb{Q}(\alpha)$. Then the numbers $1, \alpha, \alpha^2, \dots, \alpha^{d-1}$ are a $\mathbb{Q}$-basis of K, where $d = [K : \mathbb{Q}]$. In this basis, the matrix of M_α looks like

$$\begin{pmatrix} 0 & 0 & 0 & \dots & \star \\ 1 & 0 & 0 & \dots & \star \\ 0 & 1 & 0 & \dots & \star \\ \vdots & \vdots & \vdots & & \end{pmatrix}$$

where the $\star$'s in the right-most column are (up to sign) the coefficients of the minimal polynomial of α. This is a matrix of integers; so its determinant and trace are integers.

If K is larger than $\mathbb{Q}(\alpha)$, then one can check that

$$\mathrm{Nm}_{K/\mathbb{Q}}(\alpha) = \left(\mathrm{Nm}_{L/\mathbb{Q}}(\alpha)\right)^{[K:L]}, \qquad \mathrm{Tr}_{K/\mathbb{Q}}(\alpha) = [K : L] \cdot \mathrm{Tr}_{L/\mathbb{Q}}(x)$$

where $L = \mathbb{Q}(\alpha)$; and the result follows from the previous case. □

Remark 11.5 It's *not* true in general that if $x \in K$, and $\mathrm{Nm}_{K/\mathbb{Q}}(x)$ and $\mathrm{Tr}_{K/\mathbb{Q}}(y)$ are in $\mathbb{Z}$, then $x \in \mathcal{O}_K$ (although this is true if K is quadratic).

Exercise 11.6 Prove the following refinement of Proposition 10.22: for any $\alpha \in \mathcal{O}_K$, the divisibility $\alpha \mid \mathrm{Nm}_{K/\mathbb{Q}}(\alpha)$ holds in $\mathcal{O}_K$. [Hint: First reduce to the case $K = \mathbb{Q}(\alpha)$, then apply the Cayley–Hamilton theorem.]

11.2 Lattices and orders

We want to understand "how big" $\mathcal{O}_K$ is, and how it sits inside K, for an arbitrary number field K.

Definition 11.7 Let V be a finite-dimensional $\mathbb{Q}$-vector space. A *lattice* $\mathcal{L}$ in V is a subgroup of $(V, +)$ which is finitely-generated as a group. If $\mathcal{L}$ spans V as a $\mathbb{Q}$-vector space, we say $\mathcal{L}$ is *full*.

One can check (see Addendum below) that any lattice in V has to be isomorphic as a group to $\mathbb{Z}^m$ for some $m \leqslant \dim V$, with equality iff $\mathcal{L}$ is full. Moreover, a subgroup of a lattice is a lattice.

Example 11.8 For example, $\mathbb{Z}^2$ is obviously a full lattice in $\mathbb{Q}^2$. More subtly, so is $\{\begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{Z}^2 : a + b \text{ is even}\}$: it is generated by $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 2 \\ 0 \end{pmatrix}$.

Since number fields are finite-dimensional $\mathbb{Q}$-vector spaces, we can ask about lattices inside them. But a number field, unlike a general vector space, we know how to multiply things; so we can make the next definition:

Definition 11.9 An *order* in a number field K is a full lattice which is also a subring of K.

For instance, both $\mathbb{Z}$ and $\frac{1}{17}\mathbb{Z}$ are full lattices in $\mathbb{Q}$, and $\mathbb{Z}$ is an order, but $\frac{1}{17}\mathbb{Z}$ is not. Moreover, $\mathbb{Z}[i]$ is an order in $\mathbb{Q}(i)$, and both $\mathbb{Z}[\sqrt{-3}]$ and $\mathbb{Z}[\omega]$ are orders in $\mathbb{Q}(\sqrt{-3})$.

Exercise 11.10 Show that if A is an order in K, then $A \subseteq \mathcal{O}_K$.

11.3 The trace dual of a lattice

The crucial construction we'll use to understand lattices and orders in number fields is the following:

Definition 11.11 If K is a number field, and $\mathcal{L}$ is a subgroup of $(K, +)$, then the *trace dual* of $\mathcal{L}$ is defined by

$$\mathcal{L}^\vee = \{x \in K : \mathrm{Tr}_{K/\mathbb{Q}}(xy) \in \mathbb{Z} \quad \forall y \in \mathcal{L}\}.$$

Note that $\mathcal{L}^\vee$ is also a subgroup of $(K, +)$ (exercise). Moreover, taking the trace dual is *inclusion-reversing*: if $\mathcal{L} \subseteq \mathcal{M}$, then $\mathcal{L}^\vee \supseteq \mathcal{M}^\vee$.

Proposition 11.12 *If $\mathcal{L}$ is a full lattice in K, then the trace dual $\mathcal{L}^\vee$ is also a full lattice.*

Proof This is an instance of a general result (see Addendum below) applying to any finite-dimensional $\mathbb{Q}$-vector space equipped with a non-degenerate quadratic form.

To apply this in our situation, we need to check that the "trace form" $(x, y) \mapsto \mathrm{Tr}_{K/\mathbb{Q}}(xy)$ is a quadratic form (which is obvious), and that it is non-degenerate. So, suppose $x \in K$ satisfies $\mathrm{Tr}_{K/\mathbb{Q}}(xy) = 0$ for all $y \in K$. If $x \neq 0$, then we can take $y = x^{-1}$ and we have $\mathrm{Tr}(xy) = \mathrm{Tr}(1) = [K : \mathbb{Q}] \neq 0$, a contradiction. Hence we must have $x = 0$, showing that the trace form is non-degenerate. □

Exercise 11.13 Take $\mathcal{L} = \mathbb{Z}[i]$, considered as a lattice in $K = \mathbb{Q}(i)$, and calculate $\mathcal{L}^\vee$.

Proposition 11.14 *We have $\mathcal{O}_K^\vee \supseteq \mathcal{O}_K$.*

Proof Let $x \in \mathcal{O}_K$. Then for any $y \in \mathcal{O}_K$, we have $xy \in \mathcal{O}_K$ (because $\mathcal{O}_K$ is a ring), and hence $\mathrm{Tr}(xy) \in \mathbb{Z}$ by Proposition 11.4. Thus $x \in \mathcal{O}_K^\vee$. □

Corollary 11.15 *$\mathcal{O}_K$ is an order in K.*

Proof We know $\mathcal{O}_K$ is a subring, so we need to show $\mathcal{O}_K$ is a full lattice.

First we claim $\mathcal{O}_K$ *contains* a full lattice. Let $x_1, \ldots, x_d$ be a $\mathbb{Q}$-basis of $\mathcal{O}_K$. If we multiply each x_i by a non-zero integer, then the new set is still a basis, so by Proposition 10.10 we

can arrange that the x_i are all in $\mathcal{O}_K$. Thus $\mathcal{L} = \mathbb{Z}x_1 + \cdots + \mathbb{Z}x_d$ is a full lattice contained in $\mathcal{O}_K$. (Note there's no reason for it to be an order.)

Now, if $\mathcal{L}$ is a full lattice such that $\mathcal{L} \subseteq \mathcal{O}_K$, then $\mathcal{L}^\vee \supseteq \mathcal{O}_K^\vee$; and $\mathcal{L}^\vee$ is also a full lattice. Since

$$\mathcal{L}^\vee \supseteq \mathcal{O}_K^\vee \supseteq \mathcal{O}_K \supseteq \mathcal{L}.$$

So we have sandwiched $\mathcal{O}_K$ between two full lattices, $\mathcal{L}$ and $\mathcal{L}^\vee$. Since $\mathcal{O}_K \supseteq \mathcal{L}$, we know that $\mathcal{O}_K$ spans K as a $\mathbb{Q}$-vector space. Since $\mathcal{O}_K \subseteq \mathcal{L}^\vee$, we know that $\mathcal{O}_K$ is finitely-generated. Thus $\mathcal{O}_K$ is a full lattice (and hence an order). □

Remark 11.16 Note that this corollary also gives us a way of computing $\mathcal{O}_K$. Assume we know a primitive element α of K. By scaling if necessary, we can suppose $\alpha \in \mathcal{O}_K$. Then $A = \mathbb{Z}[\alpha]$ is a full lattice (and indeed an order) contained in $\mathcal{O}_K$. The quotient $A^\vee/A$ is finite (and explicitly computable); and for each element $x + A$ of this quotient, we can determine whether $x + A \in \mathcal{O}_K/A$, by calculating the minimal polynomial of x.

(This is essentially what we did in the previous chapter for quadratic fields $\mathbb{Q}(\sqrt{d})$, taking A to be the order $\mathbb{Z}[\sqrt{d}]$.)

11.4 Addendum: Some $\mathbb{Z}$-linear algebra

Just for completeness, we'll outline the proofs of a few results about subgroups of $\mathbb{Z}^n$ which we used in this chapter. The proofs in this section are **non-examinable**.

11.4.1 Subgroups of $\mathbb{Z}^n$

We begin with Theorem 4.4 of the *Algebra* module, which says the following:

Proposition 11.17 *Let G be a subgroup of the additive group $(\mathbb{Z}, +)$. Then we have*

$$G = m\mathbb{Z} = \{mz : z \in \mathbb{Z}\}$$

for a uniquely determined $m \geqslant 0$. In particular, either $G = \{0\}$ or G is isomorphic to $\mathbb{Z}$ itself.

Motivated by this, what can we say about subgroups of $\mathbb{Z}^n$, for an arbitrary $n \geqslant 1$?

Theorem 11.18 *Let H be a subgroup of $\mathbb{Z}^n$. Then there is a unique $m \in \mathbb{N}$ such that $H \cong \mathbb{Z}^m$, and we have $0 \leqslant m \leqslant n$.*

Proof of uniqueness Note that $H \cong \mathbb{Z}^m$ iff there exists a set of m elements $h_1, \ldots, h_m \in H$ which are independent generators, i.e. every $x \in H$ can be written as $x = \sum a_i h_i$ for a *unique* $(a_1, \ldots, a_m) \in \mathbb{Z}^m$.

Let W be the $\mathbb{Q}$-vector space spanned by H. Then $h_1, \ldots, h_m$ clearly span W as a $\mathbb{Q}$-vector space. They are also $\mathbb{Q}$-linearly independent, because if we had a nontrivial $\mathbb{Q}$-linear relation between them, we could clear denominators to get a nontrivial $\mathbb{Z}$-linear relation. Hence we must have $m = \dim W$, which clearly satisfies $0 \leqslant m \leqslant n$. □

Proof of existence To deduce existence, we'll use induction on n. The result is trivial for $n = 0$, so assume it holds for $n - 1$.

Given $H \subseteq \mathbb{Z}^n$, consider the "forget the last entry" map $\mathbb{Z}^n \to \mathbb{Z}^{n-1}$. The image $\bar{H}$ of H is a subgroup of $\mathbb{Z}^{n-1}$, so (by the induction hypothesis) we can find an independent generating set $\bar{h}_1, \dots, \bar{h}_r$, for some $r \leqslant n-1$. Choose arbitrary elements $h_1, \dots, h_r$ of H mapping to $\bar{h}_1, \dots, \bar{h}_r$. Then any $h \in H$ can be uniquely written as $\sum_{i=1}^{r} a_i h_i + (0, \dots, 0, x)$, for some $(a_1, \dots, a_r) \in \mathbb{Z}^r$ and $x \in \mathbb{Z}$.

Now consider the subgroup $X = \{x \in \mathbb{Z} : (0, \dots, 0, x) \in H\}$. This is a subgroup of $\mathbb{Z}$, so it must be either $\{0\}$, or $d\mathbb{Z}$ for some $d \geqslant 1$. If $X = \{0\}$, then $h_1, \dots, h_r$ are an independent generating set of H. If $X = d\mathbb{Z}$ for $d \geqslant 1$, then we set $h_{r+1} = (0, \dots, 0, d)$; then $(h_1, \dots, h_{r+1})$ are an independent generating set. □

What can we say about subgroups $H \cong \mathbb{Z}^n$ which are isomorphic to $\mathbb{Z}^n$? Of course, this doesn't imply that H is the whole of $\mathbb{Z}^n$ (as we've already seen for $n = 1$). What we can say is the following:

Theorem 11.19 *For a subgroup $H \subseteq \mathbb{Z}^n$, the following are equivalent:*

- *H is isomorphic to $\mathbb{Z}^n$;*
- *the index $[\mathbb{Z}^n : H]$ is finite.*

Proof If $[\mathbb{Z}^n : H]$ is finite, of size d say, then every element of the quotient $\mathbb{Z}^n/H$ has order dividing d ("element order divides group order"); so $dv \in H$ for every $v \in \mathbb{Z}^n$. In particular, H contains de_j for each j, and thus spans $\mathbb{Q}^n$. So it must be isomorphic to $\mathbb{Z}^n$.

Conversely, if $H \cong \mathbb{Z}^n$, then H spans $\mathbb{Q}^n$, so for each j, e_j must be a $\mathbb{Q}$-linear combination of H. Thus $\mathbb{Z}^n/H$ is an abelian group generated by finitely many elements, each of which has finite order, which is sufficient to imply that $\mathbb{Z}^n/H$ is finite.[1] □

Remark 11.20 One can show that if H is a finite-index subgroup of $\mathbb{Z}^n$, and $h_1, \dots, h_n$ is an independent generating set of H, then we have

$$[\mathbb{Z}^n : H] = |\det A|,$$

where A is the matrix with the h_i as rows.

11.4.2 Lattices in $\mathbb{Q}$-vector spaces

Now suppose V is a finite-dimensional $\mathbb{Q}$-vector space; without loss of generality $V = \mathbb{Q}^n$ for some n.

If $\mathcal{L}$ is a finitely-generated subgroup of $\mathbb{Q}^n$, then we have $\mathcal{L} \subseteq N^{-1}\mathbb{Z}^n$ for some $N \geqslant 1$ (it suffices to take the LCM of the denominators of any generating set of $\mathcal{L}$). Since multiplying by N is an isomorphism $N^{-1}\mathbb{Z}^n \cong \mathbb{Z}^n$, we conclude that $\mathcal{L}$ is isomorphic to $\mathbb{Z}^m$ for some $0 \leqslant m \leqslant n$, as before.

[1] Careful: the abelian property is needed here – there exists a famous example of an infinite non-abelian group, the *modular group*, generated by two elements of order 2 and 3 respectively.

Remark 11.21 Not all subgroups of $(V, +)$ are lattices: for instance, V itself is a subgroup of V, but it is not a lattice (except in the trivial case $V = \{0\}$).

(Exercise: can you find a *proper* subgroup of $(\mathbb{Q}, +)$ which is not a lattice?)

11.4.3 Duals of lattices

Let V be a finite-dimensional $\mathbb{Q}$-vector space, and suppose we are given a symmetric[2] bilinear form

$$\langle -, - \rangle : V \times V \to \mathbb{Q}.$$

Then, for a lattice $\mathcal{L} \subset V$, we can define

$$\mathcal{L}^\vee = \{x \in V : \langle x, y \rangle \in \mathbb{Z} \quad \forall y \in \mathcal{L}\}.$$

Let's now assume the pairing on V is *non-degenerate*, i.e. if $x \in V$ satisfies $\langle x, y \rangle = 0$ for all $y \in V$, then $x = 0$.

Proposition 11.22 *If $\mathcal{L}$ is a full lattice, then so is $\mathcal{L}^\vee$.*

Proof Let $\mathbf{v} = (v_1, \dots, v_d)$ be an (ordered) independent generating set of $\mathcal{L}$; then it is also a $\mathbb{Q}$-basis of V, since $\mathcal{L}$ is full. Let M be the matrix with (i, j) entry $\langle v_i, v_j \rangle$ (the matrix of the bilinear form).

Since the pairing $\langle -, - \rangle$ is non-degenerate, M is non-singular, so it has an inverse M^{-1}.

Let b_i be the i-th row of M^{-1}; and let $w_i = b_1 v_1 + \cdots + b_d v_d$ be the vector whose coordinates in the basis $\mathbf{v}$ are b_i. Then $\mathbf{w} = (w_1, \dots, w_d)$ is also a basis of V, and one computes that

$$\langle w_i, v_j \rangle = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Hence, if we write an arbitrary $x \in V$ as $x = \sum c_i w_i$ for some $\vec{c} \in \mathbb{Q}^d$, we have $x \in \mathcal{L}^\vee$ iff $c_i \in \mathbb{Z}$ for all i. Thus $\mathcal{L}^\vee$ is precisely the $\mathbb{Z}$-linear combinations of the basis $\mathbf{w}$, showing that it is a full lattice. □

Remark 11.23 This is related to the notion of *dual bases* from *Linear Algebra II*. More precisely, you saw in that module that a nondegenerate bilinear form defines an isomorphism from V to its dual space V^*. You also saw that for any basis $\mathbf{v} = (v_1, \dots, v_n)$ of V there is a dual basis $(\nu_1, \dots, \nu_n)$ of V^* with $\nu_i(v_j) = \delta_{ij}$. The basis $\mathbf{w}$ in the above proof, satisfying $\langle w_i, v_j \rangle = \delta_{ij}$, is given by transporting the dual basis $\boldsymbol{\nu}$ along the isomorphism $V^* \cong V$.

Exercise 11.24 Show that if $\mathcal{L}$ is a full lattice, then $\mathcal{L}^{\vee\vee} = \mathcal{L}$.

[2] This is not strictly needed, it's just for notational simplicity.

CHAPTER 12

Ideals in number fields

12.1 Ideals

Let K be a number field. We're going to study *ideals* in the ring of integers of K. (The zero ideal is an ideal, but it's not very interesting, so henceforth, when we say "ideal" we always mean *nonzero* ideal.)

Definition 12.1 (Notation for ideals) For any commutative ring R and elements $x_1, \dots, x_k$ of R, write $\langle x_1, \dots, x_k \rangle_R$ for the set $\{r_1 x_1 + \cdots + r_k x_k : r_1, \dots, r_k \in R\}$, which is an ideal of R (the ideal generated by the x_i). We omit the subscript R if it's obvious from context.

Notice that any $\alpha \in \mathcal{O}_K$ gives us an ideal – the principal ideal $\langle \alpha \rangle = \{\alpha x : x \in \mathcal{O}_K\}$. However, since integer rings aren't always PIDs, there can be more ideals which aren't of this form.

Example 12.2 Let $R = \mathbb{Z}[\sqrt{-5}] = \mathcal{O}_{\mathbb{Q}(\sqrt{-5})}$, and let I be the ideal $\langle 2, 1 - \sqrt{-5} \rangle$ of R. We claim this ideal is not principal.

Assume for contradiction that α is a generator. Then α must divide 2, so $N(\alpha) \mid N(2) = 4$; and also $N(\alpha) \mid N(1 - \sqrt{-5}) = 6$. So $N(\alpha)$ must be 1 or 2.
If $N(\alpha)$ were equal to 1, then I would be the unit ideal. But this is not possible, since every element of I has the form $x + y\sqrt{-5}$ with $x = y \bmod 2$ (exercise!), so $1 \notin I$.
Hence $N(\alpha)$ must be 2. But the equation $x^2 + 5y^2 = 2$ obviously has no solutions, so we have a contradiction. □

Exercise 12.3 Generalise the above! Show that if $d \in \mathbb{N}_+$ is square-free with $d \neq 3 \bmod 4$, $p \nmid d$ is a prime such that $\left(\frac{-d}{p}\right) = 1$, and t is a square root of $-d \bmod p$, then the ideal $\langle p, \sqrt{-d} - t \rangle$ is principal in $\mathbb{Z}[\sqrt{-d}]$ if and only if $x^2 + dy^2 = p$ has an integer solution. Can you formulate an analogue for $d = 1 \bmod 4$? What about $d < 0$?

Definition 12.4 (Product of ideals) Let I and J be ideals in $\mathcal{O}_K$. Then we define

$$IJ = \{i \cdot j : i \in I, j \in J\}.$$

You should check that ideal multiplication is compatible with element multiplication, i.e. $\langle\alpha\rangle\langle\beta\rangle = \langle\alpha\beta\rangle$. Moreover, $\langle\alpha\rangle = \langle\beta\rangle$ iff α and β are associates. So we get maps

$$(\dagger) \qquad (\mathcal{O}_K - \{0\}) \twoheadrightarrow \frac{(\mathcal{O}_K - \{0\})}{\{\text{units}\}} \hookrightarrow \{\text{nonzero ideals}\},$$

which are compatible with multiplication (and send the identity to the identity). If $\mathcal{O}_K$ is a PID (in particular if it's Euclidean), then the second map is a bijection.

The moral of the next few sections will be that there is **always** a notion of "unique prime factorisation" for *ideals*. When $\mathcal{O}_K$ is a PID, we get unique factorisation for *elements* from this using the bijectivity of the second map in (†). Conversely, when $\mathcal{O}_K$ is not a PID, we never have unique prime factorisation in $\mathcal{O}_K$; but the non-principal ideals are precisely the "extra stuff" we need to add to get unique factorisation back again.

12.2 Factoring ideals

Remember that an ideal I in any (commutative) ring A is said to be a *prime ideal* if $I \neq A$, and for all $x, y \in A$ we have $xy \in I \Rightarrow x \in I$ or $y \in I$. This obviously generalises the definition of prime *elements*: an element is prime iff the principal ideal it generates is a prime ideal.

Proposition 12.5 *Let I be a non-zero ideal in $\mathcal{O}_K$, for K a number field. Then I is prime if and only if it is* maximal*, i.e. $I \neq \mathcal{O}_K$ and there is no ideal J such that $I \supsetneq J \supsetneq \mathcal{O}_K$.*

Proof We know that I is prime iff $R = \mathcal{O}_K/I$ is an integral domain (this is just rewriting the definition).

We claim that

(a) this quotient R is *finite*,
(b) a finite integral domain is automatically a field.

To prove (a), we note that I is non-zero, so it contains a non-zero $\alpha \in \mathcal{O}_K$. Moreover, α must divide a non-zero integer C, by Theorem 10.22. Thus $C \in \mathcal{O}_K$; and $\mathcal{O}_K/C$ is finite, since $\mathcal{O}_K$ is finitely-generated and $C \neq 0$. Thus R is a quotient of a finite thing, so it's also finite[1].

To prove (b), suppose R is an integral domain and $0 \neq x \in R$. Then multiplication by x is a map $R \to R$ which is injective, by the integral-domain assumption. But an injection from a finite set to itself must be a bijection; so 1 is in the image and hence x is invertible.

To finish the proof, we note that for any commutative ring A and ideal I of A, the ideal I is maximal iff A/I is a field (exercise). So

$$(I \text{ prime}) \iff (R \text{ int. domain}) \iff (R \text{ field}) \iff (I \text{ maximal}). \qquad \square$$

Corollary 12.6 *Let $0 \neq \alpha \in \mathcal{O}_K$. Then:*

- *α is a prime element iff there is no ideal strictly containing $\langle\alpha\rangle$ except the unit ideal.*

[1] One can show using Remark 11.20 that for any $0 \neq \alpha \in \mathcal{O}_K$ we have $\#(\mathcal{O}_K/\alpha) = |\operatorname{Nm}_{K/\mathbb{Q}}(\alpha)|$.

- *α is indecomposable iff there is no* principal *ideal strictly containing $\langle\alpha\rangle$ except the unit ideal.*

Proof The first assertion is just the previous proposition applied to $\langle\alpha\rangle$. The second is obvious, since $\langle\beta\rangle \supset \langle\alpha\rangle$ iff $\beta \mid \alpha$. □

In particular, if $\mathcal{O}_K$ is a PID, then prime elements and indecomposable elements coincide (something you saw without proof in the *Algebra* module).

Theorem 12.7 (Dedekind) *Let I, J be ideals in $\mathcal{O}_K$ with $I \subseteq J$. Then there exists an ideal H such that $I = HJ$.*

This is surprisingly hard, and we're not going to prove it in this course. For a proof see Stewart & Tall.

Remark 12.8 This theorem would be false if we replaced $\mathcal{O}_K$ with a ring like $\mathbb{Z}[\sqrt{-3}]$, which isn't equal to the full ring of integers of its parent number field.

Corollary 12.9 *Multiplying by a non-zero ideal is injective: that is, if H, I, J are (nonzero!) ideals of $\mathcal{O}_K$, and $HI = HJ$, then $I = J$.*

Proof Firstly, we suppose H is principal, say $H = \langle x\rangle$. Then HI is exactly the set of elements $xi : i \in I$, and similarly HJ. Since multiplication by x is injective, it follows that $I = \{y : xy \in HI\} = \{y : xy \in HJ\} = J$.

For a general ideal H, we choose a non-zero element $x \in H$. Then $H \supseteq \langle x\rangle$, so $\langle x\rangle = H'H$ for some H'. So if $HI = HJ$ then $H'HI = H'HJ$, i.e. $\langle x\rangle I = \langle x\rangle J$, and the previous paragraph shows that $I = J$. □

Theorem 12.10 (Unique factorisation of ideals) *Any nonzero ideal of $\mathcal{O}_K$ is equal to a product of finitely many prime ideals, and its expression in this form is unique up to ordering.*

Proof For any I, there are finitely many ideals containing I, since they biject with the ideals of the finite quotient ring $\mathcal{O}_K/I$. Hence we can find one which is maximal (not contained in any other ideal). Let P be such an ideal. Then P divides I, so $I = PJ$ for some J.

Clearly J can't be equal to I, since if $I = PI$ then $\mathcal{O}_K = P$, a contradiction. So J is strictly larger than I. By induction on the size of $\mathcal{O}_K/I$, we may assume that J is a product of maximal ideals, hence so is I.

The proof of uniqueness proceeds exactly as before. □

Example 12.11 (important) Let's use what we know about unique factorization of *ideals* to understand better how unique factorisation of *elements* fails in $\mathbb{Z}[\sqrt{-5}]$. Remember that we had two different factorisations of 6 into indecomposable elements:

$$6 = 2 \cdot 3 = (1 + \sqrt{-5}) \cdot (1 - \sqrt{-5}),$$

One checks that the ideals

$$\begin{aligned} \mathfrak{p} &= \langle 2, 1 + \sqrt{-5} \rangle, \\ \mathfrak{q}_1 &= \langle 3, 1 + \sqrt{-5} \rangle, \\ \mathfrak{q}_2 &= \langle 3, 1 - \sqrt{-5} \rangle \end{aligned}$$

are all prime; but none of them can be principal, since that would contradict the indecomposability of 2 and 3 in $\mathbb{Z}[\sqrt{-5}]$.
Now, one can show (exercise!)

$$\mathfrak{p}^2 = \langle 2 \rangle, \qquad \mathfrak{q}_1\mathfrak{q}_2 = \langle 3 \rangle,$$
$$\mathfrak{p}\mathfrak{q}_1 = \langle 1 + \sqrt{-5} \rangle, \qquad \mathfrak{p}\mathfrak{q}_2 = \langle 1 - \sqrt{-5} \rangle.$$

So the (unique) factorisation of the *ideal* $\langle 6 \rangle$ is

$$\langle 6 \rangle = \mathfrak{p}^2\mathfrak{q}_1\mathfrak{q}_2,$$

and the rival factorisations of the *element* 6 into indecomposables correspond to the ways of grouping the factors into subsets whose product is principal:

$$\langle 6 \rangle = (\mathfrak{p}^2)(\mathfrak{q}_1\mathfrak{q}_2) = (\mathfrak{p}\mathfrak{q}_1)(\mathfrak{p}\mathfrak{q}_2).$$

Exercise 12.12 Compute the factorisation of $\langle 21 \rangle$ into prime ideals in $\mathbb{Z}[\sqrt{-5}]$. Hence show that there are exactly 3 distinct factorisations of 21 into indecomposable elements, up to units and re-ordering.

12.3 The class group

We're now going to cook up an algebraic object which *measures* how badly ideals can fail to be principal (and thus how badly unique factorisation fails for elements).

Definition 12.13 A *fractional ideal* of $\mathcal{O}_K$ is a subset of K of the form

$$\mathcal{O}_K x_1 + \cdots + \mathcal{O}_K x_n$$

for some $x_1, \ldots, x_n \in K$ (not all of which are zero).

Thus, a fractional ideal contained in $\mathcal{O}_K$ is just an ideal, but things like $\frac{1}{2}\mathcal{O}_K$ are also fractional ideals.

Note that one can multiply fractional ideals to get new fractional ideals; and it follows from Dedekind's theorem that every fractional ideal has an inverse. Along with some easy checks for associativity etc, this shows that fractional ideals form an abelian group.

Definition 12.14 The *class group* of K is the quotient

$$\mathrm{Cl}_K = \frac{\{\text{fractional ideals}\}}{\{\text{principal fractional ideals}\}}.$$

We'll now state one of the most important theorems in algebraic number theory:

Theorem 12.15 *For any number field K, the class group Cl_K is* finite.

We're not going to prove it in this course (see Stewart & Tall for a proof)[2]. It says that although unique factorisation can fail – because there are non-principal ideals – it only "fails finitely badly".

Example 12.16 Going back to Example 12.11, the ideal $\mathfrak{p}$ is not principal (since $x^2 + 5y^2 = 2$ has no solutions) but $\mathfrak{p}^2$ is principal, so $[\mathfrak{p}]$ is a nontrivial element of Cl_K of order 2. Since $\mathfrak{p}\mathfrak{q}_1$ and $\mathfrak{p}\mathfrak{q}_2$ are principal, all three of the ideals $\{\mathfrak{p}, \mathfrak{q}_1, \mathfrak{q}_2\}$ all lie in this nontrivial ideal class.
It turns out that this is the only non-trivial element of the class group, so $\mathrm{Cl}_K \cong C_2$.

12.4 Cyclotomic fields, and Fermat's Last Theorem

Definition 12.17 The n-th *cyclotomic field* is the number field $\mathbb{Q}(\zeta_n)$, where $\zeta_n = \exp(2\pi i/n)$.

This is indeed a number field, because $(\zeta_n)^n = 1$, so ζ_n is algebraic. One can check that the ring of integers is equal to $\mathbb{Z}[\zeta_n]$.

Theorem 12.18 (Kummer) *Let p be an odd prime, and suppose that p does not divide the order of the class group of the p-th cyclotomic field $\mathbb{Q}(\zeta_p)$. Then there are no solutions to Fermat's equation $x^n + y^n = z^n$ with n divisible by p.*

The idea of Kummer's proof was to write $y^n = x^n - z^n$ and factor this in $\mathbb{Z}[\zeta_p]$ as $(x - z)(x - \zeta_p z) \dots (x - \zeta_p^{p-1} z)$. For simplicity, suppose $xyz \neq 0 \bmod p$; then one can show that the factors on the right are pairwise coprime.

If $\mathbb{Z}[\zeta_p]$ were a PID, then – by considering prime factorisations – each of the terms must itself be a p-th power (up to units); and this eventually gives enough information to deduce that no such x, y, z exist.

Kummer realised that one can push through the same argument as long as the class group has prime-to-p order (it doesn't have to be trivial), and this seems to hold for "most" primes p – there are very few primes such that the class group is trivial, but lots for which it has prime-to-p order. So this proves Fermat's Last Theorem for a large set of exponents n, although not all of them.

[2] There is a simpler proof in the special case of imaginary quadratic fields, i.e. $\mathbb{Q}(\sqrt{-D})$ with $D > 0$, which would be a nice project for a bachelor thesis.

Remark 12.19 Several earlier mathematicians had tried to make such an argument *assuming* that unique factorisation worked in $\mathbb{Z}[\zeta_p]$, which is of course false in general. Kummer *invented* the whole machine of ideal theory and class groups in order to sort out the mess!