Trophy Hunting in the Infinite Jungle of Finitude

Janne Junnila

Compiled on December 3, 2015

[1] http://creativecommons.org/licenses/by-nc-sa/4.0/

FOREWORD

Dear fellow problem solvers, coders and mathematicians,

Welcome to this hunting trip into the jungle of mathematical objects. This time we will be venturing to the section of the forest where most of these creatures are discrete and finite.

Although a vast majority of the things we encounter are huge and dangerous beasts that we so far have no hope of defeating, the tropical forest itself is immensely beautiful. Every now and then we meet some reasonably sized species, even cute ones, that we can tame or shoot down. When this happens, we collect them as *hunting memoirs* to be recalled when we next time meet something similar. In this book several of these trophies will be presented.

One of the target audiences of the book are competitive programmers hoping to learn mathematical theory that would help them during contests. For them every competition is a hunting trip of its own, and on top of the hunting memoirs of knowledge, different kind of trophies such as glory and recognition are awarded to those who win. Of course, for most of them, the most important part is the problem solving itself and the social context around it. In any case, to make a contact with this group, the focus will be on problems that have algorithmic solutions and C++-snippets will be provided for the various algorithms developed.

Another group that would perhaps like to take part on this expedition are mathematicians that are interested in the computational side of things. Everything is written with mathematical rigour, so these people should find themselves at home.

The choice of topics has mostly been influenced by what I myself have run into. This means that there is a lot of stuff that has appeared on **Project Euler**, for instance. I have intentionally not covered basic algorithms or techniques that appear in competitive programming, so if you want to learn about shortest paths, segment trees, binary search or dynamical programming, I have collected some reading suggestions at the end of the book.

One of the goals of the book is to look for methods and general theories that makes solutions to various problems look streamlined when approached from the right angle. Sometimes in this jungle there is nothing as fun as hitting a fly with an elephant gun, and it is my aim to help you identify those guns.

As a prerequisite the reader should know a little bit of mathematics so that reading mathematical text feels comfortable. In particular the reader should be familiar with basic set theory and logic used for everyday mathematical discourse.

Now, go pack your jungle hat and machete, I wish you a fruitful safari!

Janne Junnila

PART I

NUMBER THEORY

Classical number theory deals with properties of integers. The set of all integers is usually denoted by the symbol $\mathbf{Z}$, and the set of all non-negative integers is denoted by $\mathbf{N}$.

In this first part of the book we will approach classical number theory with modern methods. When digested, ring theory provides a unified way of thinking about many topics that traditionally were hard problems of their own. This is why we will introduce the basics of commutative rings at the same time as we navigate through the foundations of elementary number theory.

The algebra developed will serve as a foundation for further topics where such tools become indispensable. The methods will be used to reason about diophantine equations (Pythagorean triples, Pell's equation, etc.) by extending the ring of integers and examining the equations in these extensions.

1 ALGEBRAIC STRUCTURE OF $\mathbf{Z}$

In this chapter we will focus on the algebraic structure of the integers. In other words we will look at the interplay of addition and multiplication, exploring topics such as greatest common divisors or prime numbers. Many of these topics may already be quite familiar to most of the readers. Therefore the emphasis is on a coherent and powerful point of view based on notions from abstract algebra. To accomplish this, we will introduce some basic theory of commutative rings as we go.

1.1 COMMUTATIVE RINGS

Let's dive straight into it.

Definition 1.1 A set R together with binary operations $+$ and $\cdot$ form a **commutative ring** if the following properties hold:

associativity – for all $a, b, c \in R$,

$$a + (b + c) = (a + b) + c,$$
$$a \cdot (b \cdot c) = (a \cdot b) \cdot c$$

identity elements – there exist elements $0 \in R$ and $1 \in R$ such that for all $r \in R$,

$$r + 0 = 0 + r = r$$
$$r \cdot 1 = 1 \cdot r = r$$

additive inverses – for every $r \in R$ there exists an element $-r \in R$ such that

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

distributivity – for all $a, b, c \in R$,

$$a \cdot (b + c) = a \cdot b + a \cdot c,$$
$$(b + c) \cdot a = b \cdot a + c \cdot a$$

commutativity – for all $a, b \in R$,

$$a + b = b + a,$$
$$a \cdot b = b \cdot a$$

A couple of syntactical notes are in order:

- We often omit the multiplication symbol $\cdot$ and write ab instead of $a \cdot b$.
- Usually we write $a - b$ instead of $a + (-b)$.

It is easy to see that the integers under the usual addition and multiplication form a commutative ring. Other examples include the rational ($\mathbf{Q}$), real ($\mathbf{R}$) and complex ($\mathbf{C}$) numbers.

One difference between $\mathbf{Z}$ and the other three rings mentioned is the following: An element $r \in R$ is called **invertible** (or a **unit**) if there exists an element $s \in R$ such that $rs = 1$. Clearly the only invertible integers are 1 and -1. Contrast to that, in the rings $\mathbf{Q}$, $\mathbf{R}$ and $\mathbf{C}$ all elements except 0 are invertible.

Definition 1.2 A non-zero (i.e. $R \neq \{0\}$) commutative ring where every non-zero element is invertible is called a **field**.

Thus $\mathbf{Q}$, $\mathbf{R}$ and $\mathbf{C}$ are all fields.

Definition 1.3 A non-zero commutative ring R is an **integral domain** if for all $r, s \in R$ the equality $rs = 0$ implies that $r = 0$ or $s = 0$.

In other words, a non-zero ring R is an integral domain if and only if the product of two non-zero elements is non-zero.

Exercise 1.4 Show that a field is an integral domain.

Exercise 1.5 Show that R is an integral domain if and only if it has the following cancellative property: if $a, x, y \in R$ and $a \neq 0$, then the equality $ax = ay$ implies that $x = y$.

We haven't yet had an example of a ring that is not an integral domain. Let us therefore close the section with the following exercise.

Exercise 1.6 Let $R = \{0, 1, a, b\}$ be a ring of 4 elements such that $1 + 1 = a$. Show that such a ring exists and is unique and that it is not an integral domain.

1.2 IDEALS AND QUOTIENT RINGS

We will next describe ideals, which are a way to extend the notion of divisibility. Assume that a and b are two integers that are both divisible by some integer $d \neq 0$. Then also $a + b$ is divisible by d. Moreover you cannot remove this factor by multiplying: ac and bc are still going to be divisible by d for any integer c. These two properties are abstracted away into the definition of an ideal.

Definition 1.7 Let R be a commutative ring. A subset $I \subset R$ is called an **ideal** if

- for all $a, b \in I$ we have $a + b \in I$,
- for all $r \in R$ and $a \in I$ we have $ra \in I$.

Example 1.8 The set $\{..., -12, -6, 0, 6, 12, ...\}$ is an ideal in $\mathbf{Z}$. It is generated by 6, and a number is divisible by 6 if and only if it lies in this set.

Definition 1.9 We let

$$\langle a_1, ..., a_n \rangle := \{r_1 a_1 + ... + r_n a_n : r_1, ..., r_n \in R\}$$

denote the ideal generated by $a_1, ..., a_n \in R$. An ideal generated by a single element is called **principal**.

Definition 1.10 An integral domain in which *every* ideal is principal is called a **principal ideal domain** (or **PID**).

Exercise 1.11 What is $\langle 3, 5 \rangle$ in $\mathbf{Z}$?

We will see in the next section that $\mathbf{Z}$ is a PID. Every field is also a PID, in fact look at the following exercise.

Exercise 1.12 Show that ideals in fields are boring.

Let $a, b \in R$. If $b \in \langle a \rangle$, we say that a **divides** b, which we also write as $a \mid b$. This is the same as saying that there exists $r \in R$ such that $b = ra$. If R is an integral domain and $a \neq 0$, such r is unique when it exists. (Why?)

One of the reasons we are interested in ideals is that we can form new rings by considering the elements *modulo* I for some ideal I.

Definition 1.13 Let $I \subset R$ be an ideal. Two elements $a, b \in R$ are said to be equivalent modulo I if $a - b \in I$. Let R/I denote the set of equivalence classes $[r]$ under this relation. Then R/I can be made into a ring by defining $[a] + [b] = [a + b]$ and $[a] \cdot [b] = [a \cdot b]$. Rings obtained this way are called **quotient rings**.

Exercise 1.14 Show that the definition makes sense. First show that $a \sim b \Leftrightarrow a - b \in I$ really is an equivalence relation. Then show that the definitions of addition and multiplication do not depend on the choice of representatives.

It is customary to write $r \equiv s \pmod{I}$ if r and s belong to the same equivalence class modulo I. In the case I is a principal ideal generated by $a \in R$, we also write $r \equiv s \pmod{a}$.

A principal ideal $\langle a \rangle$ in a ring R can also be written as aR.

Let $m \geq 1$. The ideal $\langle m \rangle$ in $\mathbf{Z}$ defines a quotient ring $\mathbf{Z}/m\mathbf{Z}$ consisting of the m elements $\{[0], [1], ..., [m-1]\}$. Two integers are equivalent modulo $\langle m \rangle$ if and only if they give the same remainder upon division by m. In C++ the representative of a modulo m in the range $0, ..., m - 1$ can thus be obtained by calculating `a%m`, assuming that a is non-negative. If a is negative, it can be calculated as `a%m + m`.

There are two important special types of ideals that we have yet to discuss.

Definition 1.15 An ideal $I \subset R$ is a **prime ideal** if

- $ab \in I$ implies $a \in I$ or $b \in I$,
- $I \neq R$.

An element $p \in R$ is called a **prime element** if $\langle p \rangle$ is a prime ideal.

Example 1.16 In $\mathbf{Z}$ the number 5 is a prime element since if ab is divisible by 5, then either a or b is divisible by 5. Note that also -5 is a prime element.

In fact the set of prime elements in $\mathbf{Z}$ is given by $\mathbf{P} \cup (-\mathbf{P})$, where

$$\mathbf{P} = \{2, 3, 5, 7, 11, 13, 17, ...\}$$

is the set of **prime numbers**, which are the positive prime elements.

Theorem 1.17 An ideal I is a prime ideal if and only if R/I is an integral domain.

Proof. Exercise. □

Definition 1.18 An ideal $I \subset R$ is a **maximal ideal** if $I \neq R$ and there does not exist ideal J such that $I \subsetneq J \subsetneq R$.

Theorem 1.19 An ideal $I \subset R$ is a maximal ideal if and only if R/I is a field.

Proof. Assume first that I is a maximal ideal. Let $[a] \in R/I$. Assume, to obtain a contradiction, that there does not exist $[r] \in R/I$ such that $[r] \cdot [a] = [1]$. Consider the ideal $\langle [a] \rangle \subset R/I$. It is not the whole R/I because $[1] \notin \langle [a] \rangle$. Let $J = \{r \in R : [r] \in \langle [a] \rangle\}$. Then J is an ideal and $I \subsetneq J \subsetneq R$, which is a contradiction.

Assume then that R/I is a field. Suppose that $I \subsetneq J \subsetneq R$ for some ideal J. Then there exists $r \in J \setminus I$. Because R/I is a field, $[r] \in R/I$ has an inverse $[s] \in R/I$ meaning that $rs - 1 \in I \subset J$. But this means that $1 \in J$, so $J = R$, a contradiction. □

Let us end the section by noting that since fields are integral domains, every maximal ideal is a prime ideal by the previous theorems.

1.3 EUCLIDEAN DOMAINS

Euclidean domains are rings for which there exists a Euclidean algorithm. For the Euclidean algorithm to work, it is necessary that division yields remainders of decreasing size. In an arbitrary ring we have to abstract away the notion of *size*, so we arrive to the following definition.

Definition 1.20 An integral domain R is a **Euclidean domain** if there exists a function $f\colon R \to \mathbf{N}$ such that for all $a \in R$ and $b \in R \setminus \{0\}$ we can find $q, r \in R$ for which

$$a = bq + r$$

and $f(r) < f(b)$.

Theorem 1.21 $\mathbf{Z}$ is a Euclidean domain.

Proof. In the definition of Euclidean domain we can simply take $f(n) = |n|$. Indeed, if $a \in \mathbf{Z}$ and $b \neq 0$, then we can write $a = qb + r$, where q is the quotient under division a/b and r is the remainder, which has a smaller absolute value than b. □

Definition 1.22 Let $a, b \in R$. An element $d \in R$ is called a **greatest common divisor** (gcd) of a and b if $d \mid a$ and $d \mid b$ and for any other such d' we have $d' \mid d$.

Dually, an element $l \in R$ is called a **least common multiple** (lcm) of a and b if $a \mid l$, $b \mid l$ and for any other such l' we have $l \mid l'$.

Example 1.23 In $\mathbf{Z}$ the numbers 5 and -5 are greatest common divisors of the numbers -15 and 35. These are actually the only greatest common divisors of the two numbers and they differ by multiplication by -1. The least common multiples of the numbers -15 and 35 are -105 and 105.

Theorem 1.24 If R is an integral domain, then two greatest common divisors d and d' of two elements a and b satisfy $d = ud'$ for some unit $u \in R$. The same holds for least common multiples.

Proof. Exercise. □

In the case of $\mathbf{Z}$, it is customary to choose $\gcd(a, b)$ and $\operatorname{lcm}(a, b)$ to be the non-negative variants.

We have the following connections from greatest common divisors and least common multiples to ideals. You should try to prove these.

- Ideals I and J are said to be **coprime** if $I + J = R$. Here $I + J = \{a + b : a \in I, b \in J\}$, which is also an ideal.
- In a PID the ideal $\langle a\rangle + \langle b\rangle$ is generated by any greatest common divisor of a and b. If $\langle a\rangle + \langle b\rangle = R$, the greatest common divisors of a and b are units and we say that they are coprime.
- If a and b are coprime, then $[a]$ is invertible in the quotient ring $R/\langle b\rangle$ and vice versa.
- If I and J are two ideals, so is $I \cap J$. In a PID the ideal $\langle a\rangle \cap \langle b\rangle$ is generated by any least common multiple of a and b.

Example 1.25 One can check that $\langle 6\rangle + \langle 15\rangle = \langle 3\rangle$ and that $\langle 6\rangle \cap \langle 15\rangle = \langle 30\rangle$.

Let's get back to Euclidean domains.

Theorem 1.26 If R is a Euclidean domain, then every pair of elements $a, b \in R\backslash\{0\}$ has a greatest common divisor $d \in R$ and there exist $x, y \in R$ such that

$$ax + by = d. \tag{1.1}$$

This equation is sometimes called *Bezout's identity.*

Proof. Let $r_0 = a$ and $r_1 = b$. Then we can find a sequence of elements $q_1, q_2, ..., q_n$ and $r_2, r_3, ..., r_n$ such that

$$\begin{aligned}
r_0 &= q_1 r_1 + r_2,\\
r_1 &= q_2 r_2 + r_3,\\
&\vdots\\
r_{n-3} &= q_{n-2} r_{n-2} + r_{n-1}\\
r_{n-2} &= q_{n-1} r_{n-1} + r_n,
\end{aligned}$$

where $r_n = 0$ and $f(r_{n-1}) < f(r_{n-2}) < \dots < f(r_2) < f(r_1)$.

Let $d := r_{n-1}$. Then d divides r_{n-2}, and it also divides r_{n-3} because $r_{n-3} = q_{n-2}r_{n-2} + r_{n-1}$. Continuing this reasoning inductively up the equation chain we see that d divides every r_i, including $r_0 = a$ and $r_1 = b$.

Conversely any common divisor d' of a and b must divide each r_k, so in particular $d' \mid d$. This means that d is a greatest common divisor. Solving $r_2 = r_0 - q_1 r_1$, substituting it to the next equation and solving r_3 and continuing like this, we will end up with a representation (1.1). □

The considerations so far show that in the case of $\mathbf{Z}$ we can find a well-defined greatest common divisor of any two integers a, b as long as at least one of them is non-zero. Just pick the positive one and denote it $\gcd(a, b)$. Thus for example $\gcd(15, 35) = 5$. Moreover, the proof above can be turned into an algorithm that computes the gcd as well as numbers $x, y \in \mathbf{Z}$ so that (1.1) is satisfied.

Algorithm 1.27 *(Extended Euclidean algorithm)* The following calculates the greatest common divisor d of a and b as well as x and y such that $ax + by = d$.

```
void eea(int64_t a, int64_t b, int64_t &d,
         int64_t &x, int64_t &y) {
    int64_t u=1, v=0, m, n, q, r;
    x=0, y=1;
    d = b;
    while(a != 0) {
        q=d/a;      r=d%a;
        m=x-u*q;    n=y-v*q;
        d=a;        a=r;
        x=u;        y=v;
        u=m;        v=n;
    }
}
```

The existence of a Euclidean algorithm is one of the greatest thing about Euclidean domains. From theoretical point of view the following is also nice.

Theorem 1.28 Any Euclidean domain is a PID.

Proof. Let R be a Euclidean domain, f its size function and $I \subset R$ an ideal. Choose $a \in I \setminus \{0\}$ for which $f(a)$ is minimal. Then clearly $\langle a \rangle \subset I$. On the other hand if $b \in I$, we can find elements $q, r \in R$ such that $b = qa + r$ and $f(r) < f(a)$. Notice that $r \in I$, so we must have $r = 0$ and therefore $b \in \langle a \rangle$. Thus $I = \langle a \rangle$. □

Now we finally see that all the ideals of $\mathbf{Z}$ are of the form $\langle m \rangle$ for some $m \in \mathbf{Z}$.

Let us close the section with the following longish (but important if you haven't seen it before) exercise.

Exercise 1.29 An equation of the form $ax + by = c$ where $a, b, c \in \mathbf{Z}$ is called a **linear diophantine equation**. Assume that a, b, c are non-zero.

- Give a necessary and sufficient condition under which there exist integers x and y that satisfy the equation.
- Give a parametrization for all the solutions.
- Implement a program solving the equation.

1.4 UNIQUE FACTORIZATION

Our next aim is to show that every $m \in \mathbf{Z}$ has a unique factorization into prime elements. Indeed, this is the case in the more general setting of so called unique factorization domains, of which PIDs are a special case.

Definition 1.30 A non-zero non-unit element $r \in R$ is **irreducible** if it is not a product of two non-invertible elements. A non-zero non-unit element that is not irreducible is called **reducible**.

The term composite *seems to be non-standard in ring theory. I like it in the context of UFDs, however.*

In $\mathbf{Z}$ the irreducible elements are the same as the prime elements. This is not the case for all commutative rings in general, although for integral domains prime elements are irreducible. In $\mathbf{Z}$ reducible elements are also called **composite**. We will use the same term also in unique factorization domains that will be defined shortly.

Theorem 1.31 In an integral domain a prime element is irreducible.

Proof. Exercise. □

The happy thing is that the converse holds in principal ideal domains, and more generally in unique factorization domains (which will be defined soon).

Theorem 1.32 In a PID an ideal generated by an irreducible element is maximal.

Proof. Let $r \in R$ be irreducible. Let $I \neq R$ be an ideal such that $\langle r \rangle \subset I$. Then because R is a PID, I is of the form $I = \langle s \rangle$, where s is necessarily not a unit. But this means that $r = us$ for some $u \in R$, which must be a unit since r is irreducible. It follows that the two ideals coincide, so $\langle r \rangle$ is maximal. □

It is an immediate corollary that in PIDs primes and irreducible elements are the same. Indeed, if r is irreducible, then $\langle r \rangle$ is maximal and thus prime.

We should also note that this means that in a PID any quotient ring $R/\langle p \rangle$ is a field if and only if p is prime/irreducible.

Definition 1.33 An integral domain R is a **unique factorization domain**, or **UFD**, if any element $r \in R$ can be written in the form

$$r = up_1p_2...p_k,$$

where $p_1, ..., p_k$ are irreducible and this product is unique up to reordering and multiplying the elements by units.

Theorem 1.34 Every PID is a UFD.

Proof. Let us start by showing that any $r \in R$ is a product of finitely many irreducibles.

First notice that R has the property that any ascending chain

$$I_1 \subset I_2 \subset I_3 \subset ...$$

of ideals must be constant from some point on. Indeed, because the chain is ascending, the set $I = \bigcup_{k=1}^{\infty} I_k$ is an ideal. Thus it is generated by some element $a \in I_n$ for large enough n. Then $I_k = I_n$ for $k \geq n$.

Let $r \in R$. If r is irreducible we are done. Otherwise we can write $r = r_0 r_1$ where neither of r_0, r_1 is a unit. We recurse to r_0 and r_1 and they are either both irreducible, or for example r_1 is reducible and we get numbers r_{10}, r_{11} such that $r_1 = r_{10} r_{11}$. Assume, towards a contradiction, that we continue expanding the terms and that the process never ends. Then there exists a binary sequence s such that

$$\langle r \rangle \subset \langle r_{s_1} \rangle \subset \langle r_{s_1 s_2} \rangle \subset \langle r_{s_1 s_2 s_3} \rangle \subset \ldots$$

As we saw above, this chain must settle, so at some point $r_{s_1 s_2 \ldots s_n}$ and $r_{s_1 s_2 \ldots s_{n+1}}$ generate the same ideal. This means that $r_{s_1 s_2 \ldots s_n} = u r_{s_1 s_2 \ldots s_{n+1}}$ for some unit u, which is a contradiction. Thus the expanding of terms ends, and every $r \in R$ admits a factorization into irreducibles.

We will conclude the proof by showing that the factorization is unique. Assume that $r = u p_1 p_2 \ldots p_n = v q_1 q_2 \ldots q_m$ where $p_1, \ldots, p_n$ and $q_1, \ldots, q_m$ are irreducibles and u and v are units. Because p_1 divides $q_1 q_2 \ldots q_n$, and because p_1 is a prime, there must exist q_i such that $p_1 \mid q_i$. Since q_i is irreducible, we see that when we cancel p_1 and q_i, we are left with a unit. We can therefore reduce to the case where there are 1 less factors on both sides and continue by induction. □

Theorem 1.35 Every irreducible element in a UFD is prime.[2]

Proof. Let $p \in R$ be an irreducible element and consider the ideal $\langle p \rangle$. Assume that $ab \in \langle p \rangle$. Then $ab = rp$ for some $r \in R$. Writing a, b and r in terms of irreducibles $a_1 \ldots a_n$, $b_1 \ldots b_m$ and $r_1 \ldots r_l$, we have for some unit u that

$$u a_1 \ldots a_n b_1 \ldots b_m = r_1 \ldots r_l p.$$

[2] One might ask why we proved this for PIDs in the first place. The reason is that we used the result in the proof that a PID is a UFD.

By unique factorization one of the $a_1 ... a_n b_1 ... b_m$ must equal p multiplied by a unit. If it is one of the a_is, then $a \in \langle p \rangle$, otherwise $b \in \langle p \rangle$. Thus $\langle p \rangle$ is a prime ideal and p is a prime element. □

1.5 PRIMES AND FACTORING IN $\mathbf{Z}$

It is time to take a break from the algebraic mumbo jumbo and see how what we have established can be used in the setting of $\mathbf{Z}$.

Since $\mathbf{Z}$ is a PID, we know that the prime elements are the same as the irreducibles. Suppose that we want to check if $m \in \mathbf{Z}$ is a prime. Without loss of generality we can assume that $m \geq 2$. Since $\mathbf{Z}$ is a UFD, we can write

$$m = p_1^{a_1} ... p_k^{a_k}$$

where $p_1 < ... < p_k$ are positive primes and $a_1, ..., a_k \geq 1$. This is called the **prime factorization** of m. If m was composite, we would either have $a_1 \geq 2$, in which case $p_1^2 \leq m$, or we would have $a_1 = 1$ and $k \geq 2$, in which case $p_1^2 < p_1 p_2 \leq m$. Thus every composite m has a prime factor of size at most $\lfloor \sqrt{m} \rfloor$. This leads to the following simple algorithm for checking whether a given number is prime.

Algorithm 1.36 *(Trial division)*

```
bool is_prime(int64_t m) {
    for(int64_t i=2;i*i<=m;i++) {
        if(m % i == 0) return false;
    }
    return true;
}
```

The running time of the algorithm is $O(\sqrt{m})$. It should be noted that (much) more efficient tests exist.

Let us next consider the problem of listing the primes up to some number N.

Algorithm 1.37 *(Sieve of Eratosthenes)* The idea is to start from 2 and rule out composite numbers whenever new primes are found. Indeed, first we mark $4, 6, 8, ...$ as composites. Then we move on to 3 and mark $3, 6, 9, 12, 15, ...$ Since we have already marked 4, we skip it and move to 5, which is our third prime. This is repeated until we reach N.

```
vector<int64_t> sieve(N+1,0);

for(int64_t p=2;p<=N;p++) {
    if(sieve[p] == 0) {
        for(int64_t i=p;i<=N;i+=p) {
            sieve[i] = p;
        }
    }
}
```

The algorithm above stores in `sieve[m]` the largest prime factor of m. This will be useful next when we consider factoring numbers. Using some asymptotic results on the distribution of primes, it is not too hard to show that the complexity of the algorithm is $O(n \log \log n)$.

Let's get straight on to factoring.

Algorithm 1.38 *(Factoring using a sieve)* Assume that we have built a factor sieve up to N and want to factor $m \leq N$. This is easily done as follows.

```
typedef pair<int64_t, int64_t> P;

vector<P> factors;

while(m != 1) {
    int64_t p = sieve[m];
    int64_t a = 0;
    while(m%p == 0) {
        m/=p;
        a++;
    }
    factors.push_back(P(p, a));
}
```

In the end `factors` will contain a pair (p_k, a_k) for each prime factor $p_k^{a_k}$ of m.

1.6 RING HOMOMORPHISMS

Definition 1.39 Let R and S be two commutative rings. A map $f\colon R \to S$ is called a **ring homomorphism** if $f(0) = 0$, $f(1) = 1$ and for all $a, b \in R$ we have

$$f(a+b) = f(a) + f(b), \text{ and}$$
$$f(ab) = f(a)f(b).$$

Exercise 1.40 Show that the map $f\colon \mathbf{C} \to \mathbf{R}^{2\times 2}$ given by

$$f(x+iy) = \begin{pmatrix} x & y \\ -y & x \end{pmatrix}$$

is a ring homomorphism from the complex numbers to the real 2×2 matrices.

Ring homomorphisms are important for the general theory of rings. For example if I is an ideal in R, then the map $\pi_I\colon R \to R/I$ given by $\pi_I(r) = [r]$ is a ring homomorphism.

Definition 1.41 Given a ring homomorphism $f\colon R \to S$, the set of elements mapping to 0 under f is denoted by $\operatorname{Ker} f$ and called the **kernel** of f.

For example the kernel of the map π_I above is simply I.

Theorem 1.42 Let $f\colon R \to S$ be a ring homomorphism. Then $\operatorname{Ker} f$ is an ideal and there exists a unique *injective* ring homomorphism $\tilde{f}\colon R/\operatorname{Ker} f \to S$ such that $f = \tilde{f} \circ \pi_{\operatorname{Ker} f}$.

Proof. Exercise. □

Note that as a special case of the above theorem if $\operatorname{Ker} f = \{0\}$ then f is injective. If a ring homomorphism f is a bijection, we say that it is a **ring isomorphism**.

Exercise 1.43 Show that the inverse mapping f^{-1} of a ring isomorphism is also a ring isomorphism.

If there exists an isomorphism between two rings R and S, then we say that R and S are **isomorphic**. One should note that in terms of ring theory isomorphic rings are indistinguishable; all their algebraic properties are the same.

The following definition and theorem make

Definition 1.44 Let $f\colon R \to S$ be a ring homomorphism. The **image** of f is the set $\operatorname{Im} f := f(R)$.

Theorem 1.45 The image of a ring homomorphism $f\colon R \to S$ is a subring of S. The rings $R/\operatorname{Ker} f$ and $\operatorname{Im} f$ are isomorphic, and an isomorphism is given by the map $\tilde{f}$ in Theorem 1.42.

Proof. Exercise. □

1.7 Chinese remainder theorem

Definition 1.46 If R and S are two commutative rings, then we can form their **product ring** $R \times S$ by taking all the tuples (r, s) with $r \in R$ and $s \in S$ and defining the ring operations component wise, i.e.

$$(r, s) + (r', s') = (r + r', s + s'),$$
$$(r, s) \cdot (r', s') = (r \cdot r', s \cdot s')$$

for all $r, r' \in R$, $s, s' \in S$.

It is easy to see that the 0 in a product ring is $(0, 0)$ and the 1 is $(1, 1)$. Product rings are not integral domains (unless one of the factors is the zero-ring and the other one is an integral domain) since $(0, 1) \cdot (1, 0) = (0, 0)$. An element (u, v) is a unit if and only if u is a unit in R and v is a unit in S.

Definition 1.47 Let I and J be ideals in R. Their product ideal IJ is the ideal generated by all products ab, $a \in I$ and $b \in J$, i.e.

$$IJ := \{r_1a_1b_1 + ... + r_na_nb_n : 1 \leq k \leq n, r_k \in R, a_k \in I, b_k \in J\}.$$

The chinese remainder theorem gives us a natural link between products of coprime ideals and products of quotient rings.

Theorem 1.48 Let I and J be two coprime ideals of a commutative ring R. Then $I \cap J = IJ$ and there exists a ring isomorphism

$$f: R/IJ \to R/I \times R/J.$$

Proof. Let us first show that $I \cap J = IJ$. If $r \in IJ$, then it is of the form $r = r_1a_1b_1 + ... + r_na_nb_n$, where $r_k \in R$, $a_k \in I$ and $b_k \in J$. It is clear that such an element belongs both to I and J. If $r \in I \cap J$, then since $I + J = R$, we have $a + b = 1$ for some $a \in I$, $b \in J$, and thus $r = ra + rb \in IJ$.

Consider the map $g: R \to R/I \times R/J$ given by

$$g(r) = (\pi_I(r), \pi_J(r))$$

for all $r \in R$. It is clear that $\operatorname{Ker} g = I \cap J = IJ$. Thus the claim will follow from the factorization theorem once we show that g is surjective. Let $x, y \in R$ be arbitrary and set $r = xb + ya$ where $a \in I$ and $b \in J$ are such that $a + b = 1$. Then $\pi_I(r) = \pi_I(x)$ and $\pi_J(r) = \pi_J(y)$. Thus the arbitrary element $(\pi_I(x), \pi_J(y))$ belongs to $\operatorname{Im} g$ and g is surjective. □

It is easy to see that by induction the above result can be generalized to n ideals $I_1, ..., I_n$ that are pairwise coprime, i.e. $I_i + I_k = R$ for all $i \neq k$. Then we get an isomorphism betwen $R/(I_1...I_n)$ and $R/I_1 \times ... \times R/I_n$. In PIDs we have the following version:

Theorem 1.49 Let R be a PID and $a_1, ..., a_n$ pairwise coprime elements of R. Then there exists an isomorphism

$$R/\langle a_1...a_n\rangle \to R/\langle a_1\rangle \times ... \times R/\langle a_n\rangle.$$

Let us specialize to the case of $\mathbf{Z}$. Let $a_1, ..., a_n \geq 1$ be pairwise coprime. Assume that we are given an element in $\mathbf{Z}/a_1\mathbf{Z} \times \cdots \times \mathbf{Z}/a_n\mathbf{Z}$ in terms of n representatives $x_1, ..., x_n \in \mathbf{Z}$, so that our element is $(\pi_{a_1}(x_1), ..., \pi_{a_n}(x_n))$.

We would like to construct a representative for the corresponding element in $\mathbf{Z}/(a_1...a_n)\mathbf{Z}$. Similarly to what was done in the proof, it is easy to see that such an element is given by

$$x = x_1(a_2...a_n)b_1 + x_2(a_1a_3...a_n)b_2 + ... + x_n(a_1a_2...a_{n-1})b_{n-1},$$

where b_k is a representative for the inverse of $\pi_{a_k}(a_1...a_{k-1}a_{k+1}...a_n)$ in $\mathbf{Z}/a_k\mathbf{Z}$.

This inverse can be found by using the extended Euclidean algorithm on a_k and $a_1...a_{k-1}a_{k+1}...a_n$ to find b_k and c_k such that

$$b_k(a_1...a_{k-1}a_{k+1}...a_n) + c_ka_k = 1.$$

We are ready to code this up.

Algorithm 1.50 *(Chinese remainder theorem)*

```
typedef pair<int64_t, int64_t> P;

// Given a vector of pairs (x_k, a_k),
// computes a number x such that x = x_k mod a_k
int64_t chinese_remainder(const vector<P> &repr) {
    int64_t prod=1;
    for(int64_t i=0;i<repr.size();i++) {
        prod*=repr[i].second;
    }
    int64_t result=0;
    for(int64_t i=0;i<repr.size();i++) {
        int64_t b,c,d;
        eea(repr[i].second, prod/repr[i].second, d, c, b);
        result+=repr[i].first*(prod/repr[i].second)*b%prod;
        result%=prod;
    }
    return result;
}
```

This concludes our study of the global structure of $\mathbf{Z}$. Indeed, we are now ready to switch our focus to the quotient rings $\mathbf{Z}/m\mathbf{Z}$.

By the chinese remainder theorem any such ring is isomorphic to a ring of the form $\mathbf{Z}/p_1^{a_1}\mathbf{Z} \times \cdots \times \mathbf{Z}/p_n^{a_n}\mathbf{Z}$, where $p_1^{a_1}\cdots p_n^{a_n}$ is the prime factorization of m. This means that it is often enough to consider the case where $m = p^k$ for some prime number p and $k \geq 1$.

2 WORKING MODULO m

2.1 EULER'S TOTIENT FUNCTION

We started our study of $\mathbf{Z}$ by finding the invertible elements, and likewise our first goal here will be to figure out the group of units of $\mathbf{Z}/m\mathbf{Z}$. The answer will be more interesting than in the case of $\mathbf{Z}$, and the required basics of group theory can be read from the first two chapters of Part III. The reader should take a look there now if he or she is not familiar with groups.

We already have several important examples of groups from the last section. If R is any commutative ring, then $(R, +)$ is a group with identity element 0. This is called the **additive group** of R. Similarly the units of R form a group under multiplication with 1 as the identity element. This is called the **multiplicative group** or **group of units** of R and we denote it by R^*. Also any ideal $I \subset R$ is a subgroup of the additive group of the ring.

For all $m \geq 1$, let $\varphi(m)$ denote the size of the group $(\mathbf{Z}/m\mathbf{Z})^*$. The function $\varphi \colon \mathbf{Z}^+ \to \mathbf{Z}^+$ is called the *Euler totient function*. Recall that if the prime factorization of m is $p_1^{a_1} \cdots p_k^{a_n}$, then there is an isomorphism

$$\mathbf{Z}/m\mathbf{Z} \cong \mathbf{Z}/p_1^{a_1}\mathbf{Z} \times \cdots \times \mathbf{Z}/p_n^{a_n}\mathbf{Z}.$$

Likewise, an element of a product ring is a unit if and only if all of its components are units, i.e. we have the isomorphism

$$(\mathbf{Z}/m\mathbf{Z})^* \cong (\mathbf{Z}/p_1^{a_1}\mathbf{Z})^* \times \cdots \times (\mathbf{Z}/p_n^{a_n}\mathbf{Z})^*.$$

Thus we immediately see that

$$\varphi(m) = \varphi(p_1^{a_1}) \ldots \varphi(p_n^{a_n}).$$

Now let p be a prime number. Recall that $0, 1, \ldots, p^a - 1$ is a full set of representatives for the ring $\mathbf{Z}/p^a\mathbf{Z}$. The ones that represent units are the ones that are coprime to p^a, i.e. the ones that are not divisible by p. There are $p^a - p^{a-1}$ such numbers, so

$$\varphi(p^a) = (p-1)p^{a-1}.$$

This lets us calculate φ for any integer for which we know its prime factorization.

Lagrange's theorem for groups tells us that $a^{\varphi(m)} = 1$ for all $a \in (\mathbf{Z}/m\mathbf{Z})^*$. In this setting this is known as *Euler's theorem.* In the case where $m = p$ is a prime number it is also called *Fermat's little theorem*, which states that $a^{p-1} = 1$ for all $a \in \mathbf{Z}/p\mathbf{Z}$, $a \neq 0$.

2.2 STRUCTURE OF THE UNIT GROUP

In this section we will identify the group structure of $(\mathbf{Z}/p^k\mathbf{Z})^*$ where p is a prime number and $k \geq 1$. We will start with the case $p = 2$, since 2 behaves differently to other primes.

Theorem 2.1 Let $k \geq 1$, then

$$(\mathbf{Z}/2^k\mathbf{Z})^* \cong \begin{cases} C_1, & \text{if } k = 1, \\ C_2, & \text{if } k = 2, \\ C_2 \times C_{2^{k-2}}, & \text{if } k \geq 3. \end{cases}$$

Proof. The cases $k = 1$ and $k = 2$ are clear by inspection. Let $k \geq 3$ and note that $\varphi(2^k) = 2^{k-1}$. It is therefore enough to check that $\{-1, 1\}$ and $\{1, 5, 5^2, ..., 5^{2^{k-2}-1}\}$ are two subgroups with trivial intersection and that the second group has 2^{k-2} elements.

Let us first show that the order of 5 is 2^{k-2}. To do this, it suffices to show that $5^{2^{k-2}} \equiv 1$ and $5^{2^{k-3}} \not\equiv 1$. By repeatedly squaring, we see that there exist odd numbers $u_0, ..., u_{k-2}$ such that

$$\begin{aligned} 5^{2^0} &= 1 + u_0 \cdot 2^2, \\ 5^{2^1} &= (1 + u_0 \cdot 2^2)^2 = 1 + u_1 \cdot 2^3, \\ 5^{2^2} &= (1 + u_1 \cdot 2^3)^2 = 1 + u_2 \cdot 2^4, \\ &\vdots \\ 5^{2^{k-3}} &= 1 + u_{k-3} \cdot 2^{k-1}, \\ 5^{2^{k-2}} &= 1 + u_{k-2} \cdot 2^k, \end{aligned}$$

which proves the claim.

Finally it is clear that $5^a \not\equiv -1$ for all a, because $5^a \equiv 1 \pmod 4$. □

Notice in particular that the group $(\mathbf{Z}/2^k\mathbf{Z})^*$ is not cyclic for any $k \geq 3$ because every element has order at most 2^{k-2}.

Let us next consider the case where $p > 2$. We will begin by showing that $(\mathbf{Z}/p\mathbf{Z})^*$ is cyclic. First two small lemmas.

Theorem 2.2 Let F be a field and P be a polynomial of degree n with coefficients from F. Then P has at most n roots in F.

Proof. Exercise. □

Theorem 2.3 Let G be a finite abelian group. If m is the least common multiple of the orders of the elements in G, then G contains an element of order m.

Proof. Let $p_1^{a_1}...p_k^{a_k}$ be the prime factorization of m. Then there exist $\tilde{g}_1, ..., \tilde{g}_k \in G$ such that $p_i^{a_i}$ divides the order of $\tilde{g}_i$. Define $g_i = \tilde{g}_i^{\operatorname{ord}(\tilde{g}_i)/(p_i^{a_i})}$. Then each g_i has order $p_i^{a_i}$. Clearly the subgroups generated by g_i intersect only at $\{1\}$, for an element belonging to the intersection of $\langle g_i \rangle$ and $\langle g_j \rangle$ has order that divides both $p_i^{a_i}$ and $p_j^{a_j}$. Therefore G contains a subgroup isomorphic to $\langle g_1 \rangle \times \cdots \times \langle g_j \rangle$, which is a cyclic group of order m. □

Theorem 2.4 Let p be an odd prime number. Then $(\mathbf{Z}/p\mathbf{Z})^* \cong C_{p-1}$.

Proof. Let m be the least common multiple of the orders of the elements in $(\mathbf{Z}/p\mathbf{Z})^*$. Then since there exists an element of order m and because its order divides $p-1$, we must have $m \leq p-1$.

On the other hand consider the polynomial $P(x) = x^m - 1$. Every element of $(\mathbf{Z}/p\mathbf{Z})^*$ is a root of P because their orders divide m. The number of roots is at most m, so we must have $p-1 \leq m$. This means that $m = p-1$ and the group must be cyclic. □

Finally, let us tackle the general case $(\mathbf{Z}/p^k\mathbf{Z})^*$ for odd prime p and $k \geq 1$.

Theorem 2.5 Let p be an odd prime and $k \geq 1$. Then $(\mathbf{Z}/p^k\mathbf{Z})^* \cong C_{(p-1)p^{k-1}}$.

Proof. The case $k = 1$ was proved above. Let $k \geq 2$. We can use a similar method as we used with $p = 2$. There we were able to start with 5 and repeatedly squared it, but here we will have to be a little bit more careful. Let g be a generator for $(\mathbf{Z}/p\mathbf{Z})^*$. Then we claim that either g^{p-1} or $(g+p)^{p-1}$ is of the form $1 + u_0 p$, where u_0 is not divisible by p. Assume that g^{p-1} is not of this form. Then by the binomial theorem

$$(g+p)^{p-1} = g^{p-1} + (p-1)g^{p-2}p + ...,$$

where the rest of the terms are divisible by p^2. Now by assumption $g^{p-1} = 1 + ap$ with a divisible by p, so we see that $(g+p)^{p-1}$ is of the wanted form. The rest of the argument goes as with $p = 2$, but instead of squaring we raise the previous number to power p repeatedly. Finishing the proof is an exercise. □

Let us collect the results of this section. If $m = p_1^{a_1} ... p_k^{a_k}$, then $(\mathbf{Z}/m\mathbf{Z})^* \cong (\mathbf{Z}/p_1^{a_1}\mathbf{Z})^* \times \cdots \times (\mathbf{Z}/p_k^{a_k}\mathbf{Z})^*$, which is cyclic if and only if m is one of $2, 4, p^k, 2p^k$ where p is an odd prime and $k \geq 1$.

2.3 PRIMITIVE ROOTS

Let m be such that $(\mathbf{Z}/m\mathbf{Z})^*$ is cyclic. Often in calculations it is handy to have a generator for this group. Such generators are also called **primitive roots**. No known algorithm exists that can deterministically find a generator.

We can however reduce the problem to finding a generator for prime m. The proof that $(\mathbf{Z}/p^k\mathbf{Z})^*$ is cyclic showed that if g is a generator for p, then either g or $g + p$ is a generator for p^k. For $m = 2p^k$ we have the following.

Theorem 2.6 Let p be an odd prime number and $k \geq 1$. If g is a generator for $(\mathbf{Z}/p^k\mathbf{Z})^*$, then $(1, g)$ is a generator for $(\mathbf{Z}/2\mathbf{Z})^* \times (\mathbf{Z}/p^k\mathbf{Z})^* \cong (\mathbf{Z}/2p^k\mathbf{Z})^*$.

Proof. This is clear. □

Thus if we think of g as an integer (a representative for an element of $(\mathbf{Z}/p^k\mathbf{Z})^*$), then the odd one among g or $g+p^k$ represents a generator for $(\mathbf{Z}/2p^k\mathbf{Z})^*$.

Let us consider now the case where $m = p$ is a prime and suppose we want to check whether g is a generator. Then we should have $\operatorname{ord}(g) = p-1$. To rule out all other possible orders, we must rule out all divisors of $p-1$. Let $q_1^{a_1}...q_k^{a_k}$ be the prime factorization of $p-1$. Every proper divisor of $p-1$ is divisible by one of the k numbers $\frac{p-1}{q_1}, ..., \frac{p-1}{q_k}$. Thus g is a primitive root if and only if for all i we have $g^{\frac{p-1}{q_i}} \neq 1$. This can be checked reasonably fast with exponentiation by squaring.

2.4 SQUARES AND SQUARE ROOTS

This section concludes the study of the structure of $\mathbf{Z}/m\mathbf{Z}$ and will focus on solving the equation $x^2 = a$ for a given $a \in \mathbf{Z}/m\mathbf{Z}$. We will first derive the law of quadratic reciprocity that can be used to determine when a solution exists, and then consider an algorithm for actually finding the x. If such an x exists, a is called a **quadratic residue** modulo m. Otherwise it is called a **quadratic non-residue**. We will also just say that a is a square in $\mathbf{Z}/m\mathbf{Z}$.

So when is a a square modulo m? By Chinese remainder theorem it is clear that if m has the prime factorization $p_1^{a_1}...p_k^{a_k}$, then a must be a square in each of $\mathbf{Z}/p_i^{a_i}$. Thus it is enough to consider the case where $m = p^k$ for some prime number p and $k \geq 1$.

Let us start with the case where $m = p$ is a prime number. If $p = 2$, then clearly both 0 and 1 are squares, so let us assume that p is an odd prime. All the squares in $\mathbf{Z}/p\mathbf{Z}$ are given by $0, 1^2, 2^2, ..., \left(\frac{p-1}{2}\right)^2$, since $x^2 = (-x)^2$ for all $x \in \mathbf{Z}/p\mathbf{Z}$. Thus there are $\frac{p+1}{2}$ squares in $\mathbf{Z}/p\mathbf{Z}$. Notice also that the product of two squares is a square and that the inverse of a square is a square. In particular if we consider the $\frac{p-1}{2}$ non-zero squares, they form a subgroup of $(\mathbf{Z}/p\mathbf{Z})^*$. The set of non-squares is the other coset, the quotient group must be isomorphic to C_2, and we have shown the following:

Theorem 2.7 Let p be a prime number and $a, b \in (\mathbf{Z}/p\mathbf{Z})^*$. Then

- if a and b are both squares or non-squares, ab is a square,
- if one of a and b is a square and the other one is a non-square, ab is a non-square.

To state the law of quadratic reciprocity, we will first define Jacobi symbols as follows.

Definition 2.8 Let $n \geq 1$ be an odd number and $a \in (\mathbf{Z}/n\mathbf{Z})^*$. Consider the map $\tau_a\colon (\mathbf{Z}/n\mathbf{Z})^* \to (\mathbf{Z}/n\mathbf{Z})^*$ given by $\tau_a(x) = ax$. The **Jacobi symbol** $\left(\frac{a}{n}\right)$ is defined to be equal to $\operatorname{sgn}(\tau_a)$, the sign of the permutation τ_a.

The following result relating the Jacobi symbol to squares modulo p is known as *Zolotarev's lemma.*

Theorem 2.9 Let p be an odd prime and $a \in (\mathbf{Z}/p\mathbf{Z})^*$. Then

$$\left(\frac{a}{p}\right) = \begin{cases} 1, & \text{if } a \text{ is a square,} \\ -1, & \text{if } a \text{ is a non-square.} \end{cases}$$

Proof. Because sgn is a homomorphism on permutations, the map

$$\varphi\colon a \mapsto \tau_a \mapsto \operatorname{sgn}(\tau_a)$$

is a homomorphism between $(\mathbf{Z}/p\mathbf{Z})^*$ and the group of two elements C_2. Notice that if g is a generator for $(\mathbf{Z}/p\mathbf{Z})^*$, then τ_g is a cycle of length $p-1$ and therefore has sign -1. This means that φ is a surjection. It follows that $(\mathbf{Z}/p\mathbf{Z})^*/\operatorname{Ker}\varphi$ is isomorphic to C_2. Since there is only one subgroup of order $\frac{p-1}{2}$ in the cyclic group of order $p-1$, $\operatorname{Ker}\varphi$ must coincide with the subgroup of squares. □

The following, known as *Euler's criterion*, is a way of calculating the Jacobi symbol when the bottom argument is an odd prime.

Theorem 2.10 Let p be an odd prime. Then $\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p}$.

Proof. Let $x \in (\mathbf{Z}/p\mathbf{Z})^*$. Clearly $x^{\frac{p-1}{2}} = \pm 1$ because $\mathbf{Z}/p\mathbf{Z}$ is a field. The map $\varphi\colon x \mapsto x^{\frac{p-1}{2}} \pmod p$ can therefore thought of as a homomorphism $(\mathbf{Z}/p\mathbf{Z})^* \to C_2$. It is a surjection because if g is a generator for $(\mathbf{Z}/p\mathbf{Z})^*$, then $g^{\frac{p-1}{2}} = -1$. Thus the kernel of φ is the unique subgroup of order $\frac{p-1}{2}$ on which $\left(\frac{x}{p}\right) = 1$. □

We will next consider how to calculate $\left(\frac{n}{m}\right)$ efficiently for any odd coprime numbers m and n. Let's first prove the law of quadratic reciprocity.

Theorem 2.11 Let m and n be odd coprime numbers. Then

$$\left(\frac{n}{m}\right)\left(\frac{m}{n}\right) = (-1)^{\frac{m-1}{2}\cdot\frac{n-1}{2}}.$$

Proof. Identify each of the rings $\mathbf{Z}/m\mathbf{Z}$, $\mathbf{Z}/n\mathbf{Z}$ and $\mathbf{Z}/mn\mathbf{Z}$ with the natural numbers $\{0, 1, ..., m-1\}$, $\{0, 1, ..., n-1\}$ and $\{0, 1, ..., mn-1\}$ respectively. Then we can give each of the rings a total order induced by the usual order on natural numbers.

Next consider the ring $\mathbf{Z}/m\mathbf{Z} \times \mathbf{Z}/n\mathbf{Z}$. We can give it two different total orders:

- The m-major order $<_m$ where $(x, y) <_m (x', y')$ if and only if $x < x'$ or $x = x'$ and $y < y'$.
- The n-major order $<_n$ where $(x, y) <_n (x', y')$ if and only if $y < y'$ or $y = y'$ and $x < x'$.

Now there exist unique maps $f_m, f_n\colon \mathbf{Z}/m\mathbf{Z} \times \mathbf{Z}/n\mathbf{Z} \to \mathbf{Z}/mn\mathbf{Z}$ that are isomorphisms from the orders $<_m$ and $<_n$ respectively to the order defined on $\mathbf{Z}/mn\mathbf{Z}$. These maps are given by

$$f_m(x, y) = nx + y \quad \text{and} \quad f_n(x, y) = x + my.$$

Let $\pi\colon \mathbf{Z}/mn\mathbf{Z} \to \mathbf{Z}/m\mathbf{Z} \times \mathbf{Z}/n\mathbf{Z}$ be the canonical ring isomorphism and define two permutations α_m and α_n on $\mathbf{Z}/m\mathbf{Z} \times \mathbf{Z}/n\mathbf{Z}$ by setting $\alpha_m = \pi \circ f_m$ and $\alpha_n = \pi \circ f_n$. We have

$$\alpha_m(x, y) = (nx + y, y) \quad \text{and} \quad \alpha_n(x, y) = (x, x + my).$$

The sign of the permutation α_m is the same as the sign of the permutation $x \mapsto nx + y$ on $\mathbf{Z}/m\mathbf{Z}$ because n is odd and α_m is constant in $\mathbf{Z}/n\mathbf{Z}$. The permutation $x \mapsto nx + y$ is the composition of the two permutations $x \mapsto nx$ and $x \mapsto x + y$. The first one has sign $\left(\frac{n}{m}\right)$ and the second one has sign 1, so $\operatorname{sgn}(\alpha_m) = \left(\frac{n}{m}\right)$ and similarly $\operatorname{sgn}(\alpha_n) = \left(\frac{m}{n}\right)$.

Consider now the permutation $\theta = \alpha_n^{-1} \circ \alpha_m = f_n^{-1} \circ f_m$. The sign of θ is the product of the signs of α_m and α_n, which is $\left(\frac{n}{m}\right)\left(\frac{m}{n}\right)$. We will now double count this sign. Notice that θ is the unique isomorphism between the total orders $<_m$ and $<_n$. To count the sign of the permutation θ, it is enough to count the number of inversions with respect to the order $<_n$. By definition we are looking for the number of pairs $(x, y), (x', y')$ such that

$$(x, y) <_n (x', y') \quad \text{and} \quad \theta(x, y) >_n \theta(x', y').$$

Now $\theta(x, y) >_n \theta(x', y')$ is equivalent with $(x, y) >_m (x', y')$, so by the definition of $<_m$ and $<_n$ we must have $x' < x$ and $y < y'$. There are $\binom{m}{2}\binom{n}{2}$ solutions to these inequalities, so

$$\left(\frac{n}{m}\right)\left(\frac{m}{n}\right) = \operatorname{sgn}(\theta) = (-1)^{\frac{m(m-1)}{2}\frac{n(n-1)}{2}} = (-1)^{\frac{m-1}{2}\frac{n-1}{2}},$$

which concludes the proof. □

What about $\left(\frac{m}{n}\right)$ when m is even? The following gives a starting point.

Theorem 2.12 Let $n \geq 1$ be odd. Then

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

Proof. Consider the permutation $\tau_2\colon x \mapsto 2x$ of $(\mathbf{Z}/n\mathbf{Z})^*$. Let us identify $\mathbf{Z}/n\mathbf{Z}$ with the set $\{0, 1, ..., n-1\}$ and give it the usual total order inherited from $\mathbf{Z}$. We can then calculate the sign of τ_2 by counting inversions. For $1 \leq x \leq \frac{n-1}{2}$ the permutation simply takes x to $2x$ and for $\frac{n+1}{2} \leq x \leq n-1$ it takes x to $2x - n$. Clearly there are no inversions for pairs (x, y) in the ranges $1 \leq x < y \leq \frac{n-1}{2}$ or $\frac{n+1}{2} \leq x < y \leq n-1$. If $1 \leq x \leq \frac{n-1}{2}$ and $\frac{n+1}{2} \leq y \leq n-1$, then $2x > 2y - n$ if and only if $y \leq x + \frac{n-1}{2}$. Thus we have

$$\sum_{x=1}^{\frac{n-1}{2}} (x + \frac{n-1}{2} - \frac{n+1}{2} + 1) = \sum_{x=1}^{\frac{n-1}{2}} x = \frac{\frac{n-1}{2}\frac{n+1}{2}}{2} = \frac{n^2-1}{8}$$

inversions in total, giving $\operatorname{sgn}(\tau_2) = (-1)^{\frac{n^2-1}{8}}$. □

The final piece in the repertoire for Jacobi symbols is that it is completely multiplicative in both the top argument and the bottom argument.

Theorem 2.13 Let $m \geq 1$ be odd and $a, b \in (\mathbf{Z}/m\mathbf{Z})^*$. Then we have

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

Similarly if $m, n \geq 1$ are odd and $a \in (\mathbf{Z}/mn\mathbf{Z})^*$, then

$$\left(\frac{a}{mn}\right) = \left(\frac{a}{m}\right)\left(\frac{a}{n}\right),$$

where on the right hand side we have identified a with its images in $(\mathbf{Z}/m\mathbf{Z})^*$ and $(\mathbf{Z}/n\mathbf{Z})^*$ under the canonical quotient maps $\mathbf{Z}/mn\mathbf{Z} \to \mathbf{Z}/m\mathbf{Z}$ and $\mathbf{Z}/mn\mathbf{Z} \to \mathbf{Z}/n\mathbf{Z}$.

Proof. The first claim is trivial by the definition of Jacobi symbol.

The second claim follows from the first one by picking odd representatives for a and using the law of quadratic reciprocity.

□

We are now in the position where we can calculate $\left(\frac{m}{n}\right)$ for any odd n efficiently by using the following method:

1. If m is larger than n, reduce it to the range $1 \leq m \leq n-1$.
2. If m is even, reduce $\left(\frac{m}{n}\right)$ to $\left(\frac{2^a}{n}\right)\left(\frac{m'}{n}\right)$ where m' is odd by using the multiplicativity.

3. Any term of the form $\left(\frac{2^a}{n}\right)$ equals $\left(\frac{2}{n}\right)^a = (-1)^{a\frac{n^2-1}{8}}$.

4. For terms of the form $\left(\frac{m}{n}\right)$ with both m and n odd we can use the law of quadratic reciprocity to swap the places of n and m, multiply by $(-1)^{\frac{m-1}{2}\frac{n-1}{2}}$ and start over from step 1.

We already know how to efficiently check for any prime p whether $a \in \mathbf{Z}/p\mathbf{Z}$ is a square (either use Euler's criterion or the above algorithm to calculate the Jacobi symbol). What is left to do to answer completely when a is a square modulo m is considering prime powers. Let us start with powers of two.

Theorem 2.14 The only squares modulo 4 are 0 and 1. Let $m = 2^s$ with $s \geq 3$. Then a is a square modulo m if and only if it is of the form $4^k(8r+1)$ for some $k \geq 0$ and $r \in \mathbf{Z}$.

Proof. Exercise. □

For odd prime powers we have the following.

Theorem 2.15 Let p be an odd prime and $s \geq 1$. Then $a \in \mathbf{Z}/p^s\mathbf{Z}$ is a square if and only if a is 0 or it is of the form $a = p^k u$, where u is a square in $(\mathbf{Z}/p\mathbf{Z})^*$ and $k \geq 0$ is even.

Proof. Exercise. □

Having classified which numbers *have* square roots modulo m, it is now time to start thinking about how to *find* them. It is enough to work in the case where $m = p^k$ is a prime power.

Let us first assume that $m = 2^k$ for some $k \geq 3$ and that $a = 8r + 1$ is odd. By induction we may assume that we have already solved $x^2 = a$ in $\mathbf{Z}/2^{k-1}\mathbf{Z}$. Pick a representative of x in $\mathbf{Z}/2^k\mathbf{Z}$, and consider x^2 and $(x + 2^{k-2})^2 = x^2 + 2^{k-1}x$. They are different since x is odd, and one of them must be a. The base case where $k = 3$ is trivial, just choose $x = \pm 1$ or $x = \pm 3$. Starting from these, the method gives four distinct square roots for each a when $k \geq 3$. If a is of the general form $4^k(8r+1)$, then we can first solve the square root of $(8r+1)$ and multiply the obtained number by 2^k to get a square root of a.

Similar ‘lifting’ works also for odd prime powers. Assume that $m = p^n$, a is not divisible by p, and we have already solved $x^2 = a$ in $\mathbf{Z}/p^{n-1}\mathbf{Z}$, pick a representative of x in $\mathbf{Z}/p^n\mathbf{Z}$ and let k be such that $x^2 + kp^{n-1} = a$ in $\mathbf{Z}/p^n\mathbf{Z}$. Then $(x + k2^{-1}x^{-1}p^{n-1})^2 = x^2 + kp^{n-1}$, so $x + k2^{-1}x^{-1}p^{n-1}$ is our solution in $\mathbf{Z}/p^n\mathbf{Z}$. In total there will be two solutions in this case since there are two solutions in $\mathbf{Z}/p\mathbf{Z}$.

Thus all that is left is to handle the base case $m = p$, where p is an odd prime. In the case $p \equiv 3 \pmod 4$ this is easy.

Theorem 2.16 If $p \equiv 3 \pmod 4$ and $\left(\frac{a}{p}\right) = 1$, then the solutions to $x^2 = a$ are given by $\pm a^{\frac{p+1}{4}}$.

Proof. Exercise. □

When $p \equiv 1 \pmod 4$ no general formula is known. We will describe below the so called *Tonelli–Shanks algorithm* for calculating a solution in this case.

Let p be an odd prime and assume that a is a square modulo p. Tonelli–Shanks begins by writing $p-1$ as $p-1 = 2^s q$, where q is odd. Let $g \in (\mathbf{Z}/p\mathbf{Z})^*$ be a non-square generator for the subgroup of order 2^s. Such a generator can be found by picking a non-square b (by testing numbers randomly – half of the candidates are non-square and thus this should take on average 2 tries) and setting $g = b^q$. (Note that $(b^q)^{2^{s-1}} = b^{\frac{p-1}{2}} = -1$, so g must have order 2^s.) Then a^q lies in the subgroup generated by g, and we can thus write

$$a^{q+1} = aa^q = ag^l$$

for some integer l. Because a is a square, also g^l must be a square, implying that l must be even. Thus if we are capable of solving the discrete logarithm problem in the 2^s element subgroup of $(\mathbf{Z}/p\mathbf{Z})^*$, then $a^{\frac{q+1}{2}} g^{-\frac{l}{2}}$ is the square root we are after and we are done.

So here is the algorithm. We will set $x_0 = a^q$ and $g_0 = g$. On each step $i \geq 0$ our goal is to recursively solve the discrete logarithm problem $g_i^{l^{(i)}} = x_i$ for $l^{(i)}$. We assume that the order of

g_i is 2^{s_i} for some $s_i \geq 0$, that x_i lies in the subgroup generated by g_i and that $l^{(i)}$ will be even, so that we may write it in binary as

$$l^{(i)} = l_1^{(i)}2^1 + l_2^{(i)}2^2 + ... + l_{s_i-1}^{(i)}2^{s_i-1}.$$

Now if $x_i = 1$, we are done since we can simply take $l^{(i)} = 0$. Otherwise we must have $s_i \geq 2$ and therefore we can let t_i be the smallest integer in the range $0 < t_i < s_i$ such that $x_i^{2^{t_i}} = 1$. Then

$$1 = g_i^{l^{(i)}2^{t_i}} = g_i^{l_1^{(i)}2^{t_i+1}+...+l_{s_i-t_i-1}^{(i)}2^{s_i-1}},$$

so we must have $l_1^{(i)} = ... = l_{s_i-t_i-1}^{(i)} = 0$. Similarly since $x_i^{2^{t_i-1}} \neq 1$, we must have

$$1 \neq g_i^{l^{(i)}2^{t_i-1}} = g_i^{l_{s_i-t_i}^{(i)}2^{s_i-1}},$$

so $l_{s_i-t_i}^{(i)} = 1$. This means in particular that

$$g_i^{-2^{s_i-t_i}}x_i = g_i^{l_{s_i-t_i+1}^{(i)}2^{s_i-t_i+1}+...+l_{s_i-1}^{(i)}2^{s_i-1}},$$

so if we let $x_{i+1} := g_i^{-2^{s_i-t_i}}x_i$ and $g_{i+1} := g_i^{2^{s_i-t_i}}$, then we may recursively solve for $l_1^{(i+1)}, ..., l_{t_i-1}^{(i+1)}$ and set $l_{s_i-t_i+j}^{(i)} = l_j^{(i+1)}$ for $1 \leq j \leq t_i - 1$ to obtain the solution for $l^{(i)}$.

In the context of Tonelli–Shanks, instead of keeping around the digits $l^{(i)}$ we simply update a variable that will contain the square root in the end. This variable, which we may call r, is first initialized to $r := a^{\frac{q+1}{2}}$. Then at the end of ith step we multiply it by $g_i^{-2^{s_i-t_i-1}}$, which takes care of the digit of $l^{(0)}$ that we found on this step. In the end r will be equal to $g^{-(l_1^{(0)}+l_2^{(0)}2+...+l_{s-1}^{(0)}2^{s-2})}a^{\frac{q+1}{2}}$, which is what we wanted.

Finally notice that we don't have to keep the generator g itself around, only its inverse. This leads us to the following code.

Algorithm 2.17 (*Tonelli–Shanks)*

```
int64_t tonelli_shanks(int64_t a, int64_t p)
```

```
{
    if(p % 4 == 3) {
        return power_mod(a, (p+1)/4, p);
    }
    int64_t odd=p-1;
    int64_t s=0;
    while(odd%2 == 0) {
        odd/=2;
        s++;
    }
    int64_t g=2;
    while(g < p) {
        if(power_mod(g, (p-1)/2, p) == p-1) break;
        g++;
    }
    g = power_mod(g, odd, p);
    int64_t tmp=power_mod(a, (odd - 1)/2, p);
    int64_t x=tmp*tmp%p*a%p;
    int64_t r=tmp*a%p;
    while(x != 1) {
        int64_t t=1;
        int64_t tmp=x*x%p;
        while(tmp != 1) {
            tmp=tmp*tmp%p;
            t++;
        }
        tmp=power_mod(g, (1 << (s-t-1)), p);
        r=r*tmp%p;
        g=tmp*tmp%p;
        x=x*g%p;
        s=t;
    }
    return r;
}
```

3 MORE ABOUT PRIMES

In this chapter we will look at a few problems specifically related to the primes in $\mathbf{Z}$, such as testing whether a number is prime or counting the number of primes less than some given number n.

3.1 MILLER–RABIN PRIMALITY TEST

In this section we will present a simple method for checking whether a given number is prime. The algorithm is probabilistic, and therefore only tells us that a number is *probably* prime.

It should however be noted that the Miller–Rabin test can easily be made into a deterministic test for ‘small’ ranges, and at the end of the section we will present a variation, which can without fail decide the primality of any number in the range $1, ..., 2^{64}$.

There are also fast deterministic primality tests for general ranges, for example the AKS primality test*. These algorithms are however out of the scope of this book.*

Let us now describe the test. Assume that p is an odd prime number and let $a \in (\mathbf{Z}/p\mathbf{Z})^*$. Then the order of a must divide $p - 1$, which we can write in the form $p - 1 = 2^s r$ where r is odd. Then $a^{2^t r} = 1$ for some smallest t in the range $0 \leq t \leq s$. If $t \geq 1$, then we see that the element $b = a^{2^{t-1} r}$ satisfies the polynomial equation $x^2 = 1$ in the field $\mathbf{Z}/p\mathbf{Z}$. This equation has -1 and 1 as its roots and since b is not 1, we must have $b = -1$. Therefore one of the following happens: Either

- $a^r = 1$, or
- $a^{2^e r} = -1$ for some $0 \leq e \leq s - 1$.

Conversely if neither of these happen, p is *not* a prime. This is the basis for the Miller–Rabin test. The numbers a that fail the test are called **witnesses of compositeness** for p.

Algorithm 3.1 *(Miller–Rabin primality test)* Let p be the number we want to test for primality. We will randomly choose elements $a \in (\mathbf{Z}/p\mathbf{Z})^*$ and see if both of the conditions above fail. In this case we will know that p is not a prime. Otherwise after suitably many elements a have been tested, we will conclude that p is probably a prime.

Notice that if $a \in \mathbf{Z}/p\mathbf{Z}$ and a does not belong to the multiplicative group, then a will automatically fail the test, so we can simply choose our a from the range $2 \leq a \leq p - 1$.

```
bool is_witness(uint64_t a, uint64_t evenpart,
                   uint64_t oddpart, uint64_t p) {
    // We have to use 128-bit integers to be able to
    // multiply two 64-bit ones.
    //
    // power_mod(a, k, p) calculates a^k modulo p
    // doing exponentiation by squaring
    uint128_t u = power_mod(a, oddpart, p);
    if(u == 1) {
        return false;
    }
    for(uint64_t j=1;j<evenpart;j*=2) {
        if(u == p-1) return false;
        u*=u; u%=p;
    }
    return true;
}

const int64_t NUMBER_OF_WITNESSES = 10;

bool is_prime_miller_rabin(uint64_t p) {
    uint64_t oddpart=p-1;
    uint64_t evenpart=1;
    while(oddpart%2 == 0) { evenpart*=2; oddpart/=2; }

    std::default_random_engine gen;
    std::uniform_int_distribution<int64_t> unif(2, p-1);

    for(uint64_t i=0;i<NUMBER_OF_WITNESSES;i++)
        if(is_witness(unif(gen), evenpart, oddpart, p))
            return false;
    return true;
}
```

One may ask how reliable the test is. It is possible to show that if k is the number of potential witnesses we test, the probability that we claim that a composite number is prime is less than 4^{-k}.

We can make a deterministic version of Miller–Rabin by choosing a fixed set of numbers a to test. Indeed at **http://miller-rabin.appspot.com** there are precalculated sets of numbers available that work for large ranges of numbers. In particular the seven number set

$$\{2, 325, 9375, 28178, 450775, 9780504, 1795265022\}$$

is claimed to work for every number of size at most 2^{64}. The algorithm has to be modified so that we test the numbers in the set in order, and if a number is divisible by p, then we report that p is prime. The set has been chosen so that the previous numbers would have rejected any number p that is composite and for which the current a is divisible by p, so that no false-positives occur. Below is an implementation of this fast primality checking algorithm.

Algorithm 3.2 *(Deterministic Miller–Rabin for 64-bit integers)*

```
bool is_prime_miller_rabin_deterministic(uint64_t p) {
    uint64_t odd=p-1;
    uint64_t even=1;
    while(odd%2 == 0) { even*=2; odd/=2; }

    const int64_t bases[7] = {2, 325, 9375, 28178, 450775,
                             9780504, 1795265022};

    for(uint64_t i=0;i<7;i++) {
        uint64_t a = bases[i]%p;
        if(a == 0) return true;
        if(is_witness(a, even, odd, p)) return false;
    }

    return true;
}
```

3.2 COUNTING PRIME NUMBERS

Let $x \geq 1$ be a real number and define the **prime-counting function** π by setting

$$\pi(x) = |\{1 \leq p \leq x : p \text{ is a prime number}\}.$$

In this section we will present a variant of the Meissel–Lehmer algorithm that can calculate $\pi(n)$ in $O(n^{2/3+\varepsilon})$ time and space. It is possible to reduce the space requirement to $O(n^{1/2+\varepsilon})$, but we have decided to keep things simple here. An optimized implementation can be found for example in the Haskell library **arithmoi** or in **sage**.

The algorithm is based on the following simple but clever inclusion-exclusion scheme. Let us denote the number of positive integers at most m and not divisible by any of the k first primes by $\phi(m,k)$. Furthermore denote by $P_j(m,k)$ the number of integers at most m and having exactly j prime factors all of which are strictly greater than the kth prime p_k. Then we clearly have

$$\sum_{j=0}^{\infty} P_j(m,k) = \phi(m,k).$$

On the other hand if we let $m = n$ and $k = \pi(\lfloor \sqrt[3]{n} \rfloor)$, then we have $P_j(n,k) = 0$ for $j \geq 3$, $P_0(n,k) = 1$ and $P_1(n,k) = \pi(n) - k$. It follows that

$$\pi(n) = \phi(n,k) + k - 1 - P_2(n,k).$$

To calculate $P_2(n,k)$ we may use the formula

$$2P_2(n,k) = \sum_{\substack{\lfloor \sqrt[3]{n} \rfloor < p \leq \lfloor \sqrt{n} \rfloor \\ p \text{ a prime}}} \Big(\pi(\frac{n}{p}) - \pi(\lfloor \sqrt[3]{n} \rfloor)\Big) + \pi(\lfloor \sqrt{n} \rfloor) - \pi(\lfloor \sqrt[3]{n} \rfloor),$$

which follows from the fact that every number $pq \leq n$ with $p, q > \sqrt[3]{n}$ will be counted twice in the sum, except when $p = q$, which the term $\pi(\lfloor \sqrt{n} \rfloor) - \pi(\lfloor \sqrt[3]{n} \rfloor)$ corrects.

The function $\phi(m,k)$ on the other hand satisfies the recursion

$$\phi(m,k) = \phi(m,k-1) - \phi(\frac{m}{p_k}, k-1).$$

The algorithm is completed by calculating $\phi(m,k)$ using this recursion and memoizing the values $\phi(m,k)$ for $m \leq \sqrt[3]{n}$ and $k \leq \pi(\sqrt[3]{n})$.

Algorithm 3.3 Here is an implementation of a function `primes_pi` that calculates the number of primes according to the method outlined above.

```
uint64_t phi(vector<vector<int64_t>> &memo,
             const vector<int64_t> &primes,
             uint64_t m, uint64_t k) {
  if(k == 0) return m;
  if(m < memo.size() && k < memo[m].size()) {
    if(memo[m][k] != -1) return memo[m][k];
    int64_t res=phi(memo, primes, m, k-1) -
                phi(memo, primes, m/primes[k-1], k-1);
    memo[m][k]=res;
    return res;
  }
  return phi(memo, primes, m, k-1) -
         phi(memo, primes, m/primes[k-1], k-1);
}

uint64_t primes_pi(uint64_t limit) {
  if(limit < 2) return 0;
  uint64_t limit2=sqrtl(limit);
  uint64_t limit3=cbrtl(limit);
  uint64_t limit23=limit/limit3;
  vector<int64_t> sieve(limit23+1, 0);
  vector<int64_t> pi(limit23+1,0);
  vector<int64_t> primeslow;
  vector<int64_t> primeshigh;
  for(int64_t p=2;p<=limit23;p++) {
    pi[p]=pi[p-1];
    if(sieve[p] == 0) {
      pi[p]++;
      if(p <= limit3) primeslow.push_back(p);
      else if(p <= limit23) primeshigh.push_back(p);
      for(int64_t i=p;i<=limit23;i+=p) sieve[i]=p;
    }
  }
  uint64_t P2=pi[limit2] - pi[limit3];
  for(uint64_t i=0;i<primeshigh.size();i++) {
    P2+=pi[limit/primeshigh[i]] - pi[limit3];
  }
  P2/=2;
  vector<vector<int64_t>> memo(limit3,
         vector<int64_t>(limit3,-1));
  return phi(memo, primeslow, limit, pi[limit3]) +
         pi[limit3] - 1 - P2;
}
```

4 GAUSSIAN INTEGERS

In this chapter we will use the ring theory we have developed in the context of **Gaussian integers**. Gaussian integers are numbers of the form $a+bi$, where a and b are integers. The set of all Gaussian integers is denoted by $\mathbf{Z}[i]$. They are useful for many tasks in classical number theory, mainly because they allow factoring s^2+t^2 as $(s+ti)(s-ti)$.

4.1 STRUCTURE OF $\mathbf{Z}[i]$

The Gaussian integers are reasonably well-behaved, since like integers, they form a Euclidean domain.

Theorem 4.1 $\mathbf{Z}[i]$ is a Euclidean domain.

Proof. Define $f\colon \mathbf{Z}[i] \to \mathbf{N}$ by setting $f(s+ti) = |z+ti|^2 = s^2+t^2$. We must show that for any $a, b \in \mathbf{Z}[i]$, $b \neq 0$, there exist $q, r \in \mathbf{Z}[i]$ such that

$$a = qb + r, \quad f(r) < f(b).$$

To do this, notice that is enough to show that there exists $q \in \mathbf{Z}[i]$ such that

$$\left| \frac{a}{b} - q \right| < 1.$$

Notice that there may be up to four nearest Gaussian integers for a given complex number. Here it is enough to choose one of them.

But it is clear that this happens, because if we choose q to be a nearest Gaussian integer to $\frac{a}{b}$ in the complex plane, then its distance to $\frac{a}{b}$ cannot be more than half of the diagonal of the unit square, which is $\frac{\sqrt{2}}{2} < 1$. □

It follows immediately that many of the good properties we are used to while working with integers apply to Gaussian integers too. In particular they are a PID and a UFD.

Let $N(z) = |z|^2$ for all $z \in \mathbf{Z}[i]$. The function $N\colon \mathbf{Z}[i] \to \mathbf{N}$ is called a **norm**. Its main property is that $N(ab) = N(a)N(b)$. Therefore if $u \in \mathbf{Z}[i]$ is a unit, we must have

$$1 = N(1) = N(uu^{-1}) = N(u)N(u^{-1}),$$

which means that $N(u) = N(u^{-1}) = 1$. Hence the only possible units in $\mathbf{Z}[i]$ are $\{1, i, -1, -i\}$, and one easily checks that these actually are units. Thus $u \in \mathbf{Z}[i]$ is a unit if and only if $N(u) = 1$.

The next big question is what are the primes in $\mathbf{Z}[i]$. For this we have the following.

Theorem 4.2 Let $p \in \mathbf{Z}[i]$. Then p a prime element if and only if it is one of the following up to multiplication by a unit:

- $1 + i$,
- $s + ti$, where $s, t \in \mathbf{Z}$ and $s^2 + t^2$ is a prime number that is 1 modulo 4, or
- q, where $q \in \mathbf{Z}$ is a prime number that is 3 modulo 4.

Proof. Assume first that $p = 1 + i$. Then p is irreducible since if $p = ab$ with $a, b \in \mathbf{Z}[i]$, we must have $2 = N(p) = N(a)N(b)$, which implies that either $N(a)$ or $N(b)$ must be 1, so either a or b is a unit.

Similarly assume that $p = s + ti$, where $s^2 + t^2$ is a prime in $\mathbf{Z}$ that is 1 modulo 4. Then if $p = ab$, we must have $s^2 + t^2 = N(p) = N(a)N(b)$, so that either $N(a)$ or $N(b)$ is 1, meaning that a or b is a unit.

Thirdly assume that $p = q$, where $q \equiv 3 \pmod 4$. Then p is irreducible since if $p = ab$ with non-units $a, b \in \mathbf{Z}[i]$, we must have $q^2 = N(a)N(b)$, so $q = N(a) = N(b)$. Notice that $N(a)$ is a sum of two squares in $\mathbf{Z}$. This is however not possible since squares are either 0 or 1 modulo 4, and $N(a) = q$ is 3 modulo 4.

Assume then that $p \in \mathbf{Z}[i]$ is a prime and write $p = s + ti$ with $s, t \in \mathbf{Z}$. Let $q_1^{\alpha_1} ... q_n^{\alpha_n}$ be the prime factorization of $N(p) = (s + ti)(s - ti)$ in $\mathbf{Z}$. We see that $p = s + ti$ must divide one of q_i. Then it follows that $N(p)|N(q_i)$, so either $N(p) = q_i$ or $N(p) = q_i^2$.

If $N(p) = s^2 + t^2 = q$ with $q \in \mathbf{Z}$ a prime, then either $q = 2$ or $q \equiv 1 \pmod 4$, giving the first and second options on the list.

If $N(p) = q^2$ with $q \in \mathbf{Z}$ a prime, then notice that $N(q) \mid N(p)$, so $u = p/q$ is a Gaussian integer with norm 1, i.e. a unit. Thus

$p = uq$ and also q is a prime in $\mathbf{Z}[i]$. This means that we cannot have $q = 2$. We also cannot have $q \equiv 1 \pmod 4$ because then $x^2 \equiv -1 \pmod q$ has a solution, which means that q divides $x^2 + 1 = (x + i)(x - i)$. If q were prime in $\mathbf{Z}[i]$, one of $\frac{x}{q} + \frac{i}{q}$ or $\frac{x}{q} - \frac{i}{q}$ would be in $\mathbf{Z}[i]$, but this is clearly not the case. The only option left is that $q \equiv 3 \pmod 4$, and this is the third option on the list. □

Another way to look at the result is that integer primes change when looked in $\mathbf{Z}[i]$, so that

– 2 becomes a unit times $(1 + i)^2$, i.e. associated to a square of a Gaussian prime,

– primes q that are 1 modulo 4 split into two Gaussian primes $s + it$ and $s - it$ such that $(s + it)(s - it) = q$,

– primes q that are 3 modulo 4 stay primes in $\mathbf{Z}[i]$.

4.2 SUMS OF SQUARES

The classification of primes in $\mathbf{Z}[i]$ immediately leads to a classification of the numbers $n \in \mathbf{N}$ that are sums of two squares. Indeed, let $n \geq 1$ and assume that its prime factorization is

$$2^{\alpha_0} p_1^{\alpha_1} ... p_k^{\alpha_k} q_1^{\beta_1} ... q_l^{\beta_l},$$

where p are primes that are 1 modulo 4 and q are primes that are 3 modulo 4. The number n is a sum of squares if and only if there exists $a \in \mathbf{Z}[i]$ such that $N(a) = a\bar{a} = n$. But by unique factorization this means that a must have a factor $1 + i$ occuring α_0 times, a split factor occuring α_i times for $1 \leq i \leq k$ and q_i occuring $\beta_i/2$ times for $1 \leq i \leq l$. This is possible if and only if all β_i are even.

It is also easy to count the number of ways an integer n can be expressed as a sum of squares. Let

$$n = 2^{\alpha_0} p_1^{\alpha_1} ... p_k^{\alpha_k} q_1^{\beta_1} ... q_l^{\beta_l},$$

where p_i are primes that are 1 modulo 4 and q are primes that are 3 modulo 4. Let $R(n)$ be the number of ways to write $n = x^2 + y^2$ where $x, y \in \mathbf{Z}$. Then if one of β_i is odd, we must have $R(n) = 0$. Otherwise we must consider all possible ways of writing $n = (x + yi)(x - yi)$. Now in any case $x + yi$ must have $q_i^{\beta_i/2}$ as a factor, so there is no choice involved there. Similarly it must have $(1 + i)^{\alpha_0}$ as a factor. Let $s + t = \alpha_i$ for some $1 \le i \le k$. To have $p_i^{\alpha_i}$ in the prime factorization of $(x + yi)(x - yi)$, we can choose $(a_i + b_i i)^s (a_i - b_i i)^t$ with $s + t = \alpha_i$ to be a factor of $x + yi$. Here we have written $p_i = (a_i + b_i i)(a_i - b_i i)$ and the exponents s and t can be chosen in $\alpha_i + 1$ different ways for each i. Finally we can multiply our number by any of the 4 units, which gives us in total

$$R(n) = \begin{cases} 0, & \text{if any } \beta_i \text{ is odd,} \\ 4(\alpha_1 + 1) ... (\alpha_k + 1), & \text{otherwise} \end{cases}$$

ways.

4·3 PYTHAGOREAN TRIPLES

A triple (a, b, c) of positive integers is called a **Pythagorean triple** if $a^2 + b^2 = c^2$. The name Pythagorean triple comes from the fact that a, b and c can be thought of as the sides of a right-angled triangle satisfying Pythagorean theorem. The most famous example of a Pythagorean triple is $(3, 4, 5)$.

A Pythagorean triple is called **primitive** if $\gcd(a, b, c) = 1$. All non-primitive triples can be obtained from a unique primitive triple by multiplying a, b and c by a common constant. Therefore we will next focus on how to generate primitive triples.

Assume that $a^2 + b^2 = c^2$ and $\gcd(a, b, c) = 1$. We can factor the equation in $\mathbf{Z}[i]$ to get

$$(a + bi)(a - bi) = c^2.$$

Now notice that $a + bi$ and $a - bi$ are coprime. Indeed, if this was not the case, then $2a = (a + bi) + (a - bi)$ and $a + bi$ would have a common factor. Because a and b are coprime, this implies that

$2 = -i(1+i)^2$ and $a+bi$ have a common factor. In particular we see that $a-bi$ has factor $1-i$, so c is divisible by 2. But this is not possible because then a and b are odd, which makes $a^2+b^2 \equiv 2 \pmod 4$ and $c^2 \equiv 0 \pmod 4$.

It follows that $a+bi$ and $a-bi$ are squares times a unit,

$$a + bi = u(m+ni)^2 = u(m^2 - n^2 + 2mni).$$

Now we may notice that $(a+bi)(a-bi) = (m^2+n^2)^2$, so $c = m^2+n^2$. Without loss of generality we may pick $u = 1$, so that $a = m^2 - n^2$ and $b = 2mn$. Then to satisfy $\gcd(a,b,c) = 1$, we must have $\gcd(m,n) = 1$ and exactly one of m, n must be odd. Thus we have shown the following.

Theorem 4.3 The primitive Pythagorean triples are of the form

$$a = m^2 - n^2, \quad b = 2mn, \quad c = m^2 + n^2,$$

where $1 \le n \le m-1$, $\gcd(m,n) = 1$, and exactly one of m, n is odd.

There is also an alternative way to generate primitive Pythagorean triples that is useful because it does not require checking for coprimality or parity of numbers. This method starts from the triple $(3, 4, 5)$ and then recursively produces new triples by applying the following three linear transformations:

$$\begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 3 \end{pmatrix}, \quad \begin{pmatrix} -1 & 2 & 2 \\ -2 & 1 & 2 \\ -2 & 1 & 3 \end{pmatrix}, \quad \begin{pmatrix} 1 & -2 & 2 \\ 2 & -1 & 2 \\ 2 & -2 & 3 \end{pmatrix}$$

The proof that this works is left as an exercise.

5 CONTINUED FRACTIONS

5.1 DEFINITIONS AND EXAMPLES

Suppose that x is a real number and let $x_0 := x$. If x_0 is not an integer, then we may write

$$x_0 = a_0 + \frac{1}{x_1},$$

where $a_0 = \lfloor x_0 \rfloor$ and $x_1 = \frac{1}{x_0 - \lfloor x_0 \rfloor} > 1$. If x_1 is not an integer, then we may continue and write

$$x_1 = a_1 + \frac{1}{x_2},$$

where $a_1 = \lfloor x_1 \rfloor$ and $x_2 = \frac{1}{x_1 - \lfloor x_1 \rfloor} > 1$. Continuing like this we get the (formal) equality

$$x = a_0 + \cfrac{1}{a_1 + \frac{1}{a_2 + \frac{1}{\cdots}}},$$

where the right hand side is called a **continued fraction with coefficients** $a_0, a_1, a_2, ...$. We will use the shorthand notation

$$[a_0, a_1, a_2, ...] := a_0 + \cfrac{1}{a_1 + \frac{1}{a_2 + \frac{1}{\cdots}}}$$

to write continued fractions. The number of coefficients can be either finite or infinite.

By the **continued fraction expansion** of a real number x we mean the continued fraction obtained by using the above procedure. In particular a_i are all integers, and for all $i \geq 1$ we have $a_i > 0$.

Exercise 5.1 Show that if x is not an integer and the continued fraction expansion of x is finite, then the last coefficient of the expansion is at least 2.

At this point continued fractions are just formal objects defined by the coefficients a_i, which we can in general assume to be arbitrary real numbers. If the continued fraction is finite it is straightforward to assign a value to it, assuming that the expression makes sense (meaning that there is no division by 0). It is less clear how we should interpret an infinite continued fraction, but we will soon show that if we truncate it to a finite number of coefficients then the sequence of such truncations converges if we assume that all of the coefficients are integers. We can then say that the value of the infinite continued fraction is the limit of its truncations.

Theorem 5.2 The continued fraction expansion of a real number x has finitely many coefficients if and only if x is rational.

Proof. It is clear that if the expansion of x has finitely many coefficients then x is rational. Let us thus assume that there exists a rational number whose continued fraction expansion has infinitely many terms. Then in particular there exists such a rational number $\frac{p}{q}$ with *the smallest possible denominator* $q > 0$. Now the first step in the expansion procedure is

$$\frac{p}{q} = \left\lfloor \frac{p}{q} \right\rfloor + \frac{1}{\frac{1}{\frac{p}{q} - \left\lfloor \frac{p}{q} \right\rfloor}},$$

so it follows that the continued fraction expansion of

$$\frac{1}{\frac{p}{q} - \left\lfloor \frac{p}{q} \right\rfloor} = \frac{q}{p - q\left\lfloor \frac{p}{q} \right\rfloor}$$

has infinitely many terms. Note however that $0 \leq p - q\left\lfloor \frac{p}{q} \right\rfloor < q$, which is a contradiction. □

Let $[a_0, a_1, a_2, ...]$ be a continued fraction. Then its nth **convergent** is defined to be the continued fraction $[a_0, a_1, a_2, ..., a_n]$. If the coefficients $a_0, ..., a_n$ are integers, it is a rational number and we will use the convention that its numerator and denominator (in lowest terms) are denoted by p_n and q_n respectively.

Theorem 5.3 The p_n and q_n satisfy the recursion relations

$$\begin{aligned} p_n &= a_n p_{n-1} + p_{n-2}, \\ q_n &= a_n q_{n-1} + q_{n-2}, \end{aligned}$$

for all $n \geq 0$ when we set

$$p_{-2} := 0, \quad q_{-2} := 1, \quad p_{-1} := 1, \quad q_{-1} := 0.$$

Proof. Our first claim is that if p_n and q_n are defined via the given recursion relations, then $\frac{p_n}{q_n} = [a_0, a_1, ..., a_n]$. The proof will proceed by induction on the length of the continued fraction and *we do not assume that* $a_0, a_1, ..., a_n$ *are integers*. It is clear that the claim holds for all continued fractions of length 1. Assume that the recursion relations hold for all continued fractions of length n. Then we can bunch together the last two coefficients of the continued fraction $[a_0, a_1, ..., a_n]$ and write it as $[a_0, a_1, ..., a_{n-2}, a_{n-1} + \frac{1}{a_n}]$, which is now a continued fraction of length n. It has the same initial segment $[a_0, a_1, ..., a_{n-2}]$ as $[a_0, a_1, ..., a_n]$ does, so the convergents up to $\frac{p_{n-2}}{q_{n-2}}$ are equal for them. By induction $[a_0, a_1, ..., a_{n-2}, a_{n-1} + \frac{1}{a_n}]$ is given by

$$\begin{aligned} \frac{(a_{n-1} + \frac{1}{a_n})p_{n-2} + p_{n-3}}{(a_{n-1} + \frac{1}{a_n})q_{n-2} + q_{n-3}} &= \frac{a_n(a_{n-1}p_{n-2} + p_{n-3}) + p_{n-2}}{a_n(a_{n-1}q_{n-2} + q_{n-3}) + q_{n-2}} \\ &= \frac{a_n p_{n-1} + p_{n-2}}{a_n q_{n-1} + q_{n-2}}, \end{aligned}$$

which shows that the claim holds for continued fractions of length $n+1$.

The second claim is that p_n and q_n are coprime integers if $a_0, a_1, ..., a_n$ are integers. That p_n and q_n are integers is obvious from the recursion. The coprimality follows from the next theorem. □

Theorem 5.4 The numbers p_n and q_n satisfy $p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1}$ for all $n \geq -1$.

Proof. It is trivial to check that $p_{-1}q_{-2} - p_{-2}q_{-1} = 1$. Assume then that $n \geq 0$ and that the claim holds for $n-1$. We have

$$\begin{aligned} p_n q_{n-1} - p_{n-1} q_n &= (a_n p_{n-1} + p_{n-2}) q_{n-1} - p_{n-1}(a_n q_{n-1} + q_{n-2}) \\ &= -(p_{n-1}q_{n-2} - p_{n-2}q_{n-1}) = (-1)^{n-1}, \end{aligned}$$

so the claim also holds for n. We are done by induction. □

An important corollary of this theorem is the following.

Theorem 5.5 The convergents of any infinite continued fraction $[a_0, a_1, a_2, ...]$ with integer coefficients such that $a_k > 0$ for $k \geq 1$ converge to some number x. Moreover every even convergent is strictly less than x, every odd convergent is strictly larger than x, and the distance to x goes to 0 monotonically.

Proof. We can rewrite the equation in Theorem 5.4 as

$$\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{(-1)^{n-1}}{q_{n-1}q_n}.$$

Since $q_n \to \infty$ monotonically, this shows that the convergents oscillate with amplitude going monotonically to 0. □

Continued fractions can be thought of as another way of representing real numbers, a little bit like the decimal number system. Like in the decimal system, the representation is essentially unique. With decimal numbers the non-uniqueness comes in the form $1 = 0.999...$, and with continued fractions it comes in the form $1 = 0 + \frac{1}{1}$. To get a unique representation with the decimal numbers we can prohibit infinite sequence of nines, and similarly with continued fractions we may require that the last coefficient is not 1, except if the last coefficient is also the first one.

Definition 5.6 We say that $[a_0, a_1, ...]$ is in the **canonical form** if

- $a_0, a_1, \ldots$ are integers,
- $a_k > 0$ for all $k \geq 1$, and
- if there are finitely many coefficients, the last coefficient is at least 2.

Theorem 5.7 There is a bijective correspondence between the real numbers and continued fractions in the canonical form given by the mapping from a number to its continued fraction expansion.

Proof. From Theorem 5.2 we already know that the rational numbers and finite continued fractions in the canonical form are in a one-to-one correspondence.

Recall from the proof of Theorem 5.3 that the convergent recursion works also for non-integers.

Let x be an irrational number. By the definition of the continued fraction expansion and the convergent recursion we have

$$x = [a_0, a_1, \ldots, a_n, x_{n+1}] = \frac{p_n x_{n+1} + p_{n-1}}{q_n x_{n+1} + q_{n-1}}$$

for some real number $x_{n+1} > 1$. From Theorem 5.4 it follows that

$$\left| x - \frac{p_n}{q_n} \right| = \left| \frac{p_{n-1} q_n - p_n q_{n-1}}{q_n (q_n x_{n+1} + q_{n-1})} \right| < \frac{1}{q_n (q_n + q_{n-1})}.$$

The right-hand side tends to 0 as n goes to ∞. This shows that there is an injection from the irrational numbers to the infinite continued fractions in the canonical form.

It remains to show that the if two infinite fractions in the canonical form have the same value, their coefficients are equal. Notice that if $x = [a_0, a_1, \ldots]$, then $\lfloor x \rfloor = a_0$ and $x_1 = \frac{1}{x - \lfloor x \rfloor} = [a_1, a_2, a_3, \ldots]$. Thus by running the continued fraction expansion process on x we see that the coefficients $a_0, a_1, \ldots$ are uniquely determined. □

It is sometimes also useful to think about the numerator and denominator of a general finite continued fraction $[a_0, a_1, a_2, \ldots, a_n]$ as polynomials in the coefficients $a_0, \ldots, a_n$. We will let $P(a_0, a_1, \ldots, a_n)$ denote the numerator and $Q(a_0, a_1, \ldots, a_n)$ the denominator. It is easy to see (compare Theorem **5.3**) that both P and Q satisfy the recursion relations

$$\begin{aligned} P(a_0, ..., a_n) &= a_n P(a_0, ..., a_{n-1}) + P(a_0, ..., a_{n-2}), \\ Q(a_0, ..., a_n) &= a_n Q(a_0, ..., a_{n-1}) + Q(a_0, ..., a_{n-2}) \end{aligned}$$

with $P() = 1$, $Q() = 0$, $P(a_0) = a_0$ and $Q(a_0) = 1$. Using these it is also straightforward to show by using induction that

$$Q(a_0, a_1, ..., a_n) = P(a_1, a_2, ..., a_n).$$

Quite interestingly the following property also holds.

Theorem 5.8 We have $P(a_0, a_1, ..., a_n) = P(a_n, a_{n-1}, ..., a_0)$.

Proof. It is easy to check that the claim holds when $n = 0, 1, 2, 3$. Assume that $n \geq 4$ and that we have shown the claim up to $n - 1$. Then by induction we have

$$\begin{aligned} P(a_0, ..., a_n) &= a_n P(a_0, ..., a_{n-1}) + P(a_0, ..., a_{n-2}) \\ &= a_n P(a_{n-1}, ..., a_0) + P(a_{n-2}, ..., a_0) \\ &= a_0 a_n P(a_{n-1}, ..., a_1) + a_n P(a_{n-1}, ..., a_2) \\ &\quad + a_0 P(a_{n-2}, ..., a_1) + P(a_{n-2}, ..., a_2) \\ &= a_0 a_n P(a_1, ..., a_{n-1}) + a_0 P(a_1, ..., a_{n-2}) \\ &\quad + a_n P(a_2, ..., a_{n-1}) + P(a_2, ..., a_{n-2}) \\ &= a_0 P(a_1, ..., a_n) + P(a_2, ..., a_n) \\ &= a_0 P(a_n, ..., a_1) + P(a_n, ..., a_2) \\ &= P(a_n, ..., a_0), \end{aligned}$$

so the claim also holds for n. □

As a word of warning, the same obviously does not hold for Q. Instead, we have

$$Q(a_0, ..., a_n) = P(a_1, ..., a_n) = P(a_n, ..., a_1) = Q(x, a_n, ..., a_1)$$

for any x.

5.2 BEST RATIONAL APPROXIMATIONS

One important application of continued fractions is in finding best rational approximations to real numbers. If x is a real number, then $\frac{a}{b}$ (where $b > 0$) is a **best rational approximation** to x if for all fractions $\frac{a'}{b'} \neq \frac{a}{b}$ with $0 < b' \leq b$ we have

$$\left| x - \frac{a}{b} \right| < \left| x - \frac{a'}{b'} \right|.$$

All such best rational approximations can be found by looking at the convergents and *semiconvergents* of the continued fraction expansion of x. A **semiconvergent** of a continued fraction $[a_0, a_1, ...]$ is a rational expression of the form

$$\frac{p_{n-1}a + p_{n-2}}{q_{n-1}a + q_{n-2}},$$

where $0 < a < a_n$ and $n \geq 1$. Notice that the case $a = 0$ corresponds to the convergent $\frac{p_{n-2}}{q_{n-2}}$ and the case $a = a_n$ corresponds to the convergent $\frac{p_n}{q_n}$. It is easy to check that the rest of the semiconvergents lie between these two extremes and get closer to x as a increases. Notice that like convergents, semiconvergents are always irreducible fractions. (Why?)

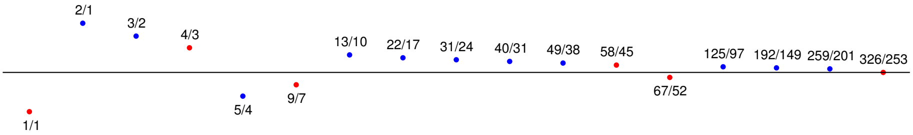

Figure 5.1 The convergents (red) and semiconvergents (blue) of $\frac{326}{253} = [1, 3, 2, 6, 1, 4]$, sorted by denominator size

Figure 5.1 illustrates the behaviour of convergents and semiconvergents. In fact the best rational approximations of $\frac{326}{253}$ can be shown to be its convergents and the semiconvergents $\frac{3}{2}$, $\frac{5}{4}$, $\frac{40}{31}$, $\frac{49}{38}$, $\frac{192}{149}$ and $\frac{259}{201}$. This is a special case of the following theorem.

Theorem 5.9 Let x be a real number with continued fraction expansion $[a_0, a_1, ...]$. All best rational approximations to x are given by

- the convergents of $[a_0, a_1, ...]$,
- the semiconvergents $\frac{p_{n-1}a+p_{n-2}}{q_{n-1}a+q_{n-2}}$ such that $\frac{a_n}{2} < a < a_n$, and
- in the case a_n is even, the semiconvergent $\frac{p_{n-1}\frac{a_n}{2}+p_{n-2}}{q_{n-1}\frac{a_n}{2}+q_{n-2}}$ is also a best rational approximation if and only if

$$\left|\frac{p_{n-1}\frac{a_n}{2}+p_{n-2}}{q_{n-1}\frac{a_n}{2}+q_{n-2}} - x\right| < \left|\frac{p_{n-1}}{q_{n-1}} - x\right|.$$

Proof. Our road to QED is the following:

(1) We start by proving that all best rational approximations are convergents or semiconvergents.

(2) Next we prove that any semiconvergent $\frac{p_{n-1}a+p_{n-2}}{q_{n-1}a+q_{n-2}}$ with $1 \le a < \frac{a_n}{2}$ is not a best rational approximation by showing that

$$\left|\frac{p_{n-1}}{q_{n-1}} - x\right| < \left|\frac{p_{n-1}a+p_{n-2}}{q_{n-1}a+q_{n-2}} - x\right|.$$

(3) Finally we prove that all semiconvergents $\frac{p_{n-1}a+p_{n-2}}{q_{n-1}a+q_{n-2}}$ with $a > \frac{a_n}{2}$ are best rational approximations by showing that

$$\left|\frac{p_{n-1}\frac{a_n+1}{2}+p_{n-2}}{q_{n-1}\frac{a_n+1}{2}+q_{n-2}} - x\right| < \left|\frac{p_{n-1}}{q_{n-1}} - x\right|.$$

> Because the semiconvergents get closer to x as a increases and because their denominators grow monotonically, all of them must be best rational approximations by part (1).

To prove (1), notice that any best rational approximation $\frac{p}{q}$ that is not a convergent or semiconvergent lies between two of them. Let

$$\frac{p_n a + p_{n-1}}{q_n a + q_{n-1}} \quad \text{and} \quad \frac{p_n(a+1) + p_{n-1}}{q_n(a+1) + q_{n-1}}$$

be two such (semi)convergents with either $n \geq 1$ and $0 \leq a < a_{n+1}$ or $n = 0$ and $1 \leq a < a_1$. We will now show that $q_n(a+1) + q_{n-1} < q$, which gives us the needed contradiction since the second (semi)convergent is closer to x than $\frac{p}{q}$ is. Clearly

$$\left|\frac{p}{q} - \frac{p_n a + p_{n-1}}{q_n a + q_{n-1}}\right| = \frac{m}{q(q_n a + q_{n-1})},$$

where $m = |p(q_n a + q_{n-1}) - q(p_n a + p_{n-1})|$. On the other hand

$$\begin{aligned}\left|\frac{p}{q} - \frac{p_n a + p_{n-1}}{q_n a + q_{n-1}}\right| &< \left|\frac{p_n(a+1) + p_{n-1}}{q_n(a+1) + q_{n-1}} - \frac{p_n a + p_{n-1}}{q_n a + q_{n-1}}\right| \\ &= \frac{|p_n q_{n-1} - p_{n-1} q_n|}{(q_n(a+1) + q_{n-1})(q_n a + q_{n-1})} \\ &= \frac{1}{(q_n(a+1) + q_{n-1})(q_n a + q_{n-1})}.\end{aligned}$$

Because $m \geq 1$, we have

$$\frac{1}{q(q_n a + q_{n-1})} < \frac{1}{(q_n(a+1) + q_{n-1})(q_n a + q_{n-1})},$$

from which the claim follows.

To show (2), notice that $\frac{p_n}{q_n}$ lies on the same side of x as $\frac{p_{n-1}a+p_{n-2}}{q_{n-1}a+q_{n-2}}$, while $\frac{p_{n-1}}{q_{n-1}}$ lies on the opposite side, so it is enough to show that

$$\left|\frac{p_{n-1}a + p_{n-2}}{q_{n-1}a + q_{n-2}} - \frac{p_n}{q_n}\right| > \left|\frac{p_{n-1}}{q_{n-1}} - \frac{p_n}{q_n}\right| = \frac{1}{q_{n-1}q_n}.$$

A short calculation reveals that the left hand side is equal to

$$\frac{a_n - a}{(q_{n-1}a + q_{n-2})q_n},$$

and after simplifying we are left with the inequality

$$a < \frac{a_n}{2} - \frac{q_{n-2}}{2q_{n-1}},$$

which is clearly true since $a < \frac{a_n}{2}$ implies that $a \leq \frac{a_n}{2} - \frac{1}{2}$.

Finally we must show (3). We will assume that the continued fraction has coefficient a_{n+1}, which implies that it is enough to show that

$$\left|\frac{p_{n-1}a + p_{n-2}}{q_{n-1}a + q_{n-2}} - \frac{p_{n+1}}{q_{n+1}}\right| < \left|\frac{p_{n-1}}{q_{n-1}} - \frac{p_{n+1}}{q_{n+1}}\right|.$$

Otherwise we can compare to $x = \frac{p_n}{q_n}$ directly, and the argument is similar to part (2) and is left for the reader to check. Now, a short calculation shows that right hand side is

$$\frac{a_{n+1}}{q_{n-1}q_{n+1}}$$

and that the left hand side is

$$\frac{a_n a_{n+1} - a a_{n+1} + 1}{(q_{n-1}a + q_{n-2})q_{n+1}}.$$

Setting $a = \frac{a_n + 1}{2}$ we get after simplifying that

$$\frac{1}{2} > \frac{1}{2a_{n+1}} - \frac{q_{n-2}}{2q_{n-1}},$$

which is clearly true. □

There is also a stronger way a rational number can be a best approximation. We say that $\frac{a}{b}$ (where $b > 0$) is a **best rational approximation of the second kind** if for all $\frac{a'}{b'} \neq \frac{a}{b}$ with $0 < b' \leq b$ we have

$$|bx - a| < |b'x - a'|.$$

Notice that this is indeed a stronger property than being a best rational approximation of the first kind, since it implies that

$$\left|x - \frac{a}{b}\right| < \frac{b'}{b}\left|x - \frac{a'}{b'}\right| \leq \left|x - \frac{a'}{b'}\right|.$$

One can also check that it is a strictly stronger property by noticing for example that $\frac{1}{3}$ is a best rational approximation of the first kind but not of second kind to $\frac{2}{5}$.

Theorem 5.10 Every best rational approximation of the second kind is a convergent. Also the converse holds, except for the trivial case $x = x_0 + \frac{1}{2}$ for some integer x_0.

Proof. The converse does not hold for $x = x_0 + \frac{1}{2}$ because x_0 and $x_0 + 1$ are equidistant from x and both have denominator 1.

Let us first show that every best rational approximation of the second kind is a convergent. We know by Theorem 5.9 that it is either a convergent or semiconvergent, so it is enough to rule out the possibility of it being a semiconvergent. Assume thus that the semiconvergent $\frac{p_{n-1}a+p_{n-2}}{q_{n-1}a+q_{n-2}}$ (where $0 < a < a_n$) is a best rational approximation of the second kind. We will show that

$$|q_{n-1}x - p_{n-1}| \leq |(q_{n-1}a + q_{n-2})x - (p_{n-1}a + p_{n-2})|,$$

which gives us a contradiction. Notice that it is enough to show that

$$\left|q_{n-1}\frac{p_n}{q_n} - p_{n-1}\right| \leq \left|(q_{n-1}a + q_{n-2})\frac{p_n}{q_n} - (p_{n-1}a + p_{n-2})\right|,$$

since $\frac{p_n}{q_n}$ is on the opposite side of x from $\frac{p_{n-1}}{q_{n-1}}$ and on the same side as $\frac{p_{n-1}a+p_{n-2}}{q_{n-1}a+q_{n-2}}$. Now the left hand side is $\frac{1}{q_n}$ and a short calculation shows that the right hand side is $\frac{a_n-a}{q_n}$, which proves the claim.

Let us now show the converse. That is, every convergent is a best rational approximation of the second kind (apart from the trivial case). By the first part it is enough to check that each convergent is a better approximation than the previous one. That is,

$$|q_n x - p_n| < |q_{n-1}x - p_{n-1}|.$$

If $x = \frac{p_n}{q_n}$, we are done. Otherwise we can instead of x compare to $\frac{p_{n+1}}{q_{n+1}}$. Notice that

$$\begin{aligned} |q_n x - p_n| &\leq \left| q_n \frac{p_{n+1}}{q_{n+1}} - p_n \right| = \frac{1}{q_{n+1}} \\ &\leq \frac{a_{n+1}}{q_{n+1}} = \left| q_{n-1} \frac{p_{n+1}}{q_{n+1}} - p_{n-1} \right| \\ &\leq |q_{n-1}x - p_{n-1}|. \end{aligned}$$

If $x = \frac{p_{n+1}}{q_{n+1}}$ it follows that $a_{n+1} \neq 1$ and the inequality in the center is strict. If $x \neq \frac{p_{n+1}}{q_{n+1}}$ then the first and last inequalites are strict by oscillation. □

Finally we would like to present the following theorem that lets us in some cases infer that if a given rational number is a good enough approximation, it must actually be a convergent.

Theorem 5.11 Let x be a real number and assume that the rational number $\frac{a}{b}$ satisfies

$$\left| x - \frac{a}{b} \right| < \frac{1}{2b^2}.$$

Then $\frac{a}{b}$ is a convergent of the continued fraction expansion of x.

Proof. By Theorem 5.10 it is enough to check that $\frac{a}{b}$ is a best rational approximation of the second kind. Assume that $\frac{p}{q} \neq \frac{a}{b}$ and

$$|qx - p| \leq |bx - a|.$$

Then we have

$$\left|x - \frac{p}{q}\right| \leq \frac{1}{q}\,|bx - a| < \frac{1}{2bq}.$$

On the other hand

$$\frac{1}{bq} \leq \left|\frac{a}{b} - \frac{p}{q}\right| < \frac{1}{2bq} + \frac{1}{2b^2} = \frac{b+q}{2b^2q},$$

which implies that $b < q$. □

5·3 QUADRATIC IRRATIONALS

Let a, b, c be integers satisfying $a \neq 0$ and $b^2 - 4ac > 0$. Then the quadratic equation $ax^2 + bx + c = 0$ has two real solutions given by

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

If $b^2 - 4ac$ is not a square, then both of these solutions are irrational and such irrational numbers are called **quadratic irrationals**.

Let x be a quadratic irrational. Then x is a root of some polynomial $az^2 + bz + c$ with integer coefficients. The other root of this polynomial is denoted by $\overline{x}$ and is called the **conjugate** of x. Notice that the definition of the conjugate does not depend on the choice of the polynomial since any such polynomial is divisible by the minimal polynomial of x.

Theorem 5.12 Let D be a square free integer and let $\mathbf{Q}(\sqrt{D})$ denote the numbers of the form $p + q\sqrt{D}$ with $p, q \in \mathbf{Q}$. Define conjugation in $\mathbf{Q}(\sqrt{D})$ by $\overline{p + q\sqrt{D}} = p - q\sqrt{D}$. Then $\mathbf{Q}(\sqrt{D})$ is a field consisting of the rational numbers and those quadratic irrationals x for which the square free part of $b^2 - 4ac$ equals D whenever $ax^2 + bx + c = 0$ for some $a, b, c \in \mathbf{Z}$. For quadratic irrationals the conjugation coincides with the conjugation defined above, and it satisfies $\overline{p+q} = \overline{p} + \overline{q}$, $\overline{pq} = \overline{p}\overline{q}$ and $\overline{p^{-1}} = \overline{p}^{-1}$ for all $p, q \in \mathbf{Q}(\sqrt{D})$.

Proof. Exercise. □

A continued fraction of the form $[a_0, a_1, ..., a_\ell, \overline{b_1, ..., b_m}]$ where the block $b_1, ..., b_m$ of coefficients repeats ad infinitum is called **periodic**. The main goal of this section is to prove the following alternative characterization of quadratic irrationals.

Theorem 5.13 A real number x is a quadratic irrational if and only if its continued fraction expansion is periodic.

The full proof of this will require a couple of steps. We will first prove a corresponding theorem for *purely periodic continued fractions* and *reduced quadratic irrationals.* The proof is then finished by a bootstrapping argument that lets us reduce the general case to this case. A continued fraction is said to be **purely periodic** if it is of the form $[\overline{b_0, b_1, ..., b_{m-1}}]$ and a quadratic irrational x is called **reduced** if it satisfies $x > 1$ and $-1 < \overline{x} < 0$.

Theorem 5.14 A continued fraction is purely periodic if and only if it is the expansion of a reduced quadratic irrational.

Proof. Let us first show that a purely periodic continued fraction is a reduced quadratic irrational.

Let $x = [\overline{b_0, b_1, ..., b_{m-1}}]$ be a purely periodic continued fraction with convergents p_n/q_n. Then in particular we have

$$x = [b_0, b_1, ..., b_{m-1}, x] = \frac{xp_{m-1} + p_{m-2}}{xq_{m-1} + q_{m-2}},$$

which shows that x is a root of the equation

$$q_{m-1}x^2 + (q_{m-2} - p_{m-1})x - p_{m-2} = 0. \tag{5.1}$$

Similarly if we reverse the coefficients and let $y = [\overline{b_{m-1}, ..., b_0}]$, then

$$\begin{aligned} y &= [b_{m-1}, b_{m-2}, ..., b_0, y] \\ &= \frac{yP(b_{m-1}, ..., b_0) + P(b_{m-1}, ..., b_1)}{yQ(b_{m-1}, ..., b_0) + Q(b_{m-1}, ..., b_1)} \\ &= \frac{yP(b_0, ..., b_{m-1}) + P(b_1, ..., b_{m-1})}{yP(b_{m-2}, ..., b_0) + P(b_{m-2}, ..., b_1)} \end{aligned}$$

$$= \frac{yP(b_0, ..., b_{m-1}) + Q(b_0, ..., b_{m-1})}{yP(b_0, ..., b_{m-2}) + Q(b_0, ..., b_{m-2})}$$

$$= \frac{yp_{m-1} + q_{m-1}}{yp_{m-2} + q_{m-2}}.$$

Thus y is a root of the equation

$$p_{m-2}y^2 + (q_{m-2} - p_{m-1})y - q_{m-1} = 0.$$

Dividing both sides by $-y^2$ shows that $-1/y$ is a root of (5.1) and thus $-1/y = \overline{x}$. Because $x > 1$ and $y > 1$ (since there can be no zeros in their continued fractions), this shows that x is a reduced quadratic irrational.

Assume then that x is a reduced quadratic irrational. We want to show that its continued fraction expansion is purely periodic. First notice that for a given non-square integer $D > 0$ there are only finitely many different reduced quadratic irrationals that satisfy a quadratic equation $az^2 + bz + c = 0$ with $a > 0$ and $b^2 - 4ac = D$. To see this, we can use Vieta's formulas to write $b = -ax - a\overline{x}$ and $c = ax\overline{x}$ where x and $\overline{x}$ are the roots of the polynomial. If x is a reduced quadratic irrational, it follows that $D = a^2(x-\overline{x})^2 > a^2$, so $|a| \leq \sqrt{D}$. Because a and c have different signs, $D = b^2 - 4ac \geq b^2$, so $|b| \leq \sqrt{D}$. Finally $|c| = \frac{|b^2-D|}{4|a|} \leq \frac{D}{4}$. Thus the coefficients a, b, c are bounded and therefore there are only finitely many such equations, thus also finitely many reduced quadratic irrationals.

Let $x_0 := x$ and choose $a_0, b_0, c_0 \in \mathbf{Z}$ so that $a_0x_0^2 + b_0x_0 + c_0 = 0$. Set $D := b_0^2 - 4a_0c_0$. Define the sequence $x_n := \frac{1}{x_{n-1} - \lfloor x_{n-1} \rfloor}$ for building the continued fraction. Notice that if we choose

$$\begin{aligned} a_n &= \lfloor x_{n-1} \rfloor^2 a_{n-1} + \lfloor x_{n-1} \rfloor b_{n-1} + c_{n-1}, \\ b_n &= 2\lfloor x_{n-1} \rfloor a_{n-1} + b_{n-1}, \\ c_n &= a_{n-1}, \end{aligned}$$

then $a_nx_n^2 + b_nx_n + c_n = 0$ and by induction

$$b_n^2 - 4a_nc_n = b_{n-1}^2 - 4a_{n-1}c_{n-1} = D.$$

By Theorem 5.12 we have $\overline{x_n} = \frac{1}{\overline{x_{n-1}} - \lfloor x_{n-1} \rfloor}$ and we see that $-1 < \overline{x_n} < 0$, so x_n is reduced. The sequence x_n must be periodic by the observations above, and therefore also the continued fraction expansion of x is periodic.

It remains to show that the expansion is *purely* periodic. We can do this by showing that if $x_{n+1} = x_{m+1}$ for some n, m, then $x_n = x_m$. Starting from indices where the periodicity holds this lets us work backwards to the beginning, showing that the periodicity holds all the time. Notice that $x_{n+1} = x_{m+1}$ implies that $x_n - \lfloor x_n \rfloor = x_m - \lfloor x_m \rfloor$. Thus $x_n \equiv x_m \pmod 1$. It is therefore enough to show that if x is a quadratic irrational, then there is exactly one k such that $x + k$ is a reduced quadratic irrational. This is however clear since the conjugate of $x + k$ is $\overline{x} + k$, which lies in the interval $(-1, 0)$ for exactly one k. □

We are now ready to prove the general case.

Proof. (*Proof of Theorem* 5.13) Let us first show that a periodic continued fraction is a quadratic irrational.

Let $x = [a_0, a_1, ..., a_\ell, \overline{b_1, ..., b_m}]$ and write $y = [\overline{b_1, ..., b_m}]$. Then by Theorem 5.14 we know that y is a quadratic irrational. It is easy to see using Theorem 5.12 that

$$x = \frac{p_\ell y + p_{\ell-1}}{q_\ell y + q_{\ell-1}}$$

is also a quadratic irrational.

Assume then that x is a quadratic irrational. Define the sequence $x_0 := x$, $x_n := \frac{1}{x_{n-1} - \lfloor x_{n-1} \rfloor}$. It is enough to show that there exists n such that x_n is a reduced quadratic irrational. Notice that for $n \geq 1$ we have $x_n > 1$, so we just have to show that for some large enough n we have $-1 < \overline{x_n} < 0$. Now for every n we have

$$x = \frac{p_{n-1} x_n + p_{n-2}}{q_{n-1} x_n + q_{n-2}},$$

where $p_{n-1}, p_{n-2}, q_{n-1}, q_{n-2}$ are integers. Thus by Theorem 5.12 we also have

$$\overline{x} = \frac{p_{n-1}\overline{x_n} + p_{n-2}}{q_{n-1}\overline{x_n} + q_{n-2}},$$

which we can solve for $\overline{x_n}$ to get

$$\overline{x_n} = -\frac{q_{n-2}}{q_{n-1}} \left(\frac{\overline{x} - \frac{p_{n-2}}{q_{n-2}}}{\overline{x} - \frac{p_{n-1}}{q_{n-1}}} \right).$$

Notice that the fraction in the parentheses has limit 1 since it tends to $\frac{\overline{x}-x}{\overline{x}-x}$. By the way the continued fractions converge it is alternately greater and less than 1 when n is large enough. Choose some n for which it is less than 1. Because the factor $\frac{q_{n-2}}{q_{n-1}}$ is always less than 1, we have that $-1 < \overline{x_n} < 0$, which finishes the proof. □

Let us close this section with a short look at the structure of the continued fraction expansions of numbers of the form $x = \sqrt{D}$. The conjugate of x is simply $\overline{x} = -\sqrt{D}$, so it is not a reduced quadratic irrational. However the number $y = \lfloor\sqrt{D}\rfloor + \sqrt{D}$ is reduced and thus we can write the continued fraction expansion of y as $[\overline{b_0, b_1, ..., b_m}]$. Now we clearly have

$$x = [b_0 - \lfloor\sqrt{D}\rfloor, b_1, ..., b_m, b_0, \overline{b_1, ..., b_m, b_0}]$$

and $b_0 = 2\lfloor\sqrt{D}\rfloor$, so the continued fraction expansion of x is of the form

$$x = [a_0, \overline{a_1, ..., a_m, 2a_0}].$$

In fact a little bit more can be said.

Theorem 5.15 The continued fraction expansion of $\sqrt{D}$ is of the form

$$\sqrt{D} = [a_0, \overline{a_1, ..., a_m, 2a_0}],$$

where $a_1, ..., a_m$ is a palindrome.

Proof. Only the fact that $a_1, ..., a_m$ is a palindrome is new. Let us use the same notation as above. In the proof of Theorem 5.14 we saw that the continued fraction $[\overline{b_m, b_{m-1}, ..., b_0}]$ corresponds to the number

$\frac{-1}{\overline{y}} = \frac{1}{\sqrt{D}-\lfloor\sqrt{D}\rfloor}$, which is simply $x_1 := \frac{1}{x-\lfloor x\rfloor}$, so it has the continued fraction $[\overline{a_1, ..., a_m, 2a_0}]$. Thus we see that $a_1 = b_m = a_m, ..., a_m = b_1 = a_1$. □

6 PELL'S EQUATION

The diophantine equation

$$x^2 - Dy^2 = 1 \tag{6.1}$$

where $D \geq 2$ is a non-square integer is called **Pell's equation**.

It can be shown that there are infinitely many solutions to (6.1), all of which can be obtained from the so called *fundamental solution*. The fundamental solution in turn can be found by looking at the continued fraction expansion of $\sqrt{D}$. In general the theory of Pell's equation is largely about quadratic irrationals and their continued fractions.

6.1 RINGS $\mathbf{Z}[\sqrt{D}]$

Let $D \geq 2$ be a non-square integer. Then the numbers of the form $a + b\sqrt{D}$ $(a, b \in \mathbf{Z})$ form a ring that we denote by $\mathbf{Z}[\sqrt{D}]$. This ring has a natural norm $N\colon \mathbf{Z}[\sqrt{D}] \to \mathbf{Z}$ given by

$$N(a + b\sqrt{D}) = a^2 - b^2 D.$$

As was the case for the norm of the Gaussian integers defined in Chapter 4, one can easily check that $N(rs) = N(r)N(s)$ for all $r, s \in \mathbf{Z}[\sqrt{D}]$. It is also easy to check that a number $r \in \mathbf{Z}[\sqrt{D}]$ is a unit if and only if $N(r) = \pm 1$.

These observations show that solving (6.1) boils down to finding the group of units in $\mathbf{Z}[\sqrt{D}]$ and among those units the ones that have norm 1. We say that a unit $a + b\sqrt{D}$ is **non-trivial** if $b \neq 0$ and **positive** if $a > 0$ and $b > 0$. Notice that all non-trivial units can be obtained from the positive units by changing the signs of a and b.

Theorem 6.1 All positive units in $\mathbf{Z}[\sqrt{D}]$ are of the form $p_n + q_n\sqrt{D}$, where p_n and q_n are the numerator and denominator of some convergent $\frac{p_n}{q_n}$ of the continued fraction expansion of $\sqrt{D}$.

Proof. Let $a + b\sqrt{D}$ be a positive unit in $\mathbf{Z}[\sqrt{D}]$. Then we have

$$|a - b\sqrt{D}||a + b\sqrt{D}| = |a^2 - b^2 D| = 1,$$

and in particular

$$|a - b\sqrt{D}| = \frac{1}{a + b\sqrt{D}} < \frac{1}{2b}$$

because $a > b$. The claim now follows from Theorem **5.11**. □

6.2 CALCULATING THE CF-EXPANSION OF $\sqrt{D}$

In this section we will present an algorithm for calculating the continued fraction expansion of $\sqrt{D}$. Besides giving us a way to eventually code a solver for the Pell's equation, the algorithm will also allow us to deduce a few extra properties about the continued fraction expansion.

In general the algorithm generates the continued fraction expansion of a quadratic irrational of the form $\frac{P+\sqrt{D}}{Q}$, where P, Q and D are integers, $D > 0$ is a non-square and Q divides $P^2 - D$. The number D will stay constant during the algorithm. Initialize

$$x_0 := \frac{P + \sqrt{D}}{Q}, \quad a_0 := \lfloor x_0 \rfloor, \quad P_0 := P, \quad Q_0 := Q.$$

Define the subsequent numbers by

$$\begin{aligned} x_n &:= \frac{1}{x_{n-1} - a_{n-1}}, \\ a_n &:= \lfloor x_n \rfloor, \\ P_n &:= Q_{n-1}a_{n-1} - P_{n-1}, \\ Q_n &:= \frac{D - P_n^2}{Q_{n-1}}. \end{aligned}$$

By definition we have $x = [a_0, a_1, ...]$. It is easy to check by using induction that the following invariants hold:

- $x_n = \frac{P_n+\sqrt{D}}{Q_n}$,
- P_n and Q_n are integers, and

– Q_n divides $P_n^2 - D$.

Assume from now on that $x = \sqrt{D}$ and that we have $P_0 = 0$, $Q_0 = 1$. We showed at the end of Section 5.3 that x_n is a reduced quadratic irrational when $n \geq 1$. In particular it follows that $x_n - \overline{x_n} > 1$, so $Q_n > 0$. This implies that a_n can be calculated using integer arithmetic by

$$a_n = \left\lfloor \frac{P_n + \sqrt{D}}{Q_n} \right\rfloor = \left\lfloor \frac{P_n + \lfloor \sqrt{D} \rfloor}{Q_n} \right\rfloor.$$

The following theorem gives an important connection between the algorithm and Pell's equation.

Theorem 6.2 For all $n \geq 0$ we have

$$p_n^2 - Dq_n^2 = (-1)^{n+1} Q_{n+1}.$$

Proof. We have

$$\begin{aligned}\sqrt{D} &= \frac{p_n x_{n+1} + p_{n-1}}{q_n x_{n+1} + q_{n-1}} \\ &= \frac{\frac{P_{n+1}+\sqrt{D}}{Q_{n+1}} p_n + p_{n-1}}{\frac{P_{n+1}+\sqrt{D}}{Q_{n+1}} q_n + q_{n-1}} \\ &= \frac{P_{n+1} p_n + \sqrt{D} p_n + Q_{n+1} p_{n-1}}{P_{n+1} q_n + \sqrt{D} q_n + Q_{n+1} q_{n-1}}.\end{aligned}$$

This implies that

$$\sqrt{D}(P_{n+1} q_n + Q_{n+1} q_{n-1}) + Dq_n = \sqrt{D} p_n + P_{n+1} p_n + Q_{n+1} p_{n-1}.$$

Equating coefficients on both sides gives

$$\begin{aligned}P_{n+1} q_n + Q_{n+1} q_{n-1} &= p_n \\ P_{n+1} p_n + Q_{n+1} p_{n-1} &= Dq_n\end{aligned}$$

Multiplying the first equation by p_n and the second one by q_n and subtracting gives us

$$p_n^2 - Dq_n^2 = Q_{n+1}(q_{n-1}p_n - p_{n-1}q_n) = (-1)^{n+1}Q_{n+1}.$$

□

Notice that in particular $p_n^2 - Dq_n^2 = \pm 1$ if and only if $Q_{n+1} = 1$. On the other hand if $Q_{n+1} = 1$, then $x_{n+1} = P_{n+1} + \sqrt{D}$, and since x_{n+1} is reduced, we must have $P_{n+1} = \lfloor\sqrt{D}\rfloor$ and $x_{n+1} = \lfloor\sqrt{D}\rfloor + \sqrt{D}$. In particular $n+1$ must be at the end of the period of the continued fraction. The converse also holds obviously.

We are ready to prove the following.

Theorem 6.3 The units in $\mathbf{Z}[\sqrt{D}]$ are generated by a single positive **fundamental unit** $p_m + q_m\sqrt{D}$ that is obtained from the mth convergent of the continued fraction expansion $[a_0, \overline{a_1, ..., a_m, 2a_0}]$ of $\sqrt{D}$.

Proof. We want to show that any unit is of the form $\pm(p_m + q_m\sqrt{D})^k$, where $k \in \mathbf{Z}$. It is enough to show that all the positive units are of the form $(p_m + q_m\sqrt{D})^k$ with $k \geq 1$.

Assume that $p + q\sqrt{D}$ is a positive unit that is not of the form $(p_m + q_m\sqrt{D})^k$. Then there exists $k \geq 0$ such that

$$(p_m + q_m\sqrt{D})^k < p + q\sqrt{D} < (p_m + q_m\sqrt{D})^{k+1}.$$

Multiplying by $(p_m - q_m\sqrt{D})^k = \frac{1}{(p_m+q_m\sqrt{D})^k}$ it follows that

$$1 < (p + q\sqrt{D})(p_m - q_m\sqrt{D})^k < p_m + q_m\sqrt{D}.$$

Let $a + b\sqrt{D} := (p + q\sqrt{D})(p_m - q_m\sqrt{D})^k$. It is clearly a unit. Moreover we have $a + b\sqrt{D} > 1$, so the inverse satisfies $0 < a - b\sqrt{D} < 1$. Adding these two inequalities gives us $2a > 1$, which implies that $a \geq 1$. Similarly $b > \frac{a-1}{\sqrt{D}} \geq 0$, so $b \geq 1$. Thus $a + b\sqrt{D}$ is a positive unit, strictly smaller than $p_m + q_m\sqrt{D}$, which is a contradiction. □

As a final point before we present a piece of code for finding the fundamental unit we note the following.

Theorem 6.4 The first time $2a_0 = 2\lfloor\sqrt{D}\rfloor$ appears as a coefficient of the continued fraction expansion of $\sqrt{D}$ marks the end of the period of $\sqrt{D} = [a_0, \overline{a_1, ..., a_m, 2a_0}]$.

Proof. Because $\frac{P+\sqrt{D}}{Q}$ is reduced, we have that $P < \sqrt{D}$. Thus $\frac{P+\sqrt{D}}{Q} \leq \lfloor\sqrt{D}\rfloor + \sqrt{D}$, which gives us

$$a_0 = \left\lfloor \frac{P+\sqrt{D}}{Q} \right\rfloor \leq 2\lfloor\sqrt{D}\rfloor$$

with equality if and only if $P = \lfloor\sqrt{D}\rfloor$ and $Q = 1$. We know that $x_{m+1} = \lfloor\sqrt{D}\rfloor + \sqrt{D}$ ends the period so we are done. □

Algorithm 6.5 The following algorithm finds the fundamental unit in $\mathbf{Z}[\sqrt{D}]$, where we assume that D is a non-square.

```
// Returns a fundamental unit for Z[sqrt(D)], i.e.
// the smallest positive pair (p, q) such that
// p^2 - D q^2 = +- 1
pair<int64_t,int64_t> fundamental_unit_of_Z_sqrt_D(int64_t D) {
    int64_t a0=isqrt(D); // floor(sqrt(D))
    int64_t a=a0;
    int64_t P=0, Q=1;
    int64_t pn1=1, qn1=0;
    int64_t pn=a0, qn=1;
    while(a != 2*a0) {
        P=Q*a - P;
        Q=(D - P*P)/Q;
        a=(P + a0)/Q;
        int64_t tmp=pn;
        pn=pn*a + pn1; pn1=tmp;
        tmp=qn;
        qn=qn*a + qn1; qn1=tmp;
    }
    return make_pair(pn1, qn1);
}
```

7

ARITHMETIC FUNCTIONS

An **arithmetic function** is simply a function $\mathbf{Z}^+ \to \mathbf{C}$, where $\mathbf{C}$ is the set of complex numbers. An arithmetic function f is called **multiplicative** if

– $f(1) = 1$, and

– $f(mn) = f(m)f(n)$ whenever m and n are coprime.

The function f is called **completely multiplicative** if the second condition holds for all m and n.

7.1 DIRICHLET RING

If f and g are two arithmetic functions, we may define their **convolution** by

$$(f * g)(n) = \sum_{d|n} f(d)g(n/d).$$

One easily checks that convolution is associative, commutative, and distributes over the usual pointwise addition of functions. Moreover, one can notice that the function

$$\delta(n) = \begin{cases} 1, & \text{if } n = 1, \\ 0, & \text{otherwise} \end{cases}$$

acts as a multiplicative identity for the convolution operation. This means that the arithmetic functions form a ring, the so called **Dirichlet ring**.

The units in Dirichlet ring are the functions for which $f(1) \neq 0$. Indeed, if f is such an function, then one can check that its inverse can be defined recursively by

$$f^{-1}(1) := \frac{1}{f(1)}, \quad f^{-1}(n) := \frac{-1}{f(1)} \sum_{d|n, d<n} f(n/d) f^{-1}(d).$$

Every multiplicative function is clearly a unit in the Dirichlet ring because $f(1) = 1$ by definition. In fact more holds.

Theorem 7.1 The multiplicative functions are a subgroup of the unit group of the Dirichlet ring.

Proof. Exercise. □

To be able to conveniently work in the Dirichlet ring, it is handy to know some arithmetic functions and their relations. Here is a collection of a few common ones.

function	**inverse**	**definition**
δ	δ	1 when $n = 1$, 0 otherwise
$\mathbf{1}$	μ	identically 1 for all $n \geq 1$
ω	–	the number of distinct prime factors of n
μ	$\mathbf{1}$	$(-1)^{\omega(n)}$ if n is squarefree, 0 otherwise
Id	$\mathrm{Id}\,\mu$	identity function, n for all $n \geq 1$
Id_k	$\mathrm{Id}_k\,\mu$	the kth power function $n \mapsto n^k$
σ_k	$\mu * (\mathrm{Id}_k\,\mu)$	the sum of kth powers of divisors of n
φ	$\mathbf{1} * (\mathrm{Id}\,\mu)$	the order of $(\mathbf{Z}/n\mathbf{Z})^*$

Table 7.1 Arithmetic functions

The reader is encouraged to try to prove at least a few of the function–inverse function pairings.

7.2 MÖBIUS INVERSION

The basic version of the *Möbius inversion formula* is simply the following: If $g = f * \alpha$ where α is invertible, then $f = g * \alpha^{-1}$. Or more explicitly: If

$$g(n) = \sum_{d|n} \alpha(n/d) f(d),$$

then

$$f(n) = \sum_{d|n} \alpha^{-1}(d) g(n/d).$$

This is most often used in the case that $\alpha = \mathbf{1}$ and $\alpha^{-1} = \mu$ when we are interested in f but g is easier to calculate.

Another version of this is the following: If α is an invertible arithmetic function, then

$$g(n) = \sum_{x=1}^{n} \alpha(x) f(\lfloor n/x \rfloor),$$

implies

$$f(n) = \sum_{x=1}^{n} \alpha^{-1}(x) g(\lfloor n/x \rfloor).$$

The proof of this is left as an exercise. Again the typical case is $\alpha = \mathbf{1}$, $\alpha^{-1} = \mu$, for which we present the following algorithm.

Algorithm 7.2 (*Fast Möbius inversion*) Suppose that we want to calculate $f(n)$, and that we know how to calculate

$$g(n) := \sum_{x=1}^{n} f(\lfloor n/x \rfloor),$$

in $O(1)$ time. A simple $O(n)$ algorithm to calculate $f(n)$ would then be to calculate first $\mu(x)$ for $1 \le x \le n$ and then use the Möbius inversion formula to calculate $f(n) = \sum_{x=1}^{n} \mu(x) g(\lfloor n/x \rfloor)$.

We can however avoid calculating the values of the Möbius function explicitly, and ultimately do better by noting that

$$\begin{aligned} f(n) &= g(n) - \sum_{x=2}^{n} f(\lfloor n/x \rfloor) \\ &= g(n) - \sum_{x=3}^{n} f(\lfloor n/x \rfloor) - g(\lfloor n/2 \rfloor) + \sum_{x=2}^{\lfloor n/2 \rfloor} f(\lfloor n/(2x) \rfloor) \\ &= g(n) - g(\lfloor n/2 \rfloor) - \sum_{3 \le x \le n, x \text{ odd}} f(\lfloor n/x \rfloor). \end{aligned}$$

Because there are only about $2\sqrt{n}$ different numbers $\lfloor n/x \rfloor$, we can calculate $f(\lfloor n/x \rfloor)$ recursively for all x in $O(n^{3/4})$ time.

```
// Calculates f(n) = (g * mu)(n)
int64_t moebius_inversion(int64_t (*g)(int64_t),
                          const int64_t n) {
    const int64_t sqrtn = integer_sqrt(n);
    // low[k] = f(k)
```

```
        vector<int64_t> low(sqrtn+1, 0);
        // high[k] = f(floor(n/(2*k + 1)))
        vector<int64_t> high(sqrtn/2+1, 0);

        for(int64_t m=0;m<=sqrtn;m++) {
            low[m] = g(m) - g(m/2);
            int64_t x = 3;
            while(x <= m) {
                int64_t nextx = m/(m/x) + 1;
                if(nextx%2 == 0) nextx++;
                low[m] -= (nextx - x)/2 * low[m/x];
                x = nextx;
            }
        }
        for(int64_t i=sqrtn/2;i>=0;i--) {
            int64_t denom = 2*i + 1;
            int64_t m = n/denom;
            high[i] = g(m) - g(m/2);
            int64_t x = 3;
            while(x <= m) {
                int64_t nextx = m/(m/x) + 1;
                if(nextx%2 == 0) nextx++;
                if(m/x <= sqrtn) {
                    high[i] -= (nextx - x)/2 * low[m/x];
                } else {
                    high[i] -= (nextx - x)/2 * high[denom*x/2];
                }
                x=nextx;
            }
        }
        return high[0];
    }
```

It should be noted that same kind of tricks (noting that there are about $2\sqrt{n}$ different numbers $\lfloor n/k \rfloor$ etc.) work for calculating many sums that are not strictly speaking Möbius inversions.

Let us denote $S_f(n) = \sum_{x=1}^{n} f(x)$ for any arithmetic function f. Notice that if $f = \alpha * g$, then we have

$$S_f(n) = \sum_{k=1}^{n} \alpha(k) S_g(\lfloor n/k \rfloor).$$

This can be used to calculate many summatory functions of arithmetic functions for which we can find well-behaved α and g.

Example 7.3 Recall that the Euler totient function φ satisfies $\varphi = \mathrm{Id} * \mu$. Therefore

$$S_\varphi(n) = \sum_{k=1}^{n} \mu(k) S_{\mathrm{Id}}(\lfloor n/k \rfloor) = \sum_{k=1}^{n} \mu(k) \frac{\left\lfloor \frac{n}{k} \right\rfloor \left(\left\lfloor \frac{n}{k} \right\rfloor + 1 \right)}{2},$$

so $S_\varphi(n) = \sum_{k=1}^{n} \varphi(k)$ can be calculated in $O(n^{3/4})$ time by using the fast Möbius Inversion algorithm given above with the function $g(n) = \frac{n(n+1)}{2}$.

PART II

ALGEBRA

In this second part we will discuss additional topics in algebra.

1 GROUPS

One way to think about group theory is to say that it is the study of invertible transformations, or *symmetries* and how they compose to form new symmetries. For example the rotations of the plane form a *group*, and combining a 30° rotation with a 45° rotation yields a 75° rotation.

Symmetries always have an inverse transform; in the above case a −30° (or 330°) rotation would cancel the 30° rotation. The trivial symmetry that does not change anything is called the identity element of the group.

See `https://en.wikipedia.org/wiki/Rubik's_Cube_group` *for more information on this specific group.*

Another example of a group could be given by all the possible transformations on the Rubik's Cube. The cube has 6 sides, each of which can be turned by 90°. The different combinations of these turns form a group with $2^{27} \cdot 3^{14} \cdot 5^3 \cdot 7^2 \cdot 11$ distinct transformations, which is also the number of possible states of the Rubik's Cube.

1.1 BASIC DEFINITIONS

The definition of an abstract group is as follows.

Definition 1.1 A set G together with a binary operation $\circ$ is a **group** if the following three properties hold:

associativity — for all $a, b, c \in G$, $a \circ (b \circ c) = (a \circ b) \circ c$

identity element — there exists an element $e \in G$ such that $e \circ g = g \circ e = g$ for all $g \in G$

inverses — for all $g \in G$ there exists an element $g^{-1} \in G$ such that $g \circ g^{-1} = g^{-1} \circ g = e$

A group $(G, \circ)$ is called **commutative** or **abelian** if $\circ$ is commutative, i.e. $g \circ h = h \circ g$ for all $g, h \in G$.

We will from now on stop using the symbol $\circ$ explicitly and just write gh in place of $g \circ h$. If the group is commutative, we can also write the group operation additively $(g + h)$. In the multiplicative notation it is often natural to use 1 to denote the identity e and similarly we usually use 0 when we work with the additive notation.

The reader is probably already familiar with many abelian groups. For example each of $\mathbf{Z}, \mathbf{Q}, \mathbf{R}, \mathbf{C}$ are groups under addition. Moreover in each of the last three all non-zero elements form a group under multiplication. All of these examples are actually special cases of the following: In Part I we defined commutative rings, and each such ring hosts two important groups. First the ring itself is a group under addition, and secondly the units of the ring form a group under multiplication. In particular for any $m \neq 0$ the additive group of the ring $\mathbf{Z}/m\mathbf{Z}$ is an example of a finite abelian group.

Example 1.2 The **dihedral group** D_n is the group of rotational and reflectional symmetries of a regular polygon with n sides. We have illustrated the results of applying the symmetries in D_6 to a colored hexagon in Figure 1.1. The first row of hexagons are simply the 6 possible rotations, and the second row is obtained by reflecting the hexagons on the first row about a vertical line.

There are two competing notations for dihedral groups: The alternative would denote the group D_n by D_{2n} since there are $2n$ elements.

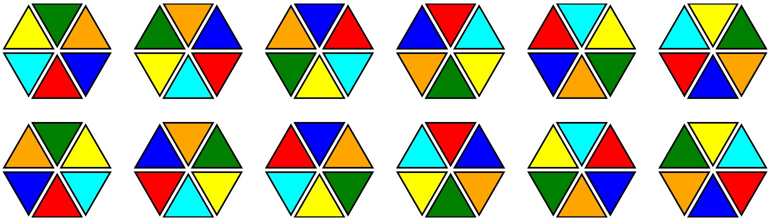

Figure 1.1 The action of D_6 on a colored hexagon.

In general the elements of D_n can be given as $1, r, r^2, ..., r^{n-1}$ and $s, sr, sr^2, ..., sr^{n-1}$, where r is a rotation by τ/n radians and s is a fixed reflection. The multiplication in D_n is then uniquely defined by the rules $s^2 = 1$, $r^n = 1$ and $srs = r^{-1}$.

Here $\tau := 2\pi$ is the perimeter of the unit circle.

Definition 1.3 Let G be a group. Then $H \subset G$ is a **subgroup** of G if H is closed under the group operation and taking inverses.

For example in the case of D_6 the sets $\{1, r, r^2, ..., r^5\}$, $\{1, s\}$ and $\{1, sr\}$ are subgroups – a patient reader can try listing all of them.

Definition 1.4 Let H be a subgroup of G. Then the sets of the form

$$gH := \{gh : h \in H\}$$

where $g \in G$ are called **left cosets** of H. Similarly the sets of the form Hg are called **right cosets**.

It is easy to check that the distinct left cosets gH are disjoint and partition G (same obviously holds for right cosets). In fact they correspond to the equivalence classes of the equivalence relation $a \sim_H b$ which we can define by

$$a \sim_H b \Leftrightarrow aH = bH \Leftrightarrow a^{-1}b \in H.$$

The set of left cosets of H in G is denoted by G/H and the set of right cosets is denoted by $H\backslash G$.

For example the left cosets of $H = \{1, sr\}$ in D_6 are

- $H = srH = \{1, sr\}$,
- $rH = sH = \{r, s\}$,
- $r^2H = sr^5H = \{r^2, sr^5\}$,
- $r^3H = sr^4H = \{r^3, sr^4\}$,
- $r^4H = sr^3H = \{r^4, sr^3\}$, and
- $r^5H = sr^2H = \{r^5, sr^2\}$.

One can notice that all of the cosets have $|H|$ elements. This is clear because $ah = ah'$ if and only if $h = h'$. Together with the fact that the cosets partition G we have shown *Lagrange's theorem.*

Theorem 1.5 Let G be a finite group and $H \subset G$ a subgroup. Then $|H|$ divides $|G|$.

In general the left and right cosets of a subgroup can differ. For example the right cosets of $\{1, sr\}$ in D_6 are $\{1, sr\}$, $\{r, sr^2\}$, $\{r^2, sr^3\}$, $\{r^3, sr^4\}$, $\{r^4, sr^5\}$, $\{r^5, s\}$, which are not the same as the left cosets we listed above. This motivates the following definition.

Definition 1.6 A subgroup $H \subset G$ for which $gH = Hg$ for every $g \in G$ is called a **normal subgroup** of G.

Notice that in the case when G is commutative, every subgroup is normal. Normal subgroups are important because they can be used to define quotient groups.

Definition 1.7 Let H be a normal subgroup of G. Then G/H can be made into a group by defining $gH \circ g'H = (gg')H$. The inverse of gH is thus $g^{-1}H$ and the identity is H. Such groups are called **quotient groups**.

The reader should check that the definition makes sense, i.e. that the multiplication is well-defined and that the obtained object actually is a group.

The subgroup $H = \{1, r^2, r^4\}$ is normal in D_6 with cosets $H = \{1, r^2, r^4\}$, $sH = \{s, sr^2, sr^4\}$, $rH = \{r, r^3, r^5\}$ and $srH = \{sr, sr^3, sr^5\}$. Thus D_6/H is a 4-element quotient group of D_6 whose multiplication table is listed in Table 1.1.

	H	sH	rH	srH
H	H	sH	rH	srH
sH	sH	H	srH	rH
rH	rH	srH	H	sH
srH	srH	rH	sH	H

Table 1.1 Multiplication table of $D_6/\{1, r^2, r^4\}$

Definition 1.8 Let G and G' be two groups. A map $f\colon G \to G'$ is called a **group homomorphism** if $f(gg') = f(g)f(g')$ for all $g, g' \in G$.

Notice that it follows that $f(1) = 1$ and $f(g^{-1}) = f(g)^{-1}$. One can also easily check that the **kernel** of f, i.e. the set

$$\operatorname{Ker} f := \{g \in G : f(g) = 1\},$$

is a normal subgroup of G and that f is injective if and only if $\operatorname{Ker} f = \{1\}$.

A bijective group homomorphism $f\colon G \to G'$ is called a **group isomorphism**. In this case the inverse map $f^{-1}\colon G' \to G$ is also a group isomorphism.

Two groups G and G' are said to be **isomorphic** if there exists a group isomorphism between them. In this case the two structures are group theoretically identical and we write $G \cong G'$. For example the quotient group $D_6/\{1, r^2, r^4\}$ above is isomorphic with the so called *Klein four-group* (denoted by V_4) that consists of 4 elements $1, a, b, c$ such that $a^2 = b^2 = c^2 = 1$, $ab = ba = c$, $bc = cb = a$ and $ca = ac = b$.

Given a normal subgroup H of G there exists a canonical surjective group homomorphism $\pi_H \colon G \to G/H$ given by $\pi_H(g) = gH$. The next theorem is a fundamental factorization result for group homomorphisms.

Theorem 1.9 Let $f \colon G \to G'$ be a group homomorphism. Then there exists a group isomorphism $\tilde{f} \colon G/\mathrm{Ker}\, f \to \mathrm{Im}\, f$ such that $f = \tilde{f} \circ \pi_{\mathrm{Ker}\, f}$.

Proof. Since $\pi_{\mathrm{Ker}\, f}$ is surjective, we may define $\tilde{f}$ by setting

$$\tilde{f}(\pi_{\mathrm{Ker}\, f}(x)) = f(x)$$

provided that we can show that whenever $\pi_{\mathrm{Ker}\, f}(x) = \pi_{\mathrm{Ker}\, f}(y)$, we have $f(x) = f(y)$. But this is clear because if x and y belong to the same coset, we have $x = yk$ for some $k \in \mathrm{Ker}\, f$, so that $f(x) = f(yk) = f(y)f(k) = f(y)$.

It remains to show that $\tilde{f}$ is injective. Now if $\tilde{f}(\pi_{\mathrm{Ker}\, f}(x)) = f(x) = 1$, then $x \in \mathrm{Ker}\, f$, so $\pi_{\mathrm{Ker}\, f}(x) = \mathrm{Ker}\, f$, which means that the kernel of $\tilde{f}$ consists of the single coset $\mathrm{Ker}\, f$. Thus $\tilde{f}$ is injective. □

For example the group homomorphism $f \colon D_6 \to V_4$ given by $f(s) = a$, $f(r) = b$ and extending uniquely to the other elements of D_6 has as its kernel $\{1, r^2, r^4\}$, and therefore $D_6/\{1, r^2, r^4\} \cong V_4$.

A group G is called **cyclic** if there exists an element $g \in G$ such that every element of g can be written as g^a for some $a \in \mathbf{Z}$. The element g is called a **generator** of G. Finite cyclic groups with n elements are denoted by C_n. The group C_n is of course isomorphic to the additive group of $\mathbf{Z}/n\mathbf{Z}$, but we will write the operation in C_n multiplicatively. Thus if we want to be concrete, we can define C_n to consist of the nth roots of unity in $\mathbf{C}$, that is the complex numbers z such that $z^n = 1$. Thus C_n is a cyclic subgroup of the multiplicative group of $\mathbf{C}$. Let us lastly note that every infinite cyclic group is clearly isomorphic to the additive group of $\mathbf{Z}$.

Notice that if G is any group and $g \in G$, then g generates a cyclic subgroup of G. If the subgroup is finite, the size of this subgroup is called the order of g and written $\operatorname{ord}(g)$. This gives us the following useful result.

Theorem 1.10 Let G be a finite group and $g \in G$. Then $g^{|G|} = 1$.

Proof. Since G is finite, $\operatorname{ord}(g) < \infty$, and by Theorem 1.5 we see that $\operatorname{ord}(g)$ divides $|G|$. Moreover $g^{\operatorname{ord}(g)} = 1$, since g is a generator of a cyclic group of order $\operatorname{ord}(g)$. Thus $g^{|G|} = (g^{\operatorname{ord}(g)})^{|G|/\operatorname{ord}(g)} = 1$.

□

We will end this section by considering direct products of groups.

Definition 1.11 If G_1 and G_2 are two groups, we can form their **direct product** $G = G_1 \times G_2$ by considering all tuples (g_1, g_2) with $g_1 \in G_1$ and $g_2 \in G_2$ and defining the group operation componentwise, i.e. $(a_1, a_2) \circ (b_1, b_2) := (a_1 b_1, a_2 b_2)$.

As an example, it is easy to see that the Klein four-group is isomorphic with the direct product $C_2 \times C_2$.

We close this section with a result that is useful when we want to try to write a group as a direct product of simpler groups.

Theorem 1.12 Let G be a group and assume that N_1 and N_2 are two normal subgroups of G such that $N_1 \cap N_2 = \{e\}$. Then $N_1 N_2$ is a subgroup of G isomorphic to $N_1 \times N_2$.

Proof. Notice that if $n_1 \in N_1$ and $n_2 \in N_2$, then we have $n_1 n_2 n_1^{-1} n_2^{-1} \in N_1 \cap N_2$ because $n_1 n_2 n_1^{-1} \in N_2$ and $n_2 n_1^{-1} n_2^{-1} \in N_1$. Thus we see that $n_1 n_2 = n_2 n_1$.

Let us now show that $N_1 N_2$ is a subgroup. Clearly $e \in N_1 N_2$. If $n_1, n_1' \in N_1$ and $n_2, n_2' \in N_2$, then $(n_1 n_2)(n_1' n_2{}') = (n_1 n_1{}')(n_2 n_2{}') \in N_1 N_2$, so $N_1 N_2$ is closed under the group operation. Moreover $(n_1 n_2)(n_1^{-1} n_2^{-1}) = (n_1 n_1^{-1})(n_2 n_2^{-1}) = e$, so $N_1 N_2$ is closed under taking inverses, too.

We can define a map $f\colon N_1 \times N_2 \to N_1N_2$ by setting $f((n_1, n_2)) = n_1n_2$. It is a homomorphism because of the commutativity proven in the first paragraph. It is clearly a surjection and if $n_1n_2 = e$, then we must have $n_1 = e = n_2$, so it is also an injection and thus an isomorphism. □

1.2 GROUP ACTIONS

Definition 1.13 Let X be a set and G a group. A map $\cdot\colon G \times X \to X$ is a (left) G-action on X if the following two axioms hold:

compatibility – $g \cdot (g' \cdot x) = (gg') \cdot x$ for all $g, g' \in G$ and $x \in X$, and

identity – $e \cdot x = x$ for all $x \in X$.

Given $x \in X$, the **orbit** of x under a G-action is $Gx := \{g \cdot x : g \in G\}$. This is just all the points we can reach from x by permuting it with elements of G. Orbits form a partition of X and the set of orbits is denoted by X/G.

Given $x \in X$, the **stabilizer** of x under a G-action is the subgroup $G_x \subset G$ containing all the elements that fix x. In symbols,

$$G_x := \{g \in G : g \cdot x = x\}.$$

That this is indeed a subgroup follows from the compatibility axiom of the action.

Given $g \in G$, the set of fixed points of g in X is denoted by

$$X^g := \{x \in X : g \cdot x = x\}.$$

If G is a finite group and X is a finite set, then we have the following two useful theorems.

Theorem 1.14 *(Orbit–stabilizer theorem)* Let G be a finite group and X a finite set. Assume that G acts on X. Then for all $x \in X$ we have

$$|Gx| = |G/G_x| = |G|/|G_x|.$$

Proof. Clearly if a and b are in the same coset of the stabilizer G_x, we must have $a \cdot x = b \cdot x$. Thus we may define a map $\varphi\colon G/G_x \to X$ given by $\varphi(gG_x) = g \cdot x$.

To prove the claim it is enough to show that φ is an injection. Assume that $\varphi(aG_x) = \varphi(bG_x)$. Then $a \cdot x = b \cdot x$. Operating by b^{-1} on both sides we get $b^{-1} \cdot a \cdot x = x$, so that $b^{-1}a \in G_x$, which implies that $aG_x = bG_x$. □

Theorem 1.15 *(Burnside's lemma)* Let G be a finite group and X a finite set. Assume that G acts on X. Then the number of orbits is given by

$$|X/G| = \frac{1}{|G|}\sum_{g\in G}|X^g|.$$

Proof. By the orbit–stabilizer theorem we have

$$|X/G| = \sum_{x\in X}\frac{1}{|Gx|} = \sum_{x\in X}\frac{|G_x|}{|G|} = \frac{1}{|G|}\sum_{g\in G}|X^g|,$$

where the last equality follows because both sums count each pair (g, x) such that $g \cdot x = x$ exactly once. □

1.3 CONJUGACY CLASSES

Let G be a group. Two group elements $g, g' \in G$ are said to be **conjugate** if there exists $h \in G$ such that $hgh^{-1} = g'$. It is easy to check that this is an equivalence relation and therefore we can partition G into **conjugacy classes** where two elements belong to the same class if and only if they are conjugate. We denote the conjugacy class of $g \in G$ by $\mathrm{Cl}(g)$.

An orthogonal notion is that of a **centralizer** of an element g defined by

$$C(g) := \{x \in G : gx = xg\},$$

that is the set of elements in G that commute with g, or equivalently, the set of elements in G that fix g upon conjugation.

It is helpful to notice that we can let G act on itself by conjugation. That is we define $g \cdot x = gxg^{-1}$ for all $g, x \in G$. Then conjugacy classes are the orbits and centralizers are the stabilizers under this action.

From the orbit–stabilizer theorem we have that

$$|G/C(g)| = |\operatorname{Cl}(g)|.$$

In particular because the conjugacy classes are disjoint, we get the **class equation**

$$|G| = \sum_{i=1}^{m} |G/C(x_i)|,$$

where we have picked exactly one element x_i from each of the m conjugacy classes of G.

2 PERMUTATIONS AND ORDERED SETS

2.1 PERMUTATION GROUP S_n

Let X be a set. A map $\sigma\colon X \to X$ that is bijective is called a **permutation** of X. In this whole chapter we are only interested in the case where X is *finite*.

Without loss of generality we may assume that $X = \{1, ..., n\}$. It is easy to see that the permutations on X form a group with composition as the group operation and the identity map $\mathrm{Id}\colon X \to X$ as the identity element. Indeed, it is clear that if $\sigma, \tau\colon X \to X$ are two permutations, then also $\tau \circ \sigma$ is a permutation. Moreover $\mathrm{Id} \circ \sigma = \sigma = \sigma \circ \mathrm{Id}$ and $\sigma^{-1} \circ \sigma = \mathrm{Id} = \sigma \circ \sigma^{-1}$, so the identity element works as expected when we take the inverse of a permutation σ to be just its inverse map σ^{-1}. This group of permutations on X is denoted by S_n.

The simplest way to represent a permutation φ is to just list the images of $1, 2, ..., n$ under the map φ. For example if $n = 3$, all the possible permutations of X would be $(1, 2, 3)$, $(1, 3, 2)$, $(2, 1, 3)$, $(2, 3, 1)$, $(3, 1, 2)$ and $(3, 2, 1)$. From this representation it is clear that there are $n! = 1 \cdot 2 \cdot ... \cdot n$ permutations in total: We have n ways to choose the image of 1, $n - 1$ ways to choose the image of 2 after the image of 1 has been chosen, $n - 2$ ways to choose the image of 3 and so on.

A **cycle** is a permutation σ for which there exists an element $x \in X$ such that all the elements moved by σ are of the form $\sigma^k(x)$ for some $k \geq 0$. If the order of σ is k, we say that σ is a k-cycle. We can use the notation

$$(x\sigma(x)\cdots\sigma^{k-1}(x))$$

to write down σ. For example if $n = 4$, then (132) would represent a 3-cycle that maps 1 to 3, 3 to 2, 2 back to 1 and keeps 4 fixed. It is easy to see that this notation is unique up to cyclic reordering of the elements. The identity permutation is the only 1-cycle and 2-cycles are also called **transpositions**.

It is clear that disjoint cycles commute, and one may easily notice that any permutation σ can be written as a product of disjoint cycles, uniquely up to

– the order of the cycles,

– the cyclic reorderings of cycles themselves, and

– the presence of cycles of length 1.

For example if $n = 5$, the permutation $(4, 3, 5, 1, 2)$ can be written as $(14)(235)$. The **cycle type** of σ is the multiset of lengths of cycles appearing in the unique representation of σ, including the cycles of length 1. For example the cycle type of the permutation $(7)(14)(235)(698)$ would be $\{1, 2, 3, 3\}$. This can also be written as $(1, 1, 2, 0, 0, ...)$ where ith coordinate gives the number of i-cycles.

Notice that any permutation can be written as a product of at most $n - 1$ transpositions. Indeed, we can represent a cycle $(x_1x_2...x_k)$ as the product

$$(x_1x_k)(x_1x_{k-1})...(x_1x_3)(x_1x_2)$$

of $k - 1$ transpositions.

Theorem 2.1 Assume that $\tau_1...\tau_m = 1$, where τ_i are transpositions. Then m is even.

Proof. Let $a, b, c, d \in X$ be distinct. Then the following equations hold:

1 $(ab)(ab) = 1$,

2 $(cd)(ab) = (ab)(cd)$,

3 $(bc)(ab) = (ac)(bc)$,

4 $(ac)(ab) = (ab)(bc)$.

Notice that if we rewrite the product $\tau_1...\tau_m$ by replacing a pair $\tau_i\tau_{i+1}$ that matches the left hand side of one of (1),(2),(3) or (4) with the corresponding right hand side, the parity of m does not change.

Now pick an element $a \in X$ that appears in some of the transpositions in the product. By using the rules (1)–(4), we may move every such a to the left-most transposition. In the end all the as must disappear since the product equals the identity permutation and we cannot be left with a single a. If we repeat this for every element that appears in the product, we will have reduced ourselves to the case $1 = 1$ without changing the parity of the number of transpositions in the product. □

Notice that as an easy corollary, if we write any $\sigma \in S_n$ as a product of transpositions, then the number of terms in the product is constant modulo 2. In particular we get a well-defined homomorphism $\operatorname{sgn}: S_n \to C_2$ by setting $\operatorname{sgn}(\tau) = -1$ for any transposition τ and extending to products of transpositions in the natural way. The number $\operatorname{sgn}(\sigma)$ is called the **sign** of the permutation σ. The permutation σ is called **even** if $\operatorname{sgn}(\sigma) = 1$ and **odd** if $\operatorname{sgn}(\sigma) = -1$.

If $n = 0$ or $n = 1$ there are no transpositions, so the image of sgn *is the trivial group.*

If $n \geq 2$, the kernel of sgn is a subgroup of index 2. This subgroup contains all the even permutations and is called the **alternating group** of order n and denoted A_n.

From the decomposition of cycles into transpositions, it is trivial to see that the sign of a k-cycle is $(-1)^{k-1}$. Thus a permutation σ is odd if and only if there is an odd number of cycles of even length.

Algorithm 2.2 *(Sign of a permutation)* The following algorithm computes the sign of a permutation by going through the cycles and alterning the sign based on the length of the cycle. The permutation is assumed to be on the set $\{1, ..., n\}$, so `perm[0]` is ignored. The function modifies `perm` in the process to keep track which cycles have been counted, but in the end the vector should be the same as in the beginning.

```
int64_t sign_of_permutation(vector<int64_t> &perm) {
    int64_t sgn=1;
    for(int64_t i=1;i<perm.size();i++) {
        if(perm[i] < 0) {
            perm[i]=-perm[i];
            continue;
        }
        int64_t j=perm[i];
        while(j != i) {
            sgn=-sgn;
            perm[j]=-perm[j];
            j=-perm[j];
        }
    }
    return sgn;
}
```

The cycle structure of the permutation also tells the least number of transpositions needed to express the permutation. Indeed for a permutation $\sigma = \sigma_1 ... \sigma_k$ where σ_i are cycles of lengths ℓ_i respectively and $\ell_1 + ... + \ell_k = n$ we can write each cycle as a product of $\ell_i - 1$ transpositions, so in total we need $(\ell_1 - 1) + ... + (\ell_k - 1) = n - k$ transpositions. We leave it as an exercise to show that this is actually optimal.

2.2 Ordered sets and inversions

Definition 2.3 Let X be a set. A binary relation $\leq$ is called a **partial order** on X if

reflexivity – $x \leq x$ for all $x \in X$,

antisymmetry – $x \leq y$ and $y \leq x$ implies $x = y$ for all $x, y \in X$,

transitivity – $x \leq y$ and $y \leq z$ implies $x \leq z$ for all $x, y, z \in X$.

If $x \leq y$ or $y \leq x$ for two elements $x, y \in X$, we say that x and y are **comparable**, otherwise they are **incomparable**. If all elements are pairwise comparable, we say that $\leq$ is a **total order** and that X together with $\leq$ is a **totally ordered set**.

Assume now that X is a finite totally ordered set and $\sigma\colon X \to X$ is a permutation on X. A pair $(x, y) \in X \times X$ is called an inversion if $x < y$ and $\sigma(x) > \sigma(y)$. Using inversions we can get yet another characterization for the sign of a permutation.

Theorem 2.4 Let m be the number of inversions of σ. Then $\operatorname{sgn}(\sigma) = (-1)^m$.

Proof. Without loss of generality we can assume that $X = \{1, ..., n\}$. For any $\sigma \in S_n$ let

$$I(\sigma) := \{(x, y) \in X \times X : x < y, \sigma(x) > \sigma(y)\}$$

be the set of inversions of σ and similarly let

$$I^c(\sigma) := \{(x, y) \in X \times X : x < y, \sigma(x) < \sigma(y)\}$$

be the set of pairs (x, y), where σ preserves the order. Consider the map $f\colon S_n \to C_2$ given by $f(\sigma) = (-1)^{|I(\sigma)|}$. Notice that f agrees with sgn on transpositions, so to prove the claim, it is enough to show that $f(\tau\sigma) = f(\tau)f(\sigma)$ for all $\tau, \sigma \in S_n$. Now

$$\begin{aligned}
|I(\tau\sigma)| &= |\{(x, y) \in I(\sigma) : (\sigma(y), \sigma(x)) \in I^c(\tau)\}| + \\
&\quad |\{(x, y) \in I^c(\sigma) : (\sigma(x), \sigma(y)) \in I(\tau)\}| \\
&= |I(\sigma)| - |\{(x, y) \in I(\sigma) : (\sigma(y), \sigma(x)) \in I(\tau)\}| + \\
&\quad |\{(x, y) \in I^c(\sigma) : (\sigma(x), \sigma(y)) \in I(\tau)\}| \\
&\equiv |I(\sigma)| + |\{(x, y) \in I(\sigma) : (\sigma(y), \sigma(x)) \in I(\tau)\}| + \\
&\quad |\{(x, y) \in I^c(\sigma) : (\sigma(x), \sigma(y)) \in I(\tau)\}| \\
&\equiv |I(\sigma)| + |I(\tau)| \pmod 2,
\end{aligned}$$

which proves the claim. □

2.3 CONJUGATES AND COMMUTING ELEMENTS

Let $\tau \in S_n$ be a permutation and let

$$\tau = (\tau_{1,1}\tau_{1,2}...\tau_{1,i_1})...(\tau_{m,1}\tau_{m,2}...\tau_{m,i_m})$$

be its cycle decomposition. Now if $\sigma \in S_n$ is any permutation, it is straightforward to see that

$$\sigma\tau\sigma^{-1} = (\sigma(\tau_{1,1})\sigma(\tau_{1,2})...\sigma(\tau_{1,i_1}))...(\sigma(\tau_{m,1})\sigma(\tau_{m,2})...\sigma(\tau_{m,i_m})).$$

Thus we see how conjugation affects the cycle structure: We simply replace each entry $\tau_{j,i}$ by $\sigma(\tau_{j,i})$. From this observation the following theorem is immeadiate.

Theorem 2.5 The conjugacy classes of S_n correspond to the different cycle types.

One way to think about conjugation in S_n is that we are given some new labels for S_n, say $l_1, ..., l_n$. Our permutation τ works on the old labels $1, ..., n$ and σ maps the old labels i to their corresponding new labels l_i. The conjugation $\sigma\tau\sigma^{-1}$ first converts the new labels to the old ones, then performs τ, and finally represents the result by using the new labels.

Let us now find the centralizer of $\tau \in S_n$. Remember that $\sigma \in S_n$ commutes with τ if and only if $\sigma\tau\sigma^{-1} = \tau$. Let us compare the cycle representations of the left and right hand sides. We have

$$(\sigma(\tau_{1,1})\sigma(\tau_{1,2})...\sigma(\tau_{1,i_1}))...(\sigma(\tau_{m,1})\sigma(\tau_{m,2})...\sigma(\tau_{m,i_m})) =$$

$$(\tau_{1,1}\tau_{1,2}...\tau_{1,i_1})...(\tau_{m,1}\tau_{m,2}...\tau_{m,i_m}).$$

Thus two elements of S_n commute if and only if σ splits into m disjoint permutations $\sigma_1, ..., \sigma_m$ such that

– each σ_j fixes all elements other than $\tau_{j,1}, ..., \tau_{j,i_j}$,

– each σ_j maps $\tau_{j,1}, ..., \tau_{j,i_j}$ to some $\tau_{k,1}, ..., \tau_{k,i_k}$ with $i_j = i_k$, preserving the cyclic order.

In particular if the cycle type of τ is $(c_1, c_2, ..., c_n)$, where c_i is the number of i-cycles, then there are $c_1!c_2!...c_n!$ ways to choose which cycles σ maps to each other and $1^{c_1}2^{c_2}...n^{c_n}$ ways to choose how to map each cycle while preserving the cyclic order. Thus the number of elements in S_n that commute with τ is $c_1!c_2!...c_n!1^{c_1}2^{c_2}...n^{c_n}$.

Another way to see this would be to recall that $|G/C(\tau)| = |\operatorname{Cl}(\tau)|$ and use a simple counting argument to show that the conjugacy class has size

$$\frac{n!}{c_1!c_2!...c_n!1^{c_1}2^{c_2}...n^{c_n}}.$$

3 LINEAR RECURRENCES

3.1 BASIC THEORY

Let k be a field and V a finite dimensional k-vector space. We let V^∞ denote the space of all sequences $(a_n)_{n=0}^\infty$ with $a_n \in V$.

Definition 3.1 A sequence $(a_n)_{n=0}^\infty \in V^\infty$ is a **linear recursive sequence** if there exist constants $c_1, ..., c_m \in k$ ($m \in \mathbf{N}$) such that

$$a_n = c_1 a_{n-1} + ... + c_m a_{n-m}$$

for all $n \geq m$.

One quite slick way of getting hold of these sequences is via the shift operator $S\colon V^\infty \to V^\infty$ defined by setting $S(a_0, a_1, ...) = (a_1, a_2, ...)$. Then $a = (a_n)_{n=0}^\infty$ is a linear recursive sequence if and only if there exist constants $c_1, ..., c_m$ such that

$$(S^m - c_1 S^{m-1} - ... - c_{m-1} S - c_m)a = 0.$$

Let $p \in k[x]$ be a polynomial. We say that p is a **characteristic polynomial** of the sequence $(a_n)_{n=0}^\infty$ if $p(S)a = 0$. If $(a_n)_{n=0}^\infty$ is a linear recursive sequence as above, then $p(x) = x^m - c_1 x^{m-1} - ... - c_m$ is a characteristic polynomial of a, so every linear recursive sequence has a characteristic polynomial.

Theorem 3.2 The characteristic polynomials of a linear recursive sequence $(a_n)_{n=0}^\infty$ form an ideal of $k[x]$, which we will denote by I_a.

Proof. If p and q are characteristic polynomials of a, then $(p+q)(S)a = p(S)a + q(S)a = 0$, so $p+q$ is a characteristic polynomial of a. Similarly if $r \in k[x]$, then $(rp)(S)a = r(S)p(S)a = 0$, so rp is a characteristic polynomial of a. □

Now because $R[x]$ is a PID, there exists a unique monic polynomial μ_a generating I_a. We say that μ_a is the **minimal polynomial** of the sequence $(a_n)_{n=0}^\infty$. The minimal polynomial divides all other characteristic polynomials of a.

Theorem 3.3 Let V and W be finite dimensional k-vector spaces and assume that $T\colon V \to W$ is a linear map. If $a \in V^\infty$ is a linear recursive sequence, then $Ta = (Ta_n)_{n=0}^\infty \in W^\infty$ is also a linear recursive sequence.

Proof. This is clear because $S(Ta) = T(Sa)$. □

The above theorem has sort of a converse.

Theorem 3.4 Let V be a finite dimensional k-vector space and $a \in V^\infty$. Assume that for all linear $\varphi\colon V \to k$ the sequence $\varphi(a) = (\varphi(a_n))_{n=0}^\infty$ is linear recursive. Then a is linear recursive.

Proof. Pick a basis $v_1, ..., v_k$ for V and let $v_1^*, ..., v_k^*$ be the dual basis. Let $I_j = I_{v_j^*(a)}$ and consider the intersection $I = \bigcap_{j=1}^k I_j$. We may pick a $p \in I$ that is not zero and satisfies $p(S)v_j^*(a) = 0$ for all $1 \leq j \leq k$. Now clearly also $p(S)a = p(S)((v_1^*(a_n)v_1)_{n=0}^\infty + ... + (v_k^*(a_n)v_k)_{n=0}^\infty) = 0$. □

Remark 3.5 As the proof indicates, in the above theorem it is enough to check the condition for a dual basis.

Exercise 3.6 Show that sums of linear recursive sequences are linear recursive. (Hint: Use the above two theorems.)

We will next see our first characterization for linear recursive sequences. In a sense they are generated by matrices.

Theorem 3.7 Let V be a k-vector space and $A\colon V \to V$ a linear map. Then for any $v \in V$ the sequence $A^n v$ is linear recursive.

Proof. Let p be the characteristic polynomial of A. Then $p(A)(A^k v) = 0$ for all $k \geq 0$, which shows that $A^n v$ is linear recursive. □

Remark 3.8 Notice that the characteristic polynomial of A is also a characteristic polynomial of the sequence $A^n v$.

Theorem 3.9 Let a be a linear recursive sequence in some k-vector space V. Then there exists a k-vector space W, a linear map $W \to W$, $w \in W$ and a linear map $\pi\colon W \to V$ such that $a_n = \pi(A^n w)$ for all $n \geq 0$.

Proof. It is enough to show this in the case where $V = k$, since in the general case it is possible to proceed coordinate wise and embed the obtained vector spaces and mappings in a bigger space. Now if $(a_n)_{n=0}^{\infty} \in k^{\infty}$ is a linear recursive sequence, it satisfies an equation of the form $a_n = c_1 a_{n-1} + ... + c_m a_{n-m}$. Pick $W = k^m$ and consider the matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ c_m & c_{m-1} & c_{m-2} & c_{m-3} & \cdots & c_1 \end{pmatrix}$$

Clearly if $v_k = (a_k, a_{k+1}, ..., a_{k+m-1})^T$ for $k \geq 0$, then $Av_k = v_{k+1}$. We can therefore choose π to be the projection on the first coordinate. □

3.2 GENERATING FUNCTIONS

In this and the following sections of the rest of this chapter we will focus on the case where the linear recursive sequence lies in k. In the view of the first section this assumption is not very restrictive, since the linear recursive sequences in higher dimensional vector spaces are composed of one dimensional ones.

We will now look at the second characterization of linear recursive sequences via generating functions.

Definition 3.10 The generating function of a sequence $(a_n)_{n=0}^{\infty} \in k^{\infty}$ is the formal power series

$$a_0 + a_1 x + a_2 x^2 + ... = \sum_{n=0}^{\infty} a_n x^n \in k[[x]].$$

The main result of this section is the following.

Theorem 3.11 The generating function of a sequence $(a_n)_{n=0}^{\infty}$ is rational if and only if $(a_n)_{n=0}^{\infty}$ is linear recursive. If this is the case, then the generating function can be written as

$$a(x) = \frac{h(x)}{x^m p(1/x)} = \frac{h(x)}{1 - c_1 x - ... - c_m x^m},$$

where $p(x) = x^m - c_1 x^{m-1} - ... - c_m$ is a characteristic function of $(a_n)_{n=0}^{\infty}$ and

$$\begin{aligned} h(x) &= a_0 + (a_1 - c_1 a_0)x + (a_2 - c_1 a_1 - c_2 a_0)x^2 + ... \\ &\qquad + (a_{m-1} - c_1 a_{m-2} - ... - c_{m-1} a_0)x^{m-1} \\ &= \sum_{k=0}^{m-1} (a_k - \sum_{j=1}^{k} c_j a_{k-j}) x^k. \end{aligned}$$

Proof. Assume first that $\frac{h(x)}{q(x)}$ is a rational function with $q(0) \neq 0$. We may without loss of generality assume that q is of the form $q(x) = 1 - c_1 x - ... - c_m x^m$. Now let $h(x) = h_0 + h_1 x + ... + h_l x^l$. We may also assume that $h_0 \neq 0$, because adding zeros in front of a linear recursive sequence keeps it still linear recursive. Consider one step of the long division:

$$\frac{h(x)}{q(x)} = h_0 + x \frac{(h_1 + c_1 h_0) + (h_2 + c_2 h_0)x + ... + (h_l + c_l h_0)x^l}{1 - c_1 x - ... - c_m x^m},$$

where we set $c_j = 0$ for $j > m$. Thus the long division maybe modeled as a linear transformation on the space k^{l+1} which maps $(x_0, ..., x_l)$ to $(x_1 + c_1 x_0, x_2 + c_2 x_0, ..., x_k + c_l x_0)$. With the initial vector $(h_0, ..., h_l)$ and projection on the first coordinate this generates a linear recursive sequence corresponding to the coefficients of the power series.

Assume then that $(a_n)_{n=0}^{\infty}$ is a linear recursive sequence and set $a(x)$ and $h(x)$ as in the statement of the theorem. It is enough to check that

$$(1 - c_1 x - ... - c_m x^m)(a_0 + a_1 x + a_2 x^2 + ...) = h(x),$$

which is straightforward to do. □

Rational generating functions for linear recursive sequences are useful in many ways. First of all they make it easy to find recurrences for sums of sequences. Second, by writing the generating function in lowest terms one can find the minimal polynomial of the sequence.

3.3 COMPONENTS OF THE RECURSION AND CLOSED FORMULAS

Generating functions lead us to the next stage of analyzing linear recursive sequences. Consider the rational function $\frac{1}{1-c_1x-...-c_mx^m}$. We may factor the denominator as $f_1(x)^{\alpha_1}...f_l(x)^{\alpha_l}$, where $f_1, ..., f_l$ are pairwise coprime and irreducible and $\alpha_j \geq 1$. After this one can do the partial fraction decomposition and obtain

$$\frac{1}{1-c_1x-...-c_mx^m} = \frac{h_1(x)}{f_1(x)^{\alpha_1}} + ... + \frac{h_l(x)}{f_l(x)^{\alpha_l}}$$

for some polynomials $h_j(x)$. This makes it possible to split a given linear recursive sequence into its irreducible components. Now let us go further and move into the splitting field of the denominator. Then our partial fraction decomposition takes the simple form

$$\frac{1}{1-c_1x-...-c_mx^m} = \frac{h_1(x)}{(x-r_1)^{\alpha_1}} + ... + \frac{h_l(x)}{(x-r_l)^{\alpha_l}},$$

where $r_1, ..., r_l$ are the roots of the denominator with multiplicities $\alpha_1, ..., \alpha_l$. From this form it is easy to see the following:

Theorem 3.12 A sequence $(a_n)_{n=0}^{\infty}$ is linear recursive if and only if it can be written as

$$a_n = g_1(n)r_1^n + ... + g_l(n)r_l^n,$$

where $g_1, ..., g_l$ are polynomials and $r_1, ..., r_l$ are the roots of the minimal polynomial of a.

An interesting consequence is that if $(a_n)_{n=0}^{\infty}$ and $(b_n)_{n=0}^{\infty}$ are linear recursive sequences, then so is $(a_nb_n)_{n=0}^{\infty}$.

PART III

COMBINATORICS

The third part of the book will focus on counting and enumerating things. One of the main instruments in modern combinatorics are generating functions. We will look at them especially in the context of so called *combinatorial species*. This is a theoretical framework that makes it easy to combine simple combinatorial objects into more complicated ones by forming equational relationships between them. Information on the objects can then be distilled via their generating functions.

1 GENERATING FUNCTIONS

Generating functions are simply a way of encoding sequences of numbers that makes it easy to manipulate and analyze certain kinds of data. They play a significant role in many areas of mathematics such as combinatorics and number theory, and they are also a useful tool for solving recurrences and doing various other tasks.

1.1 FORMAL POWER SERIES

We start by introducing so called formal power series.

Definition 1.1 Let R be a fixed commutative ring. A **formal power series** over R is an expression of the form

$$\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + ...$$

where $a_n \in R$.

Here the symbols x^n carry no meaning of their own. They serve only as a way to separate the coefficients a_n, which for us will usually be integers or rational numbers. This power series is also called the **(ordinary) generating function** for the sequence a_n.

Addition and multiplication are defined for formal power series just as if they were analytic power series. Let $A(x)$ and $B(x)$ be two formal power series with coefficients a_n and b_n respectively. Then we define the formal power series $A(x) \pm B(x)$ by

$$A(x) \pm B(x) := \sum_{n=0}^{\infty} (a_n \pm b_n) x^n$$

and $A(x)B(x)$ by

$$A(x)B(x) := \sum_{n=0}^{\infty} \sum_{k=0}^{n} (a_k b_{n-k}) x^n.$$

It is easy to check that the multiplication is distributive and in fact the formal power series form a ring that will be denoted by $R[[x]]$.

If a_0 is invertible in R, then we see that $A(x) = \sum_{n=0}^{\infty} a_n x^n$ is invertible in $R[[x]]$ and its inverse $B(x) = \sum_{n=0}^{\infty} b_n x^n$ is given by the recursive formula

$$b_0 = a_0^{-1}, \quad b_n = -\frac{1}{a_0} \sum_{k=1}^{n} a_k b_{n-k} \quad \text{(when } n \geq 1\text{)}.$$

Yet another operation is composing series, i.e. forming the series $A(B(x))$. This is again defined naturally, but we can only do this if $b_0 = 0$, because otherwise we would get a potentially infinite number of terms for each degree. For example the degree 0 coefficient we would get is $a_0 + a_1 b_0 + a_2 b_0^2 + ...$, which need not converge in R. Assuming that $b_0 = 0$, a short calculation shows that

$$A(B(x)) = \sum_{n=0}^{\infty} \sum_{\substack{k \geq 0 \\ i_1, ..., i_k \geq 1 \\ i_1 + ... + i_k = n}} a_k b_{i_1} ... b_{i_k} x^n.$$

Finally we may formally differentiate the formal power series, so we define

$$A'(x) := \sum_{n=1}^{\infty} n a_n x^{n-1}.$$

1.2 ANALYTIC FUNCTIONS

The formal power series are all fine, but the real fun starts when they happen to be Taylor series of some analytic functions. From now on we will therefore simply take $R = \mathbf{C}$.

Recall that a function $f \colon \mathbf{C} \to \mathbf{C}$ is **analytic** in a neighbourhood of 0 if it is complex differentiable in that neighbourhood. In other words, there exists an open disc $B(R) := \{z \in \mathbf{C} \colon |z| < R\}$ such that for all $a \in B(R)$ the limit

$$\lim_{z \to a} \frac{f(z) - f(a)}{z - a} = f'(z)$$

exists. It is a basic result in complex analysis that in this case f is actually differentiable infinitely many times, and we may write f in the form of a power series

$$f(x) = \sum_{n=0}^{\infty} f_n x^n$$

that converges in the disc $B(R)$. The coefficients f_n are *uniquely* determined by the formula $f_n = \frac{1}{n!} f^{(n)}(0)$, where $f^{(n)}$ is the nth derivative of f.

Whenever our formal power series represents an analytic function, we may as well work with the function itself. The uniqueness of the representation ensures that the operations (addition, multiplication, ...) we do are in one-to-one correspondence with the operations on the series.

Let us now look at some important examples of analytic functions and their power series.

First of all there is the geometric series

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n.$$

If we differentiate this k times, we get

$$\frac{1}{(1-x)^{k+1}} = \sum_{n=0}^{\infty} \binom{n+k}{k} x^n.$$

An even more general version is given by the Binomial series

$$(1+x)^{\alpha} = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n,$$

where

$$\binom{\alpha}{n} := \frac{\alpha(\alpha-1)...(\alpha-n+1)}{n!}$$

is the **generalized binomial coefficient**.

Next there is the exponential function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

Using the formula $e^{ix} = \cos(x) + i\sin(x)$ it is easy to derive the series for cos and sin:

$$\cos(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n}, \quad \sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}.$$

For logarithm we can find the series

$$\log(1-x) = -\sum_{n=1}^{\infty} \frac{x^n}{n}$$

and a few less often occuring ones around the same theme are

$$\cosh(x) = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!}, \quad \sinh(x) = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!}$$

and

$$\arctan(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1}.$$

1.3 SOLVING RECURRENCE RELATIONS AND SUMS

In this section we will apply the machinery we have developed so far to a couple of practical example problems. Let us start with the most famous one.

The **Fibonacci sequence** F_n ($n \in \mathbf{Z}$) is defined by the equations

$$F_0 = 0, \quad F_1 = 1, \quad F_n = F_{n-1} + F_{n-2}.$$

Let $F(x) = \sum_{n=0}^{\infty} F_n x^n$ be the associated generating function for the terms with non-negative indices. Now the general procedure for solving $F(x)$ in the case of a linear recursion like this goes as follows. We first multiply both sides of the recursive equation by x^n to get

$$F_n x^n = F_{n-1}x^n + F_{n-2}x^n.$$

Then we sum over n to form the (formal) power series

$$\begin{aligned} F(x) &= \sum_{n=0}^{\infty} F_{n-1}x^n + \sum_{n=0}^{\infty} F_{n-2}x^n \\ &= F_{-1} + x\sum_{n=1}^{\infty} F_{n-1}x^{n-1} + F_{-2} + F_{-1}x + x^2\sum_{n=2}^{\infty} F_{n-2}x^{n-2} \\ &= xF(x) + x + x^2F(x). \end{aligned}$$

Finally we solve for $F(x)$ to get

$$F(x) = \frac{x}{1-x-x^2}.$$

To obtain the coefficients we can use partial fraction decomposition to write

$$F(x) = \frac{x}{1-x-x^2} = \frac{x}{\varphi+\varphi^{-1}}\left(\frac{\varphi}{1-\varphi x} + \frac{\varphi^{-1}}{1+\varphi^{-1}x}\right),$$

where $\varphi := \frac{1+\sqrt{5}}{2}$ is the *Golden ratio*. Now using the formula for a geometric series we get

$$F(x) = \sum_{n=1}^{\infty} \frac{\varphi^n - (-1)^n(\varphi^{-1})^n}{\varphi+\varphi^{-1}} x^n.$$

We have the identity $\varphi + \varphi^{-1} = \sqrt{5}$.

We have thus proved *Binet's formula*

$$F_n = \frac{\varphi^n - (-\varphi)^{-n}}{\sqrt{5}}.$$

Similar techniques work for many other kind of recurrences too, and here is one more example. Let A_n be defined by

$$A_0 = 1, \quad A_n = (n-1)A_{n-1} + n \text{ when } n \geq 1.$$

Because the factor $n-1$ indicates that the sequence will grow like $n!$, we let $A(x)$ be the corresponding **exponential generating function**, which is a power series of the form

$$A(x) = \sum_{n=0}^{\infty} A_n \frac{x^n}{n!}.$$

Starting similarly as with the Fibonacci series we get

$$\begin{aligned}\sum_{n=1}^{\infty} A_n \frac{x^n}{n!} &= \sum_{n=1}^{\infty}(n-1)A_{n-1}\frac{x^n}{n!} + \sum_{n=1}^{\infty} n\frac{x^n}{n!} \\ &= x\sum_{n=1}^{\infty} A_{n-1}\frac{x^{n-1}}{(n-1)!} - \sum_{n=1}^{\infty} A_{n-1}\frac{x^n}{n!} + x\sum_{n=1}^{\infty}\frac{x^{n-1}}{(n-1)!} \\ &= xA(x) + xe^x - \sum_{n=1}^{\infty} A_{n-1}\frac{x^n}{n!}.\end{aligned}$$

The left hand side is $A(x) - A_0$, so we have

$$xA(x) - A(x) + 1 + xe^x = \sum_{n=1}^{\infty} A_{n-1}\frac{x^n}{n!}.$$

Differentiating both sides gives us

$$A(x) + xA'(x) - A'(x) + e^x + xe^x = A(x),$$

so

$$A'(x) = \frac{1+x}{1-x}e^x.$$

This does not have an antiderivative in terms of elementary functions. However, we may still extract its coefficients by looking at the product

$$A'(x) = (1+x)(1+x+x^2+...)(1+x+\frac{x^2}{2}+\frac{x^3}{3!}+...).$$

The product of the first two terms is

$$(1 + x + x^2 + ...) + (x + x^2 + x^3 + ...) = 1 + 2x + 2x^2 + ...$$

Multiplying this with the third factor gives

$$e^x + 2xe^x + 2x^2e^x + 2x^3e^x + ...,$$

whose nth coefficient is

$$\frac{1}{n!} + \frac{2}{(n-1)!} + ... + \frac{2}{(n-n)!}.$$

Thus $A'(x)$ has nth coefficient equal to

$$\frac{1 + 2n + 2n(n-1) + ... + 2n!}{n!}.$$

This means that the nth coefficient of $A(x)$ is

$$\frac{1 + 2(n-1) + 2(n-1)(n-2) + ... + 2(n-1)!}{n!},$$

which gives us

$$\begin{aligned} A_0 &= A_1 = 1, \\ A_n &= 1 + 2\sum_{j=1}^{n-1}(n-j)...(n-1) \quad (n \geq 2). \end{aligned}$$

This result is quite unsurprising if one stares at the original recursion for a while. Indeed one would have probably guessed this formula even without generating functions. Yet it should be noted that the generating function approach turned the problem into a fairly mechanical task.

2 COMBINATORIAL SPECIES

Combinatorial species is an abstract – but at the same time delightfully concrete – framework for modeling combinatorial structures. Computationally its strength arises from its strong connections to generating functions, while for modeling purposes it brings together classical combinatorial enumeration and Pólya theory.

2.1 DEFINITION

For those who know category theory, the definition of a combinatorial species is easy.

Definition 2.1 A **species** is a functor $\mathsf{F}\colon \mathbf{Bij} \to \mathbf{FinSet}$, where **Bij** is the category of finite sets and bijections, and **FinSet** is the category of finite sets and functions.

Spelled out for those that do *not* know category theory, this means that a species is a rule F that assigns to every finite set U a finite set $\mathsf{F}[U]$ and to every bijection $\sigma\colon U \to V$ a function $\mathsf{F}[\sigma]\colon \mathsf{F}[U] \to \mathsf{F}[V]$. Moreover the rule F should satisfy the following two properties:

- F respects composition, i.e. if $\sigma\colon U \to V$ and $\tau\colon V \to W$ are two bijections, then $\mathsf{F}[\tau \circ \sigma] = \mathsf{F}[\tau] \circ \mathsf{F}[\sigma]$

- F maps identity to identity, i.e. $\mathsf{F}[\mathrm{Id}_U] = \mathrm{Id}_{\mathsf{F}[U]}$

The idea is that U is a set of labels and $\mathsf{F}[U]$ is the set of all F-structures with labels drawn from U. The map $\mathsf{F}[\sigma]$ relabels the $\mathsf{F}[U]$ structures so that they become $\mathsf{F}[V]$ structures. For example if $\mathsf{F} = \mathsf{S}$ is the species of **permutations** and $U = \{1, 2, 3\}$, then

$$\mathsf{S}[U] = \{(1)(2)(3), (12)(3), (13)(2), (23)(1), (123), (132)\}.$$

Notice that it follows from the definition of species that each map $\mathsf{F}[\sigma]$ is actually a bijection.

If $\sigma\colon U \to V$ is a bijection, the map $\mathsf{S}[\sigma]$ is defined simply by mapping $\tau \in \mathsf{S}[U]$ to $\sigma\tau\sigma^{-1} \in \mathsf{S}[V]$. So if for example $\sigma\colon U \to V := \{a, b, c\}$ is defined by $\sigma(1) = a$, $\sigma(2) = b$, $\sigma(3) = c$, then $\mathsf{S}[\sigma]$ maps

$$(1)(2)(3) \mapsto (a)(b)(c), (12)(3) \mapsto (ab)(c), (13)(2) \mapsto (ac)(b),$$

$$(23)(1) \mapsto (bc)(a), (123) \mapsto (abc), (132) \mapsto (acb).$$

2.2 COMBINATORIAL EQUALITY, EMBEDDINGS AND COVERINGS

In general there are many isomorphic ways to specify a given species. For example, a permutation $\varphi\colon U \to U$ can be defined either as a bijection or as an ordered pair (F, ψ), where $F \subset U$ is the set of fixed points of φ and $\psi\colon U \setminus F \to U \setminus F$ is a derangement (a permutation without fixed points).

A **transformation** η between species F and G is a collection of maps $\eta_U\colon \mathsf{F}[U] \to \mathsf{G}[U]$, where U ranges over finite sets. Such a transformation is **natural** if the following diagram commutes for all finite sets U and V and bijections $\sigma\colon U \to V$:

$$\begin{tikzcd}
\mathsf{F}[U] \arrow[r, "{\mathsf{F}[\sigma]}"'] \arrow[d, "\eta_U"'] & \mathsf{F}[V] \arrow[d, "\eta_V"] \\
\mathsf{G}[U] \arrow[r, "{\mathsf{G}[\sigma]}"'] & \mathsf{G}[V]
\end{tikzcd}$$

A natural transformation η is said to be an isomorphism between species F and G if every η_U is a bijection. In this case we say that F and G are **combinatorially equal** and write $\mathsf{F} = \mathsf{G}$.

Finally η is said to be a **covering** if every η_U is a surjection and **embedding** if every η_U is an injection.

2.3 ASSOCIATED GENERATING FUNCTIONS

The fundamental counting series associated to a species F is its **cycle index**, defined by

$$Z_F(p_1, p_2, ...) := \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{\sigma \in S_n} |\operatorname{Fix} \mathsf{F}[\sigma]| p_1^{\sigma_1} ... p_n^{\sigma_n}.$$

Here $\operatorname{Fix} \mathsf{F}[\sigma]$ is the set of those F-structures in $\mathsf{F}[n] := \mathsf{F}[\{1, 2, ..., n\}]$ that are fixed by $\mathsf{F}[\sigma]$ and σ_j is the number of j-cycles in σ.

From the cycle index we can recover two basic generating functions associated to the species F. The first one is given by

$$F(x) := Z_F(x, 0, 0, ...) = \sum_{n=0}^{\infty} |\mathsf{F}[n]| \frac{x^n}{n!}.$$

This is an exponential generating function that simply counts the number of labeled F-structures.

The second basic generating function is given by

$$\tilde{F}(x) := Z_F(x, x^2, x^3, x^4, ...) = \sum_{n=0}^{\infty} \left(\frac{1}{n!} \sum_{\sigma \in S_n} |\operatorname{Fix} \mathsf{F}[\sigma]| \right) x^n.$$

Notice that $\tilde{F}(x)$ is an ordinary generating function since $\frac{1}{n!}$ is part of the number we are interested in.

We claim that the coefficient

$$\frac{1}{n!} \sum_{\sigma \in S_n} |\operatorname{Fix} \mathsf{F}[\sigma]|$$

is the number of *unlabeled* $\mathsf{F}[n]$-structures, which means the number of distinct *shapes* of $\mathsf{F}[n]$-structures that remain after the labels have been erased. In more technical terms, two structures are considered to have the same unlabeled structure if one can be obtained from the other one by permuting the labels. Indeed, we can think of unlabeled structures as orbits in $\mathsf{F}[n]$ under the S_n-action given by $\sigma \cdot f = \mathsf{F}[\sigma](f)$ for all $\sigma \in S_n$ and $f \in \mathsf{F}[n]$. Then what we claimed is simply Burnside's lemma.

2.4 OPERATIONS ON SPECIES

One of the main aspects of the theory of combinatorial species is its own combinatorial nature (duh). This refers to the ease at which one may combine different species to form new ones. In this section we will list some of these combining operations and look at how the generating functions of the resulting new species are calculated.

The first operation is **sum**. If F and G are two species, then we define $\mathsf{F} + \mathsf{G}$ to be the species whose structures and morphisms are

Here $\sqcup$ is the disjoint union.

– $(\mathsf{F} + \mathsf{G})[U] := \mathsf{F}[U] \sqcup \mathsf{G}[U]$,

- $(\mathsf{F}+\mathsf{G})[\sigma](x) := \begin{cases} \mathsf{F}[\sigma](x), & \text{if } x \in \mathsf{F}[U] \\ \mathsf{G}[\sigma](x), & \text{if } x \in \mathsf{G}[U] \end{cases}$

It is easy to see that the series of the new species are simply

- $Z_{F+G} = Z_F + Z_G$,
- $(F+G)(x) = F(x) + G(x)$, and
- $(\widetilde{F+G})(x) = \tilde{F}(x) + \widetilde{G}(x)$.

An intuitive way to think about a sum species $\mathsf{F}+\mathsf{G}$ is that an $(\mathsf{F}+\mathsf{G})$-structure is an F-structure OR a G-structure.

The second operation is **product**. If F and G are two species, then their product species $\mathsf{F}\cdot\mathsf{G}$ is defined by setting

- $(\mathsf{F}\cdot\mathsf{G})[U] := \bigcup\{\mathsf{F}[U_1] \times \mathsf{G}[U_2] : U = U_1 \cup U_2, U_1 \cap U_2 = \emptyset\}$,
- $(\mathsf{F}\cdot\mathsf{G})[\sigma]((f,g)) := (\mathsf{F}[\sigma|U_1](f), \mathsf{G}[\sigma|U_2](g))$ where $f \in \mathsf{F}[U_1]$ and $g \in \mathsf{G}[U_2]$.

What this means is that we look at all the possible partitions of U into two disjoint sets U_1 and U_2 and form all pairs of $\mathsf{F}[U_1]$- and $\mathsf{G}[U_2]$-structures. For the generating functions we have

- $Z_{F\cdot G} = Z_F Z_G$,
- $(F\cdot G)(x) = F(x)G(x)$,
- $(\widetilde{F\cdot G})(x) = \tilde{F}(x)\widetilde{G}(x)$.

Proving these is left as an exercise. Intuitively one can think that an $(\mathsf{F}\cdot\mathsf{G})$-structure is an F-structure AND a G-structure.

The third operation we look at is **composition**. If F and G are two species, then their composition species $\mathsf{F}\circ\mathsf{G}$ is defined by

- $(\mathsf{F}\circ\mathsf{G})[U] := \bigcup_{\pi \text{ a partition of } U} \mathsf{F}[\pi] \times \prod_{V\in\pi} \mathsf{G}[V]$,
- for $\pi = \{V_1, ..., V_m\}$ and a $(\mathsf{F}\circ\mathsf{G})$-structure $(f, g_1, ..., g_m) \in \mathsf{F}[\pi] \times \mathsf{G}[V_1] \times ... \times \mathsf{G}[V_m]$ we set

$$(\mathsf{F}\circ\mathsf{G})[\sigma](f, g_1, ..., g_m) :=$$
$$(\mathsf{F}[\sigma|\pi](f), \mathsf{G}[\sigma|V_1](g_1), ..., \mathsf{G}[\sigma|V_m](g_m)).$$

In words, an $(\mathsf{F}\circ\mathsf{G})$-structure is formed as follows: Take a partition of the given labels and construct an F-structure with the *parts* as labels. Then go through the parts and construct a G-structure on each of them.

With composition the corresponding rules for the generating functions are not quite as simple as in the case of sums and products, except for $(F \circ G)(x)$. For the cycle index we have

$$Z_{F \circ G}(p_1, p_2, ...) = Z_F(Z_G(p_1, p_2, p_3, ...), Z_G(p_2, p_4, p_6, ...), \\ Z_G(p_3, p_6, p_9, ...), Z_G(p_4, p_8, p_{12}, ...), ...),$$

from which one easily deduces

– $(F \circ G)(x) = F(G(x))$ and

– $(\widetilde{F \circ G})(x) = Z_F(\widetilde{G}(x), \widetilde{G}(x^2), \widetilde{G}(x^3), ...)$.

To maintain the light stick-to-the-point exposition, we will skip the proofs here. Intuitively one can think of an $(\mathsf{F} \circ \mathsf{G})$-structure as an F-structure OF G-structures.

2.5 BASIC SPECIES

There are two trivial species, 0 and 1. For 0 there are no structures whatsoever. For 1 there is exactly one structure for the empty set of labels and no structures for any other set of labels. The generating functions of $\mathsf{F} = 0$ and $\mathsf{G} = 1$ are $Z_F = F(x) = \widetilde{F}(x) = 0$ and $Z_G = G(x) = \widetilde{G}(x) = 1$.

In fact we can take the sum species $1 + ... + 1$ where there are n terms to get a species n which has exactly n structures on the empty set of labels and no structures on the other label sets. For $\mathsf{F} = n$ we have $Z_F = F(x) = \widetilde{F}(x) = n$. Thus $\mathbf{N}$ can be embedded in the space of species. One can check that $m \cdot n = mn$ as species, as well as $m \cdot \mathsf{F} = \mathsf{F} + ... + \mathsf{F}$ for any species F (the sum has m terms).

The next interesting species is the singleton species X which has a single structure for any label set that has exactly one element and no structures for other label sets. It follows that we have $Z_X(p_1, p_2, ...) = p_1$ and $F(x) = \widetilde{F}(x) = x$.

Let us now introduce a species that has structures for label sets of any size. The species E of sets is simply defined by $\mathsf{E}[U] = \{U\}$. Since there is exactly one structure, we have

$$E(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x \quad \text{and} \quad \widetilde{E}(x) = \sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

For the cycle index we have

$$\begin{aligned} Z_E &= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{\sigma \in S_n} p_1^{\sigma_1} ... p_n^{\sigma_n} \\ &= \sum_{n=0}^{\infty} \sum_{\sigma_1 + 2\sigma_2 + ... + n\sigma_n = n} \frac{p_1^{\sigma_1} p_2^{\sigma_2} ... p_n^{\sigma_n}}{\sigma_1! ... \sigma_n! 1^{\sigma_1} 2^{\sigma_2} ... n^{\sigma_n}} \\ &= \sum_{\sigma_1 + 2\sigma_2 + ... + n\sigma_n = 0}^{\infty} \frac{\left(\frac{p_1}{1}\right)^{\sigma_1} \left(\frac{p_2}{2}\right)^{\sigma_2} ... \left(\frac{p_n}{n}\right)^{\sigma_n}}{\sigma_1! ... \sigma_n!} \\ &= \exp\left(p_1 + \frac{p_2}{2} + \frac{p_3}{3} + \frac{p_4}{4} + ...\right), \end{aligned}$$

since there are $\frac{n!}{\sigma_1! ... \sigma_n! 1^{\sigma_1} 2^{\sigma_2} ... n^{\sigma_n}}$ different permutations of cycle type $(\sigma_1, ..., \sigma_n)$.

The species of permutations S can be defined by setting $\mathsf{S}[U] = \{\tau : U \to U : \tau \text{ is a bijection}\}$ and $\mathsf{S}[\sigma](\tau) = \sigma \circ \tau \circ \sigma^{-1}$. We have

$$\begin{aligned} Z_S &= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{\sigma \in S_n} \sigma_1! ... \sigma_n! 1^{\sigma_1} ... n^{\sigma_n} p_1^{\sigma_1} ... p_n^{\sigma_n} \\ &= \sum_{n=0}^{\infty} \sum_{\sigma_1 + 2\sigma_2 + ... + n\sigma_n = n} p_1^{\sigma_1} ... p_n^{\sigma_n} \\ &= \prod_{n=1}^{\infty} \frac{1}{1 - p_n}, \end{aligned}$$

since there are $\sigma_1! ... \sigma_n! 1^{\sigma_1} ... n^{\sigma_n}$ permutations that are fixed under conjugation by σ and in total there are $\frac{n!}{\sigma_1! ... \sigma_n! 1^{\sigma_1} ... n^{\sigma_n}}$ permutations of given cycle type. From this we easily see that $S(x) = \frac{1}{1-x}$ and $\tilde{S}(x) = \prod_{n=1}^{\infty} \frac{1}{1-x^n}$.

2.6 DERIVED SPECIES

The most common tool for investigating new species is to form combinatorial equalities. For example, if B is the species of binary trees, then we have

$$\mathsf{B} = 1 + \mathsf{X} \cdot \mathsf{B}^2,$$

meaning that B is either empty or a root together with two rooted binary trees, namely the left subtree and the right subtree. To enumerate the unlabeled binary trees we can solve the equation

$$\widetilde{B}(x) = 1 + x\widetilde{B}(x)^2$$

to get

$$\widetilde{B}(x) = \frac{1-\sqrt{1-4x}}{2x}.$$

A short computation reveals that

$$\begin{aligned}\widetilde{B}(x) &= \frac{1}{2x}\left(1 - \sum_{n=0}^{\infty}\binom{1/2}{n}(-4x)^n\right)\\ &= \frac{-1}{2x}\sum_{n=1}^{\infty}\frac{\frac{1}{2}\left(\frac{1}{2}-1\right)\dots\left(\frac{1}{2}-n+1\right)}{n!}(-4x)^n\\ &= \sum_{n=0}^{\infty}\frac{1}{n+1}\binom{2n}{n}x^n,\end{aligned}$$

so that the number of unlabeled binary trees on n vertices is given by the *Catalan number* $\frac{1}{n+1}\binom{2n}{n}$. In a similar manner we can calculate that

$$B(x) = \frac{1-\sqrt{1-4x}}{2x}$$

and

$$Z_B = \frac{1-\sqrt{1-4p_1}}{2p_1}.$$

There is a caveat that one should notice here: We never gave an explicit definition of B. Thus what we have done is simply show that if a species B satisfying $\mathsf{B} = 1 + \mathsf{X} \cdot \mathsf{B}^2$ exists, then its unlabeled generating function is given as above. (In fact the quadratic equation for $\widetilde{B}(x)$ has two solutions, and we ignored the other one on the grounds that it has a pole at 0, so its generating series has a term $\frac{1}{x}$ in it.)

One can show that in a fairly general setting such recursive equations actually admit a unique species as a solution, but we will not pursue this direction here. In applications we usually know beforehand that the species we are looking at should exist, and we are happy to just compute its generating series by utilizing combinatorial equalities such as the one given above for B.

The next species we look at is the species of cycles Cyc. It can be defined by setting $\mathsf{Cyc}[U] = \{\tau\colon U \to U : \tau \text{ is a cycle}\}$ and $\mathsf{Cyc}[\sigma](\tau) = \sigma \circ \tau \circ \sigma^{-1}$ for any bijection $\sigma\colon U \to V$.

Notice that any permutation can be written as a set of cycles. That is, we have

$$\mathsf{S} = \mathsf{E} \circ \mathsf{Cyc}.$$

In particular we have

$$\prod_{n=1}^{\infty} \frac{1}{1-p_n} = \exp\Big(Z_{\mathrm{Cyc}}(p_1, p_2, ...) + \frac{Z_{\mathrm{Cyc}}(p_2, p_4, p_6, ...)}{2} + \frac{Z_{\mathrm{Cyc}}(p_3, p_6, p_9, ...)}{3} + ...\Big),$$

and after taking logarithms on both sides this becomes

$$\sum_{n=1}^{\infty}\sum_{k=1}^{\infty}\frac{p_n^k}{k}=\sum_{n=1}^{\infty}\frac{Z_{\mathrm{Cyc}}(p_n,p_{2n},p_{3n},\dots)}{n}. \tag{2.1}$$

Let's look first at p_1. We must have

$$Z_{\mathrm{Cyc}}(p_1,p_2,\dots)=\sum_{k=1}^{\infty}\frac{p_1^k}{k}+Z_2(p_2,p_3,\dots)$$

for some series Z_2 not depending on p_1, since all the other terms on the right hand side of (2.1) do not contain p_1. Now assume by induction that we have written $Z_{\mathrm{Cyc}}(p_1,p_2,\dots)$ in the form

$$Z_{\mathrm{Cyc}}(p_1,p_2,\dots)=\sum_{n=1}^{N-1}c_n\sum_{k=1}^{\infty}\frac{p_n^k}{k}+Z_N(p_N,p_{N+1},\dots)$$

for some numbers c_n and consider p_N. All the terms that contain p_N are contained in the sum

$$\sum_{d|N}\frac{Z_{\mathrm{Cyc}}(p_d,p_{2d},\dots,p_{\frac{N}{d}\cdot d},\dots)}{d}=\sum_{d|N}\frac{1}{d}\Big(\sum_{n=1}^{N-1}c_n\sum_{k=1}^{\infty}\frac{p_{dn}^k}{k}+ Z_N(p_{dN},p_{d(N+1)},\dots)\Big).$$

In particular we see that $\sum_{d|N}\frac{c_{N/d}}{d}=1$, giving the recursion

$$c_N=1-\sum_{d|N,d\neq 1}\frac{c_{N/d}}{d}$$

and completing the induction if we choose Z_{N+1} to contain all the left-over terms once we move every term with $d>1$ on the left hand side to the right hand side. Now what is the sequence c_n? Notice that if we define $a_n=nc_n$ we get the relation $\sum_{d|n}a_{n/d}=n$, or in other words $1*a=\mathrm{Id}$, where $*$ is the Dirichlet convolution. Thus $a=\mathrm{Id}*\mu=\varphi$, where φ is the Euler totient function. We have shown that

$$Z_{\mathrm{Cyc}}(p_1, p_2, ...) = \sum_{n=1}^{\infty} \frac{\varphi(n)}{n} \log\left(\frac{1}{1-p_n}\right).$$

Looking at the unlabeled enumeration function we get the nice identity

$$\widetilde{\mathrm{Cyc}}(x) = \frac{x}{1-x} = \sum_{n=1}^{\infty} \frac{\varphi(n)}{n} \log\left(\frac{1}{1-x^n}\right),$$

and for the labeled enumeration function we have

$$\mathrm{Cyc}(x) = \log\left(\frac{1}{1-x}\right).$$

3 PÓLYA THEORY

Pólya theory is about counting objects that have symmetries. For example we might want to count in how many ways n people can be seated around a round table when different rotations of the same arrangement should be considered to be equal. In this case the rotational symmetry is easy to take into account, and the answer is of course $(n-1)!$.

The question becomes more difficult and interesting, however, if we instead ask in how many ways we can form a *necklace* with a red beads, b green beads and c blue beads such that $a+b+c=n$. Two necklaces are considered to be the same if one can be obtained from the other by rotating the necklace.

In general the symmetry is described by some group. In the case of necklaces this group is simply C_n, but we could have something more complicated instead. We might for example want to account also for reflections, in which case we would be counting *bracelets* and the group in question would be D_n. Yet another problem would be to count in how many ways we can color the 6 faces of a cube in 3 colors when all the rotational symmetries should be taken into account. The rotational symmetry group of the cube turns out to be isomorphic to S_4.

3.1 COLORED SPECIES

Colored species, more oftenly called *multisort species*, are species where the labels also carry a color (or *sort*). In the literature usually only finitely many sorts are allowed, but we will present a version without this restriction. Therefore we will use the term colored species. Formally the definitions are the following.

Definition 3.1 By a **finite colored set** U we mean a disjoint union of a sequence of finite sets U_i $(i \geq 1)$ where only finitely many of the sets U_i are non-empty. In other words $U = \bigcup_{i=1}^{\infty}(\{i\} \times U_i)$. We will think of U_i as a subset of U via the embedding $U_i \mapsto \{i\} \times U_i \subset U$ and the elements in $U_i \subset U$ are said to have color i.

Definition 3.2 A **color-respecting bijection** between two finite colored sets U and V is simply a bijection $U \to V$ that maps U_i bijectively onto V_i for all $i \geq 1$.

Definition 3.3 A **colored species** is a functor $\mathsf{F}\colon \mathbf{ColBij} \to \mathbf{FinSet}$, where **ColBij** is the category of finite colored sets and color-respecting bijections.

Notice that an ordinary species F can be thought of as a colored species for which $\mathsf{F}[U] = \emptyset$ if there are labels with other colors than 1 in U.

Let $n_1, n_2, ...$ be such that $n_i = 0$ for all but finitely many i and write

$$\mathsf{F}[n_1, n_2, ...] := \mathsf{F}[U],$$

where U is the colored finite set $U = \bigcup_{i=1}^{\infty} \{(i, 1), ..., (i, n_i)\}$. Similarly when $\sigma^{(i)} \in S_{n_i}$ $(i \geq 1)$ we write

$$\mathsf{F}[\sigma^{(1)}, \sigma^{(2)}, ...] := \mathsf{F}[\sigma],$$

where $\sigma\colon U \to U$ is the color-respecting bijection defined by setting $\sigma((i, u)) = (i, \sigma^{(i)}(u))$ for all $(i, u) \in U$.

The cycle index of a colored species F is now given by

$$Z_F((p_i^{(1)})_{i\geq1}; (p_i^{(2)})_{i\geq1}; ...) = \\ \sum_{n_1, n_2, ...} \frac{1}{n_1! n_2! ...} \sum_{\sigma^{(i)} \in S_{n_i}, i\geq1} |\operatorname{Fix} \mathsf{F}[\sigma^{(1)}, \sigma^{(2)}, ...]| \prod_{k=1}^{\infty} \prod_{j=1}^{n_k} (p_j^{(k)})^{\sigma_j^{(k)}},$$

where the first sum ranges over all sequences $n_1, n_2, ...$ such that $n_i = 0$ for all but finitely many i.

The generating function for a labeled species F is given by

$$\begin{aligned} F(x_1, x_2, ...) &= Z_F(x_1, 0, ...; x_2, 0, ...; ...) \\ &= \sum_{n_1, n_2, ...} \frac{|F[n_1, n_2, ...]|}{n_1! n_2! ...} x_1^{n_1} x_2^{n_2} .... \end{aligned}$$

and for unlabeled species we have

$$\tilde{F}(x_1, x_2, ...) = Z_F(x_1, x_1^2, x_1^3, ...; x_2, x_2^2, x_2^3, ...; ...).$$

The definitions and rules for sums and products of colored species parallel those of ordinary species and the induced operations on the generating functions are the same.

For substitution we have the following: Let F and $\mathsf{G}_1, \mathsf{G}_2, ...$ be colored species and assume that $\mathsf{G}_i[\emptyset] = \emptyset$. An $\mathsf{F} \circ (\mathsf{G}_1, \mathsf{G}_2, ...)$-structure on a colored set U is formed by taking a colored partition of U, creating an F-structure on the partition, and then assigning each part with color i a G_i-structure. The generating functions of $\mathsf{H} = \mathsf{F}(\mathsf{G}_1, \mathsf{G}_2, ...)$ are

$$\begin{aligned}
H(x_1, x_2, ...) &= F(G_1(x_1, x_2, ...), G_2(x_1, x_2, ...), ...)\\
\widetilde{H}(x_1, x_2, ...) &= Z_F(\widetilde{G}_1(x_1, x_2, ...), \widetilde{G}_1(x_1^2, x_2^2, ...), ...;\\
&\qquad \widetilde{G}_2(x_1, x_2, ...), \widetilde{G}_2(x_1^2, x_2^2, ...), ...; ...)\\
Z_H((p_i^{(1)})_{i\geq 1}; ...) &= Z_F((Z_{G_1})_1, (Z_{G_1})_2, ...; (Z_{G_2})_1, (Z_{G_2})_2, ...; ...),
\end{aligned}$$

where $(Z_{G_i})_k = Z_{G_i}(p_k^{(1)}, p_{2k}^{(1)}, ...; p_k^{(2)}, p_{2k}^{(2)}, ...; ...)$.

It is typical to denote by X_i $(i \geq 1)$ a species that only has a single structure for a single label of color i. Then for example $\mathsf{X}_1 + \mathsf{X}_2 + \mathsf{X}_3$ would be a single label of one of the three colors 1, 2 or 3. This makes it easy to start constructing colored species out of uncolored (or single-colored species) species by substitution.

Example 3.4 In the beginning we asked in how many ways we can form a necklace of a red, b green and c blue beads. This information can be distilled from the unlabeled generating function of the species $\mathsf{Cyc}(\mathsf{X}_1 + \mathsf{X}_2 + \mathsf{X}_3)$, which is simply

$$\begin{aligned}
&Z_{\mathrm{Cyc}}(x_1 + x_2 + x_3, x_1^2 + x_2^2 + x_3^2, ...) =\\
&\sum_{n=1}^{\infty} \frac{\varphi(n)}{n} \log\left(\frac{1}{1 - x_1^n - x_2^n - x_3^n}\right) =\\
&\sum_{n=1}^{\infty} \frac{\varphi(n)}{n} \sum_{k=1}^{\infty} \frac{(x_1^n + x_2^n + x_3^n)^k}{k} =
\end{aligned}$$

$$\sum_{n=1}^{\infty} \frac{\varphi(n)}{n} \sum_{k=1}^{\infty} \frac{1}{k} \sum_{i_1+i_2+i_3=k} \frac{k!}{i_1!i_2!i_3!} x_1^{ni_1} x_2^{ni_2} x_3^{ni_3}.$$

We need $ni_1 = a$, $ni_2 = b$ and $ni_3 = c$, so n must divide $\gcd(a,b,c)$. Thus the coefficient of $x_1^a x_2^b x_3^c$ is

$$\sum_{n \mid \gcd(a,b,c)} \frac{\varphi(n)}{n} \cdot \frac{\left(\frac{a+b+c}{n} - 1\right)!}{\left(\frac{a}{n}\right)! \left(\frac{b}{n}\right)! \left(\frac{c}{n}\right)!}.$$

3.2 QUOTIENT SPECIES

In this section we will look at a situation where we have a group Γ acting on some species F in a natural way. It is then possible to form a species F/Γ where the structures are Γ-orbits of F-structures.

Definition 3.5 Let Γ be a group and F a (colored) species. We say that Γ **acts naturally** on F-structures if for every finite (colored) set U there exists an action $\Gamma \times \mathsf{F}[U] \to \mathsf{F}[U]$ such that

$$g \cdot \mathsf{F}[\sigma](f) = \mathsf{F}[\sigma](g \cdot f)$$

for every (color-respecting) bijection σ, $g \in \Gamma$ and $f \in \mathsf{F}[U]$.

In other words, Γ acts naturally on F-structures if and only if its action commutes with relabeling of structures.

There is also a category theoretical way to look at this, which also justifies the use of word *natural* in the definition. If we let **Spec** be the category of species and natural transformations between them, and regard Γ as a category with one object and group elements as morphisms, then we can define a **Γ-species** to be a functor $\Gamma \to$ **Spec**. Now, if the target object of such a functor is a species F, the morphisms in Γ get mapped to natural isomorphisms $\mathsf{F} \to \mathsf{F}$, which in turn define a natural action (in the sense of Definition 3.5) of Γ on F-structures. We say that F is the **underlying species** of the Γ-species.

Definition 3.6 The quotient of a Γ-species with underlying species F is the species F/Γ defined by letting $(\mathsf{F}/\Gamma)[U] := \mathsf{F}[U]/\Gamma$ and $(\mathsf{F}/\Gamma)[\sigma](\Gamma f) := \Gamma\mathsf{F}[\sigma](f)$ for all finite (colored) sets U and (color-respecting) bijections σ.

It follows from the naturality of the action of Γ that $(\mathsf{F}/\Gamma)[\sigma](\Gamma f)$ is well-defined, i.e. does not depend on the representative f of the orbit Γf.

It is useful to define a special cycle index series for Γ-species that is parametrised also on the elements of γ.

Definition 3.7 Let F be an underlying species of a colored Γ-species. Then we define

$$Z_{\mathsf{F}}^{\Gamma}(\gamma) = \sum_{n_1,n_2,\ldots}^{\infty} \frac{1}{n_1!n_2!\ldots} \sum_{\sigma^{(i)}\in S_{n_i}} |\operatorname{Fix}(\gamma\cdot\mathsf{F}[\sigma^{(1)},\sigma^{(2)},\ldots])| \prod_{j,k} (p_j^{(k)})^{\sigma_j^{(k)}}.$$

Notice in particular that the Γ-species series retains the information on the cycle series of F, since $Z_{\mathsf{F}}^{\Gamma}(1) = Z_{\mathsf{F}}$.

Theorem 3.8 The cycle index of F/Γ is given by

$$Z_{\mathsf{F}/\Gamma} = \frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma} Z_{\mathsf{F}}^{\Gamma}(\gamma).$$

Proof. It is enough to show that

$$|\operatorname{Fix}(\mathsf{F}/\Gamma)[\sigma]| = \frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma} |\operatorname{Fix}(\gamma\cdot\mathsf{F}[\sigma])|.$$

This in turn is a direct application of the following Burnside-type lemma.

Lemma. Let Γ and S be two groups acting on a set X and assume that the actions commute, i.e. $\gamma\cdot\sigma\cdot x = \sigma\cdot\gamma\cdot x$ for all $\gamma\in\Gamma$, $\sigma\in S$ and $x\in X$. Then S acts on X/Γ by $\sigma\cdot\Gamma x = \Gamma(\sigma\cdot x)$ and

$$|\operatorname{Fix}_{X/\Gamma}(\sigma)| = \frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma} |\operatorname{Fix}_X(\gamma\cdot\sigma)|$$

for all $\sigma \in S$.

In the case at hand, both Γ and the (color-preserving) bijections act on sets of F-structures and the commutativity is satisfied by the definition of Γ-species.

It remains to prove the lemma. Notice that a given orbit Γx is fixed by $\sigma \in S$ if and only if we have $\Gamma x = \sigma \cdot (\Gamma x) = \Gamma(\sigma \cdot x)$, which happens if and only if there exists $\gamma \in \Gamma$ such that $x = \gamma \cdot \sigma \cdot x$. If this is true, then there are exactly $|\Gamma_x|$ different such γs. Thus

Recall that Γ_x is the stabilizer of x.

$$\frac{1}{|\Gamma_x|} \sum_{\gamma \in \Gamma} \mathbf{1}(x = \gamma \cdot \sigma \cdot x) = \mathbf{1}(\Gamma x \in \mathrm{Fix}_{X/\Gamma}(\sigma)).$$

It follows that

$$\begin{aligned} |\operatorname{Fix}_{X/\Gamma}(\sigma)| &= \sum_{x \in X} \frac{1}{|\Gamma x|} \mathbf{1}(\Gamma x \in \mathrm{Fix}_{X/\Gamma}(\sigma)) \\ &= \sum_{x \in X} \frac{1}{|\Gamma x|} \frac{1}{|\Gamma_x|} \sum_{\gamma \in \Gamma} \mathbf{1}(x = \gamma \cdot \sigma \cdot x) \\ &= \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} |\operatorname{Fix}_X(\gamma \cdot \sigma)|. \end{aligned}$$

Here we used the orbit–stabilizer theorem to conclude that $|\Gamma x \| \Gamma_x| = |\Gamma|$. □

For a Γ-species F we may also define the corresponding labeled and unlabeled generating functions simply by setting

- $F_\gamma(x_1, x_2, ...) := Z_F^\Gamma(x, 0, ...; x_2, 0, ...; ...)$, and
- $\tilde{F}_\gamma(x_1, x_2, ...) := Z_F^\Gamma(x_1, x_1^2, ...; x_2, x_2^2, ...; ...)$.

These count the γ-invariant structures of F. Note in particular that for the quotient species $\mathsf{G} := \mathsf{F}/\Gamma$ we have by Theorem 3.8 that

- $G(x_1, x_2, ...) = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} F_\gamma(x_1, x_2, ...)$, and
- $\widetilde{G}(x_1, x_2, ...) = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} \tilde{F}_\gamma(x_1, x_2, ...)$.

Example 3.9 We can construct the species Bra of bracelets by letting the two-element group C_2 act on Cyc by reflection and taking the quotient. By Theorem 3.8 we have

$$Z_{\mathrm{Bra}} = \frac{1}{2}\left(Z_{\mathrm{Cyc}} + Z^{\Gamma}_{\mathrm{Bra}}(\tau)\right),$$

where τ is the reflection. Now

$$Z^{\Gamma}_{\mathrm{Bra}}(\tau) = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{\sigma \in S_n} |\operatorname{Fix}(\tau \cdot \mathsf{Cyc}[\sigma])| p_1^{\sigma_1} ... p_n^{\sigma_n},$$

so we have to figure out $\operatorname{Fix}(\tau \cdot \mathsf{Cyc}[\sigma])$.

Let us denote by $(a_1 ... a_n)$ an arbitrary cycle of length n on the set $\{1, ..., n\}$. Now

$$\tau \cdot \mathsf{Cyc}[\sigma]((a_1 ... a_n)) = \tau((\sigma(a_1) ... \sigma(a_n))) = (\sigma(a_n)\sigma(a_{n-1}) ... \sigma(a_1)).$$

Thus $[a_1, ..., a_n]$ is fixed by $\tau \cdot \mathsf{Cyc}[\sigma]$ if and only if there exists $\ell \geq 0$ such that $\sigma(a_i) = a_{\ell - i}$ for all $1 \leq i \leq n$ when we take indexing modulo n.

Case 1, n is odd: In this case σ must consist of one 1-cycle and $\frac{n-1}{2}$ 2-cycles for any solutions to exist. If this is true, then we may choose a unique representation for any fixed cycle by requiring that a_1 is the fixed element. The rest of the elements in the cycle can be assigned pairwise to the $\frac{n-1}{2}$ 2-cycles in $\left(\frac{n-1}{2}\right)! 2^{\frac{n-1}{2}}$ ways. There are $\frac{n!}{\left(\frac{n-1}{2}\right)! 2^{\frac{n-1}{2}}}$ valid permutations.

Case 2a, n is even and ℓ is odd: In this case σ must consist of $\frac{n}{2}$ 2-cycles for any solutions to exist. If this is true, then we may choose a unique representation for any fixed cycle by choosing $a_1 = 1$. There are $\frac{n}{2}$ choices for ℓ and the choice $a_1 = 1$ also fixes $a_{\ell-1} = \sigma(1)$. The rest of the pairs can be chosen in $\left(\frac{n-2}{2}\right)! 2^{\frac{n-2}{2}}$ ways and thus there are $\left(\frac{n}{2}\right)! 2^{\frac{n}{2}-1}$ fixed cycles in total. There are $\frac{n!}{\left(\frac{n}{2}\right)! 2^{\frac{n}{2}}}$ valid permutations.

Case 2b, n is even and ℓ is even: In this case σ must consist of two 1 cycles and $\frac{n-2}{2}$ 2-cycles. If this is true, then we may choose a unique representation for any fixed cycle by choosing $a_1 = 1$. Since the number of fixed cycles only depends on the cycle type of σ, we may assume that $\sigma(1) \neq 1$. Then there are $\frac{n}{2} - 1$ valid choices for ℓ (we must exclude $\ell = 2$). The fixed points can be chosen in 2 ways, the pair $(a_1, a_{\ell-1})$ is fixed already and the rest of the $\frac{n-4}{2}$ pairs can be chosen in $\left(\frac{n-4}{2}\right)!2^{\frac{n-4}{2}}$ ways. Thus we have $2\left(\frac{n}{2} - 1\right)\left(\frac{n-4}{2}\right)!2^{\frac{n-4}{2}}$ fixed cycles. There are $\frac{n!}{2\left(\frac{n-2}{2}\right)!2^{\frac{n-2}{2}}}$ valid permutations.

These considerations let us derive the following formula for the the cycle index of Bra:

$$\begin{aligned} Z_{\text{Bra}} &= \frac{1}{2} Z_{\text{Cyc}} + \frac{1}{4} \sum_{n \text{ even}} (p_2^{\frac{n}{2}} + p_1^2 p_2^{\frac{n-2}{2}}) + \frac{1}{2} \sum_{n \text{ odd}} p_1 p_2^{\frac{n-1}{2}} \\ &= \frac{1}{2} Z_{\text{Cyc}} + \frac{1}{4} \frac{(1+p_1)^2}{1-p_2} - \frac{1}{4}. \end{aligned}$$

GAME THEORY

PART IV

1 IMPARTIAL GAMES

In this chapter we will look at a class of combinatorial games which are impartial in the sense that the available moves at a given game position does not depend on which player is moving.

1.1 NIMBERS

We first need a model for a game. The idea is to define an impartial game as a set of positions (games) that correspond to the possible moves in the starting position.

Definition 1.1 An **impartial game** is a finite set G such that one of the following holds:

- $G = \emptyset$
- $G = \{G_1, ..., G_n\}$ for some $n \geq 1$ and G_i are impartial games.

A normal two player play on the game corresponds to choosing a sequence $G \ni G_1 \ni G_2 ... \ni \emptyset$ and the loser is the one who cannot make a move because the sequence reached $\emptyset$.

Definition 1.2 To an impartial game G we assign a **nimber** $N(G) \in \mathbf{N}$ as follows:

$$N(G) = \begin{cases} 0, & \text{if } G = \emptyset \\ \operatorname{mex}(\{N(G_1), ..., N(G_n)\}), & \text{if } G = (G_1, ..., G_n) \end{cases}$$

Here mex(A) is the *minimum excludant* defined by

$$\operatorname{mex}(A) = \min(\mathbf{N} \setminus A)$$

for every finite $A \subset \mathbf{N}$.

Nimbers will be useful when we will consider combining games in the next two sections.

Theorem 1.3 Under perfect play the first player wins game G if and only if $N(G) \neq 0$. Equivalently G is a losing position if and only if $N(G) = 0$.

Proof. This is easy to prove by induction. The claim clearly holds for $G = \emptyset$. Assume that the claim holds for all $H \in G$. Now if

$N(G) = 0$, then $N(H) \neq 0$ for all $H \in G$, so we may only move to a winning position which makes the current position losing. Similarly if $N(G) \neq 0$ then there exists $H \in G$ such that $N(H) = 0$. Moving to H is therefore a winning move for position G. □

1.2 SUM GAMES

The sum $G + \widetilde{G}$ of two games G and $\widetilde{G}$ is a game where a player may make a move in either of the games G or $\widetilde{G}$.

Definition 1.4 The sum of two games G and $\widetilde{G}$ is defined recursively by

$$G + \widetilde{G} = \{H + \widetilde{G} : H \in G\} \cup \{G + \widetilde{H} : \widetilde{H} \in \widetilde{G}\}.$$

The following theorem is a fundamental tool in analysing many impartial games.

Theorem 1.5 If G and $\widetilde{G}$ are impartial games, then

$$N(G + \widetilde{G}) = N(G) \oplus N(\widetilde{G}),$$

where $\oplus$ is the (bitwise) XOR-operator.

Proof. By induction

$$\begin{aligned} N(G + \widetilde{G}) &= \operatorname{mex}(\{N(H + \widetilde{G}) : H \in G\} \cup \{N(G + \widetilde{H}) : \widetilde{H} \in \widetilde{G}\}) \\ &= \operatorname{mex}(\{N(H) \oplus N(\widetilde{G}) : H \in G\} \cup \{N(G) \oplus N(\widetilde{H}) : \widetilde{H} \in \widetilde{G}\}). \end{aligned}$$

It is therefore enough to show that for all $a, b \in \mathbf{N}$ we have

$$a \oplus b = \operatorname{mex}(\{a' \oplus b : a' < a\} \cup \{a \oplus b' : b' < b\}).$$

We do this also by induction: We have to show that every $c < a \oplus b$ is either of the form $a' \oplus b$ or of the form $a \oplus b'$. Looking at the largest bit of $a \oplus b \oplus c$ it is easy to see that one of $a \oplus c < b$ or $b \oplus c < a$ holds. Thus the given form is obtained either as $(b \oplus c) \oplus b$ or $a \oplus (a \oplus c)$. □

INDEX