# Chapter 4

# Algorithms for Algebraic Number Theory I

In this chapter, we give the necessary background on algebraic numbers, number fields, modules, ideals and units, and corresponding algorithms for them. Excellent basic textbooks on these subjects are, for example [Bo-Sh], [Cas-Frö], [Cohn], [Ire-Ros], [Marc], [Sam]. However, they usually have little algorithmic flavor. We will give proofs only when they help to understand an algorithm, and we urge the reader to refer to the above textbooks for the proofs which are not given.

## 4.1 Algebraic Numbers and Number Fields

### 4.1.1 Basic Definitions and Properties of Algebraic Numbers

**Definition 4.1.1.** *Let $\alpha \in \mathbb{C}$. Then $\alpha$ is called an* algebraic number *if there exists $A \in \mathbb{Z}[X]$ such that $A(\alpha) = 0$, and $A$ not identically zero. The number $\alpha$ is called an* algebraic integer *if, in addition, one can choose $A$ to be monic (i.e. with leading coefficient equal to 1).*

Then we have:

**Proposition 4.1.2.** *Let $\alpha$ be an algebraic number, and let $A$ be a polynomial with integer coefficients such that $A(\alpha) = 0$, and assume that $A$ is chosen to have the smallest degree and be primitive with $\ell(A) > 0$. Then such an $A$ is unique, is irreducible in $\mathbb{Q}[X]$, and any $B \in \mathbb{Z}[X]$ such that $B(\alpha) = 0$ is a multiple of $A$.*

*Proof.* The ring $\mathbb{Q}[X]$ is a principal ideal domain (PID), and the set of $B \in \mathbb{Q}[X]$ such that $B(\alpha) = 0$ is an ideal, hence is the ideal generated by $A$. If, in addition, $B$ has integral coefficients, Gauss's lemma (Theorem 3.2.8) implies that $B$ is a multiple of $A$ in $\mathbb{Z}[X]$. It is clear that $A$ is irreducible; otherwise $A$ would not be of smallest degree. We will call this $A$ *the* minimal polynomial of $\alpha$. □

We will use the notation $\overline{\mathbb{Q}}$ for the set of algebraic numbers, (hence $\overline{\mathbb{Q}} \subset \mathbb{C}$), $\mathbb{Z}_{\overline{\mathbb{Q}}}$ for the set of algebraic integers, and if $L$ is any subset of $\mathbb{C}$ we will set

$$\mathbb{Z}_L = \mathbb{Z}_{\overline{\mathbb{Q}}} \cap L,$$

and call it the set of integers of $L$. Note that $\overline{\mathbb{Q}}$ is *an* algebraic closure of $\mathbb{Q}$.

For example, we have $\mathbb{Z}_{\mathbb{Q}} = \mathbb{Z}$. Indeed, if $\alpha = p/q \in \mathbb{Q}$ is a root of $A \in \mathbb{Z}[X]$ with $A$ monic, we must have $q \mid \ell(A)$, hence $q = \pm 1$ so $\alpha$ is in $\mathbb{Z}$.

The first important result about algebraic numbers is as follows:

**Theorem 4.1.3.** *Let $\alpha \in \mathbb{C}$. The following four statements are equivalent.*

(1) *$\alpha$ is an algebraic integer.*
(2) *$\mathbb{Z}[\alpha]$ is a finitely generated additive Abelian group.*
(3) *$\alpha$ belongs to a subring of $\mathbb{C}$ which is finitely generated as an Abelian group.*
(4) *There exists a non-zero finitely generated additive subgroup $L$ of $\mathbb{C}$ such that $\alpha L \subset L$.*

As corollaries we have:

**Corollary 4.1.4.** *The set of algebraic integers is a ring. In particular, if $R$ is a ring, the set $\mathbb{Z}_R$ of integers of $R$ is a ring.*

**Corollary 4.1.5.** *If $\alpha \in \mathbb{C}$ is a root of a monic polynomial whose coefficients are algebraic integers (and not simply integers), then $\alpha$ is an algebraic integer.*

**Definition 4.1.6.** *Let $\alpha \in \mathbb{C}$ be an algebraic number, and $A$ its minimal polynomial. The* conjugates *of $\alpha$ are all the* $\deg(A)$ *roots of $A$ in $\mathbb{C}$.*

This notion of conjugacy is of course of fundamental importance, but what I would like to stress here is that from an algebraic point of view the conjugates are indistinguishable. For example, any algebraic identity between algebraic numbers is a simultaneous collection of conjugate identities. To give a trivial example, the identity $(1+\sqrt{2})^2 = 3+2\sqrt{2}$ implies the identity $(1-\sqrt{2})^2 = 3-2\sqrt{2}$. This remark is a generalization of the fact that an equality between two complex numbers implies the equality between their conjugates, or equivalently between their real and imaginary parts. The present example is even more striking if one looks at it from a numerical point of view: it says that the identity $(2.41421\ldots)^2 = 5.828427\ldots$ implies the identity $(0.41421\ldots)^2 = 0.171573\ldots$. Of course this is not the correct way to look at it, but the lesson to be remembered is that an algebraic number *always* comes with all of its conjugates.

### 4.1.2 Number Fields

**Definition 4.1.7.** *A* number field *is a field containing $\mathbb{Q}$ which, considered as a $\mathbb{Q}$-vector space, is finite dimensional. The number $d = \dim_{\mathbb{Q}} K$ is denoted by $[K : \mathbb{Q}]$ and called the* degree *of the number field $K$.*

We recall the following fundamental results about number fields:

**Theorem 4.1.8.** *Let $K$ be a number field of degree $n$. Then*

(1) *(Primitive element theorem) There exists a $\theta \in K$ such that*

$$K = \mathbb{Q}(\theta).$$

*Such a $\theta$ is called a primitive element. Its minimal polynomial is an irreducible polynomial of degree $n$.*

(2) *There exist exactly $n$ field embeddings of $K$ in $\mathbb{C}$, given by $\theta \mapsto \theta_i$, where the $\theta_i$ are the roots in $\mathbb{C}$ of the minimal polynomial of $\theta$. These embeddings are $\mathbb{Q}$-linear, their images $K_i$ in $\mathbb{C}$ are called the conjugate fields of $K$, and the $K_i$ are isomorphic to $K$.*

(3) *For any $i$, $K_i \subset \overline{\mathbb{Q}}$, in other words all the elements of $K_i$ are algebraic numbers and their degree divides $n$.*

The assertion made above concerning the indistinguishability of the conjugates can be clearly seen here. The choice of the conjugate field $K_i$ is a priori completely arbitrary. In many cases, this choice is already given. For example, when we speak of "the number field $\mathbb{Q}(2^{1/3})$", this is slightly incorrect, since what we mean by this is that we are considering the number field $K = \mathbb{Q}[X]/(X^3 - 2)\mathbb{Q}[X]$ *together* with the embedding $X \mapsto 2^{1/3}$ of $K$ into $\mathbb{R}$.

**Definition 4.1.9.** *The* signature *of a number field is the pair $(r_1, r_2)$ where $r_1$ is the number of embeddings of $K$ whose image lie in $\mathbb{R}$, and $2r_2$ is the number of non-real complex embeddings, so that $r_1 + 2r_2 = n$ (note that the non-real embeddings always come in pairs since if $\sigma$ is such an embedding, so is $\bar{\sigma}$, where $^-$ denotes complex conjugation). If $T$ is an irreducible polynomial defining the number field $K$ by one of its roots, the signature of $K$ will also be called the signature of $T$. Here $r_1$ (resp. $2r_2$) will be the number of real (resp. non-real) roots of $T$ in $\mathbb{C}$. When $r_2 = 0$ (resp. $r_1 = 0$) we will say that $K$ and $T$ are* totally real *(resp.* totally complex*).*

It is not difficult to determine the signature of a number field $K$, but some ways are better than others. If $K = \mathbb{Q}(\theta)$, and if $T$ is the minimal polynomial of $\theta$, we can of course compute the roots of $T$ in $\mathbb{C}$ using, for instance, the root finding Algorithm 3.6.6, and count the number of real roots. This is however quite expensive. A much better way is to use a theorem of Sturm which tells us in essence that the sequence of leading coefficients in the polynomial remainder sequence obtained by applying Euclid's algorithm or its variants to $T$ and $T'$ governs the signature. More precisely, we have the following theorem.

**Theorem 4.1.10** (Sturm). *Let $T$ be a squarefree polynomial with real coefficients. Assume that $A_0 = T$, $A_1 = T'$, and that $A_i$ is a polynomial remainder sequence such that for all $i$ with $1 \leq i \leq k$:*

$$e_i A_{i-1} = Q_i A_i - f_i A_{i+1},$$

*where the $e_i$ and $f_i$ are real and positive, and $A_{k+1}$ is a constant polynomial (non-zero since $T$ is squarefree). Set $\ell_i = \ell(A_i)$, and $d_i = \deg(A_i)$. Then, if $s$ is the number of sign changes in the sequence $\ell_0, \ell_1, \ldots, \ell_{k+1}$, and if $t$ is the number of sign changes in the sequence $(-1)^{d_0}\ell_0, (-1)^{d_1}\ell_1, \ldots, (-1)^{d_{k+1}}\ell_{k+1}$, the number of real roots of $T$ is equal to $t-s$.*

*Proof.* For any real $a$, let $s(a)$ be the number of sign changes, not counting zeros, in the sequence $A_0(a), A_1(a), \ldots, A_{k+1}(a)$. We clearly have $\lim_{a\to+\infty} s(a) = s$ and $\lim_{a\to-\infty} s(a) = t$. We are going to prove the following more general assertion: the number of roots of $T$ in the interval $]a, b]$ is equal to $s(a) - s(b)$, which clearly implies the assertion of the theorem.

First, it is clear that a sign sequence at any number $a$ cannot have two consecutive zeros, otherwise these zeros would propagate and we would have $A_{k+1} = 0$. For similar reasons, we cannot have sequences of the form $+$, $0$, $+$, or of the form $-$, $0$, $-$ since the $e_i$ and $f_i$ are positive. Now the desired formula $s(a) - s(b)$ is certainly valid if $b = a$. We will see that it stays true when $b$ increases. The quantity $s(b)$ can change only when $b$ goes through one of the roots of the $A_i$, which are finite in number. Let $x$ be a root of such an $A_i$ (maybe of several). If $\epsilon$ is sufficiently small, when $b$ goes from $x-\epsilon$ to $x$, the sign sequence corresponding to indices $i-1$, $i$ and $i+1$ goes from $+$, $\pm$, $-$ to $+$, $0$, $-$ (or from $-$, $\pm$, $+$ to $-$, $0$, $+$) when $i \geq 1$ by what has been said above (no consecutive zeros, and no sequences $+$, $0$, $+$ or $-$, $0$, $-$). Hence, there is no difference in the number of sign changes not counting zeros if $i \geq 1$. On the other hand, for $i = 0$, the sign sequence corresponding to indices 0 and 1 1 goes from $+,-$ to $0,-$, or from $-,+$ to $0,+$ since $A_1(b) < 0$ if and only if $A_0$ is decreasing (recall that $A_1$ is the derivative of $A_0$). Hence, the net change in $s(b)$ is equal to $-1$. This proves our claim and the theorem. □

From this, it is easy to derive an algorithm for computing the signature of a polynomial (hence of a number field). Such an algorithm can of course be written for any polynomial $T \in \mathbb{R}[X]$, but for number-theoretic uses $T$ will have integer coefficients, hence we should use the polynomial remainder sequence given by the sub-resultant Algorithm 3.3.1 to avoid coefficient explosion. This leads to the following algorithm.

**Algorithm 4.1.11** (Sturm). Given a polynomial $T \in \mathbb{Z}[X]$, this algorithm determines the signature $(r_1, r_2)$ of $T$ using Sturm's theorem and the sub-resultant Algorithm 3.3.1. If $T$ is not squarefree, it outputs an error message.

1. [Initializations and reductions] If $\deg(T) = 0$, output $(0,0)$ and terminate. Otherwise, set $A \leftarrow \mathrm{pp}(T)$, $B \leftarrow \mathrm{pp}(T')$, $g \leftarrow 1$, $h \leftarrow 1$, $s \leftarrow \mathrm{sign}(\ell(A))$, $n \leftarrow \deg(A)$, $t \leftarrow (-1)^{n-1}s$, $r_1 \leftarrow 1$.
2. [Pseudo division] Set $\delta \leftarrow \deg(A) - \deg(B)$. Using Algorithm 3.1.2, compute $R$ such that $\ell(B)^{\delta+1}A = BQ + R$. If $R = 0$ then $T$ was not squarefree, output

an error message and terminate the algorithm. Otherwise, if $\ell(B)>0$ or $\delta$ is odd, set $R \leftarrow -R$.

3. [Use Sturm] If $\mathrm{sign}(\ell(R)) \neq s$, set $s \leftarrow -s$, $r_1 \leftarrow r_1-1$. Then, if $\mathrm{sign}(\ell(R)) \neq (-1)^{\deg(R)}t$, set $t \leftarrow -t$, $r_1 \leftarrow r_1+1$.

4. [Finished?] If $\deg(R)=0$, output $(r_1,(n-r_1)/2)$ and terminate the algorithm. Otherwise, set $A \leftarrow B$, $B \leftarrow R/(gh^\delta)$, $g \leftarrow |\ell(A)|$, $h \leftarrow h^{1-\delta}g^\delta$, and go to step 2.

Another important notion concerning number fields is that of the Galois group of a number field. From now on, we assume that all our number fields are subfields of $\overline{\mathbb{Q}}$.

**Definition 4.1.12.** *Let $K$ be a number field of degree $n$. We say that $K$ is* Galois *(or* normal*) over $\mathbb{Q}$, or simply Galois, if $K$ is (globally) invariant by the $n$ embeddings of $K$ in $\mathbb{C}$. The set of such embeddings is a group, called the Galois group of $K$, and denoted* $\mathrm{Gal}(K/\mathbb{Q})$.

Given any number field $K$, the intersection of all subfields of $\overline{\mathbb{Q}}$ which are Galois and contain $K$ is a finite extension $K^s$ of $K$ called the Galois closure (or normal closure) of $K$ in $\overline{\mathbb{Q}}$. If $K=\mathbb{Q}(\theta)$ where $\theta$ is a root of an irreducible polynomial $T \in \mathbb{Z}[X]$, the Galois closure of $K$ can also be obtained as the splitting field of $T$, i.e. the field obtained by adjoining to $\mathbb{Q}$ all the roots of $T$. By abuse of language, even when $K$ is not Galois, we will call $\mathrm{Gal}(K^s/\mathbb{Q})$ the Galois group of the number field $K$ (or of the polynomial $T$).

A special case of the so-called "fundamental theorem of Galois theory" is as follows.

**Proposition 4.1.13.** *Let $K$ be Galois over $\mathbb{Q}$ and $x \in K$. Assume that for any $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$ we have $\sigma(x)=x$. Then $x \in \mathbb{Q}$. In particular, if in addition $x$ is an algebraic integer then $x \in \mathbb{Z}$.*

The following easy proposition shows that there are only two possibilities for the signature of a Galois extensions. Similarly, we will see (Theorem 4.8.6) that there are only a few possibilities for how primes split in a Galois extension.

**Proposition 4.1.14.** *Let $K$ be a Galois extension of $\mathbb{Q}$ of degree $n$. Then, either $K$ is totally real ($(r_1,r_2)=(n,0)$), or $K$ is totally complex ($(r_1,r_2)=(0,n/2)$ which can occur only if $n$ is even).*

The computation of the Galois group of a number field (or of its Galois closure) is in general not an easy task. We will study this for polynomials of low degree in Section 6.3.

## 4.2 Representation and Operations on Algebraic Numbers

It is very important to study the way in which algebraic numbers are represented. There are two completely different problems: that of representing algebraic numbers, and that of representing *sets* of algebraic numbers, e.g. modules or ideals. This will be considered in Section 4.7. Here we consider the problem of representing an individual algebraic number.

Essentially there are four ways to do this, depending on how the number arises. The first way is to represent $\alpha \in \overline{\mathbb{Q}}$ by its minimal polynomial $A$ which exists by Proposition 4.1.2. The three others assume that $\alpha$ is a polynomial with rational coefficients in some fixed algebraic number $\theta$. These other methods are usually preferable, since field operations in $\mathbb{Q}(\theta)$ can be performed quite simply. We will see these methods in more detail in the following sections. However, to start with, we do not always have such a $\theta$ available, so we consider the problems which arise from the first method.

### 4.2.1 Algebraic Numbers as Roots of their Minimal Polynomial

Since $A$ has $n = \deg(A)$ zeros in $\mathbb{C}$, the first question is to determine which of these zeros $\alpha$ is supposed to represent. We have seen that an algebraic number always comes equipped with all of its conjugates, so this is a problem which we must deal with. Since $\mathbb{Q}(\alpha) \simeq \mathbb{Q}[X]/(A(X)\mathbb{Q}[X])$, $\alpha$ may be represented as the class of $X$ in $\mathbb{Q}[X]/(A(X)\mathbb{Q}[X])$, which is a perfectly well defined mathematical quantity. The distinction between $\alpha$ and its conjugates, if really necessary, will then depend not on $A$ but on the specific embedding of $\mathbb{Q}[X]/(A(X)\mathbb{Q}[X])$ in $\mathbb{C}$. In other words, it depends on the numerical value of $\alpha$ as a complex number. This numerical value can be obtained by finding complex roots of polynomials, and we assume throughout that we always take sufficient accuracy to be able to distinguish $\alpha$ from its conjugates. (Recall that since the minimal polynomial of $\alpha$ is irreducible and hence squarefree, the conjugates of $\alpha$ are distinct.)

Hence, we can consider that an algebraic number $\alpha$ is represented by a pair $(A, x)$ where $A$ is the minimal polynomial of $\alpha$, and $x$ is an approximation to the complex number $\alpha$ ($x$ should be at least closer to $\alpha$ than to any of its conjugates). It is also useful to have numeric approximations to all the conjugates of $\alpha$. In fact, one can recover the minimal polynomial $A$ of $\alpha$ from this if one knows only its leading term $\ell(A)$, since if one sets $\tilde{A}(X) = \ell(A)\prod_i(X - \tilde{\alpha}_i)$, where the $\tilde{\alpha}_i$ are the approximations to the conjugates of $\alpha$, then, if they are close enough (and they must be chosen so), $A$ will be the polynomial whose coefficients are the nearest integers to the coefficients of $\tilde{A}$.

With this representation, it is clear that one can now easily work in the subfield $\mathbb{Q}(\alpha)$ generated by $\alpha$, simply by working modulo $A$.

More serious problems arise when one wants to do operations between algebraic numbers which are a priori not in this subfield. Assume for instance

that $\alpha = (X \bmod A(X))$, and $\beta = (X \bmod B(X))$, where $A$ and $B$ are primitive irreducible polynomials of respective degrees $m$ and $n$ (we omit the $\mathbb{Q}[X]$ for simplicity of notation). How does one compute the sum, difference, product and quotient of $\alpha$ and $\beta$? The simplest way to do this is to compute *resultants* of two variable polynomials. Indeed, the resultant of the polynomials $A(X-Y)$ and $B(Y)$ considered as polynomials in $Y$ alone (the coefficient ring being then $\mathbb{Q}[X]$) is up to a scalar factor equal to $P(X) = \prod_{i,j}(X - \alpha_i - \beta_j)$ where the $\alpha_i$ are the conjugates of $\alpha$, and the $\beta_j$ are the conjugates of $\beta$. Since $P$ is a resultant, it has coefficients in $\mathbb{Q}[X]$, and $\alpha+\beta$ is one of its roots, so $Q = \mathrm{pp}(P)$ is a multiple of the minimal polynomial of $\alpha + \beta$.

If $Q$ is irreducible, then it is the minimal polynomial of $\alpha + \beta$. If it is not irreducible, then the minimal polynomial of $\alpha + \beta$ is one of the irreducible factors of $Q$ which one computes by using the algorithms of Section 3.5. Once again however, it does not make sense to ask which of the irreducible factors $\alpha + \beta$ is a root of, if we do not specify embeddings in $\mathbb{C}$, in other words, numerical approximations to $\alpha$ and $\beta$. Given such approximations however, one can readily check in practice which of the irreducible factors of $Q$ is the minimal polynomial that we are looking for.

What holds for addition also holds for subtraction (take the resultant of $A(X + Y)$ and $B(Y)$), multiplication (take the resultant of $Y^m A(X/Y)$ and $B(Y)$), and division (take the resultant of $A(XY)$ with $B(Y)$).

### 4.2.2 The Standard Representation of an Algebraic Number

Let $K$ be a number field, and let $\theta_j$ $(1 \le j \le n)$ be a $\mathbb{Q}$-basis of $K$. Let $\alpha \in K$ be any element. It is clear that one can write $\alpha$ in a unique way as

$$\alpha = \frac{\sum_{j=0}^{n-1} a_j \theta_{j+1}}{d}, \text{ with } d > 0, \ a_j \in \mathbb{Z} \text{ and } \gcd(a_0, \ldots, a_{n-1}, d) = 1.$$

In the case where $\theta_j = \theta^{j-1}$ for some root $\theta$ of a monic irreducible polynomial $T \in \mathbb{Z}[X]$, the $(n+1)$-uplet $(a_0, \ldots, a_{n-1}, d) \in \mathbb{Z}^{n+1}$ will be called the *standard representation* of $\alpha$ (with respect to $\theta$). Hence, we can now assume that we know such a primitive element $\theta$. (We will see in Section 4.5 how it can be obtained.)

We must see how to do the usual arithmetic operations on these standard representations. The vector space operations on $K$ are of course trivial. For multiplication, we precompute the standard representation of $\theta^j$ for $j \le 2n-2$ in the following way: if $T(X) = \sum_{i=0}^{n} t_i X^i$ with $t_i \in \mathbb{Z}$ for all $i$ and $t_n = 1$, we have $\theta^n = \sum_{i=0}^{n-1}(-t_i)\theta^i$. If we set $\theta^{n+k} = \sum_{i=0}^{n-1} r_{k,i}\theta^i$, then the standard representation of $\theta^{n+k}$ is $(r_{k,0}, r_{k,1}, \ldots, r_{k,n-1}, 1)$ and the $r_{k,i}$ are computed by induction thanks to the formulas $r_{0,i} = -t_i$ and

$$r_{k+1,i} = \begin{cases} r_{k,i-1} - t_i r_{k,n-1} & \text{if } i \ge 1, \\ -t_0 r_{k,n-1} & \text{if } i = 0. \end{cases}$$

Now if $(a_0, \ldots, a_{n-1}, d)$ and $(b_0, \ldots, b_{n-1}, e)$ are the standard representations of $\alpha$ and $\beta$ respectively, then it is clear that

$$\alpha\beta = \frac{\sum_{k=0}^{2n-2} c_k \theta^k}{de}, \qquad \text{where } c_k = \sum_{i+j=k} a_i b_j,$$

hence

$$\alpha\beta = \frac{\sum_{k=0}^{n-1} z_k \theta^k}{de}, \qquad \text{where } z_k = c_k + \sum_{i=0}^{n-2} r_{k,i} c_{n+i}.$$

The standard representation of $\alpha\beta$ is then obtained by dividing all the $z_k$ and $de$ by $\gcd(z_0, \ldots, z_{n-1}, de)$.

Note that if we set $A(X) = \sum_{i=0}^{n-1} a_i X^i$ and $B(X) = \sum_{i=0}^{n-1} b_i X^i$, the procedure described above is equivalent to computing the remainder in the Euclidean division of $AB$ by $T$. Because of the precomputations of the $r_{k,i}$, however, it is slightly more efficient.

The problem of division is more difficult. Here, we need essentially to compute $A/B$ modulo the polynomial $T$. Hence, we need to invert $B$ modulo $T$. The simplest efficient way to do this is to use the sub-resultant Algorithm 3.3.1 to obtain $U$ and $V$ (which does not need to be computed explicitly) such that $UB + VT = d$ where $d$ is a constant polynomial. (Note that since $T$ is irreducible and $B \neq 0$, $B$ and $T$ are coprime.) Then the inverse of $B$ modulo $T$ is $\frac{1}{d}U$, and the standard representation of $\alpha/\beta$ can easily be obtained from this.

### 4.2.3 The Matrix (or Regular) Representation of an Algebraic Number

A third way to represent algebraic numbers is by the use of integral matrices. If $\theta_j$ $(1 \leq j \leq n)$ is a $\mathbb{Q}$-basis of $K$ and if $\alpha \in K$, then multiplication by $\alpha$ is an endomorphism of the $\mathbb{Q}$-vector space $K$, and we can represent $\alpha$ by the matrix $M_\alpha$ of this endomorphism in the basis $\theta_j$. This will be a matrix with rational entries, hence one can write $M_\alpha = M'/d$ where $M'$ has integral entries, $d$ is a positive integer, and the greatest common divisor of all the entries of $M'$ is coprime to $d$. This representation is of course unique, and it is clear that the map $\alpha \mapsto M_\alpha$ is an algebra homomorphism from $K$ to the algebra of $n \times n$ matrices over $\mathbb{Q}$. Thus one can compute on algebraic numbers simply by computing with the corresponding matrices. The running time is usually longer however, since more elements are involved. For example, the simple operation of addition takes $O(n^2)$ operations, while it clearly needs only $O(n)$ operations in the standard representation. The matrix representation is clearly more suited for multiplication and division. (Division is performed using the remark following Algorithm 2.2.2.)

### 4.2.4 The Conjugate Vector Representation of an Algebraic Number

The last method of representing an algebraic number $\alpha$ in a number field $K = \mathbb{Q}(\theta)$ that I want to mention, is to represent $\alpha$ by numerical approximations to its conjugates, repeated with multiplicity. More precisely, let $\sigma_j$ be the $n = \deg(K)$ distinct embeddings of $K$ in $\mathbb{C}$, ordered in the following standard way: $\sigma_1, \ldots, \sigma_{r_1}$ are the real embeddings, $\sigma_{r_1+r_2+i} = \bar{\sigma}_{r_1+i}$ for $1 \leq i \leq r_2$. If $\alpha = \sum_{i=0}^{n-1} a_i \theta^i$, then

$$\sigma_j(\alpha) = \sum_{i=0}^{n-1} a_i \sigma_j(\theta)^i,$$

and the $\sigma_j(\alpha)$ are the conjugates of $\alpha$, but in a specific order (corresponding to the choice of the ordering on the $\sigma_j$), and repeated with a constant multiplicity $n/\deg(\alpha)$. We can then represent $\alpha$ as the $(r_1+r_2)$-uplet of complex numbers

$$(\sigma_1(\alpha), \ldots, \sigma_{r_1+r_2}(\alpha)),$$

where the complex numbers $\sigma_j(\alpha)$ are given by a sufficiently good approximation. Operations on this representation are quite trivial since they are done componentwise. In particular, division, which was difficult in the other representations, becomes very simple here. Unfortunately, there is a price to pay: one must be able to go back to one of the exact representations (for example to the standard representation), and hence have good control on the roundoff errors.

For this, we precompute the inverse matrix of the matrix $\Theta = \sigma_i(\theta^{j-1})$. Then, if one knows the conjugate representation of a number $\alpha$, and an integer $d$ such that $d\alpha \in \mathbb{Z}[\theta]$, one can write $\alpha = (\sum_{j=1}^{n} a_{j-1}\theta^{j-1})/d$ where the $a_j$ are integers, and the column vector $(a_0, \ldots, a_{n-1})^t$ can be obtained as the product $d\Theta^{-1}(\sigma_1(\alpha), \ldots, \sigma_n(\alpha))^t$, and since the $a_i$ are integers, if the roundoff errors have been controlled and are not too large, this gives the $a_i$ exactly (note that in practice one can work with matrices over $\mathbb{R}$ and not over $\mathbb{C}$. The details are left to the reader).

In practice, one can ignore roundoff errors and start with quite precise numerical approximations. Then every operation except division is done using the standard representation, while for division one computes the conjugate representation of the result, converts back, and then check by exact multiplication that the roundoff errors did not accumulate to give us a wrong result. (If they did, this means that one must work with a higher precision.)

## 4.3 Trace, Norm and Characteristic Polynomial

If $\alpha$ is an algebraic number, the trace (resp. the norm) of $\alpha$ is by definition the sum (resp. the product) of the conjugates of $\alpha$. If $A(X) = \sum_{i=0}^{m} a_i X^i$ is its minimal polynomial, then we clearly have

$$\mathrm{Tr}(\alpha) = -\frac{a_{m-1}}{a_m} \qquad \text{and } \mathcal{N}(\alpha) = (-1)^m \frac{a_0}{a_m} \ ,$$

where Tr and $\mathcal{N}$ denote the trace and norm of $\alpha$ respectively. Usually however, $\alpha$ is considered as an element of a number field $K$. If $K = \mathbb{Q}(\alpha)$, then the definitions above are OK, but if $\mathbb{Q}(\alpha) \subsetneq K$, then it is necessary to modify the definitions so that Tr becomes additive and $\mathcal{N}$ multiplicative. More generally, we put:

**Definition 4.3.1.** *Let $K$ be a number field of degree $n$ over $\mathbb{Q}$, and let $\sigma_i$ be the $n$ distinct embeddings of $K$ in $\mathbb{C}$.*

(1) *The* characteristic polynomial *$C_\alpha$ of $\alpha$ in $K$ is*

$$C_\alpha(X) = \prod_{1 \le i \le n} (X - \sigma_i(\alpha)).$$

(2) *If we set*

$$C_\alpha(X) = \sum_{0 \le i \le n} (-1)^{n-i} s_{n-i}(\alpha) X^i,$$

*then $s_k(\alpha)$ is a rational number and will be called the $k^{\text{th}}$* symmetric function *of $\alpha$ in $K$.*

(3) *In particular, $s_1(\alpha)$ is called the* trace *of $\alpha$ in $K$ and denoted $\mathrm{Tr}_{K/\mathbb{Q}}(\alpha)$, and similarly $s_n(\alpha)$ is called the* norm *of $\alpha$ in $K$ and denoted $\mathcal{N}_{K/\mathbb{Q}}(\alpha)$.*

As has already been mentioned, one must be careful to distinguish the absolute trace of $\alpha$ which we have denoted $\mathrm{Tr}(\alpha)$ from the trace of $\alpha$ in the field $K$, denoted $\mathrm{Tr}_{K/\mathbb{Q}}(\alpha)$, and similarly with the norms. More precisely, we have the following proposition:

**Proposition 4.3.2.** *Let $K$ be a number field of degree $n$, $\sigma_i$ the $n$ distinct embeddings of $K$ in $\mathbb{C}$.*

(1) *If $\alpha \in K$ has degree $m$ (hence with $m$ dividing $n$), we have*

$$\mathrm{Tr}_{K/\mathbb{Q}}(\alpha) = \sum_{1 \le i \le n} \sigma_i(\alpha) = \frac{n}{m} \mathrm{Tr}(\alpha),$$

*and*

$$\mathcal{N}_{K/\mathbb{Q}}(\alpha) = \prod_{1 \le i \le n} \sigma_i(\alpha) = (\mathcal{N}(\alpha))^{n/m}.$$

(2) *For any $\alpha$ and $\beta$ in $K$ we have*

$$\mathrm{Tr}_{K/\mathbb{Q}}(\alpha+\beta) = \mathrm{Tr}_{K/\mathbb{Q}}(\alpha) + \mathrm{Tr}_{K/\mathbb{Q}}(\beta),$$

*and*

$$\mathcal{N}_{K/\mathbb{Q}}(\alpha\beta) = \mathcal{N}_{K/\mathbb{Q}}(\alpha)\mathcal{N}_{K/\mathbb{Q}}(\beta).$$

(3) *$\alpha$ is an algebraic integer if and only if $s_k(\alpha) \in \mathbb{Z}$ for all $k$ such that $1 \le k \le n$ (note that $s_0(\alpha) = 1$).*

As usual, we must find algorithms to compute traces, norms and more generally characteristic polynomials of algebraic numbers. Since we have seen four different representations of algebraic numbers (viz. by a minimal polynomial, by the standard representation, by the matrix representation and by the conjugate vector representation), there are at least that many methods to do the job. We will only sketch these methods, except when they involve fundamentally new ideas. We always assume that our number field is given as $K = \mathbb{Q}(\theta)$ where $\theta$ is an algebraic integer whose monic minimal polynomial of degree $n$ is denoted $T(X)$. We denote by $\sigma_i$ the $n$ embeddings of $K$ in $\mathbb{C}$.

In the case where $\alpha$ is represented by its minimal polynomial $A(X)$, then each of the $m = \deg(A)$ embeddings of $\mathbb{Q}(\alpha)$ in $\mathbb{C}$ lifts to exactly $n/m$ embeddings among the $\sigma_i$, hence it easily follows that

$$C_\alpha(X) = A(X)^{n/m},$$

and this immediately implies Proposition 4.3.2 (1), i.e. if we write $A(X) = \sum_{0\le i\le m} a_i X^i$, then

$$\mathrm{Tr}_{K/\mathbb{Q}}(\alpha) = -\frac{na_{m-1}}{ma_m}, \qquad \mathcal{N}_{K/\mathbb{Q}}(\alpha) = (-1)^n\left(\frac{a_0}{a_m}\right)^{n/m}.$$

In the case where $\alpha$ is given by its standard representation

$$\alpha = \frac{1}{d}\left(\sum_{0\le i\le n-1} a_i\theta^i\right),$$

the only symmetric function which is relatively easy to compute is the trace, since we can precompute the trace of $\theta^i$ using Newton's formulas as follows.

**Proposition 4.3.3.** *Let $\theta_i$ be the roots (repeated with multiplicity) of a monic polynomial $T(X) = \sum_{0\le i\le n} t_i X^i \in \mathbb{C}[X]$ of degree $n$ and set $S_k = \sum_i(\theta_i^k)$. Then*

$$S_k = -kt_{n-k} - \sum_{i=1}^{k-1} t_{n-i}S_{k-i} \qquad \textit{(where we set } t_i = 0 \textit{ for } i < 0).$$

This result is well known and its proof is left to the reader (Exercise 3).

We can however compute all the symmetric functions, i.e. the characteristic polynomial, by using resultants, as follows.

**Proposition 4.3.4.** *Let $K = \mathbb{Q}(\theta)$ be a number field where $\theta$ is a root of a monic irreducible polynomial $T(X) \in \mathbb{Z}[X]$ of degree $n$, and let*

$$\alpha = \frac{1}{d}\left(\sum_{0 \le i \le n-1} a_i \theta^i\right)$$

*be the standard representation of some $\alpha \in K$. Set $A(X) = \sum_{0 \le i \le n-1} a_i X^i$.*

*Then the characteristic polynomial $C_\alpha(X)$ of $\alpha$ is given by the formula*

$$C_\alpha(X) = d^{-n} R_Y(T(Y), dX - A(Y)),$$

*where $R_Y$ denotes the resultant taken with respect to the variable $Y$. In particular, we have*

$$\mathcal{N}_{K/\mathbb{Q}}(\alpha) = d^{-n} R(T(X), A(X)).$$

Proof. We have by definition

$$\begin{aligned} C_\alpha(X) &= \prod_i (X - \sigma_i(\alpha)) = \prod_i (X - A(\sigma_i(\theta))/d) \\ &= d^{-n} \prod_i (dX - A(\theta_i)) = d^{-n} R_Y(T(Y), dX - A(Y)), \end{aligned}$$

where the $\theta_i$ are the conjugates of $\theta$, i.e. the roots of $T$. The formula for the norm follows immediately on setting $X = 0$. □

Since the resultant can be computed efficiently by the sub-resultant Algorithm 3.3.7, used here in the UFD's $\mathbb{Z}[X]$ and $\mathbb{Z}$, we see that this proposition gives an efficient way to compute the characteristic polynomial and the norm of an algebraic number given in its standard representation.

In the case where $\alpha$ is given by numerical approximations to its conjugates, as usual we also assume that we know an integer $d$ such that $d\alpha \in \mathbb{Z}[\theta]$. Then we can compute numerically $\prod_i (X - d\sigma_i(\alpha))$, and this must have integer coefficients. Hence, if we have sufficient control on the roundoff errors and sufficient accuracy on the conjugates of $\alpha$, this enables us to compute $C_{d\alpha}(X)$ exactly, hence $C_\alpha(X) = d^{-n} C_{d\alpha}(dX)$.

Finally, we consider the case where $\alpha$ is given by its matrix representation $M_\alpha$ in the basis $1, \theta, \ldots, \theta^{n-1}$, where $dM_\alpha$ has integral coefficients for some integer $d$. Then the characteristic polynomial of $\alpha$ is simply equal to the characteristic polynomial of $M_\alpha$ (meaning always $\det(XI_n - M_\alpha)$). In particular,

the trace can be read off trivially on the diagonal coefficients, and the norm is, up to sign, equal to the determinant of $M_\alpha$.

The characteristic polynomial can be computed using one of the algorithms described Section 2.2.4, and the determinant using Algorithm 2.2.6.

In practice, it is not completely clear which representation is preferable. A reasonable choice is probably to use the standard representation and the sub-resultant algorithm. This depends on the context however, and one should always be aware of each of the four possibilities to handle algebraic numbers. Keep in mind that it is usually costly to go from one representation to another, so for a given problem the representation should be fixed.

## 4.4 Discriminants, Integral Bases and Polynomial Reduction

### 4.4.1 Discriminants and Integral Bases

We have the following basic result.

**Proposition 4.4.1.** *Let $K$ be a number field of degree $n$, $\sigma_i$ be the $n$ embeddings of $K$ in $\mathbb{C}$, and $\alpha_j$ be a set of $n$ elements of $K$. Then we have*

$$\det(\sigma_i(\alpha_j))^2 = \det(\mathrm{Tr}_{K/\mathbb{Q}}(\alpha_i\alpha_j)).$$

*This quantity is a rational number and is called the* discriminant *of the $\alpha_i$, and denoted $d(\alpha_1,\dots,\alpha_n)$. Furthermore, $d(\alpha_1,\dots,\alpha_n) = 0$ if and only if the $\alpha_j$ are $\mathbb{Q}$-linearly dependent.*

*Proof.* Consider the $n \times n$ matrix $M = (\sigma_i(\alpha_j))$. Then by definition of matrix multiplication, we have $M^tM = (a_{i,j})$ with

$$a_{i,j} = \sum_k \sigma_k(\alpha_i)\sigma_k(\alpha_j) = \mathrm{Tr}_{K/\mathbb{Q}}(\alpha_i\alpha_j).$$

Since $\det(M^t) = \det(M)$ the equality of the proposition follows. Since $\mathrm{Tr}_{K/\mathbb{Q}}(\alpha) \in \mathbb{Q}$ the discriminant is a rational number. If the $\alpha_j$ are $\mathbb{Q}$-linearly dependent, it is clear that the columns of the matrix $M$ are also (since $\mathbb{Q}$ is invariant by the $\sigma_i$). Therefore the discriminant is equal to 0. Conversely assume that the discriminant is equal to 0. This means that the kernel of the matrix $M^tM$ is non-trivial, and since this matrix has coefficients in $\mathbb{Q}$, there exists $\lambda_i \in \mathbb{Q}$ such that for every $j$, $\mathrm{Tr}(x\alpha_j) = 0$ where we have set $x = \sum_{1\le i\le n} \lambda_i\alpha_i$. If the $\alpha_j$ were linearly independent over $\mathbb{Q}$, they would generate $K$ as a $\mathbb{Q}$-vector space, and so we would have $\mathrm{Tr}(xy) = 0$ for *all* $y \in K$ with $x \neq 0$. Taking $y = 1/x$ gives $\mathrm{Tr}(1) = n = 0$, a contradiction, thus showing the proposition. □

**Remark.** We have just proved that the quadratic form $\mathrm{Tr}(x^2)$ is non-degenerate on $K$ using that $K$ is of characteristic zero (otherwise $n = 0$ may not be a contradiction). This is the definition of a *separable extension.* It is not difficult to show (see for example Proposition 4.8.11 or Exercise 5) that the signature of this quadratic form (i.e. the number of positive and negative squares after Gaussian reduction) is equal to $(r_1 + r_2, r_2)$ where as usual $(r_1, r_2)$ is the signature of the number field $K$.

Recall that we denote by $\mathbb{Z}_K$ the ring of (algebraic) integers of $K$. Then we also have:

**Theorem 4.4.2.** *The ring $\mathbb{Z}_K$ is a free $\mathbb{Z}$-module of rank $n = \deg(K)$. This is true more generally for any non-zero ideal of $\mathbb{Z}_K$.*

*Proof (Sketch).* Let $\alpha_j$ be a basis of $K$ as a $\mathbb{Q}$-vector space. Without loss of generality, we can assume that the $\alpha_j$ are algebraic integers. If $A$ is the (free) $\mathbb{Z}$-module generated by the $\alpha_j$, we clearly have $A \subset \mathbb{Z}_K$, and the formula $M^{-1} = M^{\mathrm{adj}}/\det(M)$ for the inverse of a matrix (see section 2.2.4) shows that $d\mathbb{Z}_K \subset A$, where $d$ is the discriminant of the $\alpha_j$, whence the result. (Recall that a sub-$\mathbb{Z}$-module of a free module of rank $n$ is a free module of rank less than or equal to $n$, since $\mathbb{Z}$ is a principal ideal domain, see Theorem 2.4.1.) □

It is important to note that $\mathbb{Z}$ being a PID is crucial in the above proof. Hence, if we consider *relative* extensions, Theorem 4.4.2 will a priori be true only if the base ring is also a PID, and this is not always the case.

**Definition 4.4.3.** *A $\mathbb{Z}$-basis of the free module $\mathbb{Z}_K$ will be called an* integral basis *of $K$. The discriminant of an integral basis is independent of the choice of that basis, and is called the* discriminant *of the field $K$ and is denoted by $d(K)$.*

Note that, although the two notions are closely related, the discriminant of $K$ is not in general equal to the discriminant of an irreducible polynomial defining $K$. More precisely:

**Proposition 4.4.4.** *Let $T$ be a monic irreducible polynomial of degree $n$ in $\mathbb{Z}[X]$, $\theta$ a root of $T$, and $K = \mathbb{Q}(\theta)$. Denote by $d(T)$ (resp. $d(K)$) the discriminant of the polynomial $T$ (resp. of the number field $K$).*

(1) *We have $d(1, \theta, \ldots, \theta^{n-1}) = d(T)$.*
(2) *If $f = [\mathbb{Z}_K : \mathbb{Z}[\theta]]$, we have*

$$d(T) = d(K)f^2$$

*and, in particular, $d(T)$ is a square multiple of $d(K)$.*

The proof of this is easy and left to the reader. The number $f$ will be called the *index* of $\theta$ in $\mathbb{Z}_K$.

**Proposition 4.4.5.** *The algebraic numbers $\alpha_1, \ldots, \alpha_n$ form an integral basis if and only if they are algebraic integers and if $d(\alpha_1, \ldots, \alpha_n) = d(K)$, where $d(K)$ is the discriminant of $K$.*

*Proof.* If $M$ is the matrix expressing the $\alpha_i$ on some integral basis of $K$, it is clear that $d(\alpha_1, \ldots, \alpha_n) = d(K)\det(M)^2$ and the proposition follows. □

We also have the following result due to Stickelberger:

**Proposition 4.4.6.** *Let $\alpha_1, \ldots, \alpha_n$ be algebraic integers. Then*

$$d(\alpha_1, \ldots, \alpha_n) \equiv 0 \text{ or } 1 \pmod 4.$$

*Proof.* If we expand the determinant $\det(\sigma_i(\alpha_j))$ using the $n!$ terms, we will get terms with a plus sign corresponding to permutations of even signature, and terms with a minus sign. Hence, collecting these terms separately, we can write the determinant as $P - N$ hence

$$d(\alpha_1, \ldots, \alpha_n) = (P - N)^2 = (P + N)^2 - 4PN.$$

Now clearly $P+N$ and $PN$ are symmetric functions of the $\alpha_i$, hence by Galois theory they are in $\mathbb{Q}$ and in fact in $\mathbb{Z}$ since the $\alpha_i$ are algebraic integers. This proves the proposition, since a square is always congruent to 0 or 1 mod 4. □

The determination of an explicit integral basis and of the discriminant of a number field is not an easy problem, and is one of the main tasks of this course. There is, however one case in which the result is trivial:

**Corollary 4.4.7.** *Let $T$ be a monic irreducible polynomial in $\mathbb{Z}[X]$, $\theta$ a root of $T$, and $K = \mathbb{Q}(\theta)$. Assume that the discriminant of $T$ is squarefree or is equal to $4d$ where $d$ is squarefree and not congruent to 1 modulo 4. Then the discriminant of $K$ is equal to the discriminant of $T$, and an integral basis of $K$ is given by $1, \theta, \ldots, \theta^{n-1}$.*

Since a discriminant must be congruent to 0 or 1 mod 4, this immediately follows from the above propositions. □

Unfortunately, this corollary is not of much use, since it is quite rare that the condition on the discriminant of $T$ is satisfied. We will see in Chapter 6 a complete method for finding an integral basis and hence the discriminant of a number field.

Finally, we note without proof the following consequence of the so-called "conductor-discriminant formula".

**Proposition 4.4.8.** *Let $K$ and $L$ be number fields with $K \subset L$. Then*

$$d(K)^{[L:K]} \mid d(L).$$

**Corollary 4.4.9.** *Let $K = \mathbb{Q}(\alpha)$ and $L = \mathbb{Q}(\beta)$ be two number fields, let $m = \deg(K)$, $n = \deg(L)$, $A(X)$ (resp. $B(X)$) the minimal monic polynomial of $\alpha$ (resp. $\beta$). Write $d(A)$ and $d(B)$ for the discriminants of the polynomials $A$ and $B$. Assume that $K$ is conjugate to a subfield of $L$. Then if $p$ is a prime such that $v_p(d(A))$ is odd, we must have $p^{n/m} \mid d(B)$.*

*Proof.* By Proposition 4.4.4 if $v_p(d(A))$ is odd then $p \mid d(K)$, where $d(K)$ is the discriminant of the field $K$. By the proposition we therefore have $p^{n/m} \mid d(L) \mid d(B)$, thus proving the corollary. □

### 4.4.2 The Polynomial Reduction Algorithm

We will see in Section 4.5 that it is usually not always easy to decide whether two number fields are isomorphic or not. Here we will give a heuristic approach based on the LLL algorithm and ideas of Diaz y Diaz and the author which often gives a useful answer to the following problem: given a number field $K$, can one find a monic irreducible polynomial defining $K$ which in a certain sense is as simple as possible.

Of course, if this could be done, the isomorphism problem would be completely solved. We will see in Chapters 5 and 6 that it is possible to do this for quadratic fields (in fact it is trivial in that case), and for certain classes of cubic fields, like cyclic cubic fields or pure cubic fields (see Section 6.4). In general, all one can hope for in practice is to find, maybe not *the* simplest, but *a* simple polynomial defining $K$.

A natural criterion of simplicity would be to take polynomials whose largest coefficients are as small as possible in absolute value (i.e. the $L^\infty$ norm on the coefficients), or such that the sum of the squares of the coefficients is as small as possible (the $L^2$ norm). Unfortunately, I know of no really efficient way of finding simple polynomials in this sense.

What we will in fact consider is the following "norm" on polynomials.

**Definition 4.4.10.** *Let $P \in \mathbb{C}[X]$, and let $\alpha_i$ be the complex roots of $P$ repeated with multiplicity. We define the* size *of $P$ by the formula*

$$\text{size}(P) = \sum_i |\alpha_i|^2.$$

This is not a norm in the usual mathematical sense, but it seems reasonable to say that if the size (in this sense) of a polynomial is not large, then the polynomial is simple, and its coefficients should not be too large.

More precisely, we can show (see Exercise 6) that if $P = \sum_{k=0}^{n} a_k X^k$ is a monic polynomial and if $S = \text{size}(P)$, then

$$|a_{n-k}| \leq \binom{n}{k}\left(\frac{S}{n}\right)^{k/2}.$$

Hence, the size of $P$ is related to the size of

$$\max |a_{n-k}|^{2/k}.$$

The reason we take this definition instead of an $L^p$ definition on the coefficients is that we can apply the LLL algorithm to find a polynomial of small size which defines the same number field $K$ as the one defined by a given polynomial $P$, while I do not know how to achieve this for the norms on the coefficients.

The method is as follows. Let $K$ be defined by a monic irreducible polynomial $P \in \mathbb{Z}[X]$. Using the round 2 Algorithm 6.1.8 which will be explained in Chapter 6, we compute an integral basis $\omega_1, \ldots, \omega_n$ of $\mathbb{Z}_K$. Furthermore, let $\sigma_j$ denote the $n$ isomorphisms of $K$ into $\mathbb{C}$. If we set

$$x = \sum_{i=1}^{n} x_i \omega_i$$

where the $x_i$ are in $\mathbb{Z}$, then $x$ is an arbitrary algebraic integer in $K$, hence its characteristic polynomial $M_x$ will be of the form $P_d^{n/d}$ where $P_d$ is the minimal polynomial of $x$ and $d$ the degree of $x$, and $P_d$ defines a subfield of $K$. In particular, when $d = n$, this defines an equation for $K$, and clearly all monic equations for $K$ with integer coefficients (as well as for subfields of $K$) are obtained in this way.

Now we have by definition

$$M_x(X) = \prod_{k=1}^{n}\left(X - \sum_{i=1}^{n} x_i \sigma_k(\omega_i)\right)$$

hence,

$$\text{size}(M_x) = \sum_{k=1}^{n}\left|\sum_{i=1}^{n} x_i \sigma_k(\omega_i)\right|^2$$

This is clearly a quadratic form in the $x_i$'s, and more precisely

$$\text{size}(M_x) = \sum_{i,j}\left(\sum_{1\leq k\leq n} \sigma_k(\omega_i)\overline{\sigma}_k(\omega_j)\right) x_i x_j.$$

Note that in the case where $K$ is totally real, that is when all the $\sigma_k$ are real embeddings, this simplifies to

$$\text{size}(M_x) = \sum_{i,j} \text{Tr}(\omega_i \omega_j) x_i x_j$$

which is now a quadratic form with integer coefficients which can easily be computed from the knowledge of the $\omega_i$.

In any case, whether $K$ is totally real or not, we can apply the LLL algorithm to the lattice $\mathbb{Z}^n$ and the quadratic form $\text{size}(M_x)$. The result will be a set of $n$ vectors $x$ corresponding to reasonably small values of the quadratic form (see Section 2.6 for quantitative statements), hence to polynomials $M_x$ of small size, which is what we want. Note however that the algebraic integers $x$ that we obtain in this way will often have a minimal polynomial of degree less than $n$, in other words $x$ will define a subfield of $K$. In particular, $x = 1$ is always obtained as a short vector, and this defines the subfield $\mathbb{Q}$ of $K$. Practical experiments with this method show however that there will always be at least one element $x$ of degree exactly $n$, hence defining $K$, and its minimal polynomial will hopefully be simpler than the polynomial $P$ from which we started.

However the polynomials that we obtain in this way, have sometimes greater coefficients than those of $P$. This is not too surprising since our definition of "size" of $P(X) = \sum_{0 \le k \le n} a_k X^k$ involves the size of the roots of $P$, hence of the quantities

$$|a_{n-k}|^{1/k}$$

more than the size of the coefficients themselves.

Note that as a by-product of this method, we sometimes also obtain subfields of $K$. It is absolutely not true however that we obtain all subfields of $K$ in this way. Indeed, the LLL algorithm gives us at most $n$ subfields, while the number of subfields of $K$ may be much larger.

The algorithm, which we name POLRED for polynomial reduction, is as follows (see [Coh-Diaz]).

**Algorithm 4.4.11** (POLRED). Let $K = \mathbb{Q}(\theta)$ be a number field defined by a monic irreducible polynomial $P \in \mathbb{Z}[X]$. This algorithm gives a list of polynomials defining certain subfields of $K$ (including $\mathbb{Q}$), which are often simpler than the polynomial $P$ so these can be used to define the field $K$ if they are of degree equal to the degree of $K$.

1. [Compute the maximal order] Using the round 2 Algorithm 6.1.8 of Chapter 6, compute an integral basis $\omega_1, \ldots, \omega_n$ as polynomials in $\theta$.

2. [Compute matrix] If the field $K$ is totally real (which can be easily checked using Algorithm 4.1.11), set $m_{i,j} \leftarrow \text{Tr}(\omega_i \omega_j)$ for $1 \le i, j \le n$, which will be an element of $\mathbb{Z}$.

Otherwise, using Algorithm 3.6.6, compute a reasonably accurate value of $\theta$ and its conjugates $\sigma_j(\theta)$ as the roots of $P$, then the numerical values of $\sigma_j(\omega_k)$, and finally compute a reasonably accurate approximation to

$$m_{i,j} \leftarrow \sum_{1 \le k \le n} \sigma_k(\omega_i)\overline{\sigma_k}(\omega_j)$$

(note that this will be a real number).

3. [Apply LLL] Using the LLL Algorithm 2.6.3 applied to the inner product defined by the matrix $M = (m_{i,j})$ and to the standard basis of the lattice $\mathbb{Z}^n$, compute an LLL-reduced basis $\mathbf{b}_1, \ldots, \mathbf{b}_n$.
4. [Compute characteristic polynomials] For $1 \le i \le n$, using the formulas of Section 4.3, compute the characteristic polynomial $C_i$ of the element of $\mathbb{Z}_K$ corresponding to $\mathbf{b}_i$ on the basis $\omega_1, \omega_2, \ldots, \omega_n$.
5. [Compute minimal polynomials] For $1 \le i \le n$, set $P_i \leftarrow C_i/(C_i, C_i')$ where the GCD is always normalized so as to be monic, and is computed by Euclid's algorithm. Output the polynomials $P_i$ and terminate the algorithm.

From what we have seen in Section 4.3, the characteristic polynomial $C_i$ of an element $x \in \mathbb{Z}_K$ is given by $C_i = P_i^k$, where $P_i$ is the minimal polynomial and $k$ is a positive integer, hence $C_i/(C_i, C_i') = P_i$, thus explaining step 5. In fact, to avoid ambiguities of sign which arise, it is also useful to make the following choice at the end of the algorithm. For each polynomial $P_i$, set $d_i \leftarrow \deg(P_i)$ and search for the non-zero monomial of largest degree $d$ such that $d \not\equiv d_i \pmod 2$. If such a monomial exists, make, if necessary, the change $P_i(X) \leftarrow (-1)^{d_i} P_i(-X)$ so that the sign of this monomial is negative.

Let us give an example of the use of the POLRED algorithm. This example is taken from work of M. Olivier. Consider the polynomial

$$T(X) = X^6 + 2X^5 - 7X^4 - 12X^3 + 10X^2 + 17X + 4\,.$$

Using the methods of Section 3.5, one easily shows that this polynomial is irreducible over $\mathbb{Q}$, hence defines a number field $K$ of degree 6. Furthermore, using Algorithm 3.6.6, one computes that the complex roots of $T$ are approximately equal to

$$-2.7494482169, -1.7152399972, -0.8531562311, -0.3074682781,$$
$$1.5839340557, 2.0413786677\,.$$

Using the methods of the preceding section, it is then easy to check that this field has no proper subfield apart from $\mathbb{Q}$.

From this and the classification of transitive permutation groups of degree 6 which we will see in Section 6.3, we deduce that the Galois group $G$ of the Galois closure of $K$ is isomorphic either to the alternating groups $A_5$ or $A_6$,

or to the symmetric groups $S_5$ or $S_6$. Now using the sub-resultant Algorithm 3.3.7 or Proposition 3.3.5 one computes that

$$\mathrm{disc}(T) = 11699^2$$

so by Proposition 6.3.1, we have $G \subset A_6$ hence $G$ is isomorphic either to $A_5$ or to $A_6$.

Distinguishing between the two is done by using one of the resolvent functions given in Section 6.3, and the resolvent polynomial obtained is

$$R(X) = X^6 + 3694X^5 + 1246830X^4 - 7355817976X^3 - 5140929655107X^2 + 3486026298845999X + 2593668315970494361.$$

A computation of the roots of this polynomial shows that it has an integer root $x = -673$, and the results of Section 6.3 imply that $G$ is isomorphic to $A_5$. In addition, $Q(X) = R(X)/(X + 673)$ is an irreducible fifth degree polynomial which defines a number field with the same discriminant as $K$. We have

$$Q(X) = X^5 + 3021X^4 - 786303X^3 - 6826636057X^2 - 546603588746X + 3853890514072057,$$

and the discriminant of $Q$ (which must be a square by Proposition 6.3.1) has 63 decimal digits. Now if we apply the POLRED algorithm, we obtain five polynomials, four of which define the same field as $Q$, and the polynomial with the smallest discriminant is

$$S(X) = X^5 - 2X^4 - 13X^3 + 37X^2 - 21X - 1,$$

a polynomial which is much more appealing than $Q$!

We compute that $\mathrm{disc}(S) = 11699^2$, hence this is the discriminant of the number field $K$ as well as the number field defined by the polynomial $S$.

There was a small amount of cheating in the above example: since $\mathrm{disc}(Q)$ is a 63 digit number, the POLRED algorithm, which in particular computes an integral basis of $K$ hence needs to factor $\mathrm{disc}(Q)$, may need quite a lot of time to factor this discriminant. We can however in this case "help" the POLRED algorithm by telling it that $\mathrm{disc}(Q)$ is a square, which we know a priori, but which is not usually tested for in a factoring algorithm since it is quite rare an occurrence. This is how the above example was computed in practice, and the whole computation, including typing the commands, took only a few minutes on a workstation.

We can slightly modify the POLRED algorithm so as to obtain a defining polynomial for a number field which is as canonical as possible. One possibility is as follows.

We first need a notation. If $Q(X) = \sum_{0 \le i \le n} a_i X^i$ is a polynomial of degree $n$, we set

$$v(Q) = (|\operatorname{disc}(Q)|, \operatorname{size}(Q), |a_n|, |a_{n-1}|, \ldots, |a_1|, |a_0|).$$

**Algorithm 4.4.12** (Pseudo-Canonical Defining Polynomial). Given a number field $K$ defined by a monic irreducible polynomial $P \in \mathbb{Z}[X]$ of degree $n$, this algorithm outputs another polynomial defining $K$ which is as canonical as possible.

1. [Apply POLRED] Apply the POLRED algorithm to $P$, and let $P_i$ (for $i = 1, \ldots, n$) be the $n$ polynomials which are output by the POLRED algorithm. If none of the $P_i$ are of degree $n$, output a message saying that the algorithm failed, and terminate the algorithm. Otherwise, let $\mathcal{L}$ be the set of $i$ such that $P_i$ is of degree $n$.
2. [Minimize $v(P_i)$] If $\mathcal{L}$ has a single element, let $Q$ be this element. If not, for each $i \in \mathcal{L}$ compute $v_i \leftarrow v(P_i)$ and let $v$ be the smallest $v_i$ for the lexicographic ordering of the components. Let $Q$ be any $P_i$ such that $v(P_i) = v$.
3. [Possible sign change] Search for the non-zero monomial of largest degree $d$ such that $d \not\equiv n \pmod 2$. If such a monomial exists, make, if necessary, the change $Q(X) \leftarrow (-1)^n Q(-X)$ so that the sign of this monomial is negative.
4. [Terminate] Output $Q$ and terminate the algorithm.

**Remarks.**

(1) The algorithm may fail, i.e. the POLRED algorithm may give only polynomials of degree less than $n$. That this is possible in principle has been shown by H. W. Lenstra (private communication), but in practice, on more than 100000 polynomials of various degree, I have never encountered a failure. It seems that failure is very rare.

(2) At the end of step 2 there may be several $i$ such that $v_i = v$. In that case, it may be useful to output all the possibilities (after executing step 3 on each of them) instead of only one. In practice, this is also uncommon.

(3) Although Algorithm 4.4.12 makes an effort towards finding a polynomial defining $K$ with small index $f = [\mathbb{Z}_K : \mathbb{Z}[\theta]]$, it should not be expected that it always finds a polynomial with the smallest possible index. An example is the polynomial $X^3 - X^2 - 20X + 9$ which naturally defines the cyclic cubic field with discriminant $61^2$ (see Theorem 6.4.6). Algorithm 4.4.12 finds that this is the pseudo-canonical polynomial defining the cubic field, but it has index equal to 3, while for example the polynomial $X^3 + 12X^2 - 13X + 3$ has index equal to 1. The reason for this behavior is that the notion of "size" of a polynomial is rather indirectly related to the size of the index. See also Exercise 8.

## 4.5 The Subfield Problem and Applications

Let $K = \mathbb{Q}(\alpha)$ and $L = \mathbb{Q}(\beta)$ be number fields of degree $m$ and $n$ respectively, and let $A(X), B(X) \in \mathbb{Z}[X]$ be the minimal polynomials of $\alpha$ and $\beta$ respectively. The basic subfield problem is as follows. Determine whether or not $K$ is isomorphic to a subfield of $L$, or in more down-to-earth terms whether or not some conjugate of $\alpha$ belongs to $L$. We could of course ask more precisely if $\alpha$ itself belongs to $L$, and we will see that the answer to this question follows essentially from the answer to the apparently weaker one.

We start by two fast tests. First, if $K$ is conjugate to a subfield of $L$, then the degree of $K$ clearly must divide the degree of $L$.

The second test follows from Corollary 4.4.9. We compute $d(A)$ and $d(B)$ and for each odd prime $p$ such that $v_p(d(A))$ is odd, test whether or not $p^{n/m} \mid d(B)$. Note that according to Exercise 15, it is not necessary to assume that $A$ and $B$ are monic, i.e. that $\alpha$ and $\beta$ are algebraic integers.

We could use the more stringent test $d(K)^{n/m} \mid d(L)$ using Proposition 4.4.8 directly, but this requires the computation of field discriminants, hence essentially of integral bases, and this is often lengthy. So, we do not advise using this more stringent test unless the field discriminants can be obtained cheaply.

We therefore assume that the above tests have been passed successfully. We will give three different methods for solving our problem. The first two require good approximations to the complex roots of the polynomials $A$ and $B$ (computed using for example Algorithm 3.6.6), while the third is purely algebraic, but slower.

### 4.5.1 The Subfield Problem Using the LLL Algorithm

Let $\beta$ be an arbitrary, but fixed root of the polynomial $B$ in $\mathbb{C}$. If $K$ is conjugate to a subfield of $L$, then some root $\alpha_i$ of $A$ is of the form $P(\beta)$ for some $P \in \mathbb{Q}[X]$ of degree less than $n$. In other words, the complex numbers $1, \beta, \ldots, \beta^{n-1}, \alpha_i$ are $\mathbb{Z}$-linearly dependent. To check this, use the LLL algorithm or one of its variations, as described in Section 2.7.2 on each root of $A$ (or on the root we are specifically interested in as the case may be). Then two things may happen. Either the algorithm gives a linear combination which is not very small in appearance, or it seems to find something reasonable. The reader will notice that in none of these cases have we *proved* anything. If, however, we are in the situation where LLL apparently found a nice relation, this can now be proved: assume the relation gives $\alpha_i = P(\beta)$ for some polynomial $P$ with rational coefficients. (Note that the coefficient of $\alpha_i$ in the linear combination which has been found must be non-zero, otherwise this would mean that the minimal polynomial of $\beta$ is not irreducible.) To test whether this relation is true, it is now necessary simply to check that

$$A \circ P \equiv 0 \pmod{B},$$

where $A$ and $B$ are the minimal polynomials of $\alpha$ and $\beta$ respectively. Indeed, if this is true, this means that $P(\beta)$ is a root of $A$, i.e. a conjugate of $\alpha_i$, hence is $\alpha_i$ itself since LLL told us that it was numerically very close to $\alpha_i$.

To compute $C = A \circ P \pmod{B}$, we use a form of Horner's rule for evaluating polynomials: if $A(X) = \sum_{i=0}^{m} a_i X^i$, then we set $C \leftarrow a_m$, and for $i = m-1, m-2, \ldots, 0$ we compute $C \leftarrow (a_i + P(X)C \bmod B)$.

In the implausible case where one finds that $A \circ P \not\equiv 0 \pmod{B}$, then we must again test for linear dependence with higher precision used for $\alpha_i$ and $\beta$.

**Remark.** There is a better way to test whether each conjugate $\alpha_i$ is or is not a $\mathbb{Q}$-linear combination of $1, \beta, \ldots, \beta^{n-1}$ than to apply LLL to each $\alpha_i$, each time LLL reducing an $(n+2) \times (n+1)$ matrix (or equivalently a quadratic form in $n+1$ variables). Indeed, keeping with the notations of Remark (2) at the end of Section 2.7.2, the first $n$ columns of that matrix, which correspond to the powers of $\beta$, will always be the same. Only the last column depends on $\alpha_i$. But in LLL reduction, almost all the work is spent LLL reducing the first $n$ columns, the $n+1$-st is done last. Hence, we should first LLL reduce the $(n+2) \times n$ matrix corresponding to the powers of $\beta$. Then, for each $\alpha_i$ to be tested, we can now start from the already reduced basis and just add an extra column vector, and since the first $n$ vectors are already LLL reduced, the amount of work which remains to be done to account for the last column will be very small compared to a full LLL reduction. We leave the details to the reader.

If LLL tells us that apparently there is no linear relation, then we suspect that $\alpha \notin \mathbb{Q}(\beta)$. To prove it, the best way is probably to apply one of the two other methods which we are going to explain.

### 4.5.2 The Subfield Problem Using Linear Algebra over $\mathbb{C}$

A second method is as follows (I thank A.-M. Bergé and M. Olivier for pointing it out to me.) After clearing denominators, we may as well assume that $\alpha$ and $\beta$ are algebraic integers. We then have the following.

**Proposition 4.5.1.** *With the above notations, assume that $\alpha$ and $\beta$ are algebraic integers. Then $K$ is isomorphic to a subfield of $L$ if and only if there exists an $n/m$ to one map $\phi$ from $[1,n]$ to $[1,m]$ such that for $1 \leq h < n$,*

$$s_h = \sum_{1 \leq i \leq n} \alpha_{\phi(i)} \beta_i^h \in \mathbb{Z},$$

*where the $\alpha_j$ (resp. $\beta_j$) denote the roots of $A(X)$ (resp. of $B(X)$) in $\mathbb{C}$.*

*Proof.* Assume first that $K$ is isomorphic to a subfield of $L$, i.e. that $\alpha_i = P(\beta_1)$ with $P \in \mathbb{Q}[X]$ say. Then, for every $i$, $P(\beta_i)$ is a root $\alpha_j$ of $A(X) = 0$, and

by Galois theory each $\alpha_j$ is obtained exactly $n/m$ times. Therefore the map $i \mapsto j = \phi(i)$ is $n/m$ to one. Furthermore,

$$s_h = \sum_{1 \le i \le n} \alpha_{\phi(i)} \beta_i^h = \sum_{1 \le i \le n} P(\beta_i)\beta_i^h = \mathrm{Tr}_{L/\mathbb{Q}}(P(\beta)\beta^h) \in \mathbb{Q},$$

hence $s_h \in \mathbb{Z}$ since the $\alpha_j$ and $\beta_i$ are algebraic integers.

Conversely, assume that for some $\phi$ we have $s_h \in \mathbb{Z}$ for all $h$ such that $1 \le h < n$. Note that $s_0 = (n/m)\,\mathrm{Tr}_{K/\mathbb{Q}}(\alpha) \in \mathbb{Z}$ follows automatically.

Consider the following $n \times n$ linear system:

$$\sum_{0 \le j < n} x_j \,\mathrm{Tr}_{L/\mathbb{Q}}(\beta^j \beta^h) = s_h, \qquad 0 \le h < n.$$

By Proposition 4.4.4 (1) the determinant of this system is equal to $d(B)$, hence is non-zero. Furthermore, the system has rational coefficients, so the unique solution has coefficients $x_j \in \mathbb{Q}$. If we set $P(X) = \sum_{0 \le j < n} x_j X^j$, we then have $P \in \mathbb{Q}[X]$ and $\sum_{1 \le i \le n} P(\beta_i)\beta_i^h = s_h$. It follows that the vector of the $(P(\beta_i))$ and of the $\alpha_{\phi(i)}$ are both solutions of the linear system $\sum_{1 \le i \le n} v_i \beta_i^h = s_h$, and since the $\beta_i$ are distinct this system has a unique solution, so the vectors are equal, thus proving the proposition. □

**Remarks.**

(1) The number of maps from $[1, n]$ to $[1, m]$ which are $n/m$-to-one is equal to $n!/((n/m)!)^m$ hence can be quite large, especially when $m = n$ (which corresponds to the very important *isomorphism problem*). This is to be compared to the number of trials to be done with the LLL method, which is only equal to $m$. Hence, although LLL is slow, except when $n$ is very small (say $n \le 4$), we suggest starting with the LLL method. If the answer is positive, which will in practice happen quite often, we can stop. If not, use the present method (or the purely algebraic method which is explained below).

(2) To check that $s_h \in \mathbb{Z}$ we must of course compute the roots of $A(X)$ and $B(X)$ sufficiently accurately. Now however the error estimates are trivial (compared to the ones we would need using LLL), and if $s_h$ is sufficiently far away from an integer, it is very easy to prove rigorously that it is so.

(3) We start of course by checking whether $s_1 \in \mathbb{Z}$, since this will eliminate most candidates for $\phi$.

The above leads to the following algorithm.

**Algorithm 4.5.2** (Subfield Problem Using Linear Algebra). Let $A(X)$ and $B(X)$ be primitive irreducible polynomials in $\mathbb{Z}[X]$ of degree $m$ and $n$ respectively defining number fields $K$ and $L$. This algorithm determines whether or not $K$ is isomorphic to a subfield of $L$, and if it is, gives an explicit isomorphism.

1. [Trivial check] If $m \nmid n$, output NO and terminate the algorithm.
2. [Reduce to algebraic integers] Set $a \leftarrow \ell(A)$, $b \leftarrow \ell(B)$ (the leading terms of $A$ and $B$), and set $A(X) \leftarrow a^{m-1}A(X/a)$ and $B(X) \leftarrow b^{n-1}B(X/b)$.
3. [Check discriminants] For every odd prime $p$ such that $v_p(d(A))$ is odd, check that $p^{n/m} \mid d(B)$ (where $d(A)$ and $d(B)$ are computed using Algorithm 3.3.7). If this is not the case, output NO and terminate the algorithm. If for some reason $d(K)$ and $d(L)$ are known or cheaply computed, replace these checks by the single check $d(K)^{n/m} \mid d(L)$.
4. [Compute roots] Using Algorithm 3.6.6, compute the complex roots $\alpha_i$ and $\beta_i$ of $A(X)$ and $B(X)$ to a reasonable accuracy (it may be necessary to have more accuracy in the later steps).
5. [Loop on $\phi$] For each $n/m$ to one map $\phi$ from $[1,n]$ to $[1,m]$ execute steps 6 and 7. If all the maps have been examined without termination of the algorithm, output NO and terminate the algorithm.
6. [Check $s_1 \in \mathbb{Z}$] Let $s_1 \leftarrow \sum_{1\le i\le n} \alpha_{\phi(i)}\beta_i$. If $s_1$ is not close to an integer (this is a rigorous statement, since it depends only on the chosen approximations to the roots), take the next map $\phi$ in step 5.

   Otherwise, check whether $s_h \leftarrow \sum_{1\le i\le n} \alpha_{\phi(i)}\beta_i^h$ are also close to an integer for $h = 2,\dots,n-1$. As soon as this is not the case, take the next map $\phi$ in step 5.
7. [Compute polynomial] (Here the $s_h$ are all close to integers.) Set $s_h \leftarrow \lfloor s_h \rceil$ (the nearest integer to $s_h$). Compute by induction $t_k \leftarrow \mathrm{Tr}_{L/\mathbb{Q}}(\beta_1^k)$ for $0 \le k \le 2n-2$, and using Algorithm 2.2.1 or a Gauss-Bareiss variant, find the unique solution to the linear system $\sum_{0\le j<n} x_j t_{j+h} = s_h$ for $0 \le h < n$ (note that we know that $d(B)x_j \in \mathbb{Z}$ so we can avoid rational arithmetic), and set $P(X) \leftarrow \sum_{0\le j<n} x_j X^j$.
8. [Finished?] Using the variant of Horner's rule explained in Section 4.5.1, check whether $A(P(X)) \equiv 0 \pmod{B(X)}$. If this is the case, then output YES, output also the polynomial $P(bX)/a$ which gives the isomorphism explicitly, and terminate the algorithm. Otherwise, using Algorithm 3.6.6 (or, even more simply, a few Newton iterations to obtain a higher precision) recompute the roots $\alpha_i$ and $\beta_i$ to a greater accuracy and go to step 6.

### 4.5.3 The Subfield Problem Using Algebraic Algorithms

The third solution that we give to the subfield problem is usually less efficient but has the advantage that it is guaranteed to work without worrying about complex approximations. The idea is to use Algorithm 3.6.4 which factors polynomials over number fields and the following easy proposition whose proof is left to the reader (Exercise 9).

**Proposition 4.5.3.** *Let $\alpha$ and $\beta$ be algebraic numbers with minimal polynomials $A(X)$ and $B(X)$ respectively. Set $K = \mathbb{Q}(\alpha)$, $L = \mathbb{Q}(\beta)$, and let*

*$A = \prod_{1 \le i \le g} A_i$ be a factorization of $A$ into irreducible factors in $L[X]$. There is a one-to-one correspondence between the $A_i$ of degree equal to one and the conjugates of $\alpha$ belonging to $L$. In particular, $L$ contains a subfield isomorphic to $K$ if and only if at least one of the $A_i$ is of degree equal to one.*

This immediately leads to the following algorithm. Note that we keep the same first three steps of the preceding algorithm.

**Algorithm 4.5.4** (Subfield Problem Using Factorization of Polynomials). Let $A(X)$ and $B(X)$ be primitive irreducible polynomials in $\mathbb{Z}[X]$ of degree $m$ and $n$ respectively defining number fields $K$ and $L$. This algorithm determines whether or not $K$ is isomorphic to a subfield of $L$, and if it is, gives an explicit isomorphism.

1. [Trivial check] If $m \nmid n$, output NO and terminate the algorithm.
2. [Reduce to algebraic integers] Set $a \leftarrow \ell(A)$, $b \leftarrow \ell(B)$ (the leading terms of $A$ and $B$), and set $A(X) \leftarrow a^{m-1}A(X/a)$ and $B(X) \leftarrow b^{n-1}B(X/b)$.
3. [Check discriminants] For every odd prime $p$ such that $v_p(d(A))$ is odd, check that $p^{n/m} \mid d(B)$ (where $d(A)$ and $d(B)$ are computed using Algorithm 3.3.7). If this is not the case, output NO and terminate the algorithm. If for some reason $d(K)$ and $d(L)$ are known or cheaply computed, replace these checks by the single check $d(K)^{n/m} \mid d(L)$.
4. [Factor in $L[X]$] Using Algorithm 3.6.4, let $A = \prod_{1 \le i \le g} A_i$ be a factorization of $A$ into irreducible factors in $L[X]$, where without loss of generality we may assume the $A_i$ monic.
5. [Conclude] If no $A_i$ is of degree equal to 1, then output NO otherwise output YES, and if we write $A_i = X - g_i(\beta)$ where $\beta$ is a root of $B$ such that $L = \mathbb{Q}(\beta)$, output also the polynomial $g_i(bX)/a$ which gives explicitly the isomorphism. Terminate the algorithm.

**Conclusion.** With three different algorithms to solve the subfield problem, it is now necessary to give some practical advice. These remarks are, of course, also valid for the applications of the subfield problem that we will see in the next section, such as the field isomorphism problem.

1) Start by executing steps 1 to 3 of Algorithm 4.5.2. These tests are fast and will eliminate most cases when $K$ is not isomorphic to a subfield of $L$. If these tests go through, there is now a distinct possibility that the answer to the subfield problem is yes.

2) Apply the LLL method (using the remark made at the end). This is also quite fast, and will give good results if $K$ is indeed isomorphic to a subfield of $L$. Note that sufficient accuracy should be used in computing the roots of $A(X)$ and $B(X)$ otherwise LLL may miss a dependency. If LLL fails to detect a relation, then especially if the computation has been done to high accuracy it is almost certain that $K$ is *not* isomorphic to a subfield of $L$.

An alternate method which is numerically more stable is to use Algorithm 4.5.2. However this algorithm is much slower than LLL as soon as $n$ is at all large, hence should be used only for these very small values of $n$.

3) In the remaining cases, apply Algorithm 4.5.4 which is slow but sure.

### 4.5.4 Applications of the Solutions to the Subfield Problem

Now that we have seen three methods for solving the subfield problem, we will see that this problem is basic for the solution of a number of other problems. For each of these other problems, we can then choose any method that we like to solve the underlying subfield problem.

**The Field Membership Problem.**

The first problem that we can now solve is the *field membership problem.* Given two algebraic numbers $\alpha$ and $\beta$ by their minimal polynomials $A$ and $B$ and suitable complex approximations, determine whether or not $\alpha \in \mathbb{Q}(\beta)$ and if so a polynomial $P \in \mathbb{Q}[X]$ such that $\alpha = P(\beta)$. For this, apply one of the three methods that we have studied for the subfield problem. Note that some steps may be simplified since we have chosen a specific complex root of $A(X)$. For example, if we use LLL, we simply check the linear dependence of $\alpha$ and the powers of $\beta$. If we use linear algebra, choosing a numbering of the roots such that $\alpha = \alpha_1$ and $\beta = \beta_1$, we can restrict to maps $\phi$ such that $\phi(1) = 1$. In the algebraic method on the other hand we must lengthen step 5. For every $A_i = X - g_i(\beta)$ of degree one, we compute $g_i(\beta)$ numerically (it will be a root of $A(X)$) and check whether it is closer to $\alpha$ than to any other root. If this occurs for no $i$, then the answer is NO, otherwise the answer is YES and we output the correct $g_i$.

**The Field Isomorphism Problem.**

The second problem is the *isomorphism problem.* Given two number fields $K$ and $L$ as before, determine whether or not they are isomorphic. This is of course equivalent to $K$ and $L$ having the same degree and $K$ being a subfield of $L$, so the solution to this problem follows immediately from that of the subfield problem. Since this problem is very important, we give explicitly the two algorithms corresponding to the last two methods (the LLL method can of course also be used). For still another method, see [Poh3].

**Algorithm 4.5.5** (Field Isomorphism Using Linear Algebra). Let $A(X)$ and $B(X)$ be primitive irreducible polynomials in $\mathbb{Z}[X]$ of the same degree $n$ defining number fields $K$ and $L$. This algorithm determines whether or not $K$ is isomorphic to $L$, and if it is, gives an explicit isomorphism.

1. [Reduce to algebraic integers] Set $a \leftarrow \ell(A)$, $b \leftarrow \ell(B)$ (the leading terms of $A$ and $B$), and set $A(X) \leftarrow a^{n-1}A(X/a)$ and $B(X) \leftarrow b^{n-1}B(X/b)$.
2. [Check discriminants] Compute $d(A)$ and $d(B)$ using Algorithm 3.3.7), and check whether $d(A)/d(B)$ is a square in $\mathbb{Q}$ using essentially Algorithm 1.7.3.

If this is not the case, output NO and terminate the algorithm. If for some reason $d(K)$ and $d(L)$ are known or cheaply computed, replace this check by $d(K) = d(L)$.

3. [Compute roots] Using Algorithm 3.6.6, compute the complex roots $\alpha_i$ and $\beta_i$ of $A(X)$ and $B(X)$ to a reasonable accuracy (it may be necessary to have more accuracy in the later steps).

4. [Loop on $\phi$] For each permutation $\phi$ of $[1,n]$ execute steps 5 and 6. If all the permutations have been examined without termination of the algorithm, output NO and terminate the algorithm.

5. [Check $s_1 \in \mathbb{Z}$] Let $s_1 \leftarrow \sum_{1\le i\le n} \alpha_{\phi(i)}\beta_i$. If $s_1$ is not close to an integer (this is a rigorous statement, since it depends only on the chosen approximations to the roots), take the next permutation $\phi$ in step 4.

   Otherwise, check whether $s_h \leftarrow \sum_{1\le i\le n} \alpha_{\phi(i)}\beta_i^h$ are also close to an integer for $h = 2,\dots,n-1$. As soon as this is not the case, take the next map $\phi$ in step 4.

6. [Compute polynomial] (Here the $s_h$ are all close to integers.) Set $s_h \leftarrow \lfloor s_h \rceil$ (the nearest integer to $s_h$). Compute by induction $t_k \leftarrow \mathrm{Tr}_{L/\mathbb{Q}}(\beta_1^k)$ for $0 \le k \le 2n-2$, and using Algorithm 2.2.1 or a Gauss-Bareiss variant, find the unique solution to the linear system $\sum_{0\le j<n} x_j t_{j+h} = s_h$ for $0 \le h < n$. (We know that $d(B)x_j \in \mathbb{Z}$, so we can avoid rational arithmetic.) Now set $P(X) \leftarrow \sum_{0\le j<n} x_j X^j$.

7. [Finished?] Using the variant of Horner's rule explained in Section 4.5.1, check whether $A(P(X)) \equiv 0 \pmod{B(X)}$. If this is the case, then output YES, and also output the polynomial $P(bX)/a$ which gives the isomorphism explicitly, and terminate the algorithm. Otherwise, using Algorithm 3.6.6 recompute the roots $\alpha_i$ and $\beta_i$ to a greater accuracy and go to step 5.

**Algorithm 4.5.6** (Field Isomorphism Using Polynomial Factorization). Let $A(X)$ and $B(X)$ be primitive irreducible polynomials in $\mathbb{Z}[X]$ of the same degree $n$ defining number fields $K$ and $L$. This algorithm determines whether or not $K$ is isomorphic to $L$, and if it is, gives an explicit isomorphism.

1. [Reduce to algebraic integers] Set $a \leftarrow \ell(A)$, $b \leftarrow \ell(B)$ (the leading terms of $A$ and $B$), and set $A(X) \leftarrow a^{n-1}A(X/a)$ and $B(X) \leftarrow b^{n-1}B(X/b)$.

2. [Check discriminants] Compute $d(A)$ and $d(B)$ using Algorithm 3.3.7), and check whether $d(A)/d(B)$ is a square in $\mathbb{Q}$ using a slightly modified version of Algorithm 1.7.3. If this is not the case, output NO and terminate the algorithm. If for some reason $d(K)$ and $d(L)$ are known or cheaply computed, check instead that $d(K) = d(L)$.

3. [Factor in $L[X]$] Using Algorithm 3.6.4, let $A = \prod_{1\le i\le g} A_i$ be a factorization of $A$ into irreducible factors in $L[X]$, where without loss of generality we may assume the $A_i$ monic.

4. [Conclude] If no $A_i$ has degree equal to 1, then output NO otherwise output YES, and if we write $A_i = X - g_i(\beta)$ where $\beta$ is a root of $B$ such that

$L = \mathbb{Q}(\beta)$, also output the polynomial $g_i(bX)/a$ which explicitly gives the isomorphism. Terminate the algorithm.

For the field isomorphism problem, there is a different method which works sufficiently often that it deserves to be mentioned. We have seen that Algorithm 4.4.12 gives a defining polynomial for a number field which is almost canonical. Hence, if we apply this algorithm to two polynomials $A$ and $B$, then, if the corresponding number fields are isomorphic, there is a good chance that the polynomials output by Algorithm 4.4.12 will be the same. If they are the same, this proves that the fields are isomorphic (and we can easily recover explicitly the isomorphism if desired). If not, it does not prove anything, but we can expect that they are not isomorphic. We must then apply one of the rigorous methods explained above to prove this.

**The Primitive Element Problem.**

The last application of the subfield problem that we will see is to the *primitive element problem*. This is as follows. Given algebraic numbers $\alpha_1, \ldots, \alpha_m$, set $K = \mathbb{Q}(\alpha_1, \ldots, \alpha_m)$. Then $K$ is a number field, hence it is reasonable (although not always absolutely necessary, see [Duv]) to represent $K$ by a primitive element $\theta$, i.e.

$$K = \mathbb{Q}(\alpha_1, \ldots \alpha_m) = \mathbb{Q}(\theta) \simeq \mathbb{Q}[X]/(T(X)\mathbb{Q}[X]),$$

where $T$ is the minimal polynomial of $\theta$. Hence, we need an algorithm which finds such a $T$ (which is not unique) given $\alpha_1, \ldots, \alpha_m$. We can do this by induction on $m$, and the problem boils down to the following: Given $\alpha$ and $\beta$ by their minimal polynomials $A$ and $B$ (and suitable complex approximations), find a monic irreducible polynomial $T \in \mathbb{Z}[X]$ such that

$$\mathbb{Q}(\alpha, \beta) = \mathbb{Q}(\theta), \quad \text{where } T(\theta) = 0.$$

We can use the solution to the subfield problem to solve this. According to the proof of the primitive element theorem (see [Lang1]), we can take $\theta = k\alpha + \beta$ for a small integer $k$, and $\mathbb{Q}(\alpha, \beta) = \mathbb{Q}(k\alpha + \beta)$ is equivalent to $\alpha \in \mathbb{Q}(k\alpha + \beta)$ which can be checked using one of the algorithms explained above for the field membership problem.

## 4.6 Orders and Ideals

### 4.6.1 Basic Definitions

**Definition 4.6.1.** *An* order *$R$ in $K$ is a subring of $K$ which as a $\mathbb{Z}$-module is finitely generated and of maximal rank $n = \deg(K)$ (note that we use the*

*"modern" definition of a ring, which includes the existence of the multiplicative identity 1).*

Proposition 4.1.3 shows that every element of an order $R$ is an algebraic integer, i.e. that $R \subset \mathbb{Z}_K$. We will see that the ring theory of $\mathbb{Z}_K$ is nicer than that of an arbitrary order $R$, but for the moment we let $R$ be an arbitrary order in a number field $K$. We emphasize that some of the properties mentioned here are specific to orders in number fields, and are not usually valid for general base rings.

**Definition 4.6.2.** *An* ideal *$I$ of $R$ is a sub-$R$-module of $R$, i.e. a sub-$\mathbb{Z}$-module of $R$ such that for every $r \in R$ and $i \in I$ we have $ri \in I$.*

Note that the quotient module $R/I$ has a canonical quotient ring structure. In fact we have:

**Proposition 4.6.3.** *Let $I$ be a non-zero ideal of $R$. Then $I$ is a module of maximal rank. In other words, $R/I$ is a finite ring. Its cardinality is called the* norm *of $I$ and denoted $\mathcal{N}(I)$.*

Indeed, if $i \in I$ with $i \neq 0$, then $iR \subset I \subset R$, proving the proposition. □

If $I$ is given by its HNF on a basis of $R$ (or simply by any matrix $A$), then Proposition 4.7.4 shows that the norm of $I$ is simply the absolute value of the determinant of $A$.

Ideals can be added (as modules), and the sum of two ideals is clearly again an ideal. Similarly, the intersection of two ideals is an ideal. Ideals can also be multiplied in the following way: if $I$ and $J$ are ideals, then

$$IJ = \left\{ \sum_i x_i y_i, \text{ where } x_i \in I \text{ and } y_i \in J \right\}.$$

Again, it is clear that this is an ideal. Note that we clearly have the inclusions

$$IJ \subset I \cap J \subset I \subset I + J,$$

(and similarly with $J$), and $IR = I$ for all ideals $I$. It is clearly not always true that $IJ = I \cap J$ (take $I = J = p\mathbb{Z}$ in $\mathbb{Z}$). We have however the following easy result.

**Proposition 4.6.4.** *Let $I$ and $J$ be two ideals in $R$ and assume that $I + J = R$. (It is then reasonable to say that $I$ and $J$ are* coprime.*) Then we have the equality $IJ = I \cap J$.*

*Proof.* Since $IJ \subset I \cap J$ we need to prove only the reverse inclusion. But since $I + J = R$, there exists $a \in I$ and $b \in J$ such that $a + b = 1$. If $x \in I \cap J$ it

follows that $x = ax + bx$ and clearly $ax \in IJ$ and $bx \in JI = IJ$ thus proving the proposition. □

**Definition 4.6.5.** *A* fractional ideal *$I$ in $R$ is a non-zero submodule of $K$ such that there exists a non-zero integer $d$ with $dI$ ideal of $R$. An ideal (fractional or not) is said to be a* principal ideal *if there exists $x \in K$ such that $I = xR$. Finally, $R$ is a* principal ideal domain *(PID) if $R$ is an integral domain (this is already satisfied for orders) and if every ideal of $R$ is a principal ideal.*

It is clear that if $I$ is a fractional ideal, then $I \subset R$ if and only if $I$ is an ideal of $R$, and we will then say that $I$ is an *integral ideal.*

Note that the set-theoretic inclusions seen above remain valid for fractional ideals, except for the one concerning the product. Indeed, if $I$ and $J$ are two fractional ideals, one does not even have $IJ \subset I$ in general: take $I = R$, and $J$ a non-integral ideal.

**Definition 4.6.6.** *Let $I$ be a fractional ideal of $R$. We will say that $I$ is* invertible *if there exists a fractional ideal $J$ of $R$ such that $R = IJ$. Such an ideal $J$ will be called an inverse of $I$.*

The following lemma is easy but crucial.

**Lemma 4.6.7.** *Let $I$ be a fractional ideal, and set*

$$I' = \{x \in K,\ xI \subset R\}.$$

*Then $I$ is invertible if and only if $II' = R$. Furthermore if this equality is true, then $I'$ is the unique inverse of $I$ and is denoted $I^{-1}$.*

The proof is immediate and left to the reader. □

**Remark.** It is not true in general that $\mathcal{N}(IJ) = \mathcal{N}(I)\mathcal{N}(J)$. For example, let $\omega = (1 + \sqrt{-7})/2$, take $R = \mathbb{Z} + 3\omega\mathbb{Z}$ and $I = J = 3\mathbb{Z} + 3\omega\mathbb{Z}$. Then one immediately checks that $\mathcal{N}(I) = 3$, but $\mathcal{N}(I^2) = 27$. As the following proposition shows, the equality $\mathcal{N}(IJ) = \mathcal{N}(I)\mathcal{N}(J)$ is however true when either $I$ or $J$ is an invertible ideal in $R$, and in particular, it is always true when $R = \mathbb{Z}_K$ is the maximal order of $K$ (see Section 4.6.2 for the relevant definitions).

**Proposition 4.6.8.** *Let $R$ be an order in a number field, and let $I$ and $J$ be two integral ideals of $R$. If either $I$ or $J$ is invertible, we have $\mathcal{N}(IJ) = \mathcal{N}(I)\mathcal{N}(J)$.*

*Proof.* (This proof is due to H. W. Lenstra.) Assume for example that $I$ is invertible. We will prove more generally that if $J \subset H$ where $J$ and $H$ are ideals of $R$, then $[IH : IJ] = [H : J]$. With $H = R$, this gives $[I : IJ] = [R : J]$ hence $\mathcal{N}(IJ) = [R : IJ] = [R : I][I : IJ] = \mathcal{N}(I)\mathcal{N}(J)$ thus proving the proposition.

Let us temporarily say that a pair of ideals $(J, H)$ is a *simple pair* if $[H : J] > 1$ and if there are no ideals containing $J$ and contained in $H$ apart from $H$ and $J$ themselves.

We prove the equality $[IH : IJ] = [H : J]$ by induction on $[H : J]$. For $H = J$ it is trivial, hence assume by induction that $[H : J] > 1$ and that the proposition is true for any pair of ideals such that $[H' : J'] < [H : J]$. Assume that $(J, H)$ is not a simple pair, and let $H_1$ be an ideal between $J$ and $H$ and distinct from both. By our induction hypothesis we have $[IH : IH_1] = [H : H_1]$ and $[IH_1 : IJ] = [H_1 : J]$ hence $[IH : IJ] = [H : J]$ thus proving the proposition in that case.

Assume now that $(J, H)$ is a simple pair. Then $(IJ, IH)$ is also a simple pair since $I$ is an invertible ideal (in fact multiplication by $I$ gives a one-to-one map from the set of ideals between $J$ and $H$ onto the set of ideals between $IJ$ and $IH$). Now we have the following lemma.

**Lemma 4.6.9.** *If $(J, H)$ is a simple pair, then there exists an isomorphism of $R$-modules from $H/J$ to $R/M$ for some maximal ideal $M$ of $R$. (Recall that $M$ is a maximal ideal if and only if $(M, R)$ is a simple pair.)*

Indeed, let $x \in H \setminus J$. The ideal $xR + J$ is between $J$ and $H$ but is not equal to $J$, hence $H = xR + J$. This immediately implies that the map from $R$ to $H/J$ which sends $a$ to the class of $ax$ modulo $J$ is a surjective $R$-linear map. Call $M$ its kernel, which is an ideal of $R$. Then by definition $R/M$ is isomorphic to $H/J$ and since $(J, H)$ is a simple pair it follows that $(M, R)$ is a simple pair, in other words that $M$ is a maximal ideal of $R$, thus proving the lemma. □

Resuming the proof of the proposition, we see that $H/J$ is isomorphic to $R/M$ and $IH/IJ$ is isomorphic to $R/M'$ for some maximal ideals $M$ and $M'$. By construction, $MH \subset J$ hence $MIH \subset IJ$, so $M$ annihilates $IH/IJ$ hence $M \subset M'$. Since $M$ and $M'$ are maximal ideals (or since $I$ is invertible), it follows that $M = M'$, hence that $[IH : IJ] = \mathcal{N}(M') = \mathcal{N}(M) = [H : J]$ thus showing the proposition. □

**Definition 4.6.10.** *An ideal $\mathfrak{p}$ of $R$ is called a* prime ideal *if $\mathfrak{p} \neq R$ and if the quotient ring $R/\mathfrak{p}$ is an integral domain (in other words if $xy \in \mathfrak{p}$ implies $x \in \mathfrak{p}$ or $y \in \mathfrak{p}$). The ideal $\mathfrak{p}$ is* maximal *if the quotient ring $R/\mathfrak{p}$ is a field.*

It is easy to see that an ideal $\mathfrak{p}$ is maximal if and only if $\mathfrak{p} \neq R$ and if the only ideals $I$ such that $\mathfrak{p} \subset I \subset R$ are $\mathfrak{p}$ and $R$, in other words if $(\mathfrak{p}, R)$

form a simple pair in the language used above. Furthermore, it is clear that a maximal ideal is prime. In number fields, the converse is essentially true:

**Proposition 4.6.11.** *Let $\mathfrak{p}$ be a non-zero prime ideal in $R$. Then $\mathfrak{p}$ is maximal. (Here it is essential that $R$ be an order in a number field.)*

Indeed, to say that $\mathfrak{p}$ is a prime ideal is equivalent to saying that for every $x \notin \mathfrak{p}$ the maps $y \mapsto xy$ modulo $\mathfrak{p}$ are injections from $A/\mathfrak{p}$ into itself. Since $A/\mathfrak{p}$ is finite, these maps are also bijections, hence $A/\mathfrak{p}$ is a field. □

Note that $\{0\}$ is indeed a prime ideal, but is not maximal. It will always be excluded, even when this is not explicitly mentioned.

The reason why prime ideals are called "prime" is that the prime ideals of $\mathbb{Z}$ are $\{0\}$, and the ideals $p\mathbb{Z}$ for $p$ a prime number. Prime ideals also satisfy some of the properties of prime numbers. Specifically:

**Proposition 4.6.12.** *If $\mathfrak{p}$ is a prime ideal and $\mathfrak{p} \supset I_1 \cdots I_k$, where the $I_i$ are ideals, then there exists an $i$ such that $\mathfrak{p} \supset I_i$.*

*Proof.* By induction on $k$ it suffices to prove the result for $k = 2$. Assume that $\mathfrak{p} \supset IJ$ and $\mathfrak{p} \not\supset I$ and $\mathfrak{p} \not\supset J$. Then there exists $x \in I$ such that $x \notin \mathfrak{p}$, and $y \in J$ such that $y \notin \mathfrak{p}$. Since $\mathfrak{p}$ is a prime ideal, $xy \notin \mathfrak{p}$, but clearly $xy \in IJ$, contradiction. □

If we interpret $I \supset J$ as meaning $I \mid J$, this says that if $\mathfrak{p}$ divides a product of ideals, it divides one of the factors. Although it is quite tempting to use the notation $I \mid J$, one should be careful with it since it is not true in general that $I \mid J$ implies that there exists an ideal $I'$ such that $J = II'$. As we will see, this will indeed be true if $R = \mathbb{Z}_K$, and in this case it makes perfectly good sense to use that notation.

A variant of the above mentioned phenomenon is that it is not true for general orders $R$ that every ideal is a product of prime ideals. What is always true is that every (non-zero) ideal contains a product of (non-zero) prime ideals. When $R = \mathbb{Z}_K$ however, we will see that everything we want is true at the level of ideals.

**Proposition 4.6.13.** *If $R$ is an order in a number field (or more generally a Noetherian integral domain), any non-zero integral ideal $I$ in $R$ contains a product of (non-zero) prime ideals.*

This is easily proved by Noetherian induction (see Exercise 11).

An important notion which is weaker than that of PID but almost as useful is that of a *Dedekind domain*. This is by definition a Noetherian integral domain $R$ such that every non-zero prime ideal is maximal, and which is integrally closed. This last condition means that if $x$ is a root of a monic

polynomial equation with coefficients in $R$ and if $x$ is in the field of fractions of $R$, then in fact $x \in R$. This is for example the case of $R = \mathbb{Z}$.

When $R$ is an order in a number field, all the conditions are satisfied except that $R$ must also be integrally closed. Since $R \supset \mathbb{Z}$, it is clear that if $R$ is integrally closed then $R = \mathbb{Z}_K$, and the converse is also true by Proposition 4.1.5. Hence the only order in $K$ which is a Dedekind domain is the ring of integers $\mathbb{Z}_K$. Since we know that every order $R$ is a subring of $\mathbb{Z}_K$, we will also call $\mathbb{Z}_K$ the *maximal order* of $K$.

We now specialize to the case where $R = \mathbb{Z}_K$.

### 4.6.2 Ideals of $\mathbb{Z}_K$

In this section, fix $R = \mathbb{Z}_K$. Let $\mathcal{I}(K)$ be the set of fractional ideals of $\mathbb{Z}_K$. We summarize the main properties of $\mathbb{Z}_K$-ideals in the following theorem:

**Theorem 4.6.14.**

(1) *Every fractional ideal of $\mathbb{Z}_K$ is invertible. In other words, if $I$ is a fractional ideal and if we set $I^{-1} = \{x \in K, xI \subset \mathbb{Z}_K\}$, then $II^{-1} = \mathbb{Z}_K$.*
(2) *The set of fractional ideals of $\mathbb{Z}_K$ is an Abelian group.*
(3) *Every fractional ideal $I$ can be written in a unique way as*

$$I = \prod_{\mathfrak{p}} \mathfrak{p}^{v_{\mathfrak{p}}(I)},$$

*the product being over a finite set of prime ideals, and the exponents $v_{\mathfrak{p}}(I)$ being in $\mathbb{Z}$. In particular, $I$ is an integral ideal (i.e. $I \subset \mathbb{Z}_K$) if and only if all the $v_{\mathfrak{p}}(I)$ are non-negative.*
(4) *The maximal order $\mathbb{Z}_K$ is a PID if and only if it is a UFD.*

Hence the ideals of $\mathbb{Z}_K$ behave exactly as the numbers in $\mathbb{Z}$, and can be handled in the same way. Note that (3) is much stronger than Proposition 4.6.13, but is valid only because $\mathbb{Z}_K$ is also integrally closed.

The quantity $v_{\mathfrak{p}}(I)$ is called the $\mathfrak{p}$-adic valuation of $I$ and satisfies the usual properties:

(1) $I \subset \mathbb{Z}_K \iff v_{\mathfrak{p}}(I) \geq 0$ for all prime ideals $\mathfrak{p}$.
(2) $J \subset I \iff v_{\mathfrak{p}}(I) \leq v_{\mathfrak{p}}(J)$ for all prime ideals $\mathfrak{p}$.
(3) $v_{\mathfrak{p}}(I + J) = \min(v_{\mathfrak{p}}(I), v_{\mathfrak{p}}(J))$.
(4) $v_{\mathfrak{p}}(I \cap J) = \max(v_{\mathfrak{p}}(I), v_{\mathfrak{p}}(J))$.
(5) $v_{\mathfrak{p}}(IJ) = v_{\mathfrak{p}}(I) + v_{\mathfrak{p}}(J)$.

Hence the dictionary between fractional ideals and rational numbers is as follows:

Fractional ideals $\longleftrightarrow$ (non-zero) rational numbers.

Integral ideals $\longleftrightarrow$ integers.
Inclusion $\longleftrightarrow$ divisibility (with the reverse order).
Sum $\longleftrightarrow$ greatest common divisor.
Intersection $\longleftrightarrow$ least common multiple.
Product $\longleftrightarrow$ product.

Of course, a few of these notions could be unfamiliar for rational numbers, for example the GCD, but a moment's thought shows that one can give perfectly sensible definitions.

We end this section with the notion of norm of a fractional ideal. We have seen in Proposition 4.6.3 that for an integral ideal $I$ the norm of $I$ is the cardinality of the finite ring $R/I$. As already mentioned, a corollary of Theorem 4.6.14 is that $\mathcal{N}(IJ) = \mathcal{N}(I)\mathcal{N}(J)$ for ideals $I$ and $J$ of the maximal order $R = \mathbb{Z}_K$ (recall that this is false in general if $R$ is not maximal). This allows us to extend the definition of $\mathcal{N}(I)$ to fractional ideals if desired: any fractional ideal $I$ can be written as a quotient of two integral ideals, say $I = P/Q$ (in fact by definition we can take $Q = dR$ where $d$ is an integer), and we define $\mathcal{N}(I) = \mathcal{N}(P)/\mathcal{N}(Q)$. It is easy to check that this is independent of the choice of $P$ and $Q$ and that it is still multiplicative ($\mathcal{N}(IJ) = \mathcal{N}(I)\mathcal{N}(J)$). Of course, usually it will no longer be an integer.

The notion of norm of an ideal is linked to the notion of norm of an element that we have seen above in the following way:

**Proposition 4.6.15.** *Let $x$ be a non-zero element of $K$. Then*

$$|\mathcal{N}_{K/\mathbb{Q}}(x)| = \mathcal{N}(x\mathbb{Z}_K),$$

*in other words the norm of a principal ideal of $\mathbb{Z}_K$ is equal to the absolute value of the norm (in $K$) of a generating element.*

One should never forget this absolute value. We could in fact have a nicer looking proposition (without absolute values) by using a slight extension of the notion of fractional ideal: because of Theorem 4.6.14 (3), the group of fractional ideals can be identified with the free Abelian group generated by the prime ideals $\mathfrak{p}$. Furthermore, a number field $K$ has *places*, corresponding to equivalence classes of valuations. The finite places, which correspond to non-Archimedean valuations, can be identified with the (non-zero) prime ideals of $\mathbb{Z}_K$. The other (so called infinite places) correspond to Archimedean valuations and can be identified with the embeddings $\sigma_i$ of $K$ in $\mathbb{C}$, with $\sigma$ identified with $\bar{\sigma}$ (thus giving $r_1 + r_2$ Archimedean valuations). Hence, we can consider the extended group which is the free Abelian group generated by all valuations, finite or not. One can show that to obtain a sensible definition, the coefficients of the non-real complex embeddings must be considered modulo 1, i.e. can be taken equal to 0, and the coefficients of the real embeddings must be considered modulo 2 (I do not give the justification for these claims). Hence, the group of generalized fractional ideals is

$$\mathbb{Z}[\mathcal{P}(K)] \times \{\pm 1\}^{r_1},$$

where $\mathcal{P}(K)$ is the set of non-zero prime ideals. The norm of such a generalized ideal is then the norm of its finite part multiplied by the infinite components (i.e. by a sign). Now if $x \in K$, the generalized fractional ideal associated to $x$ is, on the finite part equal to $x\mathbb{Z}_K$, and on the infinite place $\sigma_i$ (where $1 \leq i \leq r_1$) equal to the sign of $\sigma_i(x)$. It is then easy to check that these two notions of norm now correspond exactly, including sign.

The discussion above was meant as an aside, but is the beginning of the theory of adeles and ideles (see [Lang2]). In a down to earth way, we can say that most natural questions concerning number fields should treat together the Archimedean and non-Archimedean places (or primes). In addition to the present example, we have already mentioned the parallel between Propositions 4.1.14 and 4.8.6. Similarly, we will see Propositions 4.8.11 and 4.8.10. Maybe the most important consequence is that we will have to compute simultaneously class groups (i.e. the non-Archimedean part) and regulators (the Archimedean part), see Sections 4.9, 5.9 and 6.5.

## 4.7 Representation of Modules and Ideals

### 4.7.1 Modules and the Hermite Normal Form

As before, we work in a fixed number field $K$ of degree $n$, given by $K = \mathbb{Q}(\theta)$, where $\theta$ is an algebraic integer whose minimal monic polynomial is denoted $T(X)$.

**Definition 4.7.1.** *A* module *in $K$ is a finitely generated sub-$\mathbb{Z}$-module of $K$ of rank exactly equal to $n$.*

Since $\mathbb{Z}$ is a PID, such a module being torsion free and finitely generated, must be free. Let $\omega_1, \ldots, \omega_n$ be a $\mathbb{Z}$-basis of $M$. The numbers $\omega_i$ are elements of $K$, hence we can find an integer $d$ such that $d\omega_i \in \mathbb{Z}[\theta]$ for all $i$. The least such positive $d$ will be called the *denominator* of $M$ with respect to $\mathbb{Z}[\theta]$. More generally, if $R$ is another module (for example $R = \mathbb{Z}_K$), we define the denominator of $M$ with respect to $R$ as the smallest positive $d$ such that $dM \subset R$.

Note that in the context of number fields, the word "module" will always have the above meaning, in other words it will always refer to a submodule of maximal rank $n$. If as a $\mathbb{Q}$-vector space we identify $K = \mathbb{Q}(\theta)$ with $\mathbb{Q}^n$, and $\mathbb{Z}[\theta]$ with $\mathbb{Z}^n$, the above definition is the same as the one that we have given in Section 2.4.3. In particular, we can use the notions of determinant, HNF and SNF of modules.

We give the following proposition without proof.

**Proposition 4.7.2.** *Let $M$ be a module in a number field $K$ in the above sense. Then there exists an order $R$ in $K$ and a positive integer $d$ such that $dM$ is an ideal of $R$. More precisely, there is a maximal such $R$ equal to $R = \{x \in K, xM \subset M\}$, and one can take for $d$ the denominator of $M$ with respect to $R$.*

Specializing to our case the results of Section 2.4.2, we obtain:

**Theorem 4.7.3.** *Let $\alpha_1, \ldots, \alpha_n$ be $n$ $\mathbb{Z}$-linearly independent elements of $K$, and $R$ be the module which they generate. Then for any module $M$, there exists a unique basis $\omega_1, \ldots, \omega_n$ such that if we write*

$$\omega_j = \frac{1}{d}\left(\sum_{i=1}^{n} w_{i,j}\alpha_i\right),$$

*where $d$ is the denominator of $M$ with respect to $R$, then the $n \times n$ matrix $W = (w_{i,j})$ satisfies the following conditions:*

(1) *For all $i$ and $j$ the $w_{i,j}$ are integers.*
(2) *$W$ is an upper triangular matrix, i.e. $w_{i,j} = 0$ if $i > j$.*
(3) *For every $i$, we have $w_{i,i} > 0$.*
(4) *For every $j > i$ we have $0 \leq w_{i,j} < w_{i,i}$.*

The corresponding basis $(\omega_i)_{1 \leq i \leq n}$ will be called the *HNF-basis* of $M$ with respect to $R$, and the pair $(W, d)$ will be called the HNF of $M$ (with respect to $R$). If $\alpha_i = \theta^{i-1}$, we will call $W$ (or $(W, d)$) the HNF with respect to $\theta$.

We have already seen in section 2.4.3 how to test equality and inclusion of modules, how to compute the sum of two modules and the product of a module by a constant. In the context of number fields, we can also compute the *product* of two modules. This will be used mainly for ideals.

Recall that

$$MM' = \{\sum_j m_j m'_j, m_j \in M, m'_j \in M'\}.$$

It is clear that $MM'$ is again a module. To obtain its HNF, we proceed as follows: Let $\omega_1, \ldots, \omega_n$ be the basis of $M$ obtained by considering the columns of the HNF of $M$ as the coefficients of $\omega_i$ in the standard representation, and similarly for $M'$. Then the $n^2$ elements $\omega_i \omega'_j$ form a generating set of $MM'$. Hence, if we find the HNF of the $n \times n^2$ matrix formed by their coefficients in the standard representation, we will have obtained the HNF of $MM'$.

Note however that this is quite costly, since $n^2$ can be pretty large. Another method might be as follows. In the case where $M$ and $M'$ are ideals (of $\mathbb{Z}_K$ say), then $M$ and $M'$ have a $\mathbb{Z}_K$-generating set formed by two elements. In fact, one of these two elements can even be chosen in $\mathbb{Z}$ if desired. Hence it is

clear that if $\omega_1, \ldots, \omega_n$ is a $\mathbb{Z}$-basis of $M$ and $\alpha, \beta$ a $\mathbb{Z}_K$-generating set of $M'$, then $\alpha\omega_1, \ldots, \alpha\omega_n, \beta\omega_1, \ldots, \beta\omega_n$ will be a $\mathbb{Z}$-generating set of $MM'$ (note that $M$ *must* also be an ideal for this to be true). Hence we can obtain the HNF of $MM'$ more simply by finding the HNF of the $n \times 2n$ matrix formed by the coefficients of the above generating set in the standard representation.

We end this section by the following proposition, whose proof is easy and left to the reader (see Exercise 18 of Chapter 2).

**Proposition 4.7.4.** *Let $M$ be a module with denominator 1 with respect to a given $R$ (i.e. $M \subset R$), and $W = (w_{i,j})$ its HNF with respect to a basis $\alpha_1, \ldots, \alpha_n$ of $R$. Then the product of the $w_{i,i}$ (i.e. the determinant of $W$) is equal to the index $[R : M]$.*

This will be used, for example, when $R = \mathbb{Z}[\theta]$ or $R = \mathbb{Z}_K$.

### 4.7.2 Representation of Ideals

The Hermite normal form of an ideal with respect to $\theta$ has a special form, as is shown by the following theorem:

**Theorem 4.7.5.** *Let $M$ be a $\mathbb{Z}[\theta]$-module, let $(W, d)$ be its HNF with respect to the algebraic integer $\theta$, where $d$ is the denominator and $W = (w_{i,j})$ is an integral matrix in upper triangular HNF. Then for every $j$, $w_{j,j}$ divides all the elements of the $j \times j$ matrix formed by the first $j$ rows and columns. In other words, the HNF basis $\omega_1, \ldots, \omega_n$ of a $\mathbb{Z}[\theta]$-module has the form*

$$\omega_j = \frac{z_j}{d}\left(\theta^{j-1} + \sum_{1 \le i < j} h_{i,j}\theta^{i-1}\right),$$

*where the $z_j$ are positive integers such that $z_j \mid z_i$ for $i < j$, and the $h_{i,j}$ satisfy $0 \le h_{i,j} < z_i/z_j$ for $i < j$. Furthermore, $z_1$ is the smallest positive element of $dM \cap \mathbb{Z}$.*

*Proof.* Without loss of generality, we may assume $d = 1$. We prove the theorem by induction on $j$. It is trivially true for $j = 1$. Assume $j > 1$ and that it is true for $j - 1$. Consider the $(j-1)^{\text{th}}$ basis element $\omega_{j-1}$ of $M$. We have

$$\omega_{j-1} = \sum_{1 \le i < j} w_{i,j-1}\theta^{i-1}$$

hence $\theta\omega_{j-1} = w_{j-1,j-1}\theta^{j-1} + \sum_{1 \le i < j-1} w_{i,j-1}\theta^i$. Since $M$ is a $\mathbb{Z}[\theta]$-module, this must be again an element of $M$, hence it has the form $\theta\omega_{j-1} = \sum_{1 \le i \le n} a_i\omega_i$ with integers $a_i$. Now since we have a triangular basis, identification of coefficients (from $\theta^{n-1}$ downwards) shows that $a_i = 0$ for $i > j$

and that $a_j w_{j,j} = w_{j-1,j-1}$. This already shows that $w_{j,j} \mid w_{j-1,j-1}$. But by induction, we know that $w_{j-1,j-1}$ divides $w_{i',j'}$ when $i'$ and $j'$ are less than or equal to $j-1$. It follows that, modulo $w_{j-1,j-1}\mathbb{Z}[\theta]$ we have

$$0 \equiv \theta\omega_{j-1} \equiv a_j\omega_j \equiv \frac{w_{j-1,j-1}}{w_{j,j}} \sum_{1\le i\le j} w_{i,j}\theta^{i-1},$$

and this means that for every $i \le j$ we have

$$\frac{w_{j-1,j-1}}{w_{j,j}} w_{i,j} \equiv 0 \pmod{w_{j-1,j-1}},$$

which is equivalent to $w_{j,j} \mid w_{i,j}$ for $i \le j$, thus proving the theorem by induction. □

Note that the converse of this theorem is false (see Exercise 16).

Theorem 4.7.5 will be mainly used in two cases. First when $M$ is an ideal of $\mathbb{Z}_K$. The second is when $M$ is an order containing $\theta$. In that case one can say slightly more:

**Corollary 4.7.6.** *Let $R$ be an order in $K$ containing $\theta$ (hence containing $\mathbb{Z}[\theta]$). Then the HNF basis $\omega_1, \ldots, \omega_n$ of $R$ with respect to $\theta$ has the form*

$$\omega_j = \frac{1}{d_j}\left(\theta^{j-1} + \sum_{1\le i<j} h_{i,j}\theta^{i-1}\right),$$

*where the $d_j$ are positive integers such that $d_i \mid d_j$ for $i<j$, $d_1=1$, and the $h_{i,j}$ satisfy $0 \le h_{i,j} < d_j/d_i$ for $i<j$. In other words, with the notations of Theorem 4.7.5, we have $z_j \mid d$ for all $j$.*

The proof is clear once one notices that the smallest positive integer belonging to an order is 1, hence by Theorem 4.7.5 that $z_1 = d$. □

If we assume that $R = \mathbb{Z}_K$ is given by an integral basis $\alpha_1, \ldots, \alpha_n$, then the HNF matrix of an ideal $I$ with respect to this basis does *not* usually satisfy the conditions of Theorem 4.7.5. We can always assume that we have chosen $\alpha_1=1$, and in that case it is easy to show in a similar manner as above that $w_{1,1}$ is divisible by $w_{i,i}$ for all $i$, and that if $w_{i,i} = w_{1,1}$, then $w_{j,i} = 0$ for $j \ne i$. This is left as an exercise for the reader (see Exercise 17).

Hence, depending on the context, we will represent an ideal of $\mathbb{Z}_K$ by its Hermite normal form with respect to a fixed integral basis of $\mathbb{Z}_K$, or by its HNF with respect to $\theta$ (i.e. corresponding to the standard representations of the basis elements). Please note once again that the special form of the HNF described in Theorem 4.7.5 is valid only in this last case.

Whichever representation is chosen, we have seen in Sections 2.4.3 and 4.7.1 how to compute sums and products of ideals, to test equality and inclusion (i.e. divisibility). Finally, as has already been mentioned several times, the norm is the absolute value of the determinant of the matrix, and in the HNF case this is simply the product of the diagonal elements.

Note that to test whether an element of $K$ is in a given ideal is a special case of the inclusion test, since $x \in I \iff xR \subset I$. Here however it is simpler (although not so much more efficient) to solve a (triangular) system of linear equations: if $(W, d)$ is the HNF of $I$ with respect to $\theta$, then if $x = (\sum_{1 \le i \le n} x_i \theta^{i-1})/e$ is the standard representation of $x$, we must solve the equation $WA = \frac{d}{e}X$ where $X$ is the column vector of the $x_i$, and $A$ is the unknown column vector. Since $W$ is triangular, this is especially simple, and $x \in I$ if and only if $A$ has integral coefficients.

To this point, we have considered ideals mainly as $\mathbb{Z}$-modules. There is a completely different way to represent them based on the following proposition.

**Proposition 4.7.7.** *Let $I$ be an integral ideal of $\mathbb{Z}_K$.*

(1) *For any non-zero element $\alpha \in I$ there exists an element $\beta \in I$ such that $I = \alpha\mathbb{Z}_K + \beta\mathbb{Z}_K$.*
(2) *There exists a non-zero element in $I \cap \mathbb{Z}$. If we denote by $\ell(I)$ the smallest positive element of $I \cap \mathbb{Z}$, then $\ell(I)$ is a divisor of $\mathcal{N}(I) = [\mathbb{Z}_K : I]$. In particular, there exists $\beta \in I$ such that $I = \ell(I)\mathbb{Z}_K + \beta\mathbb{Z}_K$.*
(3) *If $\alpha$ and $\beta$ are in $K$, then $I = \alpha\mathbb{Z}_K + \beta\mathbb{Z}_K$ if and only if for every prime ideal $\mathfrak{p}$ we have $\min(v_\mathfrak{p}(\alpha), v_\mathfrak{p}(\beta)) = v_\mathfrak{p}(I)$ where $v_\mathfrak{p}$ denotes the $\mathfrak{p}$-adic valuation at the prime ideal $\mathfrak{p}$.*

To prove this proposition, we first prove a special case of the so-called approximation theorem valid in any Dedekind domain.

**Proposition 4.7.8.** *Let $S$ be a finite set of prime ideals of $\mathbb{Z}_K$ and $(e_i)$ a set of non-negative integers indexed by $S$. There exists a $\beta \in \mathbb{Z}_K$ such that for each $\mathfrak{p}_i \in S$ we have*

$$v_{\mathfrak{p}_i}(\beta) = e_i\,.$$

*(Note that there may exist prime ideals $\mathfrak{q}$ not belonging to $S$ such that $v_\mathfrak{q}(\beta) > 0$.)*

**Remark.** More generally, $S$ can be taken to be a set of *places* of $K$, and in particular can contain Archimedean valuations.

*Proof.* Let $r = |S|$,

$$I = \prod_{i=1}^{r} \mathfrak{p}_i^{e_i+1},$$

and for each $i$, set

$$\mathfrak{a}_i = I \cdot \mathfrak{p}_i^{-e_i-1} \ ,$$

which is still an integral ideal. It is clear that $\mathfrak{a}_1+\mathfrak{a}_2+\cdots+\mathfrak{a}_r = \mathbb{Z}_K$ (otherwise this sum would be divisible by one of the $\mathfrak{p}_i$, which is clearly impossible). Hence, let $u_i \in \mathfrak{a}_i$ such that $u_1 + u_2 + \cdots + u_r = 1$. Furthermore, for each $i$ choose $\beta_i \in \mathfrak{p}_i^{e_i} \setminus \mathfrak{p}_i^{e_i+1}$ which is possible since $\mathfrak{p}_i$ is invertible. Then I claim that

$$\beta = \sum_{i=1}^{r} \beta_i u_i$$

has the desired property. Indeed, since $\mathfrak{p}_i \mid \mathfrak{a}_j$ for $i \neq j$, it is easy to check from the definition of the $\mathfrak{a}_i$ that

$$v_{\mathfrak{p}_i}(\beta) = v_{\mathfrak{p}_i}(\beta_i u_i) = e_i$$

since $v_{\mathfrak{p}_i}(u_i) = 0$ and $v_{\mathfrak{p}_i}(\beta_i) = e_i$. Note that this is simply the proof of the Chinese remainder theorem for ideals. □

*Proof of Proposition 4.7.7.* (1) Let $\alpha\mathbb{Z}_K = \prod_{i=1}^{r} \mathfrak{p}_i^{a_i}$ be the prime ideal decomposition of the principal ideal generated by $\alpha$. Since $\alpha \in I$, we also have $I = \prod_{i=1}^{r} \mathfrak{p}_i^{e_i}$ for exponents $e_i$ (which may be equal to zero) such that $e_i \leq a_i$. According to Proposition 4.7.8 that we have just proved, there exists a $\beta$ such that $v_{\mathfrak{p}_i}(\beta) = e_i$ for $i \leq r$. This implies in particular that $I \mid \beta$, i.e. that $\beta \in I$, and furthermore if we set $I' = \alpha\mathbb{Z}_K + \beta\mathbb{Z}_K$ we have for $i \leq r$

$$v_{\mathfrak{p}_i}(I') = \min(v_{\mathfrak{p}_i}(\alpha), v_{\mathfrak{p}_i}(\beta)) = e_i$$

and if $\mathfrak{q}$ is a prime ideal which does not divide $\alpha$, $v_{\mathfrak{q}}(I') = 0$, from which it follows that $I' = \prod_{i=1}^{r} \mathfrak{p}_i^{e_i} = I$, thus proving (1).

For (2), we note that since $\mathcal{N}(I) = [\mathbb{Z}_K : I]$, any element of the Abelian quotient group $\mathbb{Z}_K/I$ is annihilated by $\mathcal{N}(I)$, in other words we have $\mathcal{N}(I)\mathbb{Z}_K \subset I$. This implies $\mathcal{N}(I) \in I \cap \mathbb{Z}$, and since any subgroup of $\mathbb{Z}$ is of the form $k\mathbb{Z}$, (2) follows.

Finally, for (3) recall that the sum of ideals correspond to taking a GCD, and that the GCD is computed by taking the minimum of the $\mathfrak{p}$-adic valuations. □

Hence every ideal has a *two element representation* $(\alpha, \beta)$ where $I = \alpha\mathbb{Z}_K + \beta\mathbb{Z}_K$, and we can take for example $\alpha = \ell(I)$. This two element representation is however difficult to handle: for the sum or product of two ideals, we get four generators over $\mathbb{Z}_K$, and we must get back to two. More generally, it is not very easy to go from the HNF (or more generally any $\mathbb{Z}$-basis $n$-element representation) to a two element representation.

There are however two cases in which that representation is useful. The first is in the case of quadratic fields ($n = 2$), and we will see this in Chapter 5. The other, which has already been mentioned in Section 4.7.1, is as follows:

we will see in Section 4.9 that prime ideals do not come out of the blue, and that in algorithmic practice most prime ideals $\mathfrak{p}$ are obtained as a two element representation $(p, x)$ where $p$ is a prime number and $x$ is an element of $\mathfrak{p}$. To go from that two element representation to the HNF form is easy, but is not desirable in general. Indeed, what one usually does with a prime ideal is to multiply it with some other ideal $I$. If $\omega_1, \ldots, \omega_n$ is a $\mathbb{Z}$-basis of $I$ (for example the basis obtained from the HNF form of $I$ on the given integral basis of $\mathbb{Z}_K$), then we can build the HNF of the product $\mathfrak{p}I$ by computing the $n \times 2n$ matrix of the generating set $p\omega_1, \ldots p\omega_n, x\omega_1, \ldots, x\omega_n$ expressed on the integral basis, and then do HNF reduction. As has already been mentioned in Section 4.7.1, this is more efficient than doing a $n \times n^2$ HNF reduction if we used both HNF representations. Note that if one really wants the HNF of $\mathfrak{p}$ itself, it suffices to apply the preceding algorithm to $I = \mathbb{Z}_K$.

Note that if $(W, d)$ (with $W = (w_{i,j})$) is the HNF of $I$ with respect to $\theta$, and if $f = [\mathbb{Z}_K : \mathbb{Z}[\theta]]$, then $\ell(I) = w_{1,1}$ and $d^n N(I) = [\mathbb{Z}_K : dI] = f \prod_{1 \le i \le n} w_{i,i}$ so

$$N(I) = d^{-n} f \prod_{1 \le i \le n} w_{i,i}.$$

Now it often happens that prime ideals are not given by a two element representation but by a larger number of generating elements. If this ideal is going to be used repeatedly, it is worthwhile to find a two element representation for it. As we have already mentioned this is not an easy problem in general, but in the special case of prime ideals we can give a reasonably efficient algorithm. This is based on the following lemma.

**Lemma 4.7.9.** *Let $\mathfrak{p}$ be a prime ideal above $p$ of norm $p^f$ ($f$ is called the residual degree of $\mathfrak{p}$ as we will see in the next section), and let $\alpha \in \mathfrak{p}$. Then we have $\mathfrak{p} = (p, \alpha) = p\mathbb{Z}_K + \alpha\mathbb{Z}_K$ if and only if $v_p(\mathcal{N}(\alpha)) = f$ or $v_p(\mathcal{N}(\alpha+p)) = f$, where $v_p$ denotes the ordinary $p$-adic valuation.*

*Proof.* This proof assumes some results and definitions introduced in the next section. Assume first that $v_p(\mathcal{N}(\alpha)) = f$. Then, since $\alpha \in \mathfrak{p}$ and $\mathcal{N}(\mathfrak{p}) = p^f$, for every prime $\mathfrak{q}$ above $p$ and different from $\mathfrak{p}$ we must have $v_\mathfrak{q}(\alpha) = 0$ otherwise $\mathfrak{q}$ would contribute more powers of $p$ to $\mathcal{N}(\alpha)$. In addition and for the same reason we must have $v_\mathfrak{p}(\alpha) = 1$. It follows that for any prime ideal $\mathfrak{q}$, $\min(v_\mathfrak{q}(p), v_\mathfrak{q}(\alpha)) = v_\mathfrak{q}(\mathfrak{p})$ and so $\mathfrak{p} = (p, \alpha)$ by Proposition 4.7.7 (3).

If $v_p(\mathcal{N}(\alpha + p)) = f$ we deduce from this that $\mathfrak{p} = p\mathbb{Z}_K + (\alpha + p)\mathbb{Z}_K$, but this is clearly also equal to $p\mathbb{Z}_K + \alpha\mathbb{Z}_K$.

Conversely, let $\mathfrak{p} = p\mathbb{Z}_K + \alpha\mathbb{Z}_K$. Then for every prime ideal $\mathfrak{q}$ above $p$ and different from $\mathfrak{p}$ we have $v_\mathfrak{q}(\alpha) = 0$, while for $\mathfrak{p}$ we can only say that $\min(v_\mathfrak{p}(p), v_\mathfrak{p}(\alpha)) = 1$.

Assume first that $v_\mathfrak{p}(\alpha) = 1$. Then clearly $v_p(\mathcal{N}(\alpha)) = v_p(\mathcal{N}(\mathfrak{p})) = f$ as desired. Otherwise we have $v_\mathfrak{p}(\alpha) > 1$, and hence $v_\mathfrak{p}(p) = 1$. But then we will have $v_\mathfrak{p}(\alpha + p) = 1$ (otherwise $v_\mathfrak{p}(p) = v_\mathfrak{p}((p + \alpha) - \alpha) > 1$), and still

$v_{\mathfrak{q}}(\alpha+p)=0$ for all other primes $\mathfrak{q}$ above $p$, and so $v_p(\mathcal{N}(\alpha+p))=f$ as before, thus proving the lemma. □

Note that the condition $v_p(\mathcal{N}(\alpha))=f$, while sufficient, is not a necessary condition (see Exercise 20).

Note also that if we write $\alpha=\sum_{1\le i\le k}\lambda_i\gamma_i$ where the $\gamma_i$ is some generating set of $\mathfrak{p}$, we may always assume that $|\lambda_i|\le p/2$ since $p\in\mathfrak{p}$. In addition, if we choose $\gamma_1=p$, we may assume that $\lambda_1=0$.

This suggests the following algorithm, which is simple minded but works quite well.

**Algorithm 4.7.10** (Two-Element Representation of a Prime Ideal). Given a prime ideal $\mathfrak{p}$ above $p$ by a system of $\mathbb{Z}$-generators $\gamma_i$ for $(1\le i\le k)$, this algorithm computes a two-element representation $(p,\alpha)$ for $\mathfrak{p}$.

We assume that one knows the norm $p^f$ of $\mathfrak{p}$ (this is always the case in practice, and in any case it can be obtained by computing the HNF of $\mathfrak{p}$ from the given generators), and that $\gamma_1=p$ (if this is not the case just add it to the list of generators).

1. [Initialize] Set $R\leftarrow 1$.
2. [Set coefficients] For $2\le i\le k$ set $\lambda_i\leftarrow R$.
3. [Compute $\alpha$ and check] Let $\alpha\leftarrow\sum_{2\le i\le k}\lambda_i\gamma_i$, $n\leftarrow\mathcal{N}(\alpha)/p^f$, where the norm is computed, for example, using the sub-resultant algorithm (see Section 4.3). If $p\nmid n$, then output $(p,\alpha)$ and terminate the algorithm. Otherwise, set $n\leftarrow\mathcal{N}(\alpha+p)/p^f$. If $p\nmid n$ then output $(p,\alpha)$ and terminate the algorithm.
4. [Decrease coefficients] Let $j$ be the largest $i\le k$ such that $\lambda_i\ne -R$ (we will always keep $\lambda_2\ge 0$ so $j$ will exist). Set $\lambda_j\leftarrow\lambda_j-1$ and for $j+1\le i\le k$ set $\lambda_i\leftarrow R$.
5. [Search for first non-zero] Let $j$ be the smallest $i\le k$ such that $\lambda_i\ne 0$. If no such $j$ exists (i.e. if all the $\lambda_i$ are equal to 0) set $R\leftarrow R+1$ and go to step 2. Otherwise go to step 3.

**Remarks.**

(1) Steps 4 and 5 of this algorithm represent a standard backtracking procedure. What we do essentially is to search for $\alpha=\sum_{2\le i\le k}\lambda_i\gamma_i$, where the $\lambda_i$ are integers between $-R$ and $R$. To avoid searching both for $\alpha$ and $-\alpha$, we add the condition that the first non-zero $\lambda$ should be positive. If the search fails, we start it again with a larger value of $R$. Of course, some time will be wasted since many old values of $\alpha$ will be recomputed, but in practice this has no real importance, and in fact $R=1$ or $R=2$ is usually sufficient. The remark made after Lemma 4.7.9 shows that the algorithm will stop with $R\le p/2$.

(2) It is often the case that one of the $\gamma_i$ for $2 \leq i \leq k$ will satisfy one of the conditions of step 3. Thus it is useful to test this before starting the backtracking procedure.

We refer to [Poh-Zas] for extensive information on the use of two-element representations.

## 4.8 Decomposition of Prime Numbers I

For simplicity, we continue to work with a number field $K$ considered as an extension of $\mathbb{Q}$, and not considered as a relative extension. Many of the theorems or algorithms which are explained in that context are still true in the more general case, but some are not. (For example, we have already seen this for the existence of integral bases.) Almost always, these generalizations fail because the ring of integers of the base field is not a PID (or equivalently a UFD).

### 4.8.1 Definitions and Main Results

The main results concerning the decomposition of primes are as follows. We always implicitly assume that the prime ideals are non-zero.

**Proposition 4.8.1.**

(1) *If $\mathfrak{p}$ is a prime ideal of $K$, then $\mathfrak{p} \cap \mathbb{Z} = p\mathbb{Z}$ for some prime number $p$.*
(2) *If $p$ is a prime number and $\mathfrak{p}$ is a prime ideal of $K$, the following conditions are equivalent:*
   (i) $\mathfrak{p} \supset p\mathbb{Z}$.
   (ii) $\mathfrak{p} \cap \mathbb{Z} = p\mathbb{Z}$.
   (iii) $\mathfrak{p} \cap \mathbb{Q} = p\mathbb{Z}$.
(3) *For any prime number $p$ we have $p\mathbb{Z}_K \cap \mathbb{Z} = p\mathbb{Z}$.*

More generally, we have $a\mathbb{Z}_K \cap \mathbb{Z} = a\mathbb{Z}$ for any integer $a$, prime or not.

**Definition 4.8.2.** *If $\mathfrak{p}$ and $p$ satisfy one of the equivalent conditions of Proposition 4.8.1 (2), we say that $\mathfrak{p}$ is a prime ideal* above *$p$, and that $p$ is* below *$\mathfrak{p}$.*

**Theorem 4.8.3.** *Let $p$ be a prime number. There exist positive integers $e_i$ such that*

$$p\mathbb{Z}_K = \prod_{i=1}^{g} \mathfrak{p}_i^{e_i},$$

*where the $\mathfrak{p}_i$ are all the prime ideals above $p$.*

**Definition 4.8.4.** *The integer $e_i$ is called the* ramification index *of $p$ at $\mathfrak{p}_i$ and is denoted $e(\mathfrak{p}_i/p)$. The degree $f_i$ of the field extension defined by*

$$f_i = [\mathbb{Z}_K/\mathfrak{p}_i : \mathbb{Z}/p\mathbb{Z}]$$

*is called the* residual degree *(or simply the* degree*) of $p$ and is denoted $f(\mathfrak{p}_i/p)$.*

Note that both $\mathbb{Z}_K/\mathfrak{p}_i$ and $\mathbb{Z}/p\mathbb{Z}$ are finite fields, and $f_i$ is the dimension of the first considered as a vector space over the second.

**Theorem 4.8.5.** *We have the following formulas:*

$$\mathcal{N}(\mathfrak{p}_i) = p^{f_i},$$

*and*

$$\sum_{i=1}^{g} e_i f_i = n = \deg(K).$$

In the case when $K/\mathbb{Q}$ is a Galois extension, the result is more specific:

**Theorem 4.8.6.** *Assume that $K/\mathbb{Q}$ is a Galois extension (i.e. that for all the embeddings $\sigma_i$ of $K$ in $\mathbb{C}$ we have $\sigma_i(K) = K$). Then, for any $p$, the ramification indices $e_i$ are equal (say to $e$), the residual degrees $f_i$ are equal as well (say to $f$), hence $efg = n$. In addition, the Galois group operates transitively on the prime ideals above $p$: if $\mathfrak{p}_i$ and $\mathfrak{p}_j$ are two ideals above $p$, there exists $\sigma$ in the Galois group such that $\sigma(\mathfrak{p}_i) = \mathfrak{p}_j$.*

**Definition 4.8.7.** *Let $p\mathbb{Z}_K = \prod_{i=1}^{g} \mathfrak{p}_i^{e_i}$ be the decomposition of a prime $p$. We will say that $p$ is* inert *if $g = 1$ and $e_1 = 1$, in other words if $p\mathbb{Z}_K = \mathfrak{p}$ (hence $f_1 = n$). We will say that $p$* splits completely *if $g = n$ (hence for all $i$, $e_i = f_i = 1$). Finally, we say that $p$ is* ramified *if there is an $e_i$ which is greater than or equal to 2 (in other words if $p\mathbb{Z}_K$ is not squarefree), otherwise we say that $p$ is* unramified. *Those prime ideals $\mathfrak{p}_i$ such that $e_i > 1$ are called the ramified prime ideals of $\mathbb{Z}_K$.*

Note that there are intermediate cases which do not deserve a special name. The fundamental theorem about ramification is as follows:

**Theorem 4.8.8.** *Let $p$ be a prime number. Then $p$ is ramified in $K$ if and only if $p$ divides the discriminant $d(K)$ of $K$ (recall that this is the discriminant of any integral basis of $\mathbb{Z}_K$). In particular, there are only a finite number of ramified primes (exactly $\omega(d(K))$, where $\omega(x)$ is the number of distinct prime divisors of an integer $x$).*

We can also define the decomposition of the "infinite prime" of $\mathbb{Q}$ in a similar manner, since we are extending valuations. The ordinary primes correspond to the non-Archimedean valuations and the real or complex embeddings correspond to the Archimedean ones. Since we are over $\mathbb{Q}$, there is only the real embedding of $\mathbb{Q}$ to lift, and (as a special case of a general definition), when the signature of $K$ is $(r_1, r_2)$, we will say that the infinite prime of $\mathbb{Q}$ lifts to a product of $r_1$ real places of $K$ times $r_2$ non-real places to the power 2. Hence, $g = r_1 + r_2$, $e_i = 1$ for $i \leq r_1$, $e_i = 2$ for $i > r_1$, and $f_i = 1$ for all $i$.

We also have the following results:

**Proposition 4.8.9.**

(1) (Hermite). *The set of isomorphism classes of number fields of given discriminant is finite.*
(2) (Minkowski). *If $K$ is a number field different from $\mathbb{Q}$, then $|d(K)| > 1$. In particular, there is at least one ramified prime in $K$.*

**Proposition 4.8.10** (Stickelberger). *If $p$ is an unramified prime in $K$ with $p\mathbb{Z}_K = \prod_{i=1}^{g} \mathfrak{p}_i$, we have*

$$\left(\frac{d(K)}{p}\right) = (-1)^{n-g}$$

*including the case $p = 2$ where $\left(\frac{d(K)}{2}\right)$ is to be interpreted as the Jacobi-Kronecker symbol (see Definition 1.4.8).*

This shows that the parity of the number of primes above $p$ (i.e. the "Möbius" function of $p$) can easily be computed.

Note that this proposition is also true for the infinite prime as given above, if we interpret the Legendre symbol as the sign of $d(K)$:

**Proposition 4.8.11.** *If $K$ is a number field with signature $(r_1, r_2)$, then the sign of the discriminant $d(K)$ is equal to $(-1)^{r_2}$.*

*Proof.* Since, up to a square, the discriminant $d(K)$ is equal to $\prod_{i<j}(\theta_i - \theta_j)^2$ (with evident notations), then a case by case examination shows that when conjugate terms are paired, all the factors become positive except for

$$\prod_{r_1 < i \leq r_1 + r_2} (\theta_i - \theta_{i+r_2})^2,$$

whose sign is $(-1)^{r_2}$ since $\theta_i - \theta_{i+r_2}$ is pure imaginary. □

**Corollary 4.8.12.** *The decomposition type of a prime number $p$ in a quadratic field $K$ of discriminant $D$ is the following: if $\left(\frac{D}{p}\right) = -1$ then $p$ is inert. If*

$\left(\frac{D}{p}\right) = 0$ *then* $p$ *is ramified (i.e.* $p\mathbb{Z}_K = \mathfrak{p}^2$*). Finally, if* $\left(\frac{D}{p}\right) = +1$*, then* $p$ *splits (completely), i.e.* $p\mathbb{Z}_K = \mathfrak{p}_1\mathfrak{p}_2$.

### 4.8.2 A Simple Algorithm for the Decomposition of Primes

We now consider a more difficult algorithmic problem, that of determining the decomposition of prime numbers in a number field. The basic theorem on the subject, which unfortunately is not completely sufficient, is as follows.

**Theorem 4.8.13.** *Let* $K = \mathbb{Q}(\theta)$ *be a number field, where* $\theta$ *is an algebraic integer, whose (monic) minimal polynomial is denoted* $T(X)$*. Let* $f$ *be the index of* $\theta$*, i.e.* $f = [\mathbb{Z}_K : \mathbb{Z}[\theta]]$*. Then for any prime* $p$ *not dividing* $f$ *one can obtain the prime decomposition of* $p\mathbb{Z}_K$ *as follows. Let*

$$T(X) \equiv \prod_{i=1}^{g} T_i(X)^{e_i} \pmod{p}$$

*be the decomposition of* $T$ *into irreducible factors in* $\mathbb{F}_p[X]$*, where the* $T_i$ *are taken to be monic. Then*

$$p\mathbb{Z}_K = \prod_{i=1}^{g} \mathfrak{p}_i^{e_i},$$

*where*

$$\mathfrak{p}_i = (p, T_i(\theta)) = p\mathbb{Z}_K + T_i(\theta)\mathbb{Z}_K.$$

*Furthermore, the residual index* $f_i$ *is equal to the degree of* $T_i$.

Since we have discussed at length in Chapter 3 algorithmic methods for finding the decomposition of polynomials in $\mathbb{F}_p[X]$, we see that this theorem gives us an excellent algorithmic method to find the decomposition of $p\mathbb{Z}_K$ when $p$ does not divide the index $f$. The hard problems start when $p \mid f$. Of course, one then could try and change $\theta$ to get a different index, if possible prime to $p$, but even this is doomed. There can exist primes, called *inessential discriminantal divisors* which divide any index, no matter which $\theta$ is chosen. It can be shown that such exceptional primes are smaller than or equal to $n-1$, so very few primes if any are exceptional. But the problem still exists: for example it is not difficult to give examples of fields of degree 3 where 2 is exceptional, see Exercise 10 of Chapter 6.

The case when $p$ divides the index is much harder, and will be studied along with an algorithm to find integral bases in Chapter 6.

*Proof of Theorem 4.8.13.* Set $f_i = \deg(T_i)$ and $\mathfrak{p}_i = p\mathbb{Z}_K + T_i(\theta)\mathbb{Z}_K$. Let us assume that we have proved the following lemma:

**Lemma 4.8.14.**

(1) *For all $i$, either $\mathfrak{p}_i = \mathbb{Z}_K$, or $\mathbb{Z}_K/\mathfrak{p}_i$ is a field of cardinality $p^{f_i}$.*
(2) *If $i \neq j$ then $\mathfrak{p}_i + \mathfrak{p}_j = \mathbb{Z}_K$.*
(3) $p\mathbb{Z}_K \mid \mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_g^{e_g}$.

Then, after reordering the $\mathfrak{p}_i$, we can assume that $\mathfrak{p}_i \neq \mathbb{Z}_K$ for $i \leq s$ and $\mathfrak{p}_i = \mathbb{Z}_K$ for $s < i \leq g$ (we will in fact see that $s = g$). Then by Lemma 4.8.14 (1), the ideals $\mathfrak{p}_i$ are prime for $i \leq s$, and since by definition they contain $p\mathbb{Z}_K$, they are above $p$ (Proposition 4.8.1). (1) also implies that the $f_i$ (for $i \leq s$) are the residual indices of $\mathfrak{p}_i$. By (2) we know that the $\mathfrak{p}_i$ for $i \leq s$ are distinct, and (3) implies that the decomposition of the ideal $p\mathbb{Z}_K$ is

$$p\mathbb{Z}_K = \prod_{i=1}^{s} \mathfrak{p}_i^{d_i} \quad \text{where } d_i \leq e_i \text{ for all } i \leq s.$$

Hence, by Theorem 4.8.5, we have $n = d_1 f_1 + \cdots + d_s f_s$. Since we also have $\deg(T) = n = e_1 f_1 + \cdots + e_g f_g$ and $d_i \leq e_i$ for all $i$, this implies that we must have $s = g$ and $d_i = e_i$ for all $i$, thus proving Theorem 4.8.13. □

*Proof of Lemma 4.8.14 (1).* Set $K_i = \mathbb{F}_p[X]/(T_i)$. Since $T_i$ is irreducible, $K_i$ is a field. Furthermore, the degree of $K_i$ over $\mathbb{F}_p$ is $f_i$, and so the cardinality of $K_i$ is $p^{f_i}$. Thus we need to show that either $\mathfrak{p}_i = \mathbb{Z}_K$ or that $\mathbb{Z}_K/\mathfrak{p}_i \simeq K_i$. Now it is clear that $\mathbb{Z}[X]/(p, T_i) \simeq K_i$, hence $(p, T_i)$ is a maximal ideal of $\mathbb{Z}[X]$. But the kernel of the natural homomorphism $\phi$ from $\mathbb{Z}[X]$ to $\mathbb{Z}_K/\mathfrak{p}_i$ which sends $X$ to $\theta$ mod $\mathfrak{p}_i$ clearly contains this ideal, hence is either $\mathbb{Z}[X]$ or $(p, T_i)$. If we show that $\phi$ is onto, this will imply that $\mathfrak{p}_i = \mathbb{Z}_K$ or $\mathbb{Z}_K/\mathfrak{p}_i \simeq \mathbb{Z}[X]/(p, T_i) \simeq K_i$, proving (1).

Now to say that $\phi$ is surjective means that $\mathbb{Z}_K = \mathbb{Z}[\theta] + \mathfrak{p}_i$. By definition, $p\mathbb{Z}_K \subset \mathfrak{p}_i$. Hence

$$[\mathbb{Z}_K : \mathbb{Z}[\theta] + \mathfrak{p}_i] \mid [\mathbb{Z}_K : \mathbb{Z}[\theta] + p\mathbb{Z}_K] = \gcd([\mathbb{Z}_K : \mathbb{Z}[\theta]], [\mathbb{Z}_K : p\mathbb{Z}_K]).$$

Since we have assumed that $p$ does not divide the index, and since $[\mathbb{Z}_K : p\mathbb{Z}_K] = p^n$, this shows that $[\mathbb{Z}_K : \mathbb{Z}[\theta] + \mathfrak{p}_i] = 1$, hence the surjectivity of $\phi$. Note that this is the only part of the whole proof of Theorem 4.8.13 which uses that $p$ does not divide the index of $\theta$.

*Proof of Lemma 4.8.14 (2).* Since $T_i$ and $T_j$ are coprime in $\mathbb{F}_p[X]$, there exist polynomials $U$ and $V$ such that $UT_i + VT_j - 1 \in p\mathbb{Z}[X]$. It follows that $U(\theta)T_i(\theta) + V(\theta)T_j(\theta) = 1 + pW(\theta)$ for some polynomial $W \in \mathbb{Z}[X]$, and this immediately implies that $1 \in \mathfrak{p}_i + \mathfrak{p}_j$, i.e. that $\mathfrak{p}_i + \mathfrak{p}_j = \mathbb{Z}_K$.

*Proof of Lemma 4.8.14 (3).* Set $\gamma_i = T_i(\theta)$, so $\mathfrak{p}_i = (p, \gamma_i)$. By distributivity, it is clear that

$$\mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_g^{e_g} \subset (p, \gamma_1^{e_1} \cdots \gamma_g^{e_g}).$$

Now I claim that $(p, \gamma_1^{e_1} \cdots \gamma_g^{e_g}) = p\mathbb{Z}_K$, from which (3) follows. Indeed, $\supset$ is trivial. Conversely we have by definition $T_1^{e_1} \cdots T_g^{e_g} - T \in p\mathbb{Z}[X]$ hence taking $X = \theta$ we obtain

$$\gamma_1^{e_1} \cdots \gamma_g^{e_g} \in p\mathbb{Z}[\theta] \subset p\mathbb{Z}_K,$$

proving our claim and the lemma. □

Note that in the general case where $p \mid f$ which will be studied in Chapter 6, the prime ideals $\mathfrak{p}_i$ above $p$ are still of the form $p\mathbb{Z}_K + T_i(\theta)\mathbb{Z}_K$, but now $T_i \in \mathbb{Q}[X]$ and does not always correspond to a factor of $T$ modulo $p$.

### 4.8.3 Computing Valuations

Once prime ideals are known in a number field $K$, we will often need to compute the $\mathfrak{p}$-adic valuation $v$ of an ideal $I$ given in its Hermite normal form, where $\mathfrak{p}$ is a prime ideal above $p$. We may, of course, assume that $I$ is an integral ideal. Then an obvious necessary condition for $v \neq 0$ is that $p \mid \mathcal{N}(I)$. Clearly this condition is not sufficient, since all primes above $p$ must "share" in some way the exponent of $p$ in $\mathcal{N}(I)$.

We assume that our prime ideal is given as $\mathfrak{p} = p\mathbb{Z}_K + \alpha\mathbb{Z}_K$ for a certain $\alpha \in \mathbb{Z}_K$. We will now describe an algorithm to compute $v_{\mathfrak{p}}(I)$, which was explained to me by H. W. Lenstra, but which was certainly known to Dedekind. It is based on the following proposition.

**Proposition 4.8.15.** *Let $R$ be an order in $K$ and $\mathfrak{p}$ a prime ideal of $R$. Then there exists $a \in K \setminus R$ such that $a\mathfrak{p} \subset R$. Furthermore, $\mathfrak{p}$ is invertible in $R$ if and only if $a\mathfrak{p} \not\subset \mathfrak{p}$, and in that case we have $\mathfrak{p}^{-1} = R + aR$.*

*Proof.* Let $x \in \mathfrak{p}$ be a non-zero element of $\mathfrak{p}$, and consider the non-zero ideal $xR$. By Proposition 4.6.13, there exist non-zero prime ideals $\mathfrak{q}_i$ such that $xR \supset \prod_{i \in E} \mathfrak{q}_i$ for some finite set $E$. Assume $E$ is chosen to be minimal in the sense that no proper subset of $E$ can have the same property. Since $\prod \mathfrak{q}_i \subset xR \subset \mathfrak{p}$, by Proposition 4.6.12 we must have $\mathfrak{q}_j \subset \mathfrak{p}$ for some $j \in E$, hence $\mathfrak{q}_j = \mathfrak{p}$ since both are maximal ideals. Set

$$\mathfrak{q} = \prod_{i \in E, i \neq j} \mathfrak{q}_i.$$

Then $\mathfrak{p}\mathfrak{q} \subset xR$ and $\mathfrak{q} \not\subset xR$ by the minimality of $E$. So choose $y \in \mathfrak{q}$ such that $y \notin xR$. Since $y\mathfrak{p} \subset xR$, the element $a = y/x$ satisfies the conditions of the proposition.

Finally, consider the ideal $\mathfrak{p} + a\mathfrak{p}$. Since it sits between the maximal ideal $\mathfrak{p}$ and $R$, it must be equal to one of the two. If it is equal to $R$, we cannot have $a\mathfrak{p} \subset \mathfrak{p}$, and since $(R + aR)\mathfrak{p} = R$, $\mathfrak{p}$ is invertible and $\mathfrak{p}^{-1} = R + aR$. If

it is equal to $\mathfrak{p}$, then $a\mathfrak{p} \subset \mathfrak{p}$, and $(R + aR)\mathfrak{p} = R\mathfrak{p}$. This implies that $\mathfrak{p}$ is not invertible since otherwise, by simplifying, we would have $R + aR = R$, hence $a \in R$. This proves the proposition. □

Knowing this proposition, it is easy to obtain an algorithm for computing a suitable value of $a$. Note that $a\mathfrak{p} \subset R$ hence $ap \in R$, so we write $a = \beta/p$ with $\beta \in R$. The conditions to be satisfied for $\beta$ are then $\beta \in R \setminus pR$ and $\beta\mathfrak{p} \subset pR$.

Let $\omega_1, \ldots, \omega_n$ be a $\mathbb{Z}$-basis of $R$, and let $\gamma_1, \ldots, \gamma_m$ be generators of $\mathfrak{p}$ (for example if $\mathfrak{p} = pR + \alpha R$ we take $\gamma_1 = p$ and $\gamma_2 = \alpha$). Then, if we write

$$\beta = \sum_{1 \le i \le n} x_i \omega_i ,$$

we want to find integers $x_i$ which are not all divisible by $p$ such that for all $j$ with $1 \le j \le m$ the coordinates of $(\sum x_i \omega_i)\gamma_j$ on the $\omega_i$ are all divisible by $p$. If we set

$$\omega_i \gamma_j = \sum_{1 \le k \le n} a_{i,j,k} \omega_k ,$$

we obtain for all $j$ and $k$

$$\sum_{1 \le i \le n} a_{i,j,k} x_i \equiv 0 \pmod{p}$$

which is a system of $mn$ equations in $n$ unknowns in $\mathbb{Z}/p\mathbb{Z}$ for which we want a non-trivial solution. Since there are many more equations than unknowns (if $m > 1$), there is, a priori, no reason for this system to have a non-trivial solution. The proposition that we have just proved shows that it does, and we can find one by standard Gaussian elimination in $\mathbb{Z}/p\mathbb{Z}$ (for example using Algorithm 2.3.1).

In the frequent special case where $m = 2$, $\gamma_1 = p$ and $\gamma_2 = \alpha$ for some $\alpha \in \mathbb{Z}_K$, the system simplifies considerably. For $j = 1$ the equations are trivial, hence we must simply solve the square linear system

$$\sum_{1 \le i \le n} a_{i,k} x_i \equiv 0 \pmod{p}$$

where $\omega_i \alpha = \sum_{1 \le k \le n} a_{i,k} \omega_k$.

From now on, we assume that $R = \mathbb{Z}_K$ so that all ideals are invertible. Let $I$ be an ideal of $\mathbb{Z}_K$ given by its HNF $(M, d)$ with respect to $\theta$, where $M$ is an $n \times n$ matrix. We want to compute $v_{\mathfrak{p}}(I)$, where $\mathfrak{p}$ is a prime ideal of $\mathbb{Z}_K$ (hence invertible). By the method explained above, we first compute $a$ such that $a \in K \setminus \mathbb{Z}_K$ and $a\mathfrak{p} \subset \mathbb{Z}_K$, and as above we set $\beta = ap \in \mathbb{Z}_K$. We may assume that $I$ is an integral ideal of $\mathbb{Z}_K$. (If $I = I'/d'$ with $I'$ an integral ideal and $d' \in \mathbb{Z}$, then clearly $v_{\mathfrak{p}}(I) = v_{\mathfrak{p}}(I') - e v_p(d')$, where $e$ is the ramification index of $\mathfrak{p}$.) Now we have the following lemma which is the *raison d'être* of $a$.

**Lemma 4.8.16.** *With the above notations, if $I$ is an integral ideal of $\mathbb{Z}_K$, then $I \subset \mathfrak{p}$ if and only if $aI \subset \mathbb{Z}_K$. In particular, $v_\mathfrak{p}(I)$ is the largest integer $v$ such that $a^v I \subset \mathbb{Z}_K$.*

*Proof.* If $I \subset \mathfrak{p}$, then $aI \subset a\mathfrak{p} \subset \mathbb{Z}_K$. Conversely, assume that $aI \subset \mathbb{Z}_K$, hence $a\mathfrak{p}I \subset \mathfrak{p}$. Since the prime ideal $\mathfrak{p}$ contains the product of the integral ideals $a\mathfrak{p}$ and $I$, Proposition 4.6.12 shows that $\mathfrak{p}$ contains one of the two. Now since $\mathfrak{p}$ is invertible, $\mathfrak{p}$ cannot contain $a\mathfrak{p}$ by the above proposition, hence $\mathfrak{p} \supset I$. The final claim about the value of $v_\mathfrak{p}(I)$ is an immediate consequence of the definitions. □

If, as above we set $a = \beta/p$ with $\beta \in \mathbb{Z}_K \setminus p\mathbb{Z}_K$, the condition $a^v I \subset \mathbb{Z}_K$ is equivalent to $\beta^v I \subset p^v \mathbb{Z}_K$. Let $(N, d)$ be the HNF of the maximal order $\mathbb{Z}_K$. By Corollary 4.7.6, we may assume that $N_{n,n} = 1$, by choosing $d = d_n$. Now since $I$ is an integral ideal, we have $dI \subset d\mathbb{Z}_K$, and $d\mathbb{Z}_K$ is represented by an integral matrix, hence $dI$ also, so the HNF with respect to $\theta$ of any integral ideal can be chosen of the form $(M, d)$ with the same $d$. Conversely, given $(M, d)$ where $M$ is an integral matrix in Hermite normal form representing a fractional ideal $I$, we can test whether $I$ is integral by checking $I + \mathbb{Z}_K = \mathbb{Z}_K$, hence by computing the HNF of a $n \times 2n$ matrix as explained in Section 2.4.3. In our situation, a better way is to compute the HNF $M'$ of $I$ with respect to the HNF basis of $\mathbb{Z}_K$ given by the matrix $N$ instead of with respect to $\theta$, where we allow $M'$ to have fractional entries. We clearly have

$$M' = N^{-1}M,$$

except that the non-diagonal entries may have to be reduced, and $I$ is an integral ideal if and only if $M'$ has integral entries.

Hence, let $(M_v, d)$ be the HNF of $\beta^v I$ with respect to $\theta$, $M'_v = N^{-1}M_v$ and set $c_v = (M'_v)_{n,n}$. Then a necessary condition for $\beta^v I$ to be contained in $p^v \mathbb{Z}_K$ is that $p^v \mid c_v$. This condition is in general not sufficient, but very often it is. For example, it is easy to show (see Exercise 21) that the condition is sufficient when $p$ does not divide the index $[\mathbb{Z}_K : \mathbb{Z}[\theta]]$, and in particular if $\mathbb{Z}_K = \mathbb{Z}[\theta]$. In the general case, we have to check the divisibility of all the coefficients of $M_v$ by $p^v$. This leads to the following algorithm.

**Algorithm 4.8.17** (Valuation at a Prime Ideal). Let $(N, d)$ be the HNF of the maximal order $\mathbb{Z}_K$, let $\mathfrak{p}$ be a prime ideal of $\mathbb{Z}_K$ above $p$ given by a generating system $\gamma_1, \ldots, \gamma_m$ over $\mathbb{Z}_K$ (for example $\gamma_1 = p$, $\gamma_2 = \alpha$ for some $\alpha \in \mathbb{Z}_K$), and let $I$ be an integral ideal of $\mathbb{Z}_K$ given by its HNF $(M, d')$. This algorithm computes the $\mathfrak{p}$-adic valuation $v_\mathfrak{p}(I)$ of the ideal $I$.

1. [Compute structure constants] Let $\omega_i$ be the HNF basis of $\mathbb{Z}_K$ corresponding to $(N, d)$. Compute the integers $a_{i,j,k}$ such that

$$\omega_i \gamma_j = \sum_{1 \le k \le n} a_{i,j,k} \omega_k$$

for $1 \leq i \leq n$ and $1 \leq j \leq m$. Note that $\omega_i\gamma_j$ is computed as a polynomial in $\theta$, and since $N$ is an upper triangular matrix it is easy to compute inductively the $a_{i,j,k}$ from $k = n$ down to $k = 1$.

2. [Compute $\beta$] Using ordinary Gaussian elimination over $\mathbb{F}_p$ or Algorithm 2.3.1, find a non-trivial solution to the system of congruences

$$\sum_{1 \leq i \leq n} a_{i,j,k}x_i \equiv 0 \pmod{p}.$$

Then set $\beta \leftarrow \sum_i x_i\omega_i$.

3. [Compute $\mathcal{N}(I)$] Set $A \leftarrow d/d'N^{-1}M$ which must be a matrix with integral entries (otherwise $I$ is not an integral ideal). Let $P$ be the product of the diagonal elements of $A$. If $p \nmid P$, output 0 and terminate the algorithm. Otherwise, set $v \leftarrow 0$.

4. [Multiply] Set $A \leftarrow \beta A$ in the following sense. Each column of $A$ corresponds to an element of $K$ in the basis $\omega_i$, and these elements are multiplied by $\beta$ and expressed again in the basis $\omega_i$, using the multiplication table for the $\omega_i$.

5. [Simple test] Using Algorithm 2.4.8, replace $A$ by its HNF. Then, if $p \nmid A_{n,n}$, output $v$ and terminate the algorithm. Otherwise, if $p$ does not divide the index $[\mathbb{Z}_K : \mathbb{Z}[\theta]] = d^n/\det(N)$, set $v \leftarrow v+1$, $A \leftarrow A/p$ (which will be integral) and go to step 4.

6. [Complete test] Set $A \leftarrow A/p$. If $A$ is not integral, output $v$ and terminate the algorithm. Otherwise, set $v \leftarrow v + 1$ and go to step 4.

Note that steps 1 and 2 depend only on the ideal $\mathfrak{p}$, hence need be done only once if many $\mathfrak{p}$-adic valuations have to be computed for the same prime ideal $\mathfrak{p}$. Hence, a reasonable way to represent a prime ideal $\mathfrak{p}$ is as a quintuplet $(p, \alpha, e, f, \beta)$. Here $p$ is the prime number over which $\mathfrak{p}$ lies, $\alpha \in \mathbb{Z}_K$ is such that $\mathfrak{p} = p\mathbb{Z}_K + \alpha\mathbb{Z}_K$, $e$ is the ramification index and $f$ the residual index of $\mathfrak{p}$, and $\beta$ is the element of $\mathbb{Z}_K$ computed by steps 1 and 2 of the above algorithm, given by its coordinates $x_i$ in the basis $\omega_i$. Note also that Proposition 4.8.15 tells us that $p\mathfrak{p}^{-1} = p\mathbb{Z}_K + \beta\mathbb{Z}_K$.

### 4.8.4 Ideal Inversion and the Different

The preceding algorithms will allow us to give a satisfactory answer to a problem which we have not yet studied, that of ideal inversion in $\mathbb{Z}_K$.

Let $I$ be an ideal of $\mathbb{Z}_K$ (which we can assume to be integral without loss of generality) given by a $\mathbb{Z}_K$-generating system $\gamma_1, \ldots, \gamma_m$. We can for example take the HNF basis of $I$ in which case $m = n$, but often $I$ will be given in a simpler way, for example by only 2 elements. We can try to mimic the first two steps of Algorithm 4.8.17 which, as remarked above, amount to computing the inverse of the prime ideal $\mathfrak{p}$.

Hence, let $\omega_1, \ldots, \omega_n$ be an integral basis of $\mathbb{Z}_K$. Then by definition of the inverse, $x \in I^{-1}$ if and only if $x\gamma_j \in \mathbb{Z}_K$ for all $j \leq k$. Fix a positive integer

$d$ belonging to $I$. Then $dx \in \mathbb{Z}_K$ so we can write $dx = \sum_{1 \le k \le n} x_k \omega_k$ with $x_k \in \mathbb{Z}$ and the condition $x \in I^{-1}$ can be written

$$\sum_{1 \le i \le n} x_k \gamma_j \omega_k \in d\mathbb{Z}_K \quad \text{for all } j.$$

If we define coefficients $u_{i,j,k} \in \mathbb{Z}$ by

$$\gamma_j \omega_k = \sum_{i=1}^{n} u_{i,j,k} \omega_i$$

we are thus led to the $nm \times n$ system of congruences $\sum_{1 \le k \le n} x_k u_{i,j,k} \equiv 0 \pmod{d}$ for all $i$ and $j$.

In the special case where $I$ is a prime ideal as in Algorithm 4.8.17, we can choose $d = p$ a prime number, and hence our system of congruences can be considered as a system of equations in the finite field $\mathbb{F}_p$, and we can apply Algorithm 2.3.1 to find a basis for the set of solutions. Here, $I$ is not a prime ideal in general, and we could try to solve the system of congruences by factoring $d$ and working modulo powers of primes. A better method is probably as follows. Introduce extra integer variables $y_{i,j}$. Then our system is equivalent to the $nm \times (n+nm)$ linear system $\sum_{1 \le k \le n} x_k u_{i,j,k} - d y_{i,j} = 0$ for all $i$ and $j$. We must find a $\mathbb{Z}$-basis of the solutions of this system, and for this we use the integral kernel Algorithm 2.7.2. The kernel will be of dimension $n$, and a $\mathbb{Z}$-basis of $dI^{-1}$ is then obtained by keeping only the first $n$ rows of the kernel (corresponding to the variables $x_k$).

In the common case where $m = n$, this algorithm involves $n^2 \times (n^2 + n)$ matrices, and this becomes large rather rapidly. Thus the algorithm is very slow as soon as $n$ is at all large, and hence we must find a better method. For this, we introduce an important notion in algebraic number theory, the different, referring to the introductory books mentioned at the beginning of this chapter for more details.

**Definition 4.8.18.** *Let $K$ be a number field. The* different $\mathfrak{d}(K)$ *of $K$ is defined as the inverse of the ideal (called the* codifferent*)*

$$\{x \in K, \quad \mathrm{Tr}_{K/\mathbb{Q}}(x\mathbb{Z}_K) \subset \mathbb{Z}\}.$$

It is clear that the different $\mathfrak{d}(K)$ is an integral ideal. What makes the different interesting in our context is the following proposition.

**Proposition 4.8.19.** *Let $(\omega_i)_{1 \le i \le n}$ be an integral basis and let $I$ be an ideal of $\mathbb{Z}_K$ given by an $n \times n$ matrix $M$ whose columns give the coordinates of a $\mathbb{Z}$-basis $(\gamma_i)_{1 \le i \le n}$ of $I$ on the chosen integral basis. Let $T = (t_{i,j})$ be the $n \times n$ matrix such that $t_{i,j} = \mathrm{Tr}_{K/\mathbb{Q}}(\omega_i \omega_j)$. Then the columns of the matrix $(M^t T)^{-1}$*

*(again considered as coordinates on our integral basis) form a $\mathbb{Z}$-basis of the ideal $I^{-1}\mathfrak{d}(K)^{-1}$.*

*Proof.* First, note that by definition of $M$, the coefficient of row $i$ and column $j$ in $M^tT$ is equal to $\mathrm{Tr}_{K/\mathbb{Q}}(\gamma_i\omega_j)$. Furthermore, if $V=(v_i)$ is a column vector, then $V$ belongs to the lattice spanned by the columns of $(M^tT)^{-1}$ if and only if $M^tTV$ has integer coefficients. This implies that for all $i$ $\mathrm{Tr}_{K/\mathbb{Q}}(\gamma_i(\sum_j v_j\omega_j)) \in \mathbb{Z}$, in other words that $\mathrm{Tr}_{K/\mathbb{Q}}(xI) \subset \mathbb{Z}$, where we have set $x = \sum_j v_j\omega_j$. Since $xI = xI\mathbb{Z}_K$, the proposition follows. □

In particular, when $I = \mathbb{Z}_K$ and $\gamma_i = \omega_i$ is an integral basis, this proposition shows that a $\mathbb{Z}$-basis of $\mathfrak{d}(K)^{-1}$ is obtained by computing the inverse of the matrix $\mathrm{Tr}_{K/\mathbb{Q}}(\omega_i\omega_j)$. Since the determinant of this matrix is by definition equal to $d(K)$, this also shows that $\mathcal{N}(\mathfrak{d}(K)) = |d(K)|$.

The following theorem is a refinement of Theorem 4.8.8 (see [Mar]).

**Theorem 4.8.20.** *The prime ideals dividing the different are exactly the ramified prime ideals, i.e. the prime ideals whose ramification index is greater than 1.*

To compute the inverse of an ideal $I$ given by a $\mathbb{Z}$-basis $\gamma_j$ represented by an $n \times n$ matrix $M$ on the integral basis as above, we thus proceed as follows. Computing $T^{-1}$ we first obtain a basis of the codifferent $\mathfrak{d}(K)^{-1}$. We then compute the *ideal* product $I\mathfrak{d}(K)^{-1}$ by Hermite reduction of an $n \times n^2$ matrix as explained in Section 4.7. If $N$ is the HNF matrix of this ideal product, then by Proposition 4.8.19, the columns $(N^tT)^{-1}$ will form a $\mathbb{Z}$-basis of the ideal $(I\mathfrak{d}(K)^{-1})^{-1}\mathfrak{d}(K)^{-1} = I^{-1}$, thus giving the inverse of $I$ after another HNF. In paractice, it is better to work only with integral ideals, and since we know that $\det(T) = d(K)$, this means that we will replace $\mathfrak{d}(K)^{-1}$ by $d(K)\mathfrak{d}(K)^{-1}$ which is an integral ideal.

This leads to the following algorithm.

**Algorithm 4.8.21** (Ideal Inversion). Given an integral basis $(\omega_i)_{1\le i\le n}$ of the ring of integers of a number field $K$ and an integral ideal $I$ given by an $n \times n$ matrix $M$ whose columns give the coordinates of a $\mathbb{Z}$-basis $\gamma_j$ of $I$ on the $\omega_i$, this algorithm computes the HNF of the inverse ideal $I^{-1}$.

1. [Compute $d(K)\mathfrak{d}(K)^{-1}$] Compute the $n \times n$ matrix $T = (t_{i,j})$ such that $t_{i,j} = \mathrm{Tr}_{K/\mathbb{Q}}(\omega_i\omega_j)$. Set $d \leftarrow \det(T)$ (this is the discriminant $d(K)$ of $K$ hence is usually available with the $\omega_i$ already). Finally, call $\delta_j$ the elements of $\mathbb{Z}_K$ whose coordinates on the $\omega_i$ are the columns of $dT^{-1}$ (thus the $\delta_j$ will be a $\mathbb{Z}$-basis of the integral ideal $d(K)\mathfrak{d}(K)^{-1}$).
2. [Compute $d(K)I\mathfrak{d}(K)^{-1}$] Let $N$ be the HNF of the $n \times n^2$ matrix whose columns are the coordinates on the integral basis of the $n^2$ products $\gamma_i\delta_j$ (the columns of $N$ will form a $\mathbb{Z}$-basis of $d(K)I\mathfrak{d}(K)^{-1}$).

3. [Compute $I^{-1}$] Set $P \leftarrow d(K)(N^tT)^{-1}$, and let $e$ be a common denominator for the entries of the matrix $P$. Let $W$ be the HNF of $eP$. Output $(W, e)$ as the HNF of $I^{-1}$ and terminate the algorithm.

The proof of the validity of the algorithm is easy and left to the reader. □

**Remarks.**

(1) If many ideal inversions are to be done in the same number field, step 1 should of course be done only once. In addition, it may be useful to find a two-element representation for the integral ideal $d(K)\mathfrak{d}(K)^{-1}$ since this will considerably speed up the ideal multiplication of step 2. Algorithm 4.7.10 cannot directly be used for that purpose since it is valid only for prime ideals, but similar algorithms exist for general ideals (see Exercise 30). In addition, if $\mathbb{Z}_K = \mathbb{Z}[\theta]$ and if $P[X]$ is the minimal monic polynomial of $\theta$, then one can prove (see Exercise 33) that $\mathfrak{d}(K)$ is the principal ideal generated by $P'(\theta)$, so the ideal multiplication of step 2 is even simpler.

(2) If we want to compute the HNF of the different $\mathfrak{d}(K)$ itself, we apply the above algorithm to the integral ideal $d(K)\mathfrak{d}(K)^{-1}$ (with $M = d(K)T^{-1}$) and multiply the resulting inverse by $d(K)$ to get $\mathfrak{d}(K)$.

Now that we know how to compute the inverse of an ideal, we can give an algorithm to compute intersections. This is based on the following formula, which is valid if $I$ and $J$ are integral ideals of $\mathbb{Z}_K$:

$$I \cap J = I \cdot J \cdot (I + J)^{-1}.$$

This corresponds to the usual formula $\mathrm{lcm}(a, b) = a \cdot b \cdot (\gcd(a, b))^{-1}$. We have seen above how to compute the HNF of sums and products of modules, and in particular of ideals, knowing the HNF of each operand. Since we have just seen an algorithm to compute the inverse of an ideal, this gives an algorithm for the intersection of two ideals.

However, a more direct (and usually better) way to compute the intersection of two ideals is described in Exercise 18.

## 4.9 Units and Ideal Classes

### 4.9.1 The Class Group

**Definition 4.9.1.** *Let $K$ be a number field and $\mathbb{Z}_K$ be the ring of integers of $K$. We say that two (fractional) ideals $I$ and $J$ of $K$ are* equivalent *if there exists $\alpha \in K^*$ such that $J = \alpha I$. The set of equivalence classes is called the* class group *of $\mathbb{Z}_K$ (or of $K$) and is denoted $Cl(K)$.*

Since fractional ideals of $\mathbb{Z}_K$ form a group it follows that $Cl(K)$ is also a group. The main theorem concerning $Cl(K)$ is that it is finite:

**Theorem 4.9.2.** *For any number field $K$, the class group $Cl(K)$ is a finite Abelian group, whose cardinality, called the* class number, *is denoted $h(K)$.*

Denote by $\mathcal{I}(K)$ the set of fractional ideals of $K$, and $\mathcal{P}(K)$ the set of principal ideals. We clearly have the exact sequence

$$1 \longrightarrow \mathcal{P}(K) \longrightarrow \mathcal{I}(K) \longrightarrow Cl(K) \longrightarrow 1.$$

The determination of the structure of $Cl(K)$ and in particular of the class number $h(K)$ is one of the main problems in algorithmic algebraic number theory. We will study this problem in the case of quadratic fields in Chapter 5 and for general number fields in Chapter 6.

Note that $h(K) = 1$ if and only if $\mathbb{Z}_K$ is a PID which in turn is if and only if $\mathbb{Z}_K$ is a UFD. Hence the class group is the obstruction to $\mathbb{Z}_K$ being a UFD.

We can also define the class group for an order in $K$ which is not the maximal order. In this case however, since not every ideal is invertible, we must slightly modify the definition.

**Definition 4.9.3.** *Let $R$ be an order in $K$ which is not necessarily maximal. We define the class group of $R$ and denote by $Cl(R)$ the set of equivalence classes of* invertible *ideals of $R$ (the equivalence relation being the same as before).*

Since all fractional ideals of $\mathbb{Z}_K$ are invertible, this does generalize the preceding definition. The class group is still a finite Abelian group whose cardinality is called the class number of $R$ and denoted $h(R)$. Furthermore, it follows immediately from the definitions that the map $I \mapsto I\mathbb{Z}_K$ from $R$-ideals to $\mathbb{Z}_K$-ideals induces a homomorphism from $Cl(R)$ to $Cl(K)$ and that this homomorphism is *surjective*. In particular, $h(R)$ is a multiple of $h(K)$.

Since the discovery of the class group in 1798 by Gauss, many results have been obtained on class groups. Our ignorance however is still enormous. For example, although widely believed to be true, it is not even known if there exist an infinite number of isomorphism classes of number fields having class number 1 (i.e. with trivial class group, or again such that $\mathbb{Z}_K$ is a PID). Numerical and heuristic evidence suggests that already for real quadratic fields $\mathbb{Q}(\sqrt{p})$ with $p$ prime and $p \equiv 1 \pmod 4$, not only should there be an infinite number of PID's, but their proportion should be around 75.446% (see [Coh-Len1], [Coh-Mar] and Section 5.10).

Class numbers and class groups arise very often in number theory. We give two examples. In the work on Fermat's last "theorem" (FLT), it was soon discovered that the obstruction to a proof was the failure of unique factorization

in the cyclotomic fields $\mathbb{Q}(\zeta_p)$ where $\zeta_p$ is a primitive $p^{\text{th}}$ root of unity (a number field of degree $p-1$, generated by the polynomial $X^{p-1}+\cdots+X+1$), where $p$ is an odd prime. It was Kummer who essentially introduced the notion of ideals, and who showed how to replace unique factorization of elements by unique factorization of ideals, which as we have seen, is always satisfied in a Dedekind domain. It is however necessary to come back to the elements themselves in order to finish the argument—that is to obtain a principal ideal. What is obtained is that $\mathfrak{a}^p$ is principal for some ideal $\mathfrak{a}$. Now, by definition of the class group, we also know that $\mathfrak{a}^h$ is principal, where $h$ is the class number of our cyclotomic field. Hence, we can deduce that $\mathfrak{a}$ itself is principal if $p$ does not divide $h$. This fortunately seems to happen quite often (for example, for 22 out of the 25 primes less than 100); this proves FLT in many cases (the so-called regular primes). One can also prove FLT in other cases by more sophisticated methods.

The second use of class groups, which we will see in more detail in Chapters 8 and 10, is for factoring large numbers. In that case one uses class groups of quadratic fields. For example, the knowledge of the class group (in fact only of the 2-Sylow subgroup) of $\mathbb{Q}(\sqrt{-N})$ is essentially equivalent to knowing the factors of $N$, hence if we can find an efficient method to compute this class group or its 2-Sylow subgroup, we obtain a method for factoring $N$. This is the basis of work initiated by Shanks ([Sha1]) and followed by many other people (see for example [Sey1], [Schn-Len] and [Bue1]).

### 4.9.2 Units and the Regulator

Recall that a unit $x$ in $K$ is an algebraic integer such that $1/x$ is also an algebraic integer, or equivalently is an algebraic integer of norm $\pm 1$.

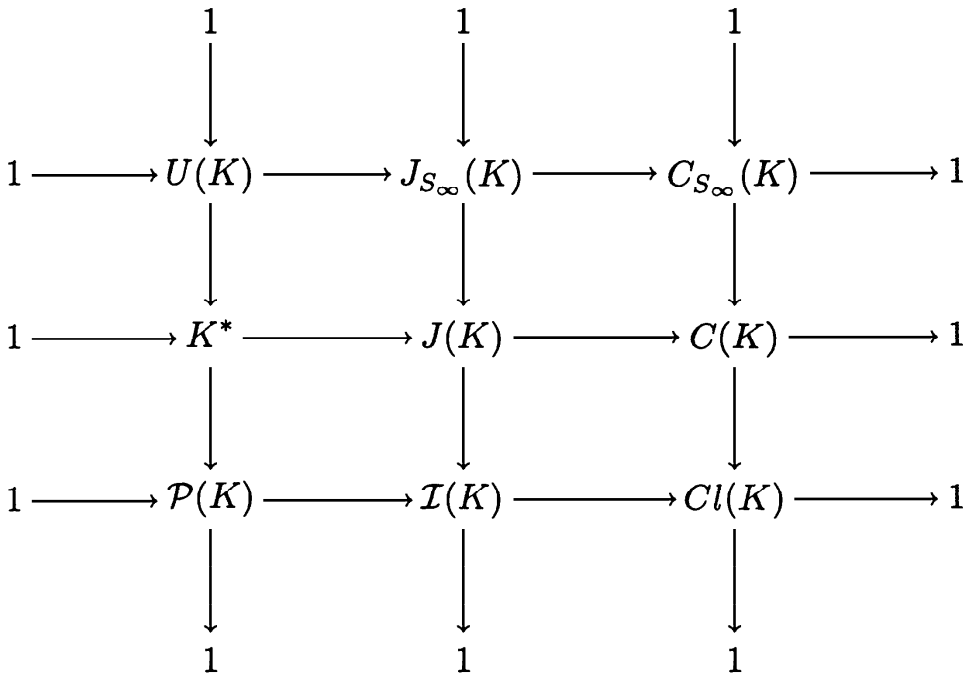
**Definition 4.9.4.** *The set of units in $K$ form a multiplicative group which we will denote by $U(K)$. The torsion subgroup of $U(K)$, i.e. the group of roots of unity in $K$, will be denoted by $\boldsymbol{\mu}(K)$.*

(Note that some people write $E(K)$ because of the German word "Einheiten" for units, but we will keep the letter $E$ for elliptic curves.)

It is clear that we have the exact sequence

$$1 \longrightarrow U(K) \longrightarrow K^* \longrightarrow \mathcal{P}(K) \longrightarrow 1,$$

where as before $\mathcal{P}(K)$ denotes the set of principal ideals in $K$. If we combine this exact sequence with the preceding one, we can complete a commutative diagram in the context of ideles, by introducing a generalization of the class group, called the idele class group $C(K)$. We will not consider these subjects in this course, but without explaining the notations (see [Lang2]) I give the diagram:




The main result concerning units is the following theorem

**Theorem 4.9.5** (Dirichlet). *Let $(r_1, r_2)$ be the signature of $K$. Then the group $U(K)$ is a finitely generated Abelian group of rank $r_1 + r_2 - 1$. In other words, we have a group isomorphism*

$$U(K) \simeq \boldsymbol{\mu}(K) \times \mathbb{Z}^{r_1+r_2-1},$$

*and $\boldsymbol{\mu}(K)$ is a finite cyclic group.*

If we set $r = r_1 + r_2 - 1$, we see that there exist units $u_1, \ldots, u_r$ such that every element $x$ of $U(K)$ can be written in a unique way as

$$x = \zeta u_1^{n_1} \cdots u_r^{n_r},$$

where $n_i \in \mathbb{Z}$ and $\zeta$ is a root of unity in $K$. Such a family $(u_i)$ will be called a *system of fundamental units* of $K$. It is not unique, but since changing a $\mathbb{Z}$-basis of $\mathbb{Z}^r$ into another involves multiplication by a matrix of determinant $\pm 1$, the absolute value of the determinant of the $u_i$ in some appropriate sense, is independent of the choice of the $u_i$, and this is what we will call the regulator of $K$. The difficulty in defining the determinant comes because the units form a multiplicative group. To use determinants, one must linearize the problem, i.e. take logarithms.

Let $\sigma_1, \ldots, \sigma_{r_1}, \sigma_{r_1+1}, \ldots, \sigma_{r_1+r_2}$ be the first $r_1 + r_2$ embeddings of $K$ in $\mathbb{C}$, where the $\sigma_i$ for $i \leq r_1$ are the real embeddings, and the other embeddings are the $\sigma_i$ and $\overline{\sigma}_i = \sigma_{r_2+i}$ for $i > r_1$.

**Definition 4.9.6.** *The* logarithmic embedding *of $K^*$ in $\mathbb{R}^{r_1+r_2}$ is the map $L$ which sends $x$ to*

$$L(x) = (\ln|\sigma_1(x)|, \ldots, \ln|\sigma_{r_1}(x)|, 2\ln|\sigma_{r_1+1}(x)|, \ldots, 2\ln|\sigma_{r_1+r_2}(x)|).$$

It is clear that $L$ is an Abelian group homomorphism. Furthermore, we clearly have $\ln|\mathcal{N}_{K/\mathbb{Q}}(x)| = \sum_{1\le i\le r_1+r_2} L_i(x)$ where $L_i(x)$ denotes the $i^{\text{th}}$ component of $L(x)$. It follows that the image of the subgroup of $K^*$ of elements of norm equal to $\pm 1$ is contained in the hyperplane $\sum_{1\le i\le r_1+r_2} x_i = 0$ of $\mathbb{R}^{r_1+r_2}$.

The first part of the following theorem is essentially a restatement of Theorem 4.9.5, and the second part is due to Kronecker (see Exercise 25).

**Theorem 4.9.7.**

(1) *The image of the group of units $U(K)$ under the logarithmic embedding is a lattice (of rank $r_1 + r_2 - 1$) in the hyperplane $\sum_{1\le i\le r_1+r_2} x_i = 0$ of $\mathbb{R}^{r_1+r_2}$.*

(2) *The kernel of the logarithmic embedding is exactly equal to the group $\mu(K)$ of the roots of unity in $K$.*

**Definition 4.9.8.** *The volume of this lattice, i.e. the absolute value of the determinant of any $\mathbb{Z}$-basis of the above defined lattice is called the* regulator *of $K$ and denoted $R(K)$. If $u_1, \ldots, u_r$ is a system of fundamental units of $K$ (where $r = r_1 + r_2 - 1$), $R(K)$ can also be defined as the absolute value of the determinant of any of the $r \times r$ matrices extracted from the $r \times (r+1)$ matrix*

$$\ln\|\sigma_j(u_i)\|_{1\le i\le r, 1\le j\le r+1},$$

*where $\|\sigma(x)\| = |\sigma(x)|$ if $\sigma$ is a real embedding and $\|\sigma(x)\| = |\sigma(x)|^2$ if $\sigma$ is a complex embedding (note that $L(x) = (\ln\|\sigma_i(x)\|)_{1\le i\le r+1}$).*

The problem of computing regulators (or fundamental units) is closely linked to the problem of computing class numbers, and is one of the other main tasks of computational algebraic number theory.

On the other hand, the problem of computing the subgroup of roots of unity $\mu(K)$ is not difficult. Note, for example, that if $r_1 > 0$ then $\mu(K) = \{\pm 1\}$ since all other roots of unity are non-real. Hence, we can assume $r_1 = 0$, and by the above theorem we must find integers $x_i$ such that for every $j$, $|\sigma_j(\sum_{1\le i\le n} x_i\omega_i)|^2 = 1$ where $\omega_i$ is an integral basis of $\mathbb{Z}_K$. If we set $\mathbf{x} = (x_1, \ldots, x_n)$, this implies that

$$Q(\mathbf{x}) = \sum_j |\sigma_j(\sum_{1\le i\le n} x_i\omega_i)|^2 = n.$$

Conversely, the inequality between arithmetic and geometric mean shows that if $\rho \in \mathbb{Z}_K \setminus \{0\}$, then

$$\sum_j |\sigma_j(\rho)|^2 \geq n(\prod_j |\sigma_j(\rho)|)^{2/n} \geq n$$

with equality if and only if all $|\sigma_j(\rho)|^2$ are equal. It follows that $n$ is the minimum non-zero value of the quadratic form $Q$ on $\mathbb{Z}^n$, and that this minimum is attained when $|\sigma_j(\rho)| = 1$ for all $j$, where $\rho = \sum_i x_i\omega_i$. Finally, Theorem 4.9.7 (2) tells us that such a $\rho$ is a root of unity (see Exercise 25). Hence, the computation of the minimal vectors of the lattice $(\mathbb{Z}^n, Q)$ using, for example, the Fincke-Pohst Algorithm 2.7.7, will give us quite rapidly the set of roots of unity in $K$. Thus we have the following algorithm.

**Algorithm 4.9.9** (Roots of Unity Using Fincke-Pohst). Let $K = \mathbb{Q}(\theta)$ be a number field of degree $n$ and $T$ the minimal monic polynomial of $\theta$ over $\mathbb{Q}$. This algorithm computes the order $w(K)$ of the group of roots of unity $\boldsymbol{\mu}(K)$ of $K$ (hence $\boldsymbol{\mu}(K)$ will be equal to the set of powers of a primitive $w(K)$-th root of unity).

1. [Initialize] Using Algorithm 4.1.11 compute the signature $(r_1, r_2)$ of $K$. If $r_1 > 0$, output $w(K) = 2$ and terminate the algorithm. Otherwise, using Algorithm 6.1.8 of Chapter 6, compute an integral basis $\omega_1, \ldots, \omega_n$ of $K$ as polynomials in $\theta$.

2. [Compute matrix] Using Algorithm 3.6.6, compute a reasonably accurate value of $\theta$ and its conjugates $\sigma_j(\theta)$ as the roots of $T$, then the numerical values of $\sigma_j(\omega_k)$. Finally, compute a reasonably accurate approximation to

$$a_{i,j} \leftarrow \sum_{1\leq k\leq n} \sigma_k(\omega_i)\overline{\sigma}_k(\omega_j)$$

   (note that this will be a real number), and let $A$ be the symmetric matrix $A = (a_{i,j})_{1\leq i,j\leq n}$.

3. [Apply Fincke-Pohst] Apply Algorithm 2.7.7 to the matrix $A$ and the constant $C = n + 0.1$.

4. [Final check] Set $s \leftarrow 0$. For each pair $(\mathbf{x}, -\mathbf{x})$ with $(x_1, \ldots, x_n)$ which is output by Algorithm 2.7.7, set $\rho \leftarrow \sum_{1\leq i\leq n} x_i\omega_i$, and set $s \leftarrow s + 1$ if $\rho$ is a root of unity (this can be checked exactly in several easy ways, see Exercise 26).

5. Output $w(K) \leftarrow 2s$ and terminate the algorithm.

**Remark.** The quadratic form $Q$ considered here is, not surprisingly, the same as the one that we used for the polynomial reduction Algorithm 4.4.11. Note however that in POLRED we only wanted small vectors in the lattice, corresponding to algebraic numbers of degree exactly equal to $n$, while here we want the smallest vectors, and they correspond in general to algebraic numbers of degree less than $n$. Note also that in practice all the vectors output in step 4 correspond to roots of unity.

We can also give an algorithm based on those of Section 4.5.3 as follows.

**Algorithm 4.9.10** (Roots of Unity Using the Subfield Problem). Let $K = \mathbb{Q}(\theta)$ be a number field of degree $n$ and $T$ the minimal monic polynomial of $\theta$ over $\mathbb{Q}$. This algorithm computes the order $w(K)$ of the group of roots of unity $\boldsymbol{\mu}(K)$ of $K$ (hence $\boldsymbol{\mu}(K)$ will be equal to the set of powers of a primitive $w(K)$-th root of unity).

1. [Initialize] Using Algorithm 4.1.11 compute the signature $(r_1, r_2)$ of $K$. If $r_1 > 0$, output $w(K) = 2$ and terminate the algorithm. Otherwise, using Algorithm 6.1.8 of Chapter 6, compute the discriminant $d(K)$ of $K$, and set $w \leftarrow 1$.

2. [Compute primes] Let $\mathcal{L}$ be the list of primes $p$ such that $(p-1) \mid n$ (since $n$ is very small, $\mathcal{L}$ can be simply obtained by trial division). Let $c$ be the number of elements of $\mathcal{L}$, and set $i \leftarrow 0$.

3. [Get next prime and exponent] Set $i \leftarrow i+1$. If $i > c$ output $w$ and terminate the algorithm. Otherwise, let $p$ be the $i$-th element in the list $\mathcal{L}$, set

$$k \leftarrow \left\lfloor \frac{v_p(d(K))}{n} + \frac{1}{p-1} \right\rfloor,$$

and set $j \leftarrow 0$.

4. [Test cyclotomic polynomials] Set $j \leftarrow j+1$. If $j > k$, go to step 3. Otherwise, applying Algorithm 4.5.4 to $A(X) = \Phi_{p^j}(X)$ and $B(X) = T(X)$ (where $\Phi_{p^j}(X) = \sum_{i=0}^{p-1} X^{ip^{j-1}}$ is the $p^j$-th cyclotomic polynomial) determine whether $K$ has a subfield isomorphic to $\mathbb{Q}(\zeta_{p^j})$ (where $\zeta_{p^j}$ is some root of $\Phi_{p^j}(X)$, i.e. a primitive $p^j$-th root of unity). If it does, go to step 4, and if not, set $w \leftarrow wp^{j-1}$ and go to step 3.

**Remarks.**

(1) The validity of the check in step 3 follows from Exercise 24 and Proposition 4.4.8. We can avoid the computation of the discriminant of $K$, and skip this step, at the expense of spending more time in step 4.

(2) We refer to [Was] or [Ire-Ros] for cyclotomic fields (which we will meet again in Chapter 9) and cyclotomic polynomials. The (general) cyclotomic polynomials can be computed either by induction or by the explicit formula

$$\Phi_m(X) = \prod_{d \mid m} (X^d - 1)^{\mu(m/d)}$$

where $\mu(n)$ is the Möbius function, but in our case this simplifies to the formula

$$\Phi_{p^j}(X) = \sum_{i=0}^{p-1} X^{ip^{j-1}}$$

used in the algorithm.

(3) Although Algorithm 4.9.10 is more pleasing to the mind, Algorithm 4.9.9 is considerably faster and should therefore be preferred in practice. Care should be taken however to be sufficiently precise in the computation of the numerical values of the coefficients of $Q$. We have given in detail Algorithm 4.9.10 to show that an exact algorithm also exists.

All the quantities that we have defined above are tied together if we view them analytically.

**Definition 4.9.11.** *Let $K$ be a number field. We define for* $\mathrm{Re}(s) > 1$ *the* Dedekind zeta function $\zeta_K(s)$ *of $K$ by the formulas*

$$\zeta_K(s) = \sum_{\mathfrak{a}} \frac{1}{\mathcal{N}(\mathfrak{a})^s} = \prod_{\mathfrak{p}} \frac{1}{1 - \frac{1}{\mathcal{N}(\mathfrak{p})^s}}$$

*where the sum is over all non-zero integral ideals of $\mathbb{Z}_K$ and the product is over all non-zero prime ideals of $\mathbb{Z}_K$.*

The equality between the two definitions follows from unique factorization into prime ideals (Theorem 4.6.14), and the convergence for $\mathrm{Re}(s) > 1$ is proved in Exercise 22.

The basic theorem concerning this function is the following.

**Theorem 4.9.12** (Dedekind). *Let $K$ be a number field of degree $n$ having $r_1$ real places and $r_2$ complex ones (so $r_1 + 2r_2 = n$). Denote by $d(K)$, $h(K)$, $R(K)$ and $w(K)$ the discriminant, class number, regulator and number of roots of unity of $K$ respectively.*

(1) *The function $\zeta_K(s)$ can be analytically continued to the whole complex plane into a meromorphic function having a single pole at $s = 1$ which is simple.*

(2) *If we set*

$$\Lambda(s) = |d(K)|^{s/2} \left(\pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\right)^{r_1+r_2} \left(\pi^{-\frac{s+1}{2}}\Gamma\left(\frac{s+1}{2}\right)\right)^{r_2} \zeta_K(s)$$

*we have the functional equation*

$$\Lambda(1-s) = \Lambda(s).$$

(3) *If we set $r = r_1 + r_2 - 1$ (which is the rank of the unit group), $\zeta_K(s)$ has a zero of order $r$ at $s = 0$ and we have*

$$\lim_{s \to 0} s^{-r}\zeta_K(s) = -h(K)R(K)w(K)^{-1}.$$

(4) *Equivalently by the functional equation, the residue of* $\zeta_K(s)$ *at* $s = 1$ *is given by*

$$\lim_{s\to 1}(s-1)\zeta_K(s) = 2^{r_1}(2\pi)^{r_2}\frac{h(K)R(K)}{w(K)\sqrt{|d(K)|}}.$$

This theorem shows one among numerous instances where $h(K)$ and $R(K)$ are inextricably linked.

**Remarks.**

(1) From this theorem it is easily shown (see Exercise 23) that if $N_K(x)$ denotes the number of integral ideals of norm less than or equal to $x$, then

$$\lim_{x\to\infty}\frac{N_K(x)}{x} = 2^{r_1}(2\pi)^{r_2}\frac{h(K)R(K)}{w(K)\sqrt{|d(K)|}}.$$

(2) It is also possible to prove the following generalization of the prime number theorem (see [Lang2]).

**Theorem 4.9.13.** *Let* $\pi_K(x)$ *(resp.* $\pi_K^{(1)}(x)$*) be the number of prime ideals (resp. prime ideals of degree 1) whose norm is less than equal to* $x$*. Then*

$$\lim_{x\to\infty}\frac{\pi_K(x)}{x/\ln(x)} = \lim_{x\to\infty}\frac{\pi_K^{(1)}(x)}{x/\ln(x)} = 1.$$

Dedekind's Theorem 4.9.12 shows that the behavior of $\zeta_K(s)$ at $s = 0$ and $s = 1$ is linked to fundamental arithmetic invariants of the number field $K$. Siegel proved that the values at negative integers are rational numbers, hence they also have some arithmetic significance. From the functional equation it is immediately clear that $\zeta_K(s)$ vanishes for all negative integers $s$ if $K$ is not totally real, and for even negative integers if $K$ is totally real. Hence, the only interesting values are the $\zeta_K(1-2m)$ for totally real fields $K$ ($r_2 = 0$) and positive integral $m$. There are special methods, essentially due to Siegel, for computing these values using the theory of Hilbert modular forms. As an example, we give the following result, which also shows the arithmetic significance of these values (see [Coh], [Zag1]).

**Theorem 4.9.14.** *Let* $K = \mathbb{Q}(\sqrt{D})$ *be a real quadratic field of discriminant* $D$*. Define* $\sigma(n)$ *to be equal to the sum of the positive divisors of* $n$ *if* $n$ *is positive, and equal to 0 otherwise. Then*

(1)
$$\zeta_K(-1) = \frac{1}{60} \sum_{s \equiv D \,(\mathrm{mod}\, 2)} \sigma\left(\frac{D - s^2}{4}\right)$$

*(this is a finite sum).*

(2) *The number $r_5(D)$ of representations of $D$ as a sum of 5 squares of elements of $\mathbb{Z}$ (counting representations with a different ordering as distinct) is given by*
$$r_5(D) = 48\left(5 - 2\left(\frac{D}{2}\right)\right)\zeta_K(-1)$$

*(this formula must be slightly modified if $D$ is not the discriminant of a real quadratic field, see [Coh2]).*

I have already mentioned how little we know about class numbers. The same can be said about regulators. For example, we can define the regulator of a number field in a $p$-adic context, essentially by replacing the real logarithms by $p$-adic ones. In that case, even an analogue of Dirichlet's theorem that the regulator does not vanish is not known. This is a famous unsolved problem known as Leopoldt's conjecture. It is known to be true for some classes of fields, for example Abelian extensions of $\mathbb{Q}$ (see [Was] Section 5.5).

We do have a theorem which gives a quantitative estimate for the product of the class number and the regulator (see [Sie], [Brau] and [Lang2]):

**Theorem 4.9.15** (Brauer-Siegel). *Let $K$ vary in a family of number fields such that $|d(K)|^{1/\deg(K)}$ tends to infinity, where $d(K)$ is the discriminant of $K$. Assume, in addition, that these fields are Galois over $\mathbb{Q}$. Then, we have the following asymptotic relation:*
$$\ln(h(K)R(K)) \sim \ln(|d(K)|^{1/2}).$$

This shows that the product $h(K)R(K)$ behaves roughly as the square root of the discriminant. The main problem with this theorem is that it is *non-effective*, meaning that nobody knows how to give explicit constants to make the $\sim$ sign disappear. For example, for imaginary quadratic fields, $r = 0$ hence $R(K) = 1$, and although the Brauer-Siegel theorem tells us that $h(K)$ tends to infinity with $|d(K)|$, and even much more, the problem of finding an explicit function $f(d)$ tending to infinity with $d$ and such that $h(K) \geq f(|d(K)|)$ is extremely difficult and was only solved recently using sophisticated methods involving elliptic curves and modular forms, by Goldfeld, Gross and Zagier ([Gol], [Gro-Zag2]).

Note that one conjectures that the theorem is still true without the hypothesis that the fields are Galois extensions. This would follow from Artin's conjecture on non-Abelian $L$-functions and on certain Generalized

Riemann Hypotheses. On the other hand, one can prove that the hypothesis on $|d(K)|^{1/\deg(K)}$ is necessary. The following is a simple corollary of the Brauer-Siegel Theorem 4.9.15:

**Corollary 4.9.16.** *Let $K$ vary over a family of number fields of fixed degree over $\mathbb{Q}$. Then, as $|d(K)| \to \infty$, we have*

$$\ln(h(K)R(K)) \sim \ln(|d(K)|^{1/2}).$$

### 4.9.3 Conclusion: the Main Computational Tasks of Algebraic Number Theory

From the preceding definitions and results, it can be seen that the main computational problems for a number field $K = \mathbb{Q}(\theta)$ are the following:

(1) Compute an integral basis of $\mathbb{Z}_K$, determine the decomposition of prime numbers in $\mathbb{Z}_K$ and $\mathfrak{p}$-adic valuations for given ideals or elements.
(2) Compute the Galois group of the Galois closure of $K$.
(3) Compute a system of fundamental units of $K$ and/or the regulator $R(K)$. Note that these two problems are not completely equivalent, since for many applications, only the approximate value of the real number $R(K)$ is desired. In most cases, by the Brauer-Siegel theorem, the fundamental units are too large even to write down, at least in a naïve manner (see Section 5.8.3 for a representation which avoids this problem).
(4) Compute the class number and the structure of the class group $Cl(K)$. It is essentially impossible to do this without also computing the regulator.
(5) Given an ideal of $\mathbb{Z}_K$, determine whether or not it is principal, and if it is, compute $\alpha \in K$ such that $I = \alpha\mathbb{Z}_K$.

In the rest of this book, we will give algorithms for these tasks, placing special emphasis on the case of quadratic fields.

Although they are all rather complex, some sophisticated versions are quite efficient. With fast computers and careful implementations, it is possible to tackle the above tasks for quadratic number fields whose discriminant has 50 or 60 decimal digits (less for general number fields). Work on this subject is currently in progress in several places.

## 4.10 Exercises for Chapter 4

1. (J. Martinet) Let $P(X) = X^4 + aX^3 + bX^2 + cX + d \in \mathbb{R}[X]$ be a squarefree polynomial. Set $D \leftarrow \mathrm{disc}(P)$, $A \leftarrow 3a^2 - 8b$, $B \leftarrow b^2 - a^2b + (3/16)a^4 + ac - 4d$. Show that the signature of $P$ is given by the following formulas. $(r_1, r_2) = (2, 1)$ iff $D < 0$, $(r_1, r_2) = (4, 0)$ iff $D > 0$, $A > 0$ and $B > 0$, and $(r_1, r_2) = (0, 2)$ iff $D > 0$ and either $A \leq 0$ or $B \leq 0$. (Hint: use Exercise 29 of Chapter 3.)

2. If $\alpha$ and $\theta$ are two algebraic numbers of degree $n$ generating the same number field $K$ over $\mathbb{Q}$, write an algorithm to find the standard representation of $\theta$ in terms of $\alpha$ knowing the standard representation of $\alpha$ in terms of $\theta$.
3. Prove Newton's formulas (i.e. Proposition 4.3.3).
4. Compute the minimal polynomial of $\alpha = 2^{1/4} + 2^{1/2}$ using several methods, and compare their efficiency.
5. Let $K$ be a number field of signature $(r_1, r_2)$. Using the canonical isomorphism

$$K \otimes \mathbb{R} \simeq \mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$$

show that the quadratic form $\mathrm{Tr}_{K/\mathbb{Q}}(x^2)$ has signature $(r_1 + r_2, r_2)$.

6. Prove that if $P = \sum_{k=0}^{n} a_k X^k$ is a monic polynomial and if $S = \mathrm{size}(P)$ in the sense of Section 4.4.2, then

$$|a_{n-k}| \leq \binom{n}{k} \left(\frac{S}{n}\right)^{k/2},$$

and that the constant is best possible if $P$ is assumed to be with complex (as opposed to integral) coefficients (hint: use a variational principle).

7. (D. Shanks.) Using for example Algorithm 4.4.11, show the following "incredible identity" $A = B$, where

$$A = \sqrt{5} + \sqrt{22 + 2\sqrt{5}}$$

and

$$B = \sqrt{11 + 2\sqrt{29}} + \sqrt{16 - 2\sqrt{29} + 2\sqrt{55 - 10\sqrt{29}}}\,.$$

See [Sha4] for an explanation of this phenomenon and other examples. See also [BFHT] and [Zip] for the general problem of radical simplification.

8. Consider modifying the POLRED algorithm as follows. Instead of the quadratic form $\mathrm{size}(P)$, we take instead

$$f(P) = \sum_{i<j} |\alpha_i - \alpha_j|^2,$$

which is still a quadratic form in the $n$ variables $x_i$ when we write $\alpha = \sum_{i=1}^{n} x_i \omega_i$. Experiment on this to compare it with POLRED, and in particular see whether it gives a larger number of proper subfields of $K$ or a smaller index.

9. Prove Proposition 4.5.3.
10. Write an algorithm which outputs all quadratic subfields of a given number field.
11. Let $R$ be a Noetherian integral domain. Show that any non-zero ideal of $R$ contains a product of non-zero prime ideals.
12. Let $d_1$ and $d_2$ be coprime integers such that $d = d_1 d_2 \in I$, where $I$ is an integral ideal in a number field $K$. Show that $I = I_1 I_2$ where $I_i = I + d_i \mathbb{Z}_K$, and show that this is false in general if $d_1$ and $d_2$ are not assumed to be coprime.
13. Let $R$ be an order in a number field, and let $I$ and $J$ be two ideals in $R$. Assume that $I$ is a maximal (i.e. non-zero prime) ideal. Show that $\mathcal{N}(I)\mathcal{N}(J) \mid \mathcal{N}(IJ)$

and that $\mathcal{N}(I^2) = \mathcal{N}(I)^2$ if and only if $I$ is invertible. (Note that these two results are not true anymore if $I$ is not assumed maximal.)

14. Let $R$ be an order in a number field. For any non-zero integral ideal of $R$, set $f(I) = [R : II']$ where as in Lemma 4.6.7 we set $I' = \{x \in K,\ xI \subset R\}$. This function can be considered as a measure of the non-invertibility of the ideal $I$.

    a) If $I$ is a maximal ideal, show that either $I$ is invertible (in which case $f(I) = 1$) or else $f(I) = \mathcal{N}(I)$.

    b) Generalizing Proposition 4.6.8, show that if $I$ and $J$ are two ideals such that $f(I)$ and $f(J)$ are coprime, we still have $\mathcal{N}(IJ) = \mathcal{N}(I)\mathcal{N}(J)$.

15. (H. W. Lenstra) Let $\alpha$ be an algebraic number which is not necessarily an algebraic integer, and let $a_n X^n + a_{n-1}X^{n-1} + \cdots + a_0$ be its minimal polynomial. Set

$$\mathbb{Z}[\alpha] = \mathbb{Z} + (a_n\alpha)\mathbb{Z} + (a_n\alpha^2 + a_{n-1}\alpha)\mathbb{Z} + \cdots .$$

    a) Show that $\mathbb{Z}[\alpha]$ is an order of $K$, and that its definition coincides with the usual one when $\alpha$ is an algebraic integer.

    b) Show that Proposition 4.4.4 (2) remains valid if $T \in \mathbb{Z}[X]$ is not assumed to be monic, if we use this generalized definition for $\mathbb{Z}[\theta]$. How should Proposition 4.4.4 (1) be modified?

16. Show that the converse of Theorem 4.7.5 is not always true, in other words if $(W, d)$ is a HNF representation of a $\mathbb{Z}$-module $M$ satisfying the properties given in the theorem, show that $M$ is not always a $\mathbb{Z}[\theta]$-module.

17. Assume that $W$ is a HNF of an ideal $I$ of $R$ with respect to a basis $\alpha_1 = 1$, $\alpha_2, \ldots, \alpha_n$ of $R$. Show that it is still true that $w_{i,i} \mid w_{1,1}$ for all $i$, and that if $w_{i,i} = w_{1,1}$ then $w_{j,i} = 0$ for $j \neq i$.

18. Show that by using Algorithms 2.4.10 or 2.7.2 instead of Algorithm 2.3.1, Algorithm 2.3.9 can be used to compute the intersection of two $\mathbb{Z}$-modules, and in particular of two ideals. Compare the efficiency of this method with that given in the text.

19. Let $\mathfrak{p}$ be a (non-zero) prime ideal in $\mathbb{Z}_K$ for some number field $K$, and assume that $\mathfrak{p}$ is not above 2. If $x \in \mathbb{Z}_K$, show that there exists a unique $\varepsilon \in \{-1, 0, +1\}$ such that

$$x^{(\mathcal{N}(\mathfrak{p})-1)/2} \equiv \varepsilon \pmod{\mathfrak{p}},$$

    where we write $x \equiv y \pmod{\mathfrak{p}}$ if $x - y \in \mathfrak{p}$. This $\varepsilon$ is called a "generalized Legendre symbol" and denoted $\left(\frac{x}{\mathfrak{p}}\right)$. Study the generalization to this symbol of the properties of the ordinary Legendre symbol seen in Chapter 1.

20. Show that the condition $v_p(\mathcal{N}(\alpha)) = f$ of Lemma 4.7.9 is not a necessary condition for $\mathfrak{p}$ to be equal to $(p, \alpha)$ (hint: decompose $\alpha$ and $p\mathbb{Z}_K$ as a product of prime ideals).

21. Using the notation of Algorithm 4.8.17, show that if the prime $p$ does not divide the index $[R : \mathbb{Z}[\theta]]$, then $p^v \mid A_{n,n}$ is equivalent to $p^v$ divides all the coefficients of the matrix $A$.

22. Let $s$ be a *real* number such that $s > 1$. Show that if $K$ is a number field of degree $n$ we have $\zeta_K(s) \leq \zeta^n(s)$ where $\zeta(s) = \zeta_{\mathbb{Q}}(s)$ is the usual Riemann zeta function, and hence that the product and series defining $\zeta_K(s)$ converge absolutely for $\mathrm{Re}(s) > 1$.

23. If $K$ is a number field, let $N_K(x)$ be the number of integral ideals of $\mathbb{Z}_K$ of norm less than or equal to $x$. Using Theorem 4.9.12, and a suitable Tauberian theorem, find the limit as $x$ tends to infinity of $N_K(x)/x$.

24. Let $K = \mathbb{Q}(\zeta_{p^k})$ where $p$ is a prime and $\zeta_m$ denotes a primitive $m$-th root of unity. One can show that $\mathbb{Z}_K = \mathbb{Z}[\zeta_{p^k}]$. Using this, compute the discriminant of the field $K$, and hence show the validity of the formula in Step 3 of Algorithm 4.9.10.

25. Let $\alpha$ be an algebraic integer of degree $d$ all of whose conjugates have absolute value 1.

    a) Show that for every positive integer $k$, the monic minimal polynomial of $\alpha^k$ in $\mathbb{Z}[X]$ has all its coefficients bounded in absolute value by $2^d$.

    b) Deduce from this that there exists only a finite number of distinct powers of $\alpha$, hence that $\alpha$ is a root of unity. (This result is due to Kronecker.)

26. Let $\rho \in \mathbb{Z}_K$ be an algebraic integer given as a polynomial in $\theta$, where $K = \mathbb{Q}(\theta)$ and $T$ is the minimal monic polynomial of $\theta$ in $\mathbb{Z}[X]$. Give algorithms to check exactly whether or not $\rho$ is a root of unity, and compare their efficiency.

27. Let $K = \mathbb{Q}[\theta]$ where $\theta$ is a root of the polynomial $X^4+1$. Show that the subgroup of roots of unity of $K$ is the group of 8-th roots of unity. Show that $1+\sqrt{2}$ is a generator of the torsion-free part of the group of units of $K$. What is the regulator of $K$? (Warning: it is not equal to $\ln(1+\sqrt{2})$).

28. Let $\mathfrak{p}$ be a (non-zero) prime ideal in $\mathbb{Z}_K$ for some number field $K$, let $e = e(\mathfrak{p}/p)$ be its ramification index, let $\mathfrak{p} = p\mathbb{Z}_K + \alpha\mathbb{Z}_K$ be a two-element representation of $\mathfrak{p}$, and finally let $v = v_{\mathfrak{p}}(\alpha)$. Let $a \geq 1$ and $b \geq 1$ be integers. By computing $\mathfrak{q}$-adic valuations for each prime ideal $\mathfrak{q}$, show that

$$p^a\mathbb{Z}_K + \alpha^b\mathbb{Z}_K = \mathfrak{p}^{\min(ae,bv)}.$$

    Deduce from this formulas for computing explicitly $\mathfrak{p}^k$ for any $k \geq 1$.

29. Let $I$ be an integral ideal in a number field $K$ and let $\ell(I)$ be the positive generator of $I \cap \mathbb{Z}$.

    a) Show that

$$\ell(I) = \prod_{p \mid \mathcal{N}(I)} p^{\max_{\mathfrak{p}\mid p}\lceil v_{\mathfrak{p}}(I)/e(\mathfrak{p}/p)\rceil}.$$

    b) Let $\alpha \in I$ be such that $(\mathcal{N}(I), \mathcal{N}(\alpha)/\mathcal{N}(I)) = 1$. Show that

$$I = \ell(I)\mathbb{Z}_K + \alpha\mathbb{Z}_K = \mathcal{N}(I)\mathbb{Z}_K + \alpha\mathbb{Z}_K$$

    (this is a partial generalization of Lemma 4.7.9).

    c) Deduce from this an algorithm for finding a two-element representation of $I$ analogous to Algorithm 4.7.10.

30. Let $K = \mathbb{Q}[\theta]$ be a number field, where $\theta$ is an algebraic integer whose minimal monic polynomial is $P(X) \in \mathbb{Z}[X]$. Assume that $\mathbb{Z}_K = \mathbb{Z}[\theta]$. Show that the different $\mathfrak{d}(K)$ is the principal ideal generated by $P'(\theta)$.

31. Let $I$ and $J$ be two integral ideals in a number field $K$ given by their HNF matrices $M_I$ and $M_J$. Assume that $I$ and $J$ are coprime, i.e. that $I + J = \mathbb{Z}_K$. Give an algorithm which finds $i \in I$ and $j \in J$ such that $i + j = 1$.

32. a) Using the preceding exercise, give an algorithm wich finds explicitly the element $\beta \in \mathbb{Z}_K$ whose existence is proven in Proposition 4.7.7.

b) Deduce from this an algorithm which finds a two-element representation $I = \alpha\mathbb{Z}_K + \beta\mathbb{Z}_K$ of an integral ideal $I$ given a non-zero element $\alpha \in I$.

c) In the case where $\alpha = \ell(I)$, compare the theoretical and practical performance of this algorithm with the one given in Exercise 29.

33. Let $\alpha$ and $\beta$ be non-zero elements of $K^*$. Show that there exist $u$ and $v$ in $\mathbb{Z}_K$ such that $\alpha\beta = u\alpha^2 + v\beta^2$, and give an algorithm for computing $u$ and $v$.

34. Modify Proposition 4.3.4 so that it is still valid when $T(X) \in \mathbb{Q}[X]$ and not necessarily monic.

# Chapter 5

# Algorithms for Quadratic Fields

## 5.1 Discriminant, Integral Basis and Decomposition of Primes

In this chapter, we consider the simplest of all number fields that are different from $\mathbb{Q}$, i.e. quadratic fields. Since $n = 2 = r_1 + 2r_2$, the signature $(r_1, r_2)$ of a quadratic field $K$ is either $(2, 0)$, in which case we will speak of *real* quadratic fields, or $(0, 1)$, in which case we will speak of *imaginary* (or complex) quadratic fields. By Proposition 4.8.11 we know that imaginary quadratic fields are those of negative discriminant, and that real quadratic fields are those with positive discriminant.

Furthermore, by Dirichlet's unit theorem, the rank of the group of units is $r_1 + r_2 - 1$, hence it can be equal to zero only in two cases: either $r_1 = 1$, $r_2 = 0$, hence $n = 1$ so $K = \mathbb{Q}$, a rather uninteresting case (see below however). Or, $r_1 = 0$ and $r_2 = 1$, hence $n = 2$, and this corresponds to imaginary quadratic fields. One reason imaginary quadratic fields are simple is that they are the only number fields (apart from $\mathbb{Q}$) with a finite number of units (almost always only 2). We consider them first in what follows. However, a number of definitions and simple results can be given uniformly.

Since a quadratic field $K$ is of degree 2 over $\mathbb{Q}$, it can be given by $K = \mathbb{Q}(\theta)$ where $\theta$ is a root of a monic irreducible polynomial of $\mathbb{Z}[X]$, say $T(X) = X^2 + aX + b$. If we set $\theta' = 2\theta + a$, then $\theta'$ is a root of $X^2 = a^2 - b = d$. Hence, $K = \mathbb{Q}(\sqrt{d})$ where $d$ is an integer, and the irreducibility of $T$ means that $d$ is not a square. Furthermore, it is clear that $\mathbb{Q}(\sqrt{df^2}) = \mathbb{Q}(\sqrt{d})$, hence we may assume $d$ squarefree. The discriminant and integral basis problem is easy.

**Proposition 5.1.1.** *Let $K = \mathbb{Q}(\sqrt{d})$ be a quadratic field with $d$ squarefree and not a square (i.e. different from 1). Let $1, \omega$ be an integral basis and $d(K)$ the discriminant of $K$. Then, if $d \equiv 1 \pmod 4$, we can take $\omega = (1 + \sqrt{d})/2$, and we have $d(K) = d$, while if $d \equiv 2$ or $3 \pmod 4$, we can take $\omega = \sqrt{d}$ and we have $d(K) = 4d$.*

This is well known and left as an exercise. Note that we can, for example, appeal to Corollary 4.4.7, which is much more general.

For several reasons, in particular to avoid making unnecessary case distinctions, it is better to consider quadratic fields as follows.

**Definition 5.1.2.** *An integer $D$ is called a* fundamental discriminant *if $D$ is the discriminant of a quadratic field $K$. In other words, $D \neq 1$ and either $D \equiv 1 \pmod 4$ and is squarefree, or $D \equiv 0 \pmod 4$, $D/4$ is squarefree and $D/4 \equiv 2$ or $3 \pmod 4$.*

If $K$ is a quadratic field of discriminant $D$, we will use the following as standard notations: $K = \mathbb{Q}(\sqrt{D})$, where $D$ is a fundamental discriminant. Hence $D = d(K)$, and an integral basis of $K$ is given by $(1, \omega)$, where

$$\omega = \frac{D + \sqrt{D}}{2},$$

and therefore $\mathbb{Z}_K = \mathbb{Z}[\omega]$.

**Proposition 5.1.3.** *If $K$ is a quadratic field of discriminant $D$, then every order $R$ of $K$ has discriminant $Df^2$ where $f$ is a positive integer called the* conductor *of the order. Conversely, if $A$ is any non-square integer such that $A \equiv 0$ or $1 \pmod 4$, then $A$ is uniquely of the form $A = Df^2$ where $D$ is a fundamental discriminant, and there exists a unique order $R$ of discriminant $A$ (and $R$ is an order of the quadratic field $\mathbb{Q}(\sqrt{D})$).*

Again this is very easy and left to the reader.

A consequence of this is that it is quite natural to consider quadratic fields together with their orders, since their discriminants form a sequence which is almost a union of two arithmetic progressions. It is however necessary to separate the positive from the negative discriminants, and for positive discriminants we should add the squares to make everything uniform. This corresponds to considering the sub-orders of the étale algebra $\mathbb{Q} \times \mathbb{Q}$ (which is not a field) as well. We will see applications of these ideas later in this chapter.

To end this section, note that Theorem 4.8.13 immediately shows how prime numbers decompose in a quadratic field:

**Proposition 5.1.4.** *Let $K = \mathbb{Q}(\sqrt{D})$ where as usual $D = d(K)$, $\mathbb{Z}_K = \mathbb{Z}[\omega]$ where $\omega = (D + \sqrt{D})/2$ its ring of integers, and let $p$ be a prime number. Then*

(1) *If $p \mid D$, i.e. if $\left(\frac{D}{p}\right) = 0$, then $p$ is ramified, and we have $p\mathbb{Z}_K = \mathfrak{p}^2$, where*

$$\mathfrak{p} = p\mathbb{Z}_K + \omega\mathbb{Z}_K$$

*except when $p = 2$ and $D \equiv 12 \pmod{16}$. In this case, $\mathfrak{p} = p\mathbb{Z}_K + (1+\omega)\mathbb{Z}_K$.*

(2) *If $\left(\frac{D}{p}\right) = -1$, then $p$ is inert, hence $\mathfrak{p} = p\mathbb{Z}_K$ is a prime ideal.*

(3) *If $\left(\frac{D}{p}\right) = 1$, then $p$ is split, and we have $p\mathbb{Z}_K = \mathfrak{p}_1\mathfrak{p}_2$, where*

$$\mathfrak{p}_1 = p\mathbb{Z}_K + \left(\omega - \frac{D+b}{2}\right)\mathbb{Z}_K \text{ and } \mathfrak{p}_2 = p\mathbb{Z}_K + \left(\omega - \frac{D-b}{2}\right)\mathbb{Z}_K,$$

*and $b$ is any solution to the congruence $b^2 \equiv D \pmod{4p}$.*

Recall that in Section 1.5 we gave an efficient algorithm to compute square roots modulo $p$. To obtain the number $b$ occurring in (3) above, it is only necessary, when $p$ is an odd prime and the square root obtained is not of the same parity as $D$, to add $p$ to it. When $p = 2$, one can always take $b = 1$ since $D \equiv 1 \pmod 8$.

## 5.2 Ideals and Quadratic Forms

Let $D$ be a non-square integer congruent to 0 or 1 modulo 4, $R$ the unique quadratic order of discriminant $D$, $(1, \omega)$ the standard basis of $R$ (i.e. with $\omega = (D + \sqrt{D})/2$) and $K$ be the unique quadratic field containing $R$ (i.e. the quotient field of $R$). We denote by $\sigma$ real or complex conjugation in $K$, i.e. the $\mathbb{Q}$-linear map sending $\sqrt{D}$ to $-\sqrt{D}$. From the general theory, we have:

**Proposition 5.2.1.** *Any integral ideal $\mathfrak{a}$ of $R$ has a unique Hermite normal form with denominator equal to 1, and with matrix*

$$\begin{pmatrix} a & b \\ 0 & c \end{pmatrix}$$

*with respect to $\omega$, where $c$ divides $a$ and $b$ and $0 \leq b < a$. In other words, $\mathfrak{a} = a\mathbb{Z} + (b + c\omega)\mathbb{Z}$. Furthermore, $a = \ell(\mathfrak{a})$ is the smallest positive integer in $\mathfrak{a}$ and $\mathcal{N}(\mathfrak{a}) = ac$.*

**Definition 5.2.2.** *We will say that an integral ideal $\mathfrak{a}$ of $R$ is* primitive *if $c = 1$, in other words if $\mathfrak{a}/n$ is not an integral ideal of $R$ for any integer $n > 1$.*

We also need some definitions about binary quadratic forms.

**Definition 5.2.3.** *A* binary quadratic form *$f$ is a function $f(x, y) = ax^2 + bxy + cy^2$ where $a$, $b$ and $c$ are integers, which is denoted more briefly by $(a, b, c)$. We say that $f$ is* primitive *if $\gcd(a, b, c) = 1$. If $f$ and $g$ are two quadratic forms, we say that $f$ and $g$ are* equivalent *if there exists a matrix $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$ (i.e. an integral matrix of determinant equal to 1), such that $g(x, y) = f(\alpha x + \beta y, \gamma x + \delta y)$.*

It is clear that equivalence preserves the discriminant $D = b^2 - 4ac$ of the quadratic form (in fact it would also be preserved by matrices of determinant equal to $-1$ but as will be seen, the use of these matrices would lead to the wrong notion of equivalence). One can also easily check that equivalence preserves primitivity. It is also clear that if $D$ is a fundamental discriminant, then any quadratic form of discriminant $D = b^2 - 4ac$ is primitive.

Note that the action of $A \in \mathrm{SL}_2(\mathbb{Z})$ is the same as the action of $-A$, hence the natural group which acts on quadratic forms (as well as on complex numbers by linear fractional transformations) is the group $\mathrm{PSL}_2(\mathbb{Z})$ where we identify $\gamma$ and $-\gamma$. By abuse of notation, we will consider an element of $\mathrm{PSL}_2(\mathbb{Z})$ as a matrix instead of an equivalence class of matrices.

We will now explain why computing on ideals and on binary quadratic forms is essentially the same. Since certain algorithms are more efficient in the context of quadratic forms, it is important to study this in detail.

As above let $D$ be a non-square integer congruent to 0 or 1 modulo 4 and $R$ be the unique order of discriminant $D$. We consider the following quotient sets.

$$F = \{(a,b,c),\ b^2 - 4ac = D\}/\Gamma_\infty$$

where $\Gamma_\infty = \left\{ \begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix},\ m \in \mathbb{Z} \right\}$ is a multiplicative group (isomorphic to the additive group of $\mathbb{Z}$) which acts on binary quadratic forms by the formula

$$\begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix} \cdot (a,b,c) = (a, b+2am, c+bm+am^2)$$

which is induced by the action of $\mathrm{SL}_2(\mathbb{Z})$.

The second set is

$$I = \{\mathfrak{a} \text{ fractional ideal of } R\}/\mathbb{Q}^*$$

where $\mathbb{Q}^*$ is understood to act multiplicatively on fractional ideals.

The third set is

$$Q = \left\{ \tau = \frac{-b+\sqrt{D}}{2a},\ a > 0 \text{ and } 4a \mid (D-b^2) \right\} / \mathbb{Z},$$

where $\mathbb{Z}$ is understood to act additively on quadratic numbers $\tau$. We also define maps as follows. If $(a,b,c)$ is a quadratic form, we set

$$\phi_{FI}(a,b,c) = \left( a\mathbb{Z} + \frac{-b+\sqrt{D}}{2}\mathbb{Z},\ \mathrm{sign}(a) \right).$$

If $\mathfrak{a}$ is a fractional ideal and $s = \pm 1$, choose a $\mathbb{Z}$-basis $(\omega_1, \omega_2)$ of $\mathfrak{a}$ with $\omega_1 \in \mathbb{Q}$ and $(\omega_2\sigma(\omega_1) - \omega_1\sigma(\omega_2))/\sqrt{D} > 0$ (this is possible by Proposition 5.2.1), and set

$$\phi_{IF}(\mathfrak{a}, s) = s\frac{\mathcal{N}(x\omega_1 - sy\omega_2)}{\mathcal{N}(\mathfrak{a})}.$$

If $\mathfrak{a}$ is a fractional ideal, choose a $\mathbb{Z}$ basis $(\omega_1, \omega_2)$ as above, and set

$$\phi_{IQ}(\mathfrak{a}) = \frac{\omega_2}{\omega_1}.$$

Finally, if $\tau = (-b + \sqrt{D})/(2a)$ is a quadratic number, set

$$\phi_{QI}(\tau) = a(\mathbb{Z} + \tau\mathbb{Z}).$$

The following theorem, while completely elementary, is fundamental to understanding the relationships between quadratic forms, ideals and quadratic numbers. We always identify the group $\mathbb{Z}/2\mathbb{Z}$ with $\pm 1$.

**Theorem 5.2.4.** *With the above notations, the maps that we have given can be defined at the level of the equivalence classes defining $F$, $I$ and $Q$, and are then set isomorphisms (which we denote in the same way). In other words, we have the following isomorphisms:*

$$F \simeq I \times \mathbb{Z}/2\mathbb{Z}, \quad I \simeq Q, \quad F \simeq Q \times \mathbb{Z}/2\mathbb{Z}.$$

*Proof.* The proof is a simple but tedious verification that everything works. We comment only on the parts which are not entirely trivial.

(1) $\phi_{FI}$ sends a quadratic form to an ideal. Indeed, if $a$ and $b$ are integers with $b \equiv D \pmod 2$, the $\mathbb{Z}$-module $a\mathbb{Z} + ((-b + \sqrt{D})/2)\mathbb{Z}$ is an ideal if and only if $4a \mid (b^2 - D)$.

(2) $\phi_{FI}$ depends only on the equivalence class modulo $\Gamma_\infty$ hence induces a map from $F$ to $I$.

(3) $\phi_{IF}$ sends a pair $(\mathfrak{a}, s)$ to an integral quadratic form. Indeed, by homogeneity, if we multiply $\mathfrak{a}$ by a suitable element of $\mathbb{Q}$, we may assume that $\mathfrak{a}$ is a primitive integral ideal. If $\omega_1 < 0$, we can also change $(\omega_1, \omega_2)$ into $(-\omega_1, -\omega_2)$. In that case, by Proposition 5.2.1 (or directly), we have $\mathcal{N}(\mathfrak{a}) = \omega_1$ and $\omega_2 - \sigma(\omega_2) = \sqrt{D}$. Finally, since $\mathfrak{a}$ is an integral ideal, $\omega_1 \mid \omega_2\sigma(\omega_2)$, and a simple calculation shows that we obtain an integral binary quadratic form of discriminant $D$.

(4) $\phi_{IF}$ does not depend on the equivalence class of $\mathfrak{a}$, nor on the choice of $\omega_1$ and $\omega_2$. Indeed, if $\omega_1$ is given, then $\omega_2$ is defined modulo $\omega_1$, and this corresponds precisely to the action of $\Gamma_\infty$ on quadratic forms.

(5) $\phi_{IF}$ and $\phi_{FI}$ are inverse maps. This is left to the reader, and is the only place where we must really use the sign$(a)$ component.

(6) I also leave to the reader the easy proof that $\phi_{IQ}$ and $\phi_{QI}$ are well defined and are inverse maps.

□

We now need to identify precisely the invertible ideals in $R$ so as to be able to work in the class group.

**Proposition 5.2.5.** *Let $\mathfrak{a} = a\mathbb{Z} + ((-b + \sqrt{D})/2)\mathbb{Z}$ be an ideal of $R$, and let $(a, b, c)$ be the corresponding quadratic form. Then $\mathfrak{a}$ is invertible in $R$ if and only if $(a, b, c)$ is primitive. In that case, we have $\mathfrak{a}^{-1} = \mathbb{Z} + ((b + \sqrt{D})/(2a))\mathbb{Z}$.*

*Proof.* From Lemma 4.6.7 we know that $\mathfrak{a}$ is invertible if and only if $\mathfrak{a}\mathfrak{b} = R$ where $\mathfrak{b} = \{z \in K,\ z\mathfrak{a} \subset R\}$. Writing $\mathfrak{a} = a\mathbb{Z} + ((-b+\sqrt{D})/2)\mathbb{Z}$, from $a \in \mathfrak{a}$ we see that such a $z$ must be the form $z = (x+y\sqrt{D})/(2a)$ with $x$ and $y$ in $\mathbb{Z}$ such that $x \equiv yD \pmod 2$. From $(-b+\sqrt{D})/2 \in \mathfrak{a}$, we obtain the congruences $bx \equiv Dy \pmod{2a}$, $x \equiv by \pmod{2a}$ and $(Dy - bx)/(2a) \equiv D(x - by)/(2a) \pmod 2$. An immediate calculation gives us $\mathfrak{b} = \mathbb{Z} + ((b+\sqrt{D})/(2a))\mathbb{Z}$ as claimed.

Now the $\mathbb{Z}$-module $\mathfrak{a}\mathfrak{b}$ is generated by the four products of the generators, i.e. by $a$, $(b+\sqrt{D})/2$, $(-b+\sqrt{D})/2$ and $-c$. We obtain immediately

$$\mathfrak{a}\mathfrak{b} = \gcd(a,b,c)\mathbb{Z} + \frac{-b+\sqrt{D}}{2}\mathbb{Z}$$

hence this is equal to $R = \mathbb{Z} + ((-b+\sqrt{D})/2)\mathbb{Z}$ if and only if $\gcd(a,b,c) = 1$, thus proving the proposition. □

**Corollary 5.2.6.** *Denote by $F_0$ the subset of classes of primitive forms in $F$, $I_0$ the subset of classes of invertible ideals in $I$ and $Q_0$ the subset of classes of primitive quadratic numbers in $Q$ (where $\tau \in Q$ is said to be primitive if $(a,b,c) = 1$ where $a$, $b$ and $c$ are as in the definition of $Q$). Then the maps $\phi_{FI}$ and $\phi_{IQ}$ also give isomorphisms:*

$$F_0 \simeq I_0 \times \mathbb{Z}/2\mathbb{Z}, \qquad I_0 \simeq Q_0, \qquad F_0 \simeq Q_0 \times \mathbb{Z}/2\mathbb{Z}.$$

Theorem 5.2.4 gives set isomorphisms between ideals and quadratic forms at the level of equivalence classes of quadratic forms modulo $\Gamma_\infty$. As we shall see, this will be useful in the real quadratic case. When considering the class group however, we need the corresponding theorem at the level of equivalence classes of quadratic forms modulo the action of the whole group $\mathrm{PSL}_2(\mathbb{Z})$. Since we must restrict to invertible ideals in order to define the class group, the above proposition shows that we will have to consider only primitive quadratic forms.

Here, it is slightly simpler to separate the case $D < 0$ from the case $D > 0$. We begin by defining the sets with which we will work.

**Definition 5.2.7.** *Let $D$ be a non-square integer congruent to 0 or 1 modulo 4, and $R$ the unique quadratic order of discriminant $D$.*

(1) *We will denote by $\mathcal{F}(D)$ the set of equivalence classes of* primitive *quadratic forms of discriminant $D$ modulo the action of* $\mathrm{PSL}_2(\mathbb{Z})$*, and in the case $D < 0$, $\mathcal{F}^+(D)$ will denote those elements of $\mathcal{F}(D)$ represented by a positive definite quadratic form (i.e. a form $(a,b,c)$ with $a > 0$).*

(2) *We will denote by $Cl(D)$ the class group of $R$, and in the case $D > 0$, $Cl^+(D)$ will denote the narrow class group of $R$, i.e. the group of equivalence classes of $R$-ideals modulo the group $\mathcal{P}^+$ of principal ideals generated by an element of positive norm.*

(3) *Finally, we will set $h(D) = |Cl(D)|$ and $h^+(D) = |Cl^+(D)|$.*

We then have the following theorems.

**Theorem 5.2.8.** *Let $D$ be a negative integer congruent to 0 or 1 modulo 4. The maps*

$$\psi_{FI}(a,b,c) = a\mathbb{Z} + \frac{-b+\sqrt{D}}{2}\mathbb{Z},$$

*and*

$$\psi_{IF}(\mathfrak{a}) = \frac{\mathcal{N}(x\omega_1 - y\omega_2)}{\mathcal{N}(\mathfrak{a})}$$

*where $\mathfrak{a} = \omega_1\mathbb{Z} + \omega_2\mathbb{Z}$ with*

$$\frac{\omega_2\sigma(\omega_1) - \omega_1\sigma(\omega_2)}{\sqrt{D}} > 0$$

*induce inverse bijections from $\mathcal{F}^+(D)$ to $Cl(D)$.*

**Theorem 5.2.9.** *Let $D$ be a non-square positive integer congruent to 0 or 1 modulo 4. The maps*

$$\psi_{FI}(a,b,c) = \left(a\mathbb{Z} + \frac{-b+\sqrt{D}}{2}\mathbb{Z}\right)\alpha,$$

*where $\alpha$ is any element of $K^*$ such that $\operatorname{sign}(\mathcal{N}(\alpha)) = \operatorname{sign}(a)$, and*

$$\psi_{IF}(\mathfrak{a}) = \frac{\mathcal{N}(x\omega_1 - y\omega_2)}{\mathcal{N}(\mathfrak{a})}$$

*where $\mathfrak{a} = \omega_1\mathbb{Z} + \omega_2\mathbb{Z}$ with*

$$\frac{\omega_2\sigma(\omega_1) - \omega_1\sigma(\omega_2)}{\sqrt{D}} > 0$$

*induce inverse bijections from $\mathcal{F}(D)$ to $Cl^+(D)$.*

*Proof.* As for Theorem 5.2.4, the proofs consist of a series of simple verifications.

(1) The map $\psi_{FI}$ is well defined on classes modulo $\mathrm{PSL}_2(\mathbb{Z})$. If $\begin{pmatrix} A & B \\ U & V \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{Z})$ acts on $(a,b,c)$, then the quantity $\tau = (-b+\sqrt{D})/(2a)$ becomes $\tau' = (V\tau - B)/(-U\tau + A)$, and $a$ becomes $a\mathcal{N}(-U\tau + A)$, hence since $\mathbb{Z} + \tau'\mathbb{Z} = (\mathbb{Z} + \tau\mathbb{Z})/(-U\tau + A)$, it follows immediately that $\psi_{FI}$ is well defined.

(2) Similarly, $\psi_{IF}$ is well defined, and we can check that it gives an integral quadratic form of discriminant $D$ as for the map $\phi_{IF}$ of Theorem 5.2.4. This form is primitive since we restrict to invertible ideals.

(3) Finally, the same verification as in the preceding theorem shows that $\psi_{IF}$ and $\psi_{FI}$ are inverse maps.

□

**Remarks.**

(1) Although we have given the bijections between classes of forms and ideals, we could, as in Theorem 5.2.4, give bijections with classes of quadratic numbers modulo the action of $\mathrm{PSL}_2(\mathbb{Z})$. This is left to the reader (Exercise 3).

(2) In the case $D < 0$, a quadratic form is either positive definite or negative definite, hence the set $F$ breaks up naturally into two disjoint pieces. The map $\psi_{FI}$ is induced by the restriction of $\phi_{FI}$ to the positive piece, and $\psi_{IF}$ is induced by $\phi_{IF}$ and forgetting the factor $\mathbb{Z}/2\mathbb{Z}$.

(3) In the case $D > 0$, there is no such natural breaking up of $F$. In this case, the maps $\phi_{FI}$ and $\phi_{IF}$ induce inverse isomorphisms between $\mathcal{F}(D)$ and

$$\mathcal{I}(D) = (\mathcal{I} \times \mathbb{Z}/2\mathbb{Z})/\tilde{\mathcal{P}},$$

where $\tilde{\mathcal{P}}$ is the quotient of $K^*$ by the subgroup of units of positive norm, and $\beta \in \tilde{\mathcal{P}}$ acts by sending $(\mathfrak{a}, s)$ to $(\beta\mathfrak{a}, s \cdot \mathrm{sign}(\mathcal{N}(\beta)))$. (Note also the exact sequence

$$1 \longrightarrow \mathcal{P}^+ \longrightarrow \tilde{\mathcal{P}} \longrightarrow \mathbb{Z}/2\mathbb{Z} \longrightarrow 1,$$

where the map to $\mathbb{Z}/2\mathbb{Z}$ is induced by the sign of the norm map.) The maps $\psi_{FI}$ and $\psi_{IF}$ are obtained by composition of the above isomorphisms with the isomorphisms between $\mathcal{I}(D)$ and $Cl^+(D)$ given as follows. The class of $(\mathfrak{a}, s)$ representing an element of $\mathcal{I}(D)$ is sent to the class of $\beta\mathfrak{a}$ in $Cl^+(D)$, where $\beta \in K^*$ is any element such that $\mathrm{sign}(\mathcal{N}(\beta)) = s$. Conversely, the class of $\mathfrak{a} \in Cl^+(D)$ is sent to the class of $(\mathfrak{a}, 1)$ in $\mathcal{I}(D)$.

Although $F$, $I$ and $Q$ are defined as quotient sets, it is often useful to use precise representatives of classes in these sets. We have already implicitly done so when we defined all the maps $\phi_{IF}$ etc ... above, but we make our choice explicit.

An element of $F$ will be represented by the unique element $(a, b, c)$ in its class chosen as follows. If $D < 0$, then $-|a| < b \leq |a|$. If $D > 0$, then $-|a| < b \leq |a|$ if $a > \sqrt{D}$, $\sqrt{D} - 2|a| < b < \sqrt{D}$ if $a < \sqrt{D}$.

An element of $I$ will be represented by the unique primitive integral ideal in its class.

An element of $Q$ will be represented by the unique element $\tau$ in its class such that $-1 < \tau + \sigma(\tau) \leq 1$, where $\sigma$ denotes (complex or real) conjugation in $K$.

The tasks that remain before us are that of computing the class group or class number, and in the real case, that of computing the fundamental unit. It is now time to separate the two cases, and in the next sections we shall examine in detail the case of imaginary quadratic fields.

## 5.3 Class Numbers of Imaginary Quadratic Fields

Until further notice, all fields which we consider will be imaginary quadratic fields. First, let us solve the problem of units. From the general theory, we know that the units of an imaginary quadratic field are the (finitely many) roots of unity inside the field. An easy exercise is to show the following:

**Proposition 5.3.1.** *Let $D < 0$ congruent to 0 or 1 modulo 4. Then the group $\mu(R)$ of units of the unique quadratic order of discriminant $D$ is equal to the group of $w(D)^{\text{th}}$ roots of unity, where*

$$w(D) = \begin{cases} 2, & \text{if } D < -4 \\ 4, & \text{if } D = -4 \\ 6, & \text{if } D = -3\,. \end{cases}$$

Let us now consider the problem of computing the class group. For this, the correspondences that we have established above between classes of quadratic forms and ideal class groups will be very useful. Usually, the ideals will be used for conceptual (as opposed to computational) proofs, and quadratic forms will be used for practical computation.

Thanks to Theorem 5.2.8, we will use interchangeably the language of ideal classes or of classes of quadratic forms. One of the advantages is that the algorithms are simpler. For example, we now consider a simple but still reasonable method for computing the class *number* of an imaginary quadratic field.

### 5.3.1 Computing Class Numbers Using Reduced Forms

**Definition 5.3.2.** *A positive definite quadratic form $(a, b, c)$ of discriminant $D$ is said to be* reduced *if $|b| \leq a \leq c$ and if, in addition, when one of the two inequalities is an equality (i.e. either $|b| = a$ or $a = c$), then $b \geq 0$.*

This definition is equivalent to saying that the number $\tau = (-b + \sqrt{D})/(2a)$ corresponding to $(a, b, c)$ as above is in the standard fundamental domain $\mathcal{D}$ of $\mathcal{H}/\operatorname{PSL}_2(\mathbb{Z})$ (where $\mathcal{H} = \{\tau \in \mathbb{C}, \operatorname{Im}(\tau) > 0\}$), defined by

$$\mathcal{D} = \left\{ \tau \in \mathcal{H}, \operatorname{Re}(\tau) \in [-\frac{1}{2}, \frac{1}{2}[, |\tau| > 1 \text{ or } |\tau| = 1 \text{ and } \operatorname{Re}(\tau) \leq 0 \right\}.$$

The nice thing about this notion is the following:

**Proposition 5.3.3.** *In every class of positive definite quadratic forms of discriminant $D < 0$ there exists exactly one reduced form. In particular $h(D)$ is equal to the number of primitive reduced forms of discriminant $D$.*

An equivalent form of this proposition is that the set $\mathcal{D}$ defined above is a fundamental domain for $\mathcal{H}/\operatorname{PSL}_2(\mathbb{Z})$.

*Proof.* Among all forms $(a,b,c)$ in a given class, consider one for which $a$ is minimal. Note that for any such form we have $c \geq a$ since $(a,b,c)$ is equivalent to $(c,-b,a)$ (change $(x,y)$ into $(-y,x)$). Changing $(x,y)$ into $(x+ky,y)$ for a suitable integer $k$ (precisely for $k = \lfloor (a-b)/(2a) \rfloor$) will not change $a$ and put $b$ in the interval $]-a,a]$. Since $a$ is minimal, we will still have $a \leq c$, hence the form that we have obtained is essentially reduced. If $c = a$, changing $(a,b,c)$ again in $(c,-b,a)$ sets $b \geq 0$ as required. This shows that in every class there exists a reduced form.

Let us show the converse. If $(a,b,c)$ is reduced, I claim that $a$ is minimal among all the forms equivalent to $(a,b,c)$. Indeed, every other $a'$ has the form $a' = am^2 + bmn + cn^2$ with $m$ and $n$ coprime integers, and the identities

$$am^2 + bmn + cn^2 = am^2\left(1 + \frac{b}{a}\frac{n}{m}\right) + cn^2 = am^2 + cn^2\left(1 + \frac{b}{c}\frac{m}{n}\right)$$

immediately imply our claim, since $|b| \leq a \leq c$. Now in fact these same identities show that the only forms equivalent to $(a,b,c)$ with $a' = a$ are obtained by changing $(x,y)$ into $(x+ky,y)$ (corresponding to $m = 1$ and $n = 0$), and this finishes the proof of the proposition. □

We also have the following lemma.

**Lemma 5.3.4.** *Let $f = (a,b,c)$ be a positive definite binary quadratic form of discriminant $D = b^2 - 4ac < 0$.*

(1) *If $f$ is reduced, we have the inequality*

$$a \leq \sqrt{|D|/3}\,.$$

(2) *Conversely, if*

$$a < \sqrt{|D|/4} \quad \text{and} \quad -a < b \leq a$$

*then $f$ is reduced.*

*Proof.* For (1) we note that if $f$ is reduced then $|D| = 4ac - b^2 \geq 4a^2 - a^2$ hence $a \leq \sqrt{|D|/3}$. For (2), we have $c = (b^2 + |D|)/(4a) \geq |D|/(4a) > a^2/a = a$, therefore $f$ is reduced. □

As a consequence, we deduce that when $D < 0$ the class number $h(D)$ of $\mathbb{Q}(\sqrt{D})$ can be obtained simply by counting reduced forms of discriminant $D$ (since in that case all forms of discriminant $D$ are primitive), using the inequalities $|b| \leq a \leq \sqrt{|D|/3}$. This leads to the following algorithm.

**Algorithm 5.3.5** ($h(D)$ Counting Reduced Forms). Given a negative discriminant $D$, this algorithm outputs the class number of quadratic forms of discriminant $D$, i.e. $h(D)$ when $D$ is a fundamental discriminant.

1. [Initialize $b$] Set $h \leftarrow 1$, $b \leftarrow D \bmod 2$ (i.e. 0 if $D \equiv 0 \pmod 4$, 1 if $D \equiv 1 \pmod 4$), $B \leftarrow \left\lfloor \sqrt{|D|/3} \right\rfloor$.
2. [Initialize $a$] Set $q \leftarrow (b^2 - D)/4$, $a \leftarrow b$, and if $a \leq 1$ set $a \leftarrow 1$ and go to step 4.
3. [Test] If $a \mid q$ then if $a = b$ or $a^2 = q$ or $b = 0$ set $h \leftarrow h + 1$, otherwise (still in the case $a \mid q$) set $h \leftarrow h + 2$.
4. [Loop on $a$] Set $a \leftarrow a + 1$. If $a^2 \leq q$ go to step 3.
5. [Loop on $b$] Set $b \leftarrow b + 2$. If $b \leq B$ go to step 2, otherwise output $h$ and terminate the algorithm.

It can easily be shown that this algorithm indeed counts reduced forms. One must be careful in the formulation of this algorithm since the extra boundary conditions which occur if $|b| = a$ or $a = c$ complicate things. It is also easy to give some cosmetic improvements to the above algorithm, but these have little effect on its efficiency.

The running time of this algorithm is clearly $O(|D|)$, but the $O$ constant is very small since very few computations are involved. Hence it is quite a reasonable algorithm to use for discriminants up to a few million in absolute value. The typical running time for a discriminant of the order of $10^6$ is at most a few seconds on modern microcomputers.

**Remark.** If we want to compute $h(D)$ for a non-fundamental discriminant $D$, we must only count primitive forms. Therefore the above algorithm must be modified by replacing the condition "if $a \mid q$" of Step 3 by "if $a \mid q$ and $\gcd(a, b, q/a) = 1$".

A better method is as follows. Write $D = D_0 f^2$ where $D_0$ is a fundamental discriminant. The general theory seen in Chapter 4 tells us that $h(D)$ is a multiple of $h(D_0)$, but in fact Proposition 5.3.12 implies the following precise formula:

$$\frac{h(D)}{w(D)} = \frac{h(D_0)}{w(D_0)} f \prod_{p \mid f} \left(1 - \frac{\left(\frac{D_0}{p}\right)}{p}\right).$$

Hence, we compute $h(D_0)$ using the above algorithm, and deduce $h(D)$ from this formula.

Reduced forms are also very useful for making *tables* of class numbers of quadratic fields or forms up to a certain discriminant bound. Although each individual computation takes time $O(|D|)$, hence for $|D| \leq M$ the time would be $O(M^2)$, it is easy to see that a simultaneous computation (needing of course $O(M)$ memory locations to hold the class numbers) takes only $O(M^{3/2})$, hence an average of $O(|D|^{1/2})$ per class number.

Since class numbers of imaginary quadratic fields occur so frequently, it is useful to have a small table available. Such a table can be found in Appendix B. Some selected values are:

- Class number 1 occurs only for $D = -3, -4, -7, -8, -11, -19, -43, -67$ and $-163$.
- Class number 2 occurs only for $D = -15, -20, -24, -35, -40, -51, -52, -88, -91, -115, -123, -148, -187, -232, -235, -267, -403, -427$.
- Class number 3 occurs only for $D = -23, -31, -59, -83, -107, -139, -211, -283, -307, -331, -379, -499, -547, -643, -883, -907$.
- Class number 4 occurs for $D = -39, -55, -56, -68, \ldots, -1555$.
- Class number 5 occurs for $D = -47, -79, -103, -127, \ldots, -2683$.
- Class number 6 occurs for $D = -87, -104, -116, -152, \ldots, -3763$.
- Class number 7 occurs for $D = -71, -151, -223, -251, \ldots, -5923$.

etc ...

Note that the first two statements concerning class numbers 1 and 2 are very difficult theorems proved in 1952 by Heegner and in 1968-1970 by Stark and Baker (see [Cox]). The general problem of determining all imaginary quadratic fields with a given class number has been solved in principle by Goldfeld-Gross-Zagier ([Gol], [Gro-Zag2]), but the explicit computations have been carried to the end only for class numbers up to 7 and all odd numbers up to 23 (see [ARW], [Wag]).

The method using reduced forms is a very simple method to implement and is eminently suitable for computing tables of class numbers or for computing class numbers of reasonable discriminant, say less than a few million in absolute value. Since it is only a simple counting process, it does not give the structure of the class group. Also, it becomes too slow for larger discriminants, therefore we must find better methods.

### 5.3.2 Computing Class Numbers Using Modular Forms

I do not intend to explain why the theory of modular forms (specifically of weight 3/2 and weight 2) is closely related to class numbers of imaginary quadratic fields, but I would like to mention formulas which enable us to compute tables of class numbers essentially as fast as the method using reduced forms. First we need a definition.

**Definition 5.3.6.** *Let $N$ be a non-negative integer. The* Hurwitz class number *$H(N)$ is defined as follows.*

(1) *If $N \equiv 1$ or $2 \pmod 4$ then $H(N) = 0$.*
(2) *If $N = 0$ then $H(N) = -1/12$.*
(3) *Otherwise (i.e. if $N \equiv 0$ or $3 \pmod 4$ and $N > 0$) we define $H(N)$ as the class number of not necessarily primitive (positive definite) quadratic forms of discriminant $-N$, except that forms equivalent to $a(x^2 + y^2)$ should be counted with coefficient $1/2$, and those equivalent to $a(x^2 + xy + y^2)$ with coefficient $1/3$.*

Let us denote by $h(D)$ the class number of *primitive* positive definite quadratic forms of discriminant $D$. (This agrees with the preceding definition when $D$ is a fundamental discriminant since in that case every form is primitive.) Next, we define $h(D) = 0$ when $D$ is not congruent to 0 or 1 modulo 4. Then we have the following lemma.

**Lemma 5.3.7.** *Let $w(D)$ be the number of roots of unity in the quadratic order of discriminant $D$, hence $w(-3) = 6$, $w(-4) = 4$ and $w(D) = 2$ for $D < -4$, and set $h'(D) = h(D)/(w(D)/2)$ (hence $h'(D) = h(D)$ for $D < -4$). Then for $N > 0$ we have*

(1)
$$H(N) = \sum_{d^2|N} h'(-N/d^2)$$

*and in particular if $-N$ is a fundamental discriminant, we have $H(N) = h(-N)$ except in the special cases $N = 3$ ($H(3) = 1/3$ and $h(-3) = 1$) and $N = 4$ ($H(4) = 1/2$ and $h(-4) = 1$).*

(2) *Conversely, we have*
$$h'(-N) = \sum_{d^2|N} \mu(d) H(N/d^2)$$

*where $\mu(d)$ is the Möbius function defined by $\mu(d) = (-1)^k$ if $d$ is equal to a product of $k$ distinct primes (including $k = 0$), and $\mu(d) = 0$ otherwise.*

*Proof.* The first formula follows immediately from the definition of $H(N)$. The second formula is a direct consequence of the Möbius inversion formula (see [H-W]). □

From this lemma, it follows that the computation of a table of the function $H(N)$ is essentially equivalent to the computation of a table of the function $h(D)$.

For $D = -N$, Algorithm 5.3.5 computes a quantity similar to $H(N)$ but without the denominator $w(-N/d^2)/2$ in the formula given above. Hence, it can be readily adapted to compute $H(N)$ itself by replacing step 3 with the following:

3'. [Test] If $a \nmid q$ go to step 4. Now if $a = b$ then if $ab = q$ set $h \leftarrow h + 1/3$ otherwise set $h \leftarrow h + 1$ and go to step 4. If $a^2 = q$, then if $b = 0$ set $h \leftarrow h + 1/2$, otherwise set $h \leftarrow h + 1$. In all other cases (i.e. if $a \neq b$ and $a^2 \neq q$) set $h \leftarrow h + 2$.

The theory of modular forms of weight 3/2 tells us that the Fourier series

$$\sum_{N=0}^{\infty} H(N) e^{2i\pi N\tau}$$

has a special behavior when one changes $\tau$ by a linear fractional transformation $\tau \mapsto \frac{a\tau+b}{c\tau+d}$ in $\mathrm{PSL}_2(\mathbb{Z})$. Combined with other results, this gives many nice recursion formulas for $H(N)$ which are very useful for practical computation.

Let $\sigma(n) = \sum_{d|n} d$ be the sum of divisors function, and define

$$\lambda(n) = \frac{1}{2}\sum_{d|n} \min(d, n/d) = \sum_{d|n, d\leq\sqrt{n}}{}' d,$$

where $\sum'$ means that if the term $d = \sqrt{n}$ is present it should have coefficient $1/2$. In addition we define $\sigma(n) = \lambda(n) = 0$ if $n$ is not integral. Then (see [Eic2], [Zag1]):

**Theorem 5.3.8** (Hurwitz, Eichler). *We have the following relations, where it is understood that the summation variable $s$ takes positive, zero or negative values:*

$$\sum_{s^2\leq 4N} H(4N - s^2) = 2\sigma(N) - 2\lambda(N),$$

*and if $N$ is odd,*

$$\sum_{s^2\leq N, s\equiv(N+1)/2 \pmod 2} H(N - s^2) = \frac{\sigma(N)}{3} - \lambda(N).$$

From a computational point of view, the second formula is better. It is used in the following way:

**Corollary 5.3.9.** *If $N \equiv 3 \pmod 4$, then*

$$H(N) = \frac{\sigma(N)}{3} - \lambda(N) - 2\sum_{1\leq s<\sqrt{N/4}} H(N - 4s^2),$$

*and if $N \equiv 0 \pmod 4$, then*

$$H(N) = \frac{\sigma(N+1)}{6} - \frac{\lambda(N+1)}{2} - \sum_{1\leq s\leq(\sqrt{N+1}-1)/2} H(N - 4s(s+1)).$$

This corollary allows us to compute a table of class numbers up to any given bound $M$ in time $O(M^{3/2})$, hence is comparable to the method using reduced forms. It is slightly simpler to implement, but has the disadvantage that individual class numbers cannot be computed without knowing the preceding ones. It has an advantage, however, in that the computation of a block of class numbers can be done simply using the table of the lower ones, while this

cannot be done with the reduced forms technique, at least without wasting a lot of time.

**Remark.** The above theorem is similar to Theorem 4.9.14 and can be proved similarly. While $\zeta_K(-1)$ is closely linked to $r_5(D)$ when $D > 0$, $\zeta_K(0)$ (or essentially $h(D)$) is closely linked to $r_3(-D)$ when $D < 0$. More precisely we have (see [Coh2]):

**Proposition 5.3.10.** *Let $D < -4$ be the discriminant of an imaginary quadratic field $K$. Then the number $r_3(|D|)$ of representations of $|D|$ as a sum of 3 squares of elements of $\mathbb{Z}$ (counting representations with a different ordering as distinct) is given by*

$$r_3(|D|) = -24\left(1 - \left(\frac{D}{2}\right)\right)\zeta_K(0) = 12\left(1 - \left(\frac{D}{2}\right)\right)h(D).$$

*(This formula must be slightly modified if $D$ is not the discriminant of an imaginary quadratic field, see [Coh2].)*

### 5.3.3 Computing Class Numbers Using Analytic Formulas

It would carry us too far afield to enter into the details of the analytic theory of $L$-functions, hence we just recall a few definitions and results.

**Proposition 5.3.11** (Dirichlet). *Let $D$ be a negative discriminant (not necessarily fundamental), and define*

$$L_D(s) = \sum_{n \geq 1} \left(\frac{D}{n}\right) n^{-s}.$$

*This series converges for $\mathrm{Re}(s) > 1$, and defines an analytic function which can be analytically continued to the whole complex plane to an entire function. If in addition $D$ is a fundamental discriminant, this function satisfies the functional equation*

$$\Lambda_D(1-s) = \Lambda_D(s),$$

*where we have set*

$$\Lambda_D(s) = \left|\frac{D}{\pi}\right|^{(s+1)/2} \Gamma\left(\frac{s+1}{2}\right) L_D(s).$$

The link with class numbers is the following result also due to Dirichlet:

**Proposition 5.3.12.** *If $D$ is a negative discriminant (not necessarily fundamental), then*

$$L_D(1) = \frac{2\pi h(D)}{w(D)\sqrt{|D|}}$$

*and in particular $L_D(1) = \pi h(D)/\sqrt{|D|}$ if $D < -4$.*

Note that these results are special cases of Theorem 4.9.12 since it immediately follows from Proposition 5.1.4 that if $K = \mathbb{Q}(\sqrt{D})$, then

$$\zeta_K(s) = \zeta(s)L_D(s).$$

The series $L_D(1)$ is only conditionally convergent, hence it is not very reasonable to compute $L_D(1)$ directly using Dirichlet's theorem. A suitable transformation of the series however gives the following:

**Corollary 5.3.13.** *If $D < -4$ is a fundamental discriminant, then*

$$h(D) = \frac{1}{D}\sum_{1\le r<|D|} r\left(\frac{D}{r}\right) = \frac{1}{2-\left(\frac{D}{2}\right)}\sum_{1\le r<|D|/2}\left(\frac{D}{r}\right).$$

This formula is aesthetically very pleasing, and it can be transformed into even simpler expressions. It is unfortunately totally useless from a computational point of view since one must compute $D$ terms each involving the computation (admittedly rather short) of a Kronecker symbol. Hence, the execution time would be $O(|D|^{1+\epsilon})$, worse than the preceding methods.

A considerable improvement can be obtained if we also use the functional equation. This leads to a formula which is less pleasing, but which is much more efficient:

**Proposition 5.3.14.** *Let $D < -4$ be a fundamental discriminant. Then*

$$h(D) = \sum_{n\ge 1}\left(\frac{D}{n}\right)\left(\operatorname{erfc}\left(n\sqrt{\frac{\pi}{|D|}}\right) + \frac{\sqrt{|D|}}{\pi n}e^{-\pi n^2/|D|}\right),$$

*where*

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\,dt$$

*is the complementary error function.*

Note that the function $\operatorname{erfc}(x)$ can be computed efficiently using the following formulas.

**Proposition 5.3.15.**

(1) *We have for all $x$*

$$\operatorname{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}}\sum_{k\ge 0}(-1)^k\frac{x^{2k+1}}{k!(2k+1)},$$

*and this should be used when $x$ is small, say $x \le 2$.*

(2) *We have for all* $x > 0$

$$\operatorname{erfc}(x) = \frac{e^{-x^2}}{x\sqrt{\pi}} \left( 1 - \cfrac{1/2}{2 + X - \cfrac{1 \cdot 3/2}{4 + X - \cfrac{2 \cdot 5/2}{6 + X - \ddots}}} \right),$$

*where* $X = x^2 - 1/2$, *and this should be used for* $x$ *large, say* $x \geq 2$.

**Implementation Remark.** When implementing these formulas it is easy to make a mistake in the computation of $\operatorname{erfc}(x)$, and tables of this function are not always at hand. One good check is of course that the value found for $h(D)$ must be close to an integer, and for small $D$ equal to the values found by the slower methods. Another check is that, although we have given the most rapidly convergent series for $h(D)$ which can be obtained from the functional equation, we can get a one parameter family of formulas:

$$h(D) = \sum_{n \geq 1} \left(\frac{D}{n}\right) \left( \operatorname{erfc}\left( n\sqrt{\frac{\pi}{A|D|}} \right) + \frac{\sqrt{|D|}}{\pi n} e^{-\pi n^2 A/|D|} \right).$$

The sum of the series must be independent of $A > 0$.

The above results show that the series given in Proposition 5.3.14 for $h(D)$ converges exponentially, and since $h(D)$ is an integer it is clear that the computation time of $h(D)$ by this method is $O(|D|^{1/2+\epsilon})$ for any $\epsilon > 0$, however with a large $O$ constant. In fact it is not difficult to show the following precise result:

**Corollary 5.3.16.** *With the same notations as in Proposition 5.3.14,* $h(D)$ *is the closest integer to the* $n$*-th partial sum of the series of Proposition 5.3.14 for* $h(D)$, *where* $n = \left\lfloor \sqrt{|D| \ln |D|/(2\pi)} \right\rfloor$.

Hence, we see that this method is considerably faster than the two preceding methods, at least for sufficiently large discriminants. In addition, it is possible to avoid completely the computation of the higher transcendental function erfc, and this makes the method even more attractive (See Exercise 28).

It is reasonable to compute class numbers of discriminants having 12 to 15 digits by this method, but not much more. We must therefore find still better methods. In addition, we still have not given any method for computing the class *group*.

## 5.4 Class Groups of Imaginary Quadratic Fields

It was noticed by Shanks in 1968 that if one tries to obtain the class group structure and not only the class number, this leads to an algorithm which is much faster than the preceding algorithms, in average time $O(|D|^{1/4+\epsilon})$ or even $O(|D|^{1/5+\epsilon})$ if the Generalized Riemann Hypothesis is true, for any $\epsilon > 0$. Hence not only does one get much more information, i.e. the whole group structure, but even if one is interested only in the class number, this is a much better method.

Before entering into the details of the algorithm, we will describe a method introduced (for this purpose) by Shanks and which is very useful in many group-theoretic and similar contexts.

### 5.4.1 Shanks's Baby Step Giant Step Method

We first explain the general idea. Let $G$ be a finite Abelian group and $g$ an element of $G$. We want to compute the order of $g$ in $G$, i.e. the smallest positive integer $n$ such that $g^n = 1$, where we denote by 1 the identity element of $G$. One way of doing this is simply to compute $g$, $g^2$, $g^3$, ..., until one gets 1. This clearly takes $O(n)$ group operations. In certain cases, it is impossible to do much better. In most cases however, one knows an upper bound, say $B$ on the number $n$, and in that case one can do much better, using Shanks's baby-step giant-step strategy. One proceeds as follows. Let $q = \lceil \sqrt{B} \rceil$. Compute $1, g, \ldots, g^{q-1}$, and set $g_1 = g^{-q}$. Then if the order $n$ of $g$ is written in the form $n = aq + r$ with $0 \leq r < q$, by the choice of $q$ we must also have $a \leq q$. Hence, for $a = 1, \ldots, q$ we compute $g_1^a$ and check whether or not it is in our list of $g^r$ for $r < q$. If it is, we have $g^{aq+r} = 1$, hence $n$ is a divisor of $aq + r$, and the exact order can easily be obtained by factoring $aq + r$, at least if $aq + r$ is of factorable size (see Chapter 10). This method clearly requires only $O\left(B^{1/2}\right)$ group operations, and this number is much smaller than $O(n)$ if $B$ is a reasonable upper bound.

There is however one pitfall to avoid in this algorithm: we need to search (at most $q$ times) if an element belongs to a list having $q$ elements. If this is done naïvely, this will take $O(q^2) = O(B)$ comparisons, and even if group operations are much slower than comparisons, this will ultimately dominate the running time and render useless the method. To avoid this, we can first *sort* the list of $q$ elements, using a $O(q \ln q)$ sorting method such as heapsort (see [Knu3]). A search in a sorted list will then take only $O(\ln q)$ comparisons, bringing the total time down to $O(q \ln q)$. We can also use hashing techniques (see [Knu3] again).

This simple instance of Shanks's method involves at most $q$ "giant steps" (i.e. multiplication by $g_1$), each of size $q$. Extra information on $n$ can be used to improve the efficiency of the algorithm. We give two basic examples. Assume that in addition to an upper bound $B$, we also know a lower bound $C$, say, so that $C \leq n \leq B$. Then, by starting our list with $g^C$ instead of $g^0 = 1$, we

can reduce both the maximum number of giant steps and the size of the giant steps (and of the list) to $\lceil\sqrt{B-C}\rceil$.

As a second example, assume that we know that $n$ satisfies some congruence condition $n \equiv n_0 \pmod{b}$. Then it is easily seen that one can reduce the size and number of giant steps to $\lceil\sqrt{B/b}\rceil$.

Shanks's method is usually used not only to find the order of an element of the group $G$, but the order of the group itself. If $g$ is a generator of $G$, the preceding algorithm does the trick. In general however this will not be the case, and in addition $G$ may be non-cyclic (although cyclic groups occur much more often than one expects, see Section 5.10). In this case we must use the whole group structure, and not only one cyclic part. To do this, we can use the following algorithm.

**Algorithm 5.4.1** (Shanks's Baby-Step Giant-Step Method). Given that one can compute in $G$, and the inequalities $B/2 < C \leq h \leq B$ on the order $h$ of $G$, this algorithm finds $h$. We denote by 1 the identity element of $G$ and by $\cdot$ the product operation in $G$. The variables $S$ and $L$ will represent subsets of $G$.

1. [Initialize] Set $h \leftarrow 1$, $C_1 \leftarrow C$, $B_1 \leftarrow B$, $S \leftarrow \{1\}$, $L \leftarrow \{1\}$.
2. [Take a new $g$] (Here we know that the order of $G$ is a multiple of $h$). Choose a new random $g \in G$, $q \leftarrow \lceil\sqrt{B_1 - C_1}\rceil$.
3. [Compute small steps] Set $x_0 \leftarrow 1$, $x_1 \leftarrow g^h$ and if $x_1 = 1$ set $n \leftarrow 1$ and go to step 6. Otherwise, for $r = 2$ to $r = q-1$ set $x_r \leftarrow x_1 \cdot x_{r-1}$. For each $r$ with $0 \leq r < q$ set $S_{1,r} \leftarrow x_r \cdot S$, $S_1 \leftarrow \bigcup_{0\leq r<q} S_{1,r}$, and sort $S_1$ so that a search in $S_1$ is easy. If during this computation one finds $1 \in S_{1,r}$ for $r > 0$, set $n \leftarrow r$ (where $r$ is the smallest) and go to step 6. Otherwise, set $y \leftarrow x_1 \cdot x_{q-1}$, $z \leftarrow x_1^{C_1}$, $n \leftarrow C_1$.
4. [Compute giant steps] For each $w \in L$, set $z_1 \leftarrow z \cdot w$ and search for $z_1$ in the sorted list $S_1$. If $z_1$ is found and $z_1 \in S_{1,r}$, set $n \leftarrow n - r$ and go to step 6.
5. [Continue] Set $z \leftarrow y \cdot z$, $n \leftarrow n + q$. If $n \leq B_1$ go to step 4. Otherwise output an error message stating that the order of $G$ is larger than $B$ and terminate the algorithm.
6. [Initialize order] Set $n \leftarrow hn$.
7. [Compute the order of $g$ mod $L \cdot S$] (Here we know that $g^n \in L \cdot S$). For each prime $p$ dividing $n$, do the following: set $S_1 \leftarrow g^{n/p} \cdot S$ and sort $S_1$. If there exists a $z \in L$ such that $z \in S_1$, set $n \leftarrow n/p$ and go to step 7.
8. [Finished?]. Set $h \leftarrow hn$. If $h \geq C$ then output $h$ and terminate the algorithm. Otherwise, set $B_1 \leftarrow \lfloor B_1/n \rfloor$, $C_1 \leftarrow \lceil C_1/n \rceil$, $q \leftarrow \lceil\sqrt{n}\rceil$, $S \leftarrow \bigcup_{0\leq r<q} g^r \cdot S$, $y \leftarrow g^q$, $L \leftarrow \bigcup_{0\leq a\leq q} y^a \cdot L$ and go to step 2.

This is of course a probabilistic algorithm. The correctness of the result depends in an essential way on the correctness of the bounds $C$ and $B$. Since during the algorithm the order of $G$ is always a multiple of $h$, and since

$C > B/2$, the stopping criterion $h \geq C$ in step 8 is correct (any multiple of $h$ larger than $h$ would be larger than $B$). In practice however we may not be so lucky as to have a lower bound $C$ such that $C > B/2$. In that case, one cannot easily give any stopping criteria, and my advice is to stop as soon as $h$ has not changed after 10 passes through step 8. Note however that this is no longer an algorithm, since nothing guarantees the correctness of the result.

Note that if $g_i$ are elements of $G$ of respective orders $e_i$, then the exponent of $G$ is a multiple of the least common multiple (LCM) of the $e_i$. Hence, if one expects the exponent of the group to be not too much lower than the order $h$, one can use a much simpler method in which one simply computes the LCM of sufficiently many random elements of $G$, and then taking the multiple of this LCM which is between the given bounds $C$ and $B$. For this to succeed, the bounds have to be close enough. In practice, it is advised to first use this method to get a tentative order, then to use the rigorous algorithm given above to prove it, since a knowledge of the exponent of $G$ can clearly be used to improve the efficiency of Algorithm 5.4.1.

Let us explain why Algorithm 5.4.1 works. Let $H$ be the true order of $G$. Consider the first $g$. We have $g^H = 1$, and if we write $H - C = aq - r$ with $0 \leq r < q$ and $q = \lceil \sqrt{B-C} \rceil$, then also $a - 1 < (H-C)/q \leq (B-C)/q \leq q$ hence $a \leq q$. This implies that we have an equality of the form

$$g^C \cdot (g^q)^a = g^r$$

with $0 \leq r < q$ and $1 \leq a \leq q$. This is detected in step 4 of the algorithm, where we have $x_r = g^r$, $y = g^q$ and $z_1 = g^C \cdot (g^q)^a$. When we arrive in step 6 we know that $g^n = 1$ with $n = C + aq - r$, hence the order of $g$ is a divisor of $n$, and step 7 is the standard method for computing the order of an element in a group.

After that, $h$ is set to the order of $g$, and by a similar baby step giant step construction, $S$ and $L$ are constructed so that $S \cdot L = < g >$, the subgroup generated by $g$. We also know that the order $H$ of $G$ is a multiple of $h$. Hence, for a new $g_1$, instead of writing $g_1^H = 1$ and $H - C = aq - r$ we will write $(g_1^h)^{H_1} \in < g >$ and $H_1 - C_1 = aq_1 - r_1$, where $H_1 = H/h$ is known to be between $C_1 = \lceil C/h \rceil$ and $B_1 = \lfloor B/h \rfloor$, whence the modifications given in the algorithm when we start with a new $g$. □

Note that as we have already mentioned, it is essential to do some kind of ordering on the $x_r$ in step 3, otherwise the search time in step 4 would dominate the total time. In practical implementations, the best method is probably not to sort completely, but to use hashing techniques (see [Knu3]).

The expected running time of this algorithm is $O((B-C)^{1/2})$ group operations, and this is usually $O(B^{1/2+\epsilon})$ for all $\epsilon > 0$. For obvious reasons, the method above is called Shanks's baby-step giant-step method, and it can be profitably used in many contexts. For example, it can be used to compute class numbers and class groups (see Algorithm 5.4.10), regulators (see Algorithm 5.8.5), or the number of points of an elliptic curve over a finite field (see Algorithm 7.4.12).

We must now explain how to obtain the whole group structure. Call $g_1$, ..., $g_k$ the elements of $G$ which are chosen in step 2. Then when a match is found in step 3 or 4, we must record not only the exponent of $g$ which occurs, but the specific exponents of the preceding $g_i$. In other words, one must keep track of the multi-index exponents in the lists $L$ and $S$. If at step $i$ we have a relation of the form $g_1^{k_{1,i}} \cdots g_{i-1}^{k_{i-1,i}} g_i^{k_{i,i}} = 1$, with $g = g_i$ and $k_{i,i} = n$ after step 7 in the notation of the algorithm, we then consider the matrix $K = (k_{i,j})$ where we set $k_{i,j} = 0$ if $i > j$. Then we compute the Smith normal form of this matrix using Algorithm 2.4.14, and if $d_i$ are the diagonal elements of the Smith normal form, we have

$$G \simeq \bigoplus_{1 \le i \le k} (\mathbb{Z}/d_i\mathbb{Z}),$$

i.e. the group structure of $G$.

### 5.4.2 Reduction and Composition of Quadratic Forms

Before being able to apply the above algorithm (or any other algorithm using the group structure) to the class group, it is absolutely essential to be able to compute in the class group. As already mentioned, we could do this by using HNF computations on ideals. Although theoretically equivalent, it is more practical however to work on classes of quadratic forms. In Theorem 5.2.8 we have seen that the set of classes of quadratic forms is in a natural bijection with the class group. Hence, we can easily transport this group structure so as to give a group structure to classes of quadratic forms. This operation, introduced by Gauss in 1798 is called *composition* of quadratic forms. Also, since we will want to work with a class of forms, we will have a *reduction procedure* which, given any quadratic form, will give us the unique reduced form in its class. I refer the reader to [Len1] and [Bue] for more details on this subject.

The reduction algorithm is a variant of Euclid's algorithm:

**Algorithm 5.4.2** (Reduction of Positive Definite Forms). Given a positive definite quadratic form $f = (a, b, c)$ of discriminant $D = b^2 - 4ac < 0$, this algorithm outputs the unique reduced form equivalent to $f$.

1. [Initialize] If $-a < b \le a$ go to step 3.
2. [Euclidean step] Let $b = 2aq + r$ with $0 \le r < 2a$ be the Euclidean division of $b$ by $2a$. If $r > a$, set $r \leftarrow r - 2a$ and $q \leftarrow q + 1$. (In other words, we want $b = 2aq + r$ with $-a < r \le a$.) Then set $c \leftarrow c - \frac{1}{2}(b + r)q$, $b \leftarrow r$.
3. [Finished?] If $a > c$ set $b \leftarrow -b$, exchange $a$ and $c$ and go to step 2. Otherwise, if $a = c$ and $b < 0$, set $b \leftarrow -b$. Output $(a, b, c)$ and terminate the algorithm.

The proof of the validity of this algorithm follows from the proof of Proposition 5.3.3. Note that in step 2 we could have written $c \leftarrow c - bq + aq^2$, but writing it the way we have done avoids one multiplication per loop.

This algorithm has exactly the same behavior as Euclid's algorithm which we have analyzed in Chapter 1, hence is quite fast. In fact, we have the following.

**Proposition 5.4.3.** *The number of Euclidean steps in Algorithm 5.4.2 is at most equal to*

$$2 + \left\lceil \lg\left(\frac{a}{\sqrt{|D|}}\right)\right\rceil.$$

*Proof.* Consider the form $(a,b,c)$ at the beginning of step 3. Note first that if $a > \sqrt{|D|}$, then

$$c = \frac{b^2+|D|}{4a} \leq \frac{a^2+a^2}{4a} = \frac{a}{2},$$

hence, since in step 3 $a$ and $c$ are exchanged, $a$ decreases by a factor at least equal to 2. Hence, after at most $\lceil \lg(a/\sqrt{|D|})\rceil$ steps, we obtain at the beginning of step 3 a form with $a < \sqrt{|D|}$. Now we have the following lemma.

**Lemma 5.4.4.** *Let $(a,b,c)$ is a positive definite quadratic form of discriminant $D = b^2 - 4ac < 0$ such that $-a < b \leq a$ and $a < \sqrt{|D|}$. Then either $(a,b,c)$ is already reduced, or the form $(c,r,s)$ where $-b = 2cq + r$ with $-c < r \leq c$ obtained by one reduction step of Algorithm 5.4.2 will be reduced.*

*Proof.* If $(a,b,c)$ is already reduced, there is nothing to prove. Assume it is not. Since $-a < b \leq a$, this means that $a > c$ or $a = c$ and $b < 0$. This last case is trivial since at the next step we obtain the reduced form $(a,-b,a)$. Hence, assume $a > c$. If $-c < -b \leq c$, then $q = 0$ and so $(c,r,s) = (c,-b,a)$ is reduced. If $a \geq 2c$, then $c < \sqrt{|D|/4}$, and hence $(c,r,s)$ is reduced by Lemma 5.3.4. So we may assume $c < a < 2c$ and $-b \leq -c$ or $-b > c$. Since $|b| \leq a$, it follows that in the Euclidean division of $-b$ by $2c$ we must have $q = \pm 1$, the sign being the sign of $-b$. Now we have $s = a - bq + cq^2$, hence when $q = \pm 1$, $s = a \mp b + c \geq c$ since $|b| \leq a$. This proves that $(c,r,s)$ is reduced, except perhaps when $s = c$. In that case however we must have $a = \pm b$, hence $a = b$ so $b > 0$, $q = -1$ and $r = 2c - b \geq 0$. Therefore $(c,r,s)$ is also reduced in this case. This proves the lemma, and hence Proposition 5.4.3. □

We will now consider composition of forms. Although the group structure on ideal classes carries over only to classes of quadratic forms via the maps $\phi_{FI}$ and $\phi_{IF}$ defined in Section 5.2, we can define an operation between forms, which we call composition, which becomes a group law only at the level of classes modulo $\mathrm{PSL}_2(\mathbb{Z})$. Hence we will usually work on the level of forms.

Let $(a_1,b_1,c_1)$ and $(a_2,b_2,c_2)$ be two quadratic forms with the *same* discriminant $D$, and consider the corresponding ideals

$$I_k = a_k\mathbb{Z} + \frac{-b_k+\sqrt{D}}{2}\mathbb{Z} \qquad (k = 1,2)$$

given by the map $\phi_{FI}$ of Theorem 5.2.4. We have the following lemma

**Lemma 5.4.5.** *Let $I_1$ and $I_2$ be two ideals as above, set $s = (b_1 + b_2)/2$, $d = \gcd(a_1, a_2, s)$, and let $u$, $v$, $w$ be integers such that $ua_1 + va_2 + ws = d$. Then we have*

$$I_1 \cdot I_2 = d\left(A\mathbb{Z} + \frac{-B + \sqrt{D}}{2}\mathbb{Z}\right),$$

*where*

$$A = d_0\frac{a_1a_2}{d^2}, \qquad B = b_2 + \frac{2a_2}{d}(v(s - b_2) - wc_2)$$

*and $d_0 = 1$ if at least one of the forms $(a_1, b_1, c_1)$ or $(a_2, b_2, c_2)$ is primitive and in general $d_0 = \gcd(a_1, a_2, s, c_1, c_2, n)$ where $n = (b_1 - b_2)/2$.*

*Proof.* The ideal $I_3 = I_1 \cdot I_2$ is generated as a $\mathbb{Z}$-module by the four products of the generators of $I_1$ and $I_2$, i.e. by $g_1 = a_1a_2$, $g_2 = (-a_1b_2 + a_1\sqrt{D})/2$, $g_3 = (-a_2b_1 + a_2\sqrt{D})/2$ and $g_4 = ((b_1b_2 + D)/2 - s\sqrt{D})/2$. Now by Proposition 5.2.1 we know that we can write

$$I_3 = C\left(A\mathbb{Z} + \frac{-B + \sqrt{D}}{2}\mathbb{Z}\right)$$

for some integers $A$, $B$ and $C$. It is clear that $C$ is the smallest positive coefficient of $\sqrt{D}/2$ in $I_3$, hence is equal to the GCD of $a_1$, $a_2$ and $s$, so $C = d$ as stated. If one of the forms is primitive, or equivalently by Proposition 5.2.5 if one of the ideals is invertible, then by Proposition 4.6.8, we have $\mathcal{N}(I_3) = \mathcal{N}(I_1)\mathcal{N}(I_2) = a_1a_2$ and since $\mathcal{N}(I_3) = AC^2$ we have $A = a_1a_2/d^2$. (By Exercise 14 of Chapter 4, this will in fact still be true if $\gcd(a_1, b_1, c_1, a_2, b_2, c_2) = 1$, which is a slightly stronger condition than $d_0 = 1$.) This will also follow from the more general result where we make no assumptions of primitivity.

Let us directly determine the value of $AC$, i.e. the least positive integer belonging to $I_3$. Any element of $I_3$ being of the form $u_1g_1 + u_2g_2 + u_3g_3 + u_4g_4$ for integers $u_i$, the set $I_3 \cap \mathbb{Z}$ is the set of such elements with $u_2a_1 + u_3a_2 - u_4s = 0$. Using Exercise 11, the general solution to this is given by $u_2 = a_2/(a_1, a_2)\nu - s/(a_1, s)\mu$, $u_3 = s/(a_2, s)\lambda - a_1/(a_1, a_2)\nu$, $u_4 = a_2/(a_2, s)\lambda - a_1/(a_1, s)\mu$ for integers $\lambda$, $\mu$, $\nu$. After a short calculation, we see that $I_3 \cap \mathbb{Z} = e\mathbb{Z}$ where

$$e = \gcd\left(\frac{a_1a_2c_1}{(a_2, s)}, \frac{a_1a_2c_2}{(a_1, s)}, \frac{a_1a_2n}{(a_1, a_2)}, a_1a_2\right).$$

Another computation (see Exercise 8) shows that

$$e = \frac{a_1a_2}{(a_1, a_2, s)}\gcd(a_1, a_2, s, c_1, c_2, n)$$

thus giving the claimed value for $A = e/C = e/d$. Since $b_1 = s + n$ and $b_2 = s - n$, it is clear that if one of the forms is primitive then $d_0 = \gcd(a_1, a_2, s, c_1, c_2, n) = 1$ thus proving the statement made above.

Finally, if $d = ua_1 + va_2 + ws$, one possible value of $B$ is clearly

$$B = \frac{ua_1b_2 + va_2b_1 + w(b_1b_2 + D)/2}{d} = \frac{db_2 + va_2(b_1 - b_2) - 2a_2c_2w}{d},$$

thus proving the lemma. □

Note that if one writes $I_i = a_i(\mathbb{Z} + \tau_i\mathbb{Z})$, then we can reformulate the above lemma by saying that (with the same definitions of $d$, $u$, $v$ and $w$) we have $a_3 = a_1a_2d_0/d$ and $\tau_3 = (d/d_0)(u\tau_2 + v\tau_1 + w\tau_1\tau_2)$.

This leads to the following basic definition of the composite of two forms.

**Definition 5.4.6.** *Let $f_1 = (a_1, b_1, c_1)$ and $f_2 = (a_2, b_2, c_2)$ be two quadratic forms of the same discriminant $D$. Set $s = (b_1 + b_2)/2$, $n = (b_1 - b_2)/2$ and let $u$, $v$, $w$ and $d$ be such that*

$$ua_1 + va_2 + ws = d = \gcd(a_1, a_2, s)$$

*(obtained by two applications of Euclid's extended algorithm), and let $d_0 = \gcd(d, c_1, c_2, n)$. We define the composite of the two forms $f_1$ and $f_2$ as the form*

$$(a_3, b_3, c_3) = \left(d_0\frac{a_1a_2}{d^2},\ b_2 + \frac{2a_2}{d}(v(s - b_2) - wc_2),\ \frac{b_3^2 - D}{4a_3}\right).$$

*modulo the action of $\Gamma_\infty$, i.e. viewed as a form in the set $F$ introduced in Section 5.2.*

Since composition comes from the product of ideals, using the isomorphism given in Section 5.2, it is clear that the class in $F$ of $(a_3, b_3, c_3)$ does not depend on the particular choices of $u$, $v$ and $w$. This can of course also be checked directly (see Exercise 12). Note that if we do not take the class modulo $\Gamma_\infty$, the result is not at all canonical. Therefore when we speak of composition of quadratic forms we will always implicitly assume that we are working modulo the action of $\Gamma_\infty$, i.e. in the set $F$, and not on quadratic forms themselves.

To obtain the reduced composite of two forms, it is usually necessary to reduce the form obtained by composition. By abuse of language, in the case of negative discriminants we will also call this reduced form the composite of the two forms. (In the case of positive discriminants, there is in general more than one reduced form equivalent to a given form, hence this abuse of language is not permitted.)

Although the raw formulas given in the definition can be used directly, they can be improved by careful rearrangements. This leads to the following

algorithm, due to Shanks [Sha1]. Since imprimitive forms are almost never used, for the sake of efficiency we will restrict to the case of primitive forms. Note also that the composite of two primitive forms is still primitive (Exercise 9).

**Algorithm 5.4.7** (Composition of Positive Definite Forms). Given two *primitive* positive definite quadratic forms $f_1 = (a_1, b_1, c_1)$ and $f_2 = (a_2, b_2, c_2)$ with the same discriminant, this algorithm computes the composite $f_3 = (a_3, b_3, c_3)$ of $f_1$ and $f_2$.

1. [Initialize] If $a_1 > a_2$ exchange $f_1$ and $f_2$. Then set $s \leftarrow \frac{1}{2}(b_1+b_2)$, $n \leftarrow b_2-s$.
2. [First Euclidean step] If $a_1 \mid a_2$, set $y_1 \leftarrow 0$ and $d \leftarrow a_1$. Otherwise, using Euclid's extended algorithm compute $(u, v, d)$ such that $ua_2 + va_1 = d = \gcd(a_2, a_1)$, and set $y_1 \leftarrow u$.
3. [Second Euclidean step] If $d \mid s$, set $y_2 \leftarrow -1$, $x_2 \leftarrow 0$ and $d_1 \leftarrow d$. Otherwise, using Euclid's extended algorithm compute $(u, v, d_1)$ such that $us+vd = d_1 = \gcd(s, d)$, and set $x_2 \leftarrow u$, $y_2 \leftarrow -v$.
4. [Compose] Set $v_1 \leftarrow a_1/d_1$, $v_2 \leftarrow a_2/d_1$, $r \leftarrow (y_1y_2n - x_2c_2 \bmod v_1)$, $b_3 \leftarrow b_2+2v_2r$, $a_3 \leftarrow v_1v_2$, $c_3 \leftarrow (c_2d_1+r(b_2+v_2r))/v_1$ (or $c_3 \leftarrow (b_3^2-D)/(4a_3)$), then reduce the form $f = (a_3, b_3, c_3)$ using Algorithm 5.4.2, output the result and terminate the algorithm.

Note that this algorithm should be implemented as written: in step 2 we first consider the special case $a_1 \mid a_2$ because it occurs very often (at least each time one squares a form, and this is the most frequent operation when one raises a form to a power.) Therefore, it should be considered separately for efficiency's sake, although the general Euclidean step would give the same result. Similarly, in step 3 it often happens that $d \mid s$ because $d = 1$ also occurs quite often. Finally, note that the computation of $c_3$ in step 4 can be done using any of the two formulas given.

The generalization of this algorithm to imprimitive forms is immediate (see Exercise 10).

Since we have $|b_3| \le a_3 \le \sqrt{|D|/3}$ and since $c_3$ can be computed from $a_3$ and $b_3$, it seems plausible that one can make most of the computations in Algorithm 5.4.7 using numbers only of size $O(\sqrt{|D|})$ and not $O(D)$ or worse. That this is the case was noticed comparatively recently by Shanks and published only in 1989 [Sha2]. The improvement is considerable since in multi-precision situations it may gain up to a factor of 4, while in the case where $\sqrt{|D|}$ is single precision while $D$ is not, the gain is even larger.

This modified algorithm (called NUCOMP by Shanks) was modified again by Atkin [Atk1]. As mentioned above, squaring of a form is important and simpler, so Atkin gives two algorithms, one for duplication and one for composition.

**Algorithm 5.4.8** (NUDUPL). Given a primitive positive definite quadratic form $f = (a, b, c)$ of discriminant $D$, this algorithm computes the square $f^2 = f_2 = (a_2, b_2, c_2)$ of $f$. We assume that the constant $L = \lfloor |D/4|^{1/4} \rfloor$ has been precomputed.

1. [Euclidean step] Using Euclid's extended algorithm, compute $(u, v, d_1)$ such that $ub + va = d_1 = \gcd(b, a)$. Then set $A \leftarrow a/d_1$, $B \leftarrow b/d_1$, $C \leftarrow (-cu \bmod A)$, $C_1 \leftarrow A - C$ and if $C_1 < C$, set $C \leftarrow -C_1$.
2. [Partial reduction] Execute Sub-algorithm PARTEUCL$(A, C)$ below (this is an extended partial Euclidean algorithm).
3. [Special case] If $z = 0$, set $g \leftarrow (Bv_3 + c)/d$, $a_2 \leftarrow d^2$, $c_2 \leftarrow v_3^2$, $b_2 \leftarrow b + (d + v_3)^2 - a_2 - c_2$, $c_2 \leftarrow c_2 + gd_1$, reduce the form $f_2 = (a_2, b_2, c_2)$, output the result and terminate the algorithm.
4. [Final computations] Set $e \leftarrow (cv + Bd)/A$, $g \leftarrow (ev_2 - B)/v$ (these divisions are both exact and $v = 0$ has been dealt with in step 3), then $b_2 \leftarrow ev_2 + vg$. Then, if $d_1 > 1$, set $b_2 \leftarrow d_1 b_2$, $v \leftarrow d_1 v$, $v_2 \leftarrow d_1 v_2$. Finally, in order, set $a_2 \leftarrow d^2$, $c_2 \leftarrow v_3^2$, $b_2 \leftarrow b_2 + (d + v_3)^2 - a_2 - c_2$, $a_2 \leftarrow a_2 + ev$, $c_2 \leftarrow c_2 + gv_2$, reduce the form $f_2 = (a_2, b_2, c_2)$, output the result and terminate the algorithm.

**Sub-algorithm** PARTEUCL$(a, b)$. This algorithm does an extended partial Euclidean algorithm on $a$ and $b$, but uses the variables $v$ and $v_2$ instead of $u$ and $v_1$ in Algorithm 1.3.6.

1. [Initialize] Set $v \leftarrow 0$, $d \leftarrow a$, $v_2 \leftarrow 1$, $v_3 \leftarrow b$, $z \leftarrow 0$.
2. [Finished?] If $|v_3| > L$ go to step 3. Otherwise, if $z$ is odd, set $v_2 \leftarrow -v_2$ and $v_3 \leftarrow -v_3$. Terminate the sub-algorithm.
3. [Euclidean step] Let $d = qv_3 + t_3$ be the Euclidean division of $d$ by $v_3$ with $0 \le t_3 < |v_3|$. Set $t_2 \leftarrow v - qv_2$, $v \leftarrow v_2$, $d \leftarrow v_3$, $v_2 \leftarrow t_2$, $v_3 \leftarrow t_3$, $z \leftarrow z + 1$ and go to step 2.

I have given the gory details in steps 3 and 4 of Algorithm 5.4.8 just to show how a careful implementation can save time: the formula for $b_2$ in step 4 could have simply been written $b_2 \leftarrow b_2 + 2dv_3$. This would involve one multiplication and 2 additions. Since we need the quantities $d^2$ and $v_3^2$ for $a_2$ and $c_2$ anyway, the way we have written the formula involves 3 additions and one squaring. By a suitable implementation of a method analogous to the splitting method for polynomials explained in Chapter 3, this will be faster than 2 additions and one multiplication. Of course the gain is slight and the lazy reader may implement this in the more straightforward way, but it should be remembered that we are programming a basic operation in a group which will be used a large number of times, so any gain, even small, is worth taking.

Note also that the final reduction of $f_2$ will be very short, usually one or two Euclidean steps at most.

The proof of the validity of the algorithm is not difficult (see [Sha2]) and is left to the reader. It can also be checked that all the iterations (Euclid and reductions) are done on numbers less than $O(\sqrt{|D|})$, and that only a small and fixed number of operations are done on larger numbers.

Let us now look at the general algorithm for composition.

**Algorithm 5.4.9** (NUCOMP). Given two primitive positive definite quadratic forms with the same discriminant $f_1 = (a_1, b_1, c_1)$ and $f_2 = (a_2, b_2, c_2)$, this algorithm computes the composite $f_3 = (a_3, b_3, c_3)$ of $f_1$ and $f_2$. As in NUDUPL (Algorithm 5.4.8) we assume already precomputed the constant $L = \lfloor |D/4|^{1/4} \rfloor$. Note that the values of $a_1$ and $a_2$ may get changed, so they should be preserved if needed.

1. [Initialize] If $a_1 < a_2$ exchange $f_1$ and $f_2$. Then set $s \leftarrow \frac{1}{2}(b_1 + b_2)$, $n \leftarrow b_2 - s$.
2. [First Euclidean step] Using Euclid's extended algorithm, compute $(u, v, d)$ such that $ua_2 + va_1 = d = \gcd(a_1, a_2)$. If $d = 1$, set $A \leftarrow -un$, $d_1 \leftarrow d$ and go to step 5. If $d \mid s$ but $d \neq 1$, set $A \leftarrow -un$, $d_1 \leftarrow d$, $a_1 \leftarrow a_1/d_1$, $a_2 \leftarrow a_2/d_1$, $s \leftarrow s/d_1$ and go to step 5.
3. [Second Euclidean step] (here $d \nmid s$) Using Euclid's extended algorithm again, compute $(u_1, v_1, d_1)$ such that $u_1 s + v_1 d = d_1 = \gcd(s, d)$. Then, if $d_1 > 1$, set $a_1 \leftarrow a_1/d_1$, $a_2 \leftarrow a_2/d_1$, $s \leftarrow s/d_1$ and $d \leftarrow d/d_1$.
4. [Initialization of reduction] Compute $l \leftarrow -u_1(uc_1 + vc_2) \bmod d$ by first reducing $c_1$ and $c_2$ (which are large) modulo $d$ (which is small), doing the operation, and reducing again. then set $A \leftarrow -u(n/d) + l(a_1/d)$.
5. [Partial reduction] Set $A \leftarrow (A \bmod a_1)$, $A_1 \leftarrow a_1 - A$ and if $A_1 < A$ set $A \leftarrow -A_1$, then execute Sub-algorithm PARTEUCL$(a_1, A)$ above.
6. [Special case] If $z = 0$, set $Q_1 \leftarrow a_2 v_3$, $Q_2 \leftarrow Q_1 + n$, $f \leftarrow Q_2/d$, $g \leftarrow (v_3 s + c_2)/d$, $a_3 \leftarrow da_2$, $c_3 \leftarrow v_3 f + g d_1$, $b_3 \leftarrow 2Q_1 + b_2$, reduce the form $f_3 = (a_3, b_3, c_3)$, output the result and terminate the algorithm.
7. [Final computations] Set $b \leftarrow (a_2 d + nv)/a_1$, $Q_1 \leftarrow bv_3$, $Q_2 \leftarrow Q_1 + n$, $f \leftarrow Q_2/d$, $e \leftarrow (sd + c_2 v)/a_1$, $Q_3 \leftarrow ev_2$, $Q_4 \leftarrow Q_3 - s$, $g \leftarrow Q_4/v$ (the case $v = 0$ has been dealt with in step 6), and if $d_1 > 1$ set $v_2 \leftarrow d_1 v_2$, $v \leftarrow d_1 v$. Finally, set $a_3 \leftarrow db + ev$, $c_3 \leftarrow v_3 f + g v_2$, $b_3 \leftarrow Q_1 + Q_2 + d_1(Q_3 + Q_4)$, reduce the form $f_3 = (a_3, b_3, c_3)$, output the result and terminate the algorithm.

Note that all the divisions which are performed in this algorithm are exact, and that the final reduction step, as in NUDUPL, will be very short, usually one or two Euclidean steps at most. As for NUDUPL, we leave to the reader the proof of the validity of this algorithm.

**Implementation Remark.** We have used the basic Algorithm 1.3.6 as a template for Sub-algorithm PARTEUCL. In practice, when dealing with multi-precision numbers, it is preferable to use one of its variants such as Algorithm 1.3.7 or 1.3.8.

### 5.4.3 Class Groups Using Shanks's Method

From the Brauer-Siegel theorem, we know that the class number $h(D)$ of an imaginary quadratic field grows roughly like $|D|^{1/2}$. This means that the baby-step giant-step algorithm given above allows us to compute $h(D)$ in time $O(|D|^{1/4+\epsilon})$, which is much better than the preceding methods. In fact, suitably implemented, one can reasonably expect to compute class numbers and class groups of discriminants having up to 20 or 25 decimal digits. For taking powers of the quadratic forms one should use the powering algorithm of Section 1.2, using if possible NUDUPL for the squarings and NUCOMP for general composition, or else using Shanks less optimized but simpler Algorithm 5.4.7. To be able to use the baby-step giant-step Algorithm 5.4.1 however, we need bounds for the class number $h(D)$. Now rigorous and explicit bounds are difficult to obtain, even assuming the GRH. Hence, we will push our luck and give only *tentative* bounds. Of course, this completely invalidates the rigor of the algorithm. To be sure that the result is correct, one should start with proven bounds like $C = 0$ and $B = \frac{1}{\pi}\sqrt{|D|}\ln|D|$ (see Exercise 27), however the performance is much worse.

Now the series giving $L_D(1)$ is only conditionally convergent, as is the corresponding Euler product

$$L_D(s) = \prod_p \left(1 - \frac{\left(\frac{D}{p}\right)}{p^s}\right)^{-1}.$$

However this Euler product is faster to compute to a given accuracy, since only the primes are needed. Hence, to start Shanks's algorithm, we take a large prime number bound $P$ (say $P = 2^{18}$), and guess that, for $D < -4$, $h(D)$ will be close to

$$\tilde{h} = \left\lfloor \frac{\sqrt{|D|}}{\pi} \prod_{p \le P} \left(1 - \frac{\left(\frac{D}{p}\right)}{p}\right)^{-1} \right\rfloor.$$

Assuming GRH, one can show that

$$h(D) - \tilde{h} = O(\tilde{h} P^{-1/2} \ln(P|D|)),$$

and one can give explicit values for the $O$ constant. In practice, Shanks noticed experimentally that the relative error is around $1/1000$ when $P = 2^{17}$. Hence, if we use these numerical bounds combined with the baby-step giant-step method, we will correctly compute $h(D)$ unless the exponent of the group is very small compared to the order.

A very important speedup in computing $h(D)$ by Shanks's method is obtained by noticing that the inverse for composition of the form $(a, b, c)$ is the form $(a, -b, c)$, hence requires no calculation. Hence, one can double the size of the giant steps (by setting $y \leftarrow x_1^{2q}$ instead of $y \leftarrow x_1^q$ in step 3 of Algorithm 5.4.1). Therefore the optimal value for $q$ is no longer $\sqrt{B - C}$ but rather $\sqrt{(B - C)/2}$.

Finally, note that during the computation of the Euler product leading to $\tilde{h}$, we will also have found the primes $p$ for which $\left(\frac{D}{p}\right) = 1$. For the first few such $p$, we compute the square root $b_p$ of $D$ mod $4p$ by a simple modification of Algorithm 1.5.1, and we store the forms $(p, b_p, c_p)$ where $c_p = (b_p^2 - D)/(4p)$. These will be used as our "random" $x$ in step 2 of the algorithm.

Putting all these ideas together leads to the following method:

**Heuristic Algorithm 5.4.10** ($h(D)$ Using Baby-Step Giant-Step). If $D < -4$ is a discriminant, this algorithm tries to compute $h(D)$ using a simpleminded version of Shanks's baby-step giant-step method. We denote by $\cdot$ the operation of composition of quadratic forms, and by 1 the unit element in the class group. We choose a small bound $b$ (for example $b = 10$).

1. [Compute Euler product] For $P = \max(2^{18}, |D|^{1/4})$, compute the product

$$Q \leftarrow \left\lfloor \frac{\sqrt{|D|}}{\pi} \prod_{p \le P} \left(1 - \frac{\left(\frac{D}{p}\right)}{p}\right)^{-1} \right\rfloor.$$

   Then set $B \leftarrow \lfloor Q(1 + 1/(2\sqrt{P})) \rfloor$, $C \leftarrow \lceil Q(1 - 1/(2\sqrt{P})) \rceil$. For the first $b$ values of $p$ such that $\left(\frac{D}{p}\right) = 1$, compute $b_p$ such that $b_p^2 \equiv D \pmod{4p}$ using Algorithm 1.5.1 (and modifying the result to get the correct parity). Set $f_p \leftarrow (p, b_p, (b_p^2 - D)/(4p))$.

2. [Initialize] Set $e \leftarrow 1$, $c \leftarrow 0$, $B_1 \leftarrow B$, $C_1 \leftarrow C$, $Q_1 \leftarrow Q$.

3. [Take a new $g$] (Here we know that the exponent of $Cl(D)$ is a multiple of $e$). Set $g \leftarrow f_p$ for the first new $f_p$, and set $c \leftarrow c + 1$, $q \leftarrow \lceil \sqrt{(B_1 - C_1)/2} \rceil$.

4. [Compute small steps] Set $x_0 \leftarrow 1$, $x_1 \leftarrow g^e$ then for $r = 2$ to $r = q - 1$ set $x_r \leftarrow x_1 \cdot x_{r-1}$. If, during this computation one finds $x_r = 1$, then set $n \leftarrow r$ and go to step 7. Otherwise, sort the $x_r$ so that searching among them is easy, and set $y \leftarrow x_1 \cdot x_{q-1}$, $y \leftarrow y^2$, $z \leftarrow x_1^{Q_1}$, $n \leftarrow Q_1$ .

5. [Compute giant steps] Search for $z$ or $z^{-1}$ in the sorted list of $x_r$ for $0 \le r < q$ (recall that if $z = (a, b, c)$, $z^{-1} = (a, -b, c)$). If a match $z = x_r$ is found, set $n \leftarrow n - r$ and go to step 7. If a match $z^{-1} = x_r$ is found, set $n \leftarrow n + r$ and go to step 7.

6. [Continue] Set $z \leftarrow y \cdot z$, $n \leftarrow n + 2q$. If $n \le B_1$ go to step 5. Otherwise output an error message stating that the order of $G$ is larger than $B$ and terminate the algorithm.

7. [Compute the order of $g$] (Here we know that $g^{en} = x_1^n = 1$). For each prime $p$ dividing $n$, do the following: if $x_1^{n/p} = 1$, then set $n \leftarrow n/p$ and go to step 7.

8. [Finished?] (Here $n$ is the exact order of $x_1$). Set $e \leftarrow en$. If $e > B - C$, then set $h \leftarrow e\lfloor B/e \rfloor$, output $h$ and terminate the algorithm. If $c \ge b$ output a message saying that the algorithm fails to find an answer and terminate the algorithm. Otherwise set $B_1 \leftarrow \lfloor B_1/n \rfloor$, $C_1 \leftarrow \lceil C_1/n \rceil$ and go to step 3.

This is *not* an algorithm, in the sense that the output may be false. One should compute the whole group structure using Algorithm 5.4.1 to be sure that the result is valid. It almost always gives the right answer however, and thus should be considered as a first step.

## 5.5 McCurley's Sub-exponential Algorithm

We now come to an algorithm discovered in 1988 by McCurley [McCur, Haf-McCur1] and which is much faster than the preceding algorithms for large discriminants. Several implementations of this algorithm have been done, for example by Düllmann, ([Buc-Dül]) and it is now reasonable to compute the class group for a discriminant of 50 decimal digits. Such examples have been computed by Düllmann and Atkin.

Incidentally, unlike almost all other algorithms in this book, little has been done to optimize the algorithm that we give, and there is plenty of room for (serious) improvements. This is, in fact, a subject of active research.

### 5.5.1 Outline of the Algorithm

Before giving the details of the algorithm, let us give an outline of the main ideas. First, instead of trying to obtain the class number and class group "from below", by finding relations $x^e = 1$, and hence *divisors* of the class number, we will find it "from above", i.e. by finding *multiples* of the class number.

Let $\mathcal{P}$ be a finite set of primes $p$ such that $\left(\frac{D}{p}\right) = 1$ for all $p \in \mathcal{P}$. Then, as in Shanks's method, we can find reduced forms $f_p = (p, b_p, c_p)$, which we will call *prime forms*, for each $p \in \mathcal{P}$. Now, assuming GRH, one can prove that there exists a constant $c$ which can be computed effectively such that if $\mathcal{P}$ contains all the primes $p$ such that $\left(\frac{D}{p}\right) = 1$ and $p \leq c\ln^2|D|$, then the classes of the forms $f_p$ for $p \in \mathcal{P}$ generate the class group. This means that if we set $n = |\mathcal{P}|$, the map

$$\begin{aligned} \phi : \mathbb{Z}^n &\longrightarrow Cl(D) \\ (x_p)_{p\in\mathcal{P}} &\mapsto \prod_{p\in\mathcal{P}} f_p^{x_p} \end{aligned}$$

is a surjective group homomorphism. Hence, the kernel $\Lambda$ of $\phi$ is a sublattice of $\mathbb{Z}^n$, and we have

$$\mathbb{Z}^n/\Lambda \simeq Cl(D) \qquad \text{and} \qquad |\det(\Lambda)| = h(D),$$

denoting by $\det(\Lambda)$ the determinant of any $\mathbb{Z}$-basis of $\Lambda$. The lattice $\Lambda$ is the lattice of relations among the $f_p$. If one finds any system of $n$ independent elements in this lattice, it is clear that the determinant of this system will

be a multiple of the determinant of $\Lambda$, hence of $h(D)$. This is how we obtain multiples of the class number.

Now there remains the question of obtaining (many) relations between the $f_p$. To do this, one uses the following lemma:

**Lemma 5.5.1.** *Let $(a,b,c)$ be a primitive positive definite quadratic form of discriminant $D<0$, and $a=\prod_p p^{v_p}$ be the prime decomposition of $a$. Then we have up to equivalence:*

$$(a,b,c)=\prod_p f_p^{\epsilon_p v_p},$$

*where $f_p=(p,b_p,c_p)$ is the prime form corresponding to $p$, and $\epsilon_p=\pm 1$ is defined by the congruence*

$$b\equiv \epsilon_p b_p \pmod{2p}.$$

In fact, all the possible choices for the $\epsilon_p$ correspond exactly to the possible square roots $b$ of $D$ mod $4a$, with $b$ defined modulo $2a$.

*Proof.* This lemma follows immediately from the raw formulas for composition that we have given in Section 5.4.2. In terms of ideals, using the correspondence given by Theorem 5.2.8, if $I=\psi_{FI}(\bar{f})$, the factorization of $a=\mathcal{N}(I)$ corresponds to a factorization $I=\prod \mathfrak{p}^{v_p}$ where $\mathfrak{p}$ is an ideal above $p\mathbb{Z}_K$, and $\epsilon_p$ must be chosen as stated so that $\mathfrak{p}\supset I$. □

This leads immediately to the following idea for generating relations in $\Lambda$: choose random integer exponents $e_p$, and compute the reduced form $(a,b,c)$ equivalent to $\prod_{p\in\mathcal{P}} f_p^{e_p}$. If all the factors of $a$ are in $\mathcal{P}$, we keep the form $(a,b,c)$, otherwise we take other random exponents. If the form is kept, we will have the relation

$$\prod_{p\in\mathcal{P}} f_p^{e_p-\epsilon_p v_p}=1,$$

giving the element

$$(e_p-\epsilon_p v_p)_{p\in\mathcal{P}}\in\Lambda\subset\mathbb{Z}^n.$$

Continuing in this way, one may reasonably hope to generate $\Lambda$ if $\mathcal{P}$ has been chosen large enough, and this is indeed what one proves, under suitable hypotheses.

The crucial point is the choice of $\mathcal{P}$. We will take

$$\mathcal{P}=\left\{p\le P,\left(\frac{D}{p}\right)\neq -1\right\}$$

for a suitable $P$, but one must see how large this $P$ must be to optimize the algorithm. If $P$ is chosen too small, numbers $a$ produced as above will almost

never factor into primes less than $P$. If $P$ is too large, then the factoring time of $a$ becomes prohibitive, as does the memory required to keep all the relations and the $f_p$. To find the right compromise, one must give the algorithm in much greater detail and analyze its behavior. This is done in [Haf-McCur1], where it is shown that $P$ should be taken of the order of $L(|D|)^{\alpha}$, where $L(x)$ is a very important function defined by

$$L(x) = e^{\sqrt{\ln x \ln \ln x}},$$

and $\alpha$ depends on the particular implementation, one possible value being $1/\sqrt{8}$. We will meet this very important function $L(x)$ again in Chapter 10 in connection with modern factoring methods.

In addition we must have $P \geq c \ln^2 |D|$ so that (assuming GRH) the classes of prime forms $f_p$ with $p \in \mathcal{P}$ generate the class group. Unfortunately, at present, the best known bound for the constant $c$, due to Bach, is 6, although practical experience shows that this is much too pessimistic. (In fact it is believed that $O(\ln^{1+\epsilon} |D|)$ generators should suffice for any $\epsilon > 0$). Hence, we will choose

$$P = \max\left(6 \ln^2 |D|, L(|D|)^{1/\sqrt{8}}\right).$$

Note that, although the $\ln^2$ function grows asymptotically much more slowly than the $L(|D|)$ function, in practice the constants 6 and $1/\sqrt{8}$ will make the $\ln^2$ term dominate. More precisely, the $L(|D|)$ term will start to dominate only for discriminants having at least 103 digits, well outside the range of practical applicability of this method. Even if one could reduce the constant 6 to 1, the $\ln^2$ term would still dominate for numbers having up to 70 digits.

Let $n$ be the number of $p \in \mathcal{P}$. To give a specific numerical example, for $D$ of the order of $-10^{40}$, with the above formula $P$ will be around 50900, and $n$ around 2600, while if $D$ is of the order of $-10^{50}$, $P$ will be around 79500 and $n$ around 3900. Since we will be handling determinants of $n \times n$ matrices, many problems become serious, in particular the storage problems, though they are perhaps still manageable. In any case, the computational load becomes very great. In particular, for matrices of this size it is essential to use special techniques adapted to the type of matrices which we have, i.e. sparse matrices. Since we are over $\mathbb{Z}$ and not over a field, the use of methods such as Wiedemann's *coordinate recurrence method* (see [Wie]) is possible only through the use of the Chinese remainder theorem, and is quite painful. An easier approach is to use "intelligent Hermite reduction", analogous to the intelligent Gaussian elimination technique used by LaMacchia and Odlyzko (see [LaM-Odl]). This method has been implemented by Düllmann ([Buc-Dül]) and by Cohen, Diaz y Diaz and Olivier ([CohDiOl]), and is described below.

### 5.5.2 Detailed Description of the Algorithm

We first make a few remarks.

The first important remark is that although one should generate random relations using Lemma 5.5.1, one may hope to obtain a non-trivial relation as soon as $\prod_p p^{e_p} > \sqrt{|D|/3}$ since the resulting form obtained by multiplication without reduction will not be reduced. Hence, instead of taking the whole of $\mathcal{P}$ to compute the products, we take a much smaller subset $\mathcal{P}_0$ not containing any prime dividing $D$ and such that

$$\prod_{p \in \mathcal{P}_0} p > \sqrt{|D|/3}\,.$$

Then $\mathcal{P}_0$ will be *very* small, typically of cardinality 10 or 20, even for discriminants in the 40 to 50 digit range. In fact, by the prime number theorem, the cardinality of $\mathcal{P}_0$ should be of the order of $\ln|D|/\ln\ln|D|$. For similar reasons, although the exponents $e_p$ should be chosen randomly up to $|D|$ as McCurley's analysis shows, in practice it suffices to take very small random exponents, say $1 \leq e_p \leq 20$.

A second remark is that, even if we use intelligent Hermite reduction as will be described, the size of the matrix involved will be very large. Hence, we must try to make it smaller even before we start the reduction. One way to do this is to decide to take a lower value of $P$, say one corresponding to the constant $c = 1$ (i.e. the split primes of norm less than $\ln^2|D|$ instead of $6\ln^2|D|$). This would probably work, but even under the GRH the result may be false since we may not have enough generators. There is however one way out of this. For every prime $q$ such that $\ln^2|D| < q < 6\ln^2|D|$, let $g_q$ be a reduced form equivalent to $f_q \prod_{p \in \mathcal{P}_0} f_p^{e_p}$ with small random exponents $e_p$ as before. If $g_q = (a, b, c)$, then, if $a$ factors over our factor base $\mathcal{P}$, since $q$ is quite large, with a little luck after a few trials we will find an $a$ which not only factors, but whose prime factors are all less than $q$. This means that $f_q$ belongs to the subgroup generated by the other $f_p$'s, hence can be discarded as a generator of the class group. Doing this for all the $q > \ln^2|D|$ is fast and does not involve any matrix handling, and in effect reduces the problem to taking the constant 1 instead of 6 in the definition of $P$, giving much smaller matrices. Note that the constant 1 which we have chosen is completely arbitrary, but it must not be chosen too small, otherwise it will become very difficult to eliminate the big primes $q$. In practice, values between 0.5 and 2 seem reasonable.

These kind of ideas can be pushed further. Instead of taking products using only powers of forms $f_p$ with $p \in \mathcal{P}_0$, we can systematically multiply such a relation by a prime $q$ larger than the ones in $\mathcal{P}_0$, with the hope that this extra prime will still occur non-trivially in the resulting relation.

A third remark is that ambiguous forms (i.e. whose square is principal) have to be treated specially in the factor base, since only the parity of the exponents will count. (This is why we have excluded primes dividing $D$ in

$\mathcal{P}_0$.) In fact, it would be better to add the free relations $f_p^2 = 1$ for all $p \in \mathcal{P}$ dividing $D$. On the other hand, when $D$ is not a fundamental discriminant, one must exclude from $\mathcal{P}$ the primes $p$ dividing $D$ to a power higher than the first (except for $p = 2$ which one keeps if $D/4$ is congruent to 2 or 3 modulo 4). For our present exposition, such primes will be called *bad*, the others *good*.

**Algorithm 5.5.2** (Sub-Exponential Imaginary Class Group). If $D < 0$ is a discriminant, this algorithm computes the class number $h(D)$ and the class group $Cl(D)$. As before, in practice we work with binary quadratic forms. We also choose a positive real constant $b$.

1. [Compute primes and Euler product] Set $m \leftarrow b\ln^2|D|$, $M \leftarrow L(|D|)^{1/\sqrt{8}}$, $P \leftarrow \lfloor \max(m, M) \rfloor$

$$\mathcal{P} \leftarrow \left\{ p \leq P, \left(\frac{D}{p}\right) \neq -1 \text{ and } p \text{ good} \right\}$$

and compute the product

$$B \leftarrow \left\lfloor \frac{\sqrt{|D|}}{\pi} \prod_{p \leq P} \left(1 - \frac{\left(\frac{D}{p}\right)}{p}\right)^{-1} \right\rfloor.$$

2. [Compute prime forms] Let $\mathcal{P}_0$ be the set made up of the smallest primes of $\mathcal{P}$ not dividing $D$ such that $\prod_{p \in \mathcal{P}_0} p > \sqrt{|D|/3}$. For the primes $p \in \mathcal{P}$ do the following. Compute $b_p$ such that $b_p^2 \equiv D \pmod{4p}$ using Algorithm 1.5.1 (and modifying the result to get the correct parity). If $b_p > p$, set $b_p \leftarrow 2p - b_p$. Set $f_p \leftarrow (p, b_p, (b_p^2 - D)/(4p))$. Finally, let $n$ be the number of primes $p \in \mathcal{P}$.
3. [Compute powers] For each $p \in \mathcal{P}_0$ and each integer $e$ such that $1 \leq e \leq 20$ compute and store the unique reduced form equivalent to $f_p^e$. Set $k \leftarrow 0$.
4. [Generate random relations] Let $f_q$ be the primeform number $k + 1 \bmod n$ in the factor base. Choose random $e_p$ between 1 and 20, and compute the unique reduced form $(a, b, c)$ equivalent to

$$f_q \prod_{p \in \mathcal{P}_0} f_p^{e_p}$$

until $v_q(a) \neq 1$ (note that the $f_p^{e_p}$ have already been computed in step 3). Set $e_p \leftarrow 0$ if $p \notin \mathcal{P}_0$ then $e_q \leftarrow e_q + 1$.
5. [Factor $a$] Factor $a$ using trial division. If a prime factor of $a$ is larger than $P$, do not continue the factorization and go to step 4. Otherwise, if $a = \prod_{p \leq P} p^{v_p}$, set $k \leftarrow k + 1$, and for $i \leq n$

$$a_{i,k} \leftarrow e_{p_i} - \epsilon_{p_i} v_{p_i}$$

where $\epsilon_{p_i} = +1$ if $(b \bmod 2p_i) \leq p_i$, $\epsilon_{p_i} = -1$ otherwise.

6. [Enough relations?] If $k < n + 10$ go to step 4.
7. [Be honest] For each prime $q$ such that $P < q \leq 6\ln^2|D|$ do the following. Choose random $e_p$ between 1 and 20 (say) and compute the primeform $f_q$ corresponding to $q$ and the unique reduced form $(a, b, c)$ equivalent to $f_q \prod_{p \in \mathcal{P}_0} f_p^{e_p}$. If $a$ does not factor into primes less than $q$, choose other exponents $e_p$ and continue until $a$ factors into such primes. Then go on to the next prime $q$ until the list is exhausted.
8. [Simple HNF] Perform a preliminary simple Hermite reduction on the $n \times k$ matrix $A = (a_{i,j})$ as described below, thus obtaining a much smaller matrix $A_1$.
9. [Compute determinant] Using standard Gaussian elimination techniques, compute the determinant of the lattice generated by the columns of the matrix $A_1$ modulo small primes $p$. Then compute the determinant $d$ exactly using the Chinese remainder theorem and Hadamard's inequality (see also Exercise 13). If the matrix is not of rank equal to its number of rows, get 5 more relations (in steps 4 and 5) and go to step 8.
10. [HNF reduction] Using Algorithm 2.4.8 compute the Hermite normal form $H = (h_{i,j})$ of the matrix $A_1$ using modulo $d$ techniques. Then, for every $i$ such that $h_{i,i} = 1$, suppress row and column $i$. Let $W$ be the resulting matrix.
11. [Finished?] Let $h \leftarrow \det(W)$ (i.e. the product of the diagonal elements). If $h \geq B\sqrt{2}$, get 5 more relations (in steps 4 and 5) and go to step 8. (It will not be necessary to recompute the whole HNF, but only to take into account the last 5 columns.) Otherwise, output $h$ as the class number.
12. [Class group] Compute the Smith normal form of $W$ using Algorithm 2.4.14. Output those diagonal elements $d_i$ which are greater than 1 as the invariants of the class group (i.e. $Cl(D) = \bigoplus \mathbb{Z}/d_i\mathbb{Z}$) and terminate the algorithm.

**Implementation Remarks.**

(1) The constant $b$ used in step 1 is important mainly to control the size of the final matrix $A$ on which we are going to work. As mentioned above however, $b$ must not be chosen too small, otherwise we will have a lot of trouble in the factoring stages. Practice shows that values between 0.5 and 2.0 are quite reasonable.

With such a choice of $b$, we could of course avoid step 7 entirely since it seems highly implausible that the class group is not generated by the first $0.5\ln^2|D|$ primeforms. Including step 7, however, makes the correctness of the result depend only on the GRH and nothing else. Note also that strictly speaking the above algorithm could run indefinitely, either because it does not find enough relations, or because the condition of step 7 is never satisfied for some prime $q$. In practice this never occurs.

(2) The simple Hermite reduction which is needed in step 8 is the following. We first scan all the rows of the $n \times k$ matrix A to detect if some have a

single $\pm 1$, the other coefficients being equal to zero. If this is the case and we find that $a_{i,j} = \pm 1$ is the only non-zero element of its row, we exchange rows $i$ and $n$ and columns $j$ and $k$, and scan the matrix formed by the first $n-1$ rows and $k-1$ columns. We continue in this way until no such rows are found. We are now reduced to the study of a $(n-s) \times (k-s)$ matrix $A'$, where $s$ is the number of rows found.

In the second stage, we scan $A'$ for rows having only 0 and $\pm 1$. In this case, simple arithmetic is needed to eliminate the $\pm 1$ as one does in ordinary HNF reduction, and, in particular, one may hope to work entirely with ordinary (as opposed to multi-precision) integers. The second stage ends when either all rows have been scanned, or if a coefficient exceeds half the maximal possible value for ordinary integers.

In a third and last stage before starting the modulo $d$ HNF reduction of step 10, we can proceed as follows (see [Buc-Dül]). We apply the ordinary HNF reduction Algorithm 2.4.5 keeping track of the size of the coefficients which are encountered. In this manner, we Hermite-reduce a few rows (corresponding to the index $j$ in Algorithm 2.4.5) until some coefficient becomes in absolute value larger than a given bound (for example as soon as a coefficient does not fit inside a single-precision number). If the first non-Hermite-reduced row has index $j$, we use the MLLL Algorithm 2.6.8 or an all-integer version on the matrix formed by the first $j$ rows. The effect of this will be to decrease the size of the coefficients, and since as in Hermite reduction only column operations are involved, the LLL reduction is allowed. We now start again Hermite-reducing a few rows using Algorithm 2.4.5, and we continue until either the matrix is completely reduced, or until the LLL reduction no longer improves matters (i.e. the partial Hermite reduction reduced no row at all).

After these reductions are performed, practical experience shows that the size of the matrix will have been considerably reduced, and this is essential since otherwise the HNF reduction would have to be performed on matrices having up to several thousand rows and columns, and this is almost impossible in practice.

(3) If Hermite reduction is performed carefully as described above, by far the most costly part of the algorithm is the search for relations. This part can be considerably improved by using the *large prime variation* idea common to many modern factoring methods (see Remark (2) in Section 10.1) as follows. In step 5, all $a$ with a prime factor greater than $P$ will be rejected. But assume that all prime factors of $a$ are less than or equal to $P$, except one prime factor $p_a$ which is larger. The corresponding quadratic form cannot be used directly without increasing the value of $P$. But assume that for two values of $a$, i.e. for two quadratic forms $f = (a, b, c)$ and $g = (a', b', c')$, the large prime $p_a$ is the same. Then either the form $fg^{-1}$ or the form $fg$ (depending on whether $b' \equiv b \pmod{p_a}$ or not) will give us a relation in which no primes larger than $P$ will occur, hence a useful relation. The coincidence of two values of $p_a$ will not be a rare phenomenon, and for

large discriminants the improvement will be considerable. See Exercise 14 for some hints on how to implement the large prime variation.

(4) Note that the '10' and '5' which occur in the algorithm are quite arbitrary, but are usually sufficient in practice. Note also that the correctness of the result is guaranteed only if one assumes GRH. Hence, this is a conditional algorithm, but in a much more precise sense than Algorithm 5.4.10.

(5) In step 5, we need to factor $a$ using trial division. Now $a$ can be as large as $\sqrt{|D|/3}$, hence $a$ may have more than 20 digits in the region we are aiming for, and factoring by trial division may seem too costly. We have seen however that $M$ is a few thousand at most in this region, so using trial divisors up to $M$ is reasonable. We can improve on this by using the early abort strategy which will be explained in Chapter 10.

(6) Step 9 requires computing a determinant using the Chinese remainder theorem (although as seen in Exercise 13 we can also compute it directly). This means that we first compute it modulo sufficiently many small primes. Then, by using the Chinese remainder Algorithm 1.3.12, we can obtain it modulo the product of these primes. Finally, Hadamard's inequality (Proposition 2.2.4) gives us an upper bound on the result. Hence, if the product of our primes is greater than twice this upper bound, we find the value of the determinant exactly. We have already mentioned this method in Section 4.3 for computing norms of algebraic integers.

The Hadamard bound may, however, be extremely large, and in that case it is preferable to proceed as follows. We take many more extra relations than needed (say 100 instead of 10) and we must assume that we will obtain the class number itself and not a multiple of it. Then the quantity $B\sqrt{2}$ is an upper bound for the determinant and can be used instead of the Hadamard bound. Once the class group is obtained, we must then check that it is correct, and this can be done without too much difficulty (or we can stop and assume that the result is correct).

(7) Finally, the main point of this method is, of course, its speed since under reasonable hypotheses one can prove that the expected asymptotic average running time is

$$O(L(|D|)^{\alpha})$$

with $\alpha = \sqrt{2}$, and perhaps even $\alpha = \sqrt{9/8}$. This is much faster than any of the preceding methods. Furthermore, it can be hoped that one can bring down the constant $\alpha$ to 1. This seems to be the limit of what one can expect to achieve on the subject for the following reason. Many fast factoring methods are known, using very different methods. To mention just a few, there is one using the 2-Sylow subgroup of the class group, one using elliptic curves (ECM), and a sieve type method (MPQS). All these methods have a common expected running time of the order of $O(L(N))$. In 1989, the discovery of the number field sieve lowered this running time to $O(e^{\ln(N)^{1/3+\epsilon}})$ (see Chapter 10), but this becomes better than the preceding methods for special numbers having more than 100 digits, and for general numbers having more than (perhaps) 130 digits,

hence does not concern us here. Since computing the class group is at least as difficult as factoring, one cannot expect to find a significantly faster method than McCurley's algorithm without fundamentally new ideas. It is plausible, however, that using ideas from the number field sieve would give an $O(e^{\ln(N)^{1/3+\epsilon}})$ algorithm, but nobody knows how to do this at the time of this writing. In practice, using Section 6.5, we may speedup Algorithm 5.5.2 by finding some of the relations using the basic number field sieve idea (see remark (3) after Algorithm 6.5.9).

### 5.5.3 Atkin's Variant

A variant of the above algorithm has been proposed by Atkin. It has the advantage of being faster, but the disadvantage of not always giving the class group. Atkin's idea is as follows.

Instead of taking $P_0$, which is already a small subset of the factor base of prime forms, to generate the relations, we choose a *single form* $f$. Of course, there is now no reason for $f$ to generate the class group, but at least when the discriminant is prime this often happens, as tables and the heuristics of [Coh-Len1] show (see Section 5.10).

We then determine the order of $f$ in the class group, using a method which is more efficient than the baby-step giant-step Algorithm 5.4.1 for large discriminants, since it is also a sub-exponential algorithm. The improvement comes, as in McCurley's algorithm, from the use of a factor base. (The philosophy being that any number-theoretic algorithm which can be made to efficiently use factor bases automatically becomes sub-exponential thanks to the theorem of Canfield-Erdős-Pomerance 10.2.1 that we will see in Chapter 10.)

To compute the order of $f$, we start with the same two steps as Algorithm 5.5.2. In particular, we set $n$ equal to the number of primeforms in our factor base.

We now compute the reduced forms equivalent to $f$, $f^2$, $f^3$, ... For each such form $(a, b, c)$ we execute step 5 of Algorithm 5.5.2, i.e. we check whether the form factors on our factor base, and if it does, we keep the corresponding relation.

We continue in this way until exactly $n+1$ relations have been obtained, i.e. one more than the cardinality of the factor base. Let $e_1$, $e_2$, ... , $e_{n+1}$ be the exponents of $f$ for which we have obtained a relation. Since we have now an $n \times (n+1)$ matrix with integral entries, there exists a non-trivial linear relation between the columns with integral coefficients, and this relation can be obtained by simple linear algebra, *not* by using number-theoretic methods such as Hermite normal form computations which are much slower. We can for example use a special case of Algorithm 2.3.1.

Now, if $C_i$ is column number $i$ of our matrix, for $1 \le i \le n+1$, and if $x_i$ are the coefficients of our relation, so that $\sum_{1 \le i \le n+1} x_i C_i = 0$, then clearly

$$f^N = 1 \ , \quad \text{where} \quad N = \sum_{1 \le i \le n+1} x_i e_i .$$

This is exactly the kind of relation that one obtains by using the baby-step giant-step method, but the running time can be shown to be sub-exponential as in McCurley's algorithm.

The relation may of course be trivial, i.e. we may have $N = 0$. This happens rarely however. Furthermore, if it does happen, we may have at our disposal more independent relations between the columns of our $n \times (n+1)$ matrix, which are also given by Algorithm 2.3.1. If not, we take higher powers of $f$ until we obtain a non-trivial relation.

As soon as we have a non-zero $N$ such that $f^N = 1$, we can compute the exact order of $f$ in the class group as in Algorithm 5.4.1, after having factored $N$. Of course, this factorization may not be easy, but $N$ is probably of similar size as the class number, hence about $\sqrt{|D|}$, so even if $D$ has 60 digits, we probably will have to factor a number having around 30 digits, which is not too difficult.

If $e$ is the exact order of $f$, we know that $e$ divides the class number. If $e$ already satisfies the lower bound inequalities given by the Euler product, that is if

$$e > \frac{1}{\sqrt{2}} \frac{\sqrt{|D|}}{\pi} \prod_{p \le P} \left(1 - \frac{\left(\frac{D}{p}\right)}{p}\right)^{-1} ,$$

then assuming GRH, we must have $e = h(D)$, and the class group is cyclic and generated by $f$. When it applies, this gives a faster method to compute the class number and class group than McCurley's algorithm. If the inequality is not satisfied, we can proceed with another form, as in Algorithm 5.4.1. The details are left to the reader.

Note that according to tables and the heuristic conjectures of [Coh-Len1] (see Section 5.10), the odd part of the class group should very often be cyclic (probability greater than 97%). Hence, if the discriminant $D$ is prime, so that the class number is odd, there is a very good chance that $Cl(D)$ is cyclic. Furthermore, the number of generators of a cyclic group with $h$ elements is $\phi(h)$, and this is also quite large, so there is a good chance that our randomly chosen $f$ will generate the class group.

The implementation details of Atkin's algorithm are left to the reader (see Exercise 15).

## 5.6 Class Groups of Real Quadratic Fields

We now consider the problem of computing the class group and the regulator of a real quadratic field $K = \mathbb{Q}(\sqrt{D})$, and more generally of the unique real quadratic order of discriminant $D$. We will consider the problem of computing the regulator in Section 5.7, so we assume that we already have computed the regulator which we will denote by $R(D)$.

### 5.6.1 Computing Class Numbers Using Reduced Forms

Thanks to Theorem 5.2.9, we still have a correspondence between the narrow ideal class group and equivalence classes of quadratic forms of the same discriminant $D$. It is not difficult to have a correspondence with the ideal class group itself.

**Proposition 5.6.1.** *If $D$ is a non-square positive integer congruent to 0 or 1 modulo 4, the maps $\psi_{FI}$ and $\psi_{IF}$ of Theorem 5.2.9 induce inverse isomorphisms between $Cl(D)$ and the quotient set of $\mathcal{F}(D)$ obtained by identifying the class of $(a,b,c)$ with the class of $(-a,b,-c)$.*

The proof is easy and left to the reader (Exercise 18).

The big difference between forms of negative and positive discriminant however is that, although one can define the notion of a reduced form (differently from the negative case), there will in general not exist only one reduced form per equivalence class, but several, which are naturally organized in a cycle structure.

**Definition 5.6.2.** *Let $f = (a,b,c)$ be a quadratic form with positive discriminant $D$. We say that $f$ is* reduced *if we have*

$$\left|\sqrt{D} - 2|a|\right| < b < \sqrt{D}.$$

The justification for this definition, as well as for the definition in the case of negative discriminants, is given in Exercise 16.

Note immediately the following proposition.

**Proposition 5.6.3.** *Let $(a,b,c)$ be a quadratic form with positive discriminant $D$. Then*

(1) *If $(a,b,c)$ is reduced, then $|a|$, $b$ and $|c|$ are less than $\sqrt{D}$ and $a$ and $c$ are of opposite signs.*
(2) *More precisely, if $(a,b,c)$ is reduced, we have $|a| + |c| < \sqrt{D}$.*
(3) *Finally, $(a,b,c)$ is reduced if and only if $\left|\sqrt{D} - 2|c|\right| < b < \sqrt{D}$.*

*Proof.* The result for $b$ is trivial, and since $ac = (b^2 - D)/4 < 0$ it is clear that $a$ and $c$ are of opposite signs. Now we have

$$|a| + |c| - \sqrt{D} = \frac{D - 4|a|\sqrt{D} + 4a^2 - b^2}{4|a|} = \frac{(\sqrt{D} - 2|a|)^2 - b^2}{4|a|},$$

hence by definition of reduced we have $|a| + |c| - \sqrt{D} < 0$, which implies (2) and hence (1).

To prove (3), we note that we have the identity

$$2|c| - \sqrt{D} = \frac{(\sqrt{D} - |a|)^2 - a^2 - b^2}{2|a|},$$

hence if $\epsilon = \pm 1$, we have

$$b - \epsilon(2|c| - \sqrt{D}) = \frac{(\sqrt{D} + \epsilon b)(b + \epsilon(2|a| - \sqrt{D}))}{2|a|}$$

which is positive by definition. Since $a$ and $c$ play symmetrical roles, this proves (3) and hence the proposition. □

If $\tau = (-b + \sqrt{D})/(2|a|)$ is the quadratic number associated to the form $(a, b, c)$ as in Section 5.2, it is not difficult to show that $(a, b, c)$ is reduced if and only if $0 < \tau < 1$ and $-\sigma(\tau) > 1$.

We now need a reduction algorithm on quadratic forms of positive discriminant. It is useful to give a preliminary definition:

**Definition 5.6.4.** *Let $D > 0$ be a discriminant. If $a \neq 0$ and $b$ are integers, we define $r(b, a)$ to be the unique integer $r$ such that $r \equiv b \pmod{2a}$ and $-|a| < r \leq |a|$ if $|a| > \sqrt{D}$, $\sqrt{D} - 2|a| < r < \sqrt{D}$ if $|a| < \sqrt{D}$. In addition, we define the* reduction operator *$\rho$ on quadratic forms $(a, b, c)$ of discriminant $D > 0$ by*

$$\rho(a, b, c) = \left(c, r(-b, c), \frac{r(-b, c)^2 - D}{4c}\right).$$

The reduction algorithm is then simply as follows.

**Algorithm 5.6.5** (Reduction of Indefinite Quadratic Forms). Given a quadratic form $f = (a, b, c)$ with positive discriminant $D$, this algorithm finds a reduced form equivalent to $f$.

1. [Iterate] If $(a, b, c)$ is reduced, output $(a, b, c)$ and terminate the algorithm. Otherwise, set $(a, b, c) \leftarrow \rho(a, b, c)$ and go to step 1.

We must show that this algorithm indeed produces a reduced form after a finite number of iterations. In fact, we have the following stronger result:

**Proposition 5.6.6.**

(1) *The number of iterations of $\rho$ which are necessary to reduce a form $(a,b,c)$ is at most $2+\lceil \lg(|c|/\sqrt{D})\rceil$.*
(2) *If $f=(a,b,c)$ is a reduced form, then $\rho(a,b,c)$ is again a reduced form.*
(3) *The reduced forms equivalent to $f$ are exactly the forms $\rho^n(f)$, for $n$ sufficiently large (i.e. $n$ greater than or equal to the least $n_0$ such that $\rho^{n_0}(f)$ is reduced) and are finite in number.*

*Proof.* The proof of (1) is similar in nature to that of Proposition 5.4.3. Set $\rho(f)=(a',b',c')$. I first claim that if $|c|>\sqrt{D}$ then $|c'|\le |c|/2$. Indeed, in that case $|r(-b,c)|\le |c|$, hence

$$|c'| = \frac{|r(-b,c)^2-D|}{4|c|} \le \frac{2c^2}{4|c|} \le \frac{|c|}{2}$$

since $D<c^2$. So, after at most $\lceil \lg(|c|/\sqrt{D})\rceil$ iterations, we will end up with a form where $|c|<\sqrt{D}$. As in the imaginary case one can then check that the form is almost reduced, in the sense that after another iteration of $\rho$ we will have $|a|$, $|b|$ and $|c|$ less than $\sqrt{D}$, and then either the form is reduced, or it will be after one extra iteration. The details are left as an exercise for the reader.

For (2), note that if $(a,b,c)$ is reduced, then

$$r(-b,c) = -b+2|c|\left\lfloor \frac{b+\sqrt{D}}{2|c|}\right\rfloor,$$

since this is clearly in the interval $[\sqrt{D}-2|c|,\sqrt{D}]$. If $|c|<\sqrt{D}/2$, this implies that $\rho(a,b,c)$ is reduced by definition. If $|c|>\sqrt{D}/2$, it is clear that

$$r(-b,c) = -b+2|c| > 2|c|-\sqrt{D} = |\sqrt{D}-2|c||,$$

proving again that $\rho(a,b,c)$ is reduced.

Finally, to prove (3), set $\sigma(a,b,c)=(c,b,a)$. Using again Proposition 5.6.3 (3), it is clear that $\sigma$ is an involution on reduced forms. Furthermore, one checks immediately that $\rho\sigma$ and $\sigma\rho$ are both involutions on the set of reduced forms, thus proving that $\rho$ is a permutation of this set, the inverse of $\rho$ being $\rho^{-1}=\sigma\rho\sigma$.

Another way to see this is to check directly that the inverse of $\rho$ on reduced forms is given explicitly by

$$\rho^{-1}(a,b,c) = \left(\frac{r(-b,a)^2-D}{4a}, r(-b,a), a\right),$$

and $\rho^{-1}$ can be used instead of $\rho$ to reduce a form, although one must take care that for non-reduced forms, it will *not* be the inverse of $\rho$ since $\rho$ is not one-to-one. □

We can summarize Proposition 5.6.6 by saying that if we start with any form $f$, the sequence $\rho^n(f)$ is ultimately periodic, and we arrive inside the period exactly when the form is reduced.

Finally, note that it follows from Proposition 5.6.3 that the set of reduced forms of discriminant $D$ has cardinality at most $D$ (the possible number of pairs $(a,b)$), but a closer analysis shows that its cardinality is $O(D^{1/2}\ln D)$.

It follows from the above discussion and results that in every equivalence class of quadratic forms of discriminant $D>0$, there is not only one reduced form, but a cycle of reduced forms (cycling under the operation $\rho$), and so the class number is the number of such *cycles*.

It is not necessary to formally write an algorithm analogous to Algorithm 5.3.5 for computing the class number using reduced forms. We make a list of all the reduced forms of discriminant $D$ by testing among all pairs $(a,b)$ such that $|a|<\sqrt{D}$, $|\sqrt{D}-2|a||<b<\sqrt{D}$ and $b\equiv D \pmod 2$, those for which $b^2-D$ is divisible by $4a$. Then we count the number of orbits under the permutation $\rho$, and the result is the narrow class number $h^+(D)$. If, in addition, we identify the forms $(a,b,c)$ and $(-a,b,-c)$, then, according to Proposition 5.6.1 we obtain the class number $h(D)$ itself.

As for Algorithm 5.3.5, this is an algorithm with $O(D)$ execution time, so is feasible only for discriminants up to $10^6$, say. Hence, as in the imaginary case, it is necessary to find better methods.

For future reference, let us determine the exact correspondence between the action of $\rho$ and the continued fraction expansion of a quadratic irrationality.

In Section 5.2 we have defined maps $\phi_{FI}$ and $\phi_{IQ}$, and by composition, Theorem 5.2.4 tells us that the map $\phi_{FQ}$ from $I$ to $Q\times\mathbb{Z}/2\mathbb{Z}$ defined by

$$\phi_{FQ}(a,b,c)=\left(\frac{-b+\sqrt{D}}{2|a|},\operatorname{sign}(a)\right)$$

is an isomorphism. (Note the absolute value of $a$, coming from the necessity of choosing an oriented basis for our ideals.)

From this, one checks immediately that if $f=(a,b,c)$ is reduced, and if $\phi_{FQ}(f)=(\tau,s)$, then

$$\phi_{FQ}(\rho(f))=\left(\frac{1}{\tau}-\left\lfloor\frac{1}{\tau}\right\rfloor,-s\right),$$

where by abuse of notation we still use the notation $\phi_{FQ}$ for the map at the level of forms and not at the level of classes of forms modulo $\Gamma_\infty$.

For $\rho^{-1}$ we define

$$\psi_{FQ}(a,b,c)=\left(\frac{b+\sqrt{D}}{2|a|},\operatorname{sign}(a)\right).$$

Then, if $f = (a, b, c)$ is reduced and $\psi_{FQ}(f) = (\tau', s)$, we have

$$\psi_{FQ}(\rho^{-1}(f)) = \left(\frac{1}{\tau' - \lfloor \tau' \rfloor}, -s\right).$$

Thus the action of $\rho$ and $\rho^{-1}$ on reduced forms correspond exactly to the continued fraction expansion of $\tau$ and $\tau' = -\sigma(\tau)$ respectively, with in addition a $\pm 1$ variable which gives the parity of the number of reduction steps.

In addition, since $\rho$ and $\rho^{-1}$ are inverse maps on reduced forms, we obtain as a corollary of Proposition 5.6.6 the following.

**Corollary 5.6.7.** *Let $\tau = (-b + \sqrt{D})/(2|a|)$ corresponding to a* reduced *quadratic form $(a, b, c)$. Then the continued fraction expansion of $\tau$ is purely periodic, and the period of the continued fraction expansion of $-\sigma(\tau) = (b + \sqrt{D})/(2|a|)$ is the reverse of that of $\tau$.*

### 5.6.2 Computing Class Numbers Using Analytic Formulas

We will follow closely Section 5.3.3. The definition of $L_D(s)$ is the same, but the functional equation is slightly different:

**Proposition 5.6.8.** *Let $D$ be a positive fundamental discriminant, and define*

$$L_D(s) = \sum_{n \geq 1} \left(\frac{D}{n}\right) n^{-s}.$$

*This series converges for* $\mathrm{Re}(s) > 1$, *and defines an analytic function which can be analytically continued to the whole complex plane to an entire function satisfying*

$$\Lambda_D(1 - s) = \Lambda_D(s),$$

*where we have set*

$$\Lambda_D(s) = \left(\frac{D}{\pi}\right)^{s/2} \Gamma\left(\frac{s}{2}\right) L_D(s).$$

Note that the special case $D = 1$ of this proposition (which is excluded since it is not the discriminant of a quadratic field) is still true if one adds the fact that the function has a simple pole at $s = 1$. In that case, we simply recover the usual functional equation of the Riemann zeta function. The link with the class number and the regulator is as follows. (Recall that the regulator $R(D)$ is in our case the logarithm of the unique generator greater than 1 of the torsion free part of the unit group.)

**Proposition 5.6.9.** *If $D$ is a positive fundamental discriminant, then*

$$L_D(1) = \frac{2h(D)R(D)}{\sqrt{D}}.$$

Note that as in the imaginary case, these results are special cases of Theorem 4.9.12 using the identity $\zeta_K(s) = \zeta(s)L_D(s)$ for $K = \mathbb{Q}(\sqrt{D})$.

Also, as in the imaginary case, it is not very reasonable to compute $L_D(1)$ directly from this formula since its defining series converges so slowly. However, a suitable reordering of the series gives the following:

**Corollary 5.6.10.** *If $D$ is a positive fundamental discriminant, then*

$$h(D)R(D) = -\sum_{r=1}^{\lfloor (D-1)/2 \rfloor} \left(\frac{D}{r}\right) \ln \sin\left(\frac{r\pi}{D}\right).$$

As usual, this kind of formula, although a finite sum, is useless from a computational point of view, and is worse than the method of reduced forms, although maybe slightly simpler to program. If we also use the functional equation we obtain a considerable improvement, leading to a complicated but much more efficient formula:

**Proposition 5.6.11.** *If $D$ is a positive fundamental discriminant, then*

$$2h(D)R(D) = \sum_{n \geq 1} \left(\frac{D}{n}\right)\left(\frac{\sqrt{D}}{n} \operatorname{erfc}\left(n\sqrt{\frac{\pi}{D}}\right) + E_1\left(\frac{\pi n^2}{D}\right)\right),$$

*where* $\operatorname{erfc}(x)$ *is the complementary error function (see Propositions 5.3.14 and 5.3.15), and* $E_1(x)$ *is the exponential integral function defined by*

$$E_1(x) = \int_x^\infty \frac{e^{-t}}{t}\, dt.$$

Note that the function $E_1(x)$ can be computed efficiently using the following formulas.

**Proposition 5.6.12.**

(1) *We have for all $x$*

$$E_1(x) = -\gamma - \ln(x) + \sum_{k \geq 1} (-1)^{k-1} \frac{x^k}{k!k},$$

*where $\gamma = 0.57721566490153286\ldots$ is Euler's constant, and this should be used when $x$ is small, say $x \leq 4$.*

(2) *We have for all* $x > 0$

$$E_1(x) = \frac{e^{-x}}{x}\left(1 - \cfrac{1}{2 + x - \cfrac{1 \cdot 2}{4 + x - \cfrac{2 \cdot 3}{6 + x - \ddots}}}\right),$$

*and this should be used for* $x$ *large, say* $x \geq 4$.

**Implementation Remark.** The remark made after Proposition 5.3.15 is also valid here, the general formula being here

$$2h(D)R(D) = \sum_{n \geq 1} \left(\frac{D}{n}\right)\left(\frac{AD}{n}\operatorname{erfc}\left(n\sqrt{\frac{\pi}{AD}}\right) + E_1\left(\frac{\pi n^2 A}{D}\right)\right).$$

These results show that the series given in Proposition 5.6.11 converges exponentially, and since $h(D)$ is an integer and $R(D)$ has been computed beforehand, it is clear that the computation time of $h(D)$ by this method is $O(D^{1/2+\epsilon})$ for any $\epsilon > 0$. As in the case $D < 0$ it would be easy to give an upper bound for the number of terms that one must take in the series. This is left as an exercise for the reader. See also Exercise 28 for a way to avoid computing the transcendental functions erfc and $E_1$.

### 5.6.3 A Heuristic Method of Shanks

An examination of the heuristic conjectures of [Coh-Len1] (see Section 5.10) shows that one must expect that, on average, the class number $h(D)$ will be quite small for positive discriminants, in contrast to the case of negative discriminants. Hence, one can use the following method, which is of course not an algorithm, but has a very good chance of giving the correct result quite quickly.

**Heuristic Algorithm 5.6.13** (Class Number for $D > 0$). Given a positive fundamental discriminant $D$, this algorithm computes a value which has a pretty good chance of being equal to the class number $h(D)$. As always, we assume that the regulator $R(D)$ has already been computed. We denote by $p_i$ the $i^{\text{th}}$ prime number.

1. [Regulator small?] If $R(D) < D^{1/4}$, then output a message saying that the algorithm will probably not work, and terminate the algorithm.
2. [Initialize] Set $h_1 \leftarrow \sqrt{D}/(2R(D))$, $h \leftarrow 0$, $c \leftarrow 0$, $k \leftarrow 0$.
3. [Compute block] Set

$$h_1 \leftarrow h_1 \prod_{500k<i\leq 500(k+1)} \left(1 - \frac{\left(\frac{D}{p_i}\right)}{p_i}\right)^{-1},$$

$m \leftarrow \lfloor h_1 \rceil$, $k \leftarrow k+1$.

4. [Seems integral?] If $|m - h_1| > 0.1$ set $c \leftarrow 0$ and go to step 3.
5. [Seems constant?] If $m \neq h$, set $h \leftarrow m$ and $c \leftarrow 1$ and go to step 3. Otherwise, set $c \leftarrow c+1$. If $c \leq 5$ go to step 3, otherwise output $h$ as the tentative class number and terminate the algorithm.

The reason for the frequent success of this algorithm is clear. Although we use the slowly convergent Euler product for $L_D(1)$, if the regulator is not too small, the integer $m$ computed in step 3 has a reasonable chance of being equal to the class number. The heuristic criterion that we use, due to Shanks, is that if the Euler product is less than 0.1 away from the same integer $h$ for 6 consecutive blocks of 500 prime numbers, we assume that $h$ is the class number. In fact, assuming GRH, this heuristic method can be made completely rigorous. I refer to [Mol-Wil] for details. In practice it works quite well, except of course for the quite rare cases in which the regulator is too small.

We still have not given any method for computing the structure of the class group. Before considering this point, we now consider the question of computing the regulator of a real quadratic field.

## 5.7 Computation of the Fundamental Unit and of the Regulator

As we have seen, reduced forms are grouped into $h(D)$ cycles under the permutation $\rho$. We will see that one can define a distance between forms which, in particular, has the property that the length of each cycle is the same, and equal to the regulator. Note that this is absolutely *not* true for the naïve length defined as the number of forms.

### 5.7.1 Description of the Algorithms

The action of $\rho$ and $\rho^{-1}$ corresponding to the continued fraction expansion of the quadratic irrationals $\tau$ and $-\sigma(\tau)$ respectively, it is clear that we must be able to compute the fundamental unit and the regulator from these expansions. From Corollary 5.6.7, we know that one of these expansions will be reverse of the other, so we can choose as we like between the two.

It is slightly simpler to use the expansion of $-\sigma(\tau)$, and this leads to the following algorithm whose validity will be proved in the next section. Note that in this algorithm we assume $a > 0$, but it is easy to modify it so that it stays valid in general (Exercise 20).

**Algorithm 5.7.1** (Fundamental Unit Using Continued Fractions). Given a quadratic irrational $\tau = (-b + \sqrt{D})/(2a)$ where $4a \mid (D - b^2)$ and $a > 0$, corresponding to a *reduced* form $(a, b, (b^2 - D)/(4a))$, this algorithm computes the fundamental unit $\varepsilon$ of $\mathbb{Q}(\sqrt{D})$ using the ordinary continued fraction expansion of $-\sigma(\tau)$.

1. [Initialize] Set $u_1 \leftarrow -b$, $u_2 \leftarrow 2a$, $v_1 \leftarrow 1$, $v_2 \leftarrow 0$, $p \leftarrow b$ and $q \leftarrow 2a$. Precompute $d \leftarrow \lfloor \sqrt{D} \rfloor$.

2. [Euclidean step] Set $A \leftarrow \lfloor (p+d)/q \rfloor$, then in that order, set $p \leftarrow Aq - p$ and $q \leftarrow (D - p^2)/q$. Finally, set $t \leftarrow Au_2 + u_1$, $u_1 \leftarrow u_2$, $u_2 \leftarrow t$, $t \leftarrow Av_2 + v_1$, $v_1 \leftarrow v_2$, and $v_2 \leftarrow t$.

3. [End of period?] If $q = 2a$ and $p \equiv b \pmod{2a}$, set $u \leftarrow |u_2/a|$, $v \leftarrow |v_2/a|$ (both divisions being exact), output $\varepsilon \leftarrow (u + v\sqrt{D})/2$, and terminate the algorithm. Otherwise, go to step 2.

As will be proved in the next section, the result of this algorithm is the fundamental unit, independently of the initial reduced form. Hence, the simplest solution is to start with the unit reduced form, i.e. with $\tau = (-b+\sqrt{D})/2$ and $b = d$ if $d \equiv D \pmod 2$, $b = d - 1$ otherwise, where as in the algorithm $d = \lfloor \sqrt{D} \rfloor$.

Also, note that the form corresponding to $(p + \sqrt{D})/q$ at step $i$ is

$$((-1)^i q/2, p, (-1)^i (p^2 - D)/(2q)).$$

If we had wanted the exact action of $\rho^{-1}$, we would have to put $q \leftarrow (p^2 - D)/q$ instead of $q \leftarrow (D - p^2)/q$ in step 2 of the algorithm, and then $q$ would alternate in sign instead of always being positive.

Now the continued fraction expansion of the quadratic irrational corresponding to the unit reduced form is not only periodic, but in fact symmetric. This is true more generally for forms belonging to *ambiguous* cycles, i.e. forms whose square lie in the principal cycle (see Exercise 22). Hence, it is possible to divide by two the number of iterations in Algorithm 5.7.1. This leads to the following algorithm, whose proof is left to the reader.

**Algorithm 5.7.2** (Fundamental Unit). Given a fundamental discriminant $D > 0$, this algorithm computes the fundamental unit of $\mathbb{Q}(\sqrt{D})$.

1. [Initialize] Set $d \leftarrow \lfloor \sqrt{D} \rfloor$. If $d \equiv D \pmod 2$, set $b \leftarrow d$ otherwise set $b \leftarrow d - 1$. Then set $u_1 \leftarrow -b$, $u_2 \leftarrow 2$, $v_1 \leftarrow 1$, $v_2 \leftarrow 0$, $p \leftarrow b$ and $q \leftarrow 2$.

2. [Euclidean step] Set $A \leftarrow \lfloor (p + d)/q \rfloor$, $t \leftarrow p$ and $p \leftarrow Aq - p$. If $t = p$ and $v_2 \neq 0$, then go to step 4, otherwise set $t \leftarrow Au_2 + u_1$, $u_1 \leftarrow u_2$, $u_2 \leftarrow t$, $t \leftarrow Av_2 + v_1$, $v_1 \leftarrow v_2$, and $v_2 \leftarrow t$, $t \leftarrow q$, $q \leftarrow (D - p^2)/q$.

3. [Odd period?] If $q = t$ and $v_2 \neq 0$, set $u \leftarrow |(u_1 u_2 + D v_1 v_2)/q|$, $v \leftarrow |(u_1 v_2 + u_2 v_1)/q|$ (both divisions being exact), output $\varepsilon \leftarrow (u + v\sqrt{D})/2$ and terminate the algorithm. Otherwise, go to step 2.

4. [Even period] Set $u \leftarrow |(u_2^2 + v_2^2 D)/q|$, $v \leftarrow |2u_2v_2/q|$ (both divisions being exact), output $\varepsilon \leftarrow (u + v\sqrt{D})/2$ and terminate the algorithm.

The performance of both these algorithms is quite reasonable for discriminants up to $10^6$. It can be proved that the number of steps is $O(D^{1/2+\epsilon})$ for all $\epsilon > 0$. Furthermore, all the computations on $p$ and $q$ are done with numbers less than $2\sqrt{D}$, hence of reasonable size. The main problem is that the fundamental unit itself has coefficients $u$ and $v$ which are of unreasonable size. One can show that $\ln u$ and $\ln v$ can be as large as $\sqrt{D}$. Hence, although the number of steps is $O(D^{1/2+\epsilon})$, this does not correctly reflect the practical execution time, since multi-precision operations become predominant. In fact, it is easy to see that the only bound one can give for the execution time itself is $O(D^{1+\epsilon})$.

The problem is therefore not so much in computing the numbers $u$ and $v$, which do not make much sense when they are so large, but in computing the regulator itself to some reasonable accuracy, since after all, this is all we need in the class number formula. It would seem that it is not possible to compute $R(D)$ without computing $\varepsilon$ exactly, but luckily this is not the case, and there is a variant of Algorithm 5.7.2 (or 5.7.1) which gives the regulator instead of the fundamental unit. This variant uses floating point numbers, which must be computed to sufficient accuracy (but not unreasonably so: double precision, i.e. 15 decimals, is plenty). The advantage is that no numbers will become large.

### 5.7.2 Analysis of the Continued Fraction Algorithm

To do this, we must analyze the behavior of the continued fraction algorithm, and along the way we will prove the validity of Algorithm 5.7.1. We assume for the sake of simplicity that $a > 0$ (hence $c < 0$), although the same analysis holds in general.

Call $p_i$, $q_i$, $A_i$, $u_{1,i}$, $u_{2,i}$, $v_{1,i}$, $v_{2,i}$ the quantities occurring in step $i$ of the algorithm, where the initializations correspond to step 0, and set for $i \geq -1$, $a_i = u_{1,i+1}$, $b_i = v_{1,i+1}$. Then we can summarize the recursion implicit in the algorithm by the following formulas:

For all $i \geq 0$, $u_{1,i} = a_{i-1}$, $u_{2,i} = a_i$, $v_{1,i} = b_{i-1}$, $v_{2,i} = b_i$. Furthermore:

$p_0 = b$, $q_0 = 2a$, $a_{-1} = -b$, $a_0 = 2a$, $b_{-1} = 1$, $b_0 = 0$ (recall that $a = 1$ in Algorithm 5.7.2), and for $i \geq 0$:

$A_i = \lfloor (p_i + d)/q_i \rfloor$, $p_{i+1} = A_i q_i - p_i$, $q_{i+1} = (D - p_{i+1}^2)/q_i$, $a_{i+1} = A_i a_i + a_{i-1}$, $b_{i+1} = A_i b_i + b_{i-1}$.

By the choice of $b$, we know that $q_0 \mid D - p_0^2$, and if by induction we assume that all the above quantities are integers and that $q_i \mid D - p_i^2$, one sees that $D - p_{i+1}^2 \equiv D - p_i^2 \equiv 0 \pmod{q_i}$, hence $q_{i+1}$ is an integer. In addition, we clearly have $q_{i+1} \mid D - p_{i+1}^2$ since the quotient is simply $q_i$, thus proving our claim by induction. We also have $q_{i+1} - q_{i-1} = (D - p_{i+1}^2)/q_i - (D - p_i^2)/q_i = (p_i - p_{i+1})(p_i + p_{i+1})/q_i$, hence we obtain the formula

$$q_{i+1} = q_{i-1} - A_i(p_{i+1} - p_i),$$

which is in general computationally simpler than the formula used in the algorithms.

That the algorithms above correspond to the continued fraction expansion of $(b+\sqrt{D})/(2a)$ (where in Algorithm 5.7.2 it is understood that we take $a = 1$) is quite clear. Set $\zeta_i = (p_i + \sqrt{D})/q_i$. Then we have $\zeta_0 = (b + \sqrt{D})/(2a)$, $A_i = \lfloor \zeta_i \rfloor$, and hence

$$\frac{1}{\zeta_i - \lfloor \zeta_i \rfloor} = \frac{q_i}{p_i - A_i q_i + \sqrt{D}} = \frac{A_i q_i - p_i + \sqrt{D}}{(D - (A_i q_i - p_i)^2)/q_i} = \zeta_{i+1},$$

thus giving the above formulas.

This is of course nothing other than the translation of the formula giving $\psi_{FQ}(\rho^{-1}(f))$ in terms of $\psi_{FQ}(f)$.

Note that in practice the computations on the pair $(p, q)$ should be done in the following way: use three extra variables $r$ and $p_1$, $q_1$. Replace steps 1 and 2 of Algorithm 5.7.2 by

1'. [Initialize] Set $d \leftarrow \lfloor \sqrt{D} \rfloor$. If $d \equiv D \pmod 2$, set $b \leftarrow d$ otherwise set $b \leftarrow d - 1$. Then set $u_1 \leftarrow -b$, $u_2 \leftarrow 2$, $v_1 \leftarrow 1$, $v_2 \leftarrow 0$, $p \leftarrow b$ and $q \leftarrow 2$, $q_1 \leftarrow (D - p^2)/q$.

2'. [Euclidean step] Let $p + d = qA + r$ with $0 \leq r < q$ be the Euclidean division of $p + d$ by $q$, and set $p_1 \leftarrow p$, $p \leftarrow d - r$. If $p_1 = p$ and $v_2 \neq 0$, then go to step 4, otherwise set $t \leftarrow Au_2 + u_1$, $u_1 \leftarrow u_2$, $u_2 \leftarrow t$, $t \leftarrow Av_2 + v_1$, $v_1 \leftarrow v_2$, and $v_2 \leftarrow t$, $t \leftarrow q$, $q \leftarrow q_1 - A(p - p_1)$, $q_1 \leftarrow t$.

This has the same effect as steps 1 and 2 of Algorithm 5.7.2, but avoids one division in each loop. Note that this method can also be used in general.

Now that we have seen that we are computing the continued fraction expansion of $(b + \sqrt{D})/(2a)$, we must study the behavior of the sequences $a_i$ and $b_i$. This is summarized in the following proposition.

**Proposition 5.7.3.** *With the above notations, we have*

(1)
$$\frac{a_{i+1} + b_{i+1}\sqrt{D}}{a_i + b_i\sqrt{D}} = \frac{p_{i+1} + \sqrt{D}}{q_i},$$

(2)
$$a_i b_{i-1} - a_{i-1} b_i = (-1)^i 2a,$$

(3)
$$a_i^2 - b_i^2 D = (-1)^i 2a q_i,$$

(4)
$$a_i a_{i-1} - b_i b_{i-1} D = (-1)^{i-1} 2a p_i,$$

(5)
$$\sqrt{D} = \frac{a_i\zeta_i + a_{i-1}}{b_i\zeta_i + b_{i-1}},$$

*where as before* $\zeta_i = (p_i + \sqrt{D})/q_i$.

*Proof.* Denote real conjugation $\sqrt{D} \mapsto -\sqrt{D}$ in the field $\mathbb{Q}(\sqrt{D})$ by $\sigma$, and set $\rho_i = (p_i + \sqrt{D})/q_{i-1}$. Then $\rho_{i+1} = A_i - \sigma(\zeta_i)$ and since $\zeta_{i+1} = 1/(\zeta_i - A_i)$ we have by applying $\sigma$,

$$\sigma(\zeta_{i+1}) = 1/(\sigma(\zeta_i) - A_i) = -1/\rho_{i+1}.$$

Therefore $\rho_{i+1} = A_i - \sigma(\zeta_i) = A_i + 1/\rho_i$. On the other hand, to be compatible with the recursions, we must define $q_{-1} = (D - b^2)/(2a)$. Thus we see that $\rho_0 = 2a/(\sqrt{D} - b)$ (which comes also from the formula $\rho_i = -1/\sigma(\zeta_i)$). If we set $\alpha_i = a_i + b_i\sqrt{D}$, the recursions show that $\alpha_{i+1} = A_i\alpha_i + \alpha_{i-1}$. Therefore if we set $\beta_i = \alpha_i/\alpha_{i-1}$, we have $\beta_{i+1} = A_i + 1/\beta_i$, and this is the same recursion satisfied by $\rho_i$. Since we have $\beta_0 = 2a/(\sqrt{D} - b) = \rho_0$, this shows that $\beta_i = \rho_i$ for all $i$, thus showing (1).

Formula (2) is a standard formula in continued fraction expansions: we have the matrix recursion

$$\begin{pmatrix} a_{i+1} & b_{i+1} \\ a_i & b_i \end{pmatrix} = \begin{pmatrix} A_i & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a_i & b_i \\ a_{i-1} & b_{i-1} \end{pmatrix},$$

hence formula (2) follows trivially on taking determinants and noticing that $a_0b_{-1} - a_{-1}b_0 = 2a$.

To prove (3), we take the norm (with respect to $\mathbb{Q}(\sqrt{D})/\mathbb{Q}$) of formula (1). We obtain:

$$\frac{a_{i+1}^2 - b_{i+1}^2 D}{a_i^2 - b_i^2 D} = \frac{p_{i+1}^2 - D}{q_i^2} = -\frac{q_{i+1}}{q_i},$$

hence by multiplying out we obtain

$$a_i^2 - b_i^2 D = (-1)^i q_i \frac{a_0^2 - b_0^2 D}{q_0} = (-1)^i 2aq_i,$$

showing (3).

Finally, to prove (4) we take the trace (with respect to $\mathbb{Q}(\sqrt{D})/\mathbb{Q}$) of formula (1). We obtain:

$$\frac{a_{i+1} + b_{i+1}\sqrt{D}}{a_i + b_i\sqrt{D}} + \frac{a_{i+1} - b_{i+1}\sqrt{D}}{a_i - b_i\sqrt{D}} = \frac{2p_{i+1}}{q_i},$$

hence grouping and using (3) we get

$$\frac{2a_{i+1}a_i - 2b_{i+1}b_iD}{(-1)^i 2aq_i} = \frac{2p_{i+1}}{q_i},$$

and this proves (4).

Formula (5) follows easily from (1) and its proof is left to the reader. □

**Corollary 5.7.4.** *Set $c = (b^2 - D)/(4a)$, so $D = b^2 - 4ac$. Define two sequences $c_i$ and $d_i$ by $c_{-1} = 0$, $c_0 = 1$, $c_{i+1} = A_i c_i + c_{i-1}$, and $d_{-1} = -2c$, $d_0 = b$ and $d_{i+1} = A_i d_i + d_{i-1}$. Then the five formulas of Proposition 5.7.3 hold with $(a, a_i, b_i)$ replaced by $(c, d_i, c_i)$.*

The proof is easy and left to the reader.

Now for simplicity let us consider the case of Algorithm 5.7.1. Let $i = k$ be the stage at which we stop, i.e. for which $q_k = 2a$ and $p_k \equiv b \pmod{2a}$. Then we output $\varepsilon = (|a_k| + |b_k|\sqrt{D})/|2a|$. We are going to show that this is indeed the correct result. First, I claim that $\varepsilon$ is a unit. Indeed, notice that using (3), the norm of $\varepsilon$ is equal to $(-1)^k$. Hence, to show that $\varepsilon$ is a unit, it is only necessary to show that it is an algebraic integer. Moreover, since its norm is equal to $\pm 1$, hence integral, we must only show that the trace of $\varepsilon$ is integral, i.e. that $a_k \equiv 0 \pmod{a}$.

For this, we use the sequence $c_i$ defined in Corollary 5.7.4. It is clear that we have $a_i = 2ac_i - bb_i$. From Proposition 5.7.3 (3) an easy computation gives

$$b_k(cb_k - bc_k) = a((-1)^k \frac{q_k}{2a} - c_k^2) \equiv 0 \pmod{a},$$

since $q_k = 2a$. Similarly, since $p_k \equiv b \pmod{2a}$, from (4) a similar computation gives

$$b_k(cb_{k-1} - bc_{k-1}) = a\left((-1)^{k-1}\frac{p_k - b}{2a} - c_k c_{k-1}\right) \equiv 0 \pmod{a}.$$

If we set $\delta_i = cb_i - bc_i$, it is clear by induction that

$$\gcd(\delta_k, \delta_{k-1}) = \gcd(\delta_{k-1}, \delta_{k-2}) = \cdots = \gcd(\delta_0, \delta_{-1}) = \gcd(b, c).$$

From the two congruences proved above and the existence of $u$ and $v$ such that $u\delta_k + v\delta_{k-1} = \gcd(b, c)$, it follows that

$$b_k \gcd(b, c) \equiv 0 \pmod{a}.$$

But since $D$ is a fundamental discriminant, the quadratic form $(a, b, c)$ is primitive, hence $\gcd(a, b, c) = 1 = \gcd(\gcd(b, c), a)$, so we obtain $b_k \equiv 0 \pmod{a}$, hence also $a_k = 2ac_k - bb_k \equiv 0 \pmod{a}$ as was to be shown.

Now that we know that $\varepsilon$ is a unit, we will show it is the fundamental unit. Since clearly $\varepsilon > 1$, this will follow from the following more general result. We say that an algebraic integer $\alpha$ is *primitive* if for any integer $n$, $\alpha/n$ is an algebraic integer only for $n = \pm 1$. Then we have:

**Proposition 5.7.5.** *Let us keep all the above notations. Let $N \geq 1$ be a squarefree integer such that $\gcd(a, N) = 1$. Assume that $2|a|N < \sqrt{D}$.*

*Then the solutions $(A, B)$ of the Diophantine equation*

$$A^2 - B^2 D = \pm 4N, \quad \text{with } A > 0,\ B > 0 \text{ and } \frac{A + B\sqrt{D}}{2} \text{ primitive}$$

*are given by $(A, B) = (|a_n/a|, |b_n/a|)$, for every $n$ such that $q_n = 2|a|N$ and $p_n \equiv b \pmod{2a}$.*

*Proof.* We have proven above that $\varepsilon$ was an algebraic *integer* using only $q_k \equiv 0 \pmod{2a}$ and not precisely the value $q_k = 2a$. This shows that if the conditions of Proposition 5.7.5 are satisfied, we will have $a \mid a_n$ and $a \mid b_n$, and since by Proposition 5.7.3 (3) we have $a_n^2 - b_n^2 D = \pm 2aq_n = \pm 4a^2 N$, the pair $(A, B) = (|a_n/a|, |b_n/a|)$ is indeed a solution to our Diophantine equation with $A > 0$, $B > 0$, and since $\mathcal{N}((a_n + b_n\sqrt{D})/(2a)) = \pm N$ and that $N$ is squarefree, $(A + B\sqrt{D})/2$ is primitive.

We must now show the converse. Assume that $A^2 - B^2 D = 4sN$ with $s = \pm 1$. Let $\tau' = -\sigma(\tau) = (b + \sqrt{D})/(2a)$ as in Algorithm 5.7.1. Then an easy calculation gives

$$\left| \tau' - \frac{A + bB}{2aB} \right| = \left| \frac{4N}{2aB^2 |\sqrt{D} + A/B|} \right|.$$

Now, $A/B = \sqrt{D \pm 4N/B^2} \geq \sqrt{D - 4N/B^2}$, hence

$$\left| \tau' - \frac{(A + bB)/2}{aB} \right| \leq \frac{4N}{2|a|B^2(\sqrt{D} + \sqrt{D - 4N/B^2})}.$$

We also have the following lemma whose proof is left to the reader. (See [H-W] for a slightly weaker version, but the proof is the same, see Exercise 21.)

**Lemma 5.7.6.** *If $p$ and $q$ are integers such that*

$$\left| \tau' - \frac{p}{q} \right| \leq \frac{1}{q(\max(2q - 1, 2))}$$

*then $p/q$ is a convergent in the continued fraction expansion of $\tau'$.*

Consider first the case $|a| > 1$. One easily checks that $4a^2 N^2 - 4N/B^2 > (2|a|N - 2N/B)^2$ is equivalent to $2|a|BN > N + 1$ which is clearly true. Hence, since $\sqrt{D} > 2|a|N$, we have $\sqrt{D} + \sqrt{D - 4N/B^2} > 4|a|N - 2N/B$ and therefore

$$\left| \tau' - \frac{(A + bB)/2}{aB} \right| < \frac{1}{|a|B(2|a|B - 1)}.$$

Since $b \equiv D \pmod 2$ and $A \equiv BD \pmod 2$, $(A + bB)/2$ is an integer, and so we can apply the lemma. This shows that $\frac{(A+bB)/2}{aB}$ is a convergent to $\tau'$. A similar proof applies to the case $|a| = 1$, except when $B = 1$. But in

the case $|a| = B = 1$, we have $D - 2\sqrt{D} < A^2 < D + 2\sqrt{D}$ hence either $\sqrt{D}-1 < A < \sqrt{D}+1$, and hence $|\tau' - (A+b)/2| < 1/2$ and we can conclude as before that $(A+b)/2$ is a convergent, or else $D-2\sqrt{D} < A^2 < D-2\sqrt{D}+1$ which implies that $\tau' - 1 < (b+A)/2 < \tau'$ hence $(b+A)/2 = \lfloor \tau' \rfloor$ is also a convergent to $\tau'$.

By definition, the convergents to $\tau'$ are $c_n/b_n$, and the equation $(A + bB)/(2aB) = c_n/b_n$ is equivalent to $A/B = a_n/b_n$.

Now we have the following lemma:

**Lemma 5.7.7.** *We have for all $i$,*

$$\left(\frac{p_i - b}{2}, \frac{q_i}{2}, a\right) = (b_i, a),$$

*and*

$$\frac{a_i + b_i\sqrt{D}}{2(b_i, a)}$$

*is a primitive algebraic integer.*

*Proof.* We know that

$$b_i(cb_i - bc_i) = (-1)^i \frac{q_i}{2} - ac_i^2 \quad \text{and} \quad b_i(cb_{i-1} - bc_{i-1}) = (-1)^{i-1}\frac{p_i - b}{2} - ac_ic_{i-1}$$

hence as above $b_i(b, c) \equiv 0 \pmod{((p_i - b)/2, q_i/2, a)}$, and since $(a, b, c) = 1$, we obtain $((p_i - b)/2, q_i/2, a) \mid (b_i, a)$. Conversely, the same relations show immediately that $(b_i, a) \mid ((p_i - b)/2, q_i/2, a)$, thus giving the first formula of the lemma. For the second, we note that $a_i = 2ac_i - bb_i$, hence $(b_i, a) \mid a_i$, and since by Proposition 5.7.3 (3) $a_i^2 - b_i^2 D = (-1)^i 2aq_i$, we see that $4(b_i, a)^2 \mid a_i^2 - b_i^2 D$ since we have proved that $(b_i, a) \mid q_i/2$, and these two divisibility conditions show that $\alpha = (a_i + b_i\sqrt{D})/(2(b_i, a))$ is an algebraic integer.

Let us show that it is primitive. Note first that since $a_i = 2ac_i - bb_i$ and $(c_i, b_i) = 1$, we have $(a_i, b_i) = (b_i, 2a)$. This shows that if we write $\alpha = (A' + B'\sqrt{D})/2$, we have $(A', B') = (b_i, 2a)/(b_i, a)$ and therefore $(A', B') \mid 2$. If $D \equiv 1 \pmod 4$, it can easily be seen that this is the only required condition for primitivity. If $D \equiv 0 \pmod 4$, we must show that $A'$ is even and that $(A'/2, B') = 1$. In this case however, $b \equiv D \equiv 0 \pmod 2$, hence $a_i/2 = ac_i - (b/2)b_i$ showing that $A' = a_i/(b_i, a)$ is even, and $(a_i/2, b_i) = (a, b_i)$ so $(A'/2, B') = 1$ as was to be shown. □

Now that we have this lemma, we can finish the proof of Proposition 5.7.5. We have shown that $A/B = a_n/b_n$, and since $(A + B\sqrt{D})/2$ was assumed primitive, we obtain from the lemma the equalities $A = |a_n|/(b_n, a)$, $B = |b_n|/(b_n, a)$. Plugging this in the Diophantine equation gives, using Proposition 5.7.3 (3), $\pm 4N = 2aq_n/(b_n, a)^2$ or in other words since it is clear by induction that $aq_i > 0$ for all $i$:

$$N = \frac{a}{(b_n, a)} \frac{q_n/2}{(b_n, a)}.$$

Since we have assumed $(N, a) = 1$, it follows that $a/(b_n, a) = \pm 1$, so that $a \mid b_n$, hence also $a \mid a_n$, and hence $q_n = 2aN$, thus finishing the proof of Proposition 5.7.5. □

Although we have proved a lot, we are still not finished. We need to show that we do indeed obtain the fundamental unit and not a power of it for every reduced $(a, b, c)$, and not simply for $2|a| < \sqrt{D}$. To do this, it would be necessary to relax the condition $2|a|N < \sqrt{D}$ to $|a|N < \sqrt{D}$ for instance, but this is false as can easily be seen (take for example $D = 136$, $(a, b, c) = (5, 6, -5)$ and $N = 2$. This is only a random example). In the special case $N = 1$ however, which is the case we are most interested in, we can prove our claim by using the symmetry between $a$ and $c$, i.e. by also using Corollary 5.7.4. First, we note the proposition which is symmetric to Proposition 5.7.5.

**Proposition 5.7.8.** *Let us keep all the above notations, and, in particular, those of Corollary 5.7.4. Let $N \geq 1$ be a squarefree integer such that $\gcd(c, N) = 1$ and $2|c|N < \sqrt{D}$.*

*Then the solutions $(A, B)$ of the Diophantine equation*

$$A^2 - B^2 D = \pm 4N, \quad \text{with } A > 0,\ B > 0 \text{ and } \frac{A + B\sqrt{D}}{2} \text{ primitive}$$

*are given by $(A, B) = (|d_n/c|, |c_n/c|)$, for every $n$ such that $q_n = 2|c|N$ and $p_n \equiv -b \pmod{2c}$.*

The proof is identical to that of Proposition 5.7.5, but uses the formulas of Corollary 5.7.4 instead of those of Proposition 5.7.3. □

Now we can prove:

**Proposition 5.7.9.** *The conclusion of Proposition 5.7.5 is valid for $N = 1$, with the only needed condition being that $(a, b, c)$ is a reduced quadratic form.*

*Proof.* If $|a| < \sqrt{D}/2$, then the result follows from Proposition 5.7.5. Assume now $|a| > \sqrt{D}/2$. By Proposition 5.6.3 (2), we have $|c| < \sqrt{D}/2$, hence we can apply Proposition 5.7.8. We obtain $(A, B) = (|d_n/c|, |c_n/c|)$ for an $n$ such that $p_n \equiv -b \pmod{2c}$ and $q_n = 2|c|$. This implies that $p_{n+1} = A_n q_n - p_n \equiv b \pmod{2c}$ and furthermore, by definition of $A_n$, that $\sqrt{D} - 2|c| < p_{n+1} < \sqrt{D}$. Hence, since $|c| < \sqrt{D}/2$ and $(a, b, c)$ is reduced, we have $p_{n+1} = b$, so $q_{n+1} = 2|a|$. Now from Proposition 5.7.3 and Corollary 5.7.4, we obtain immediately that

$$\frac{d_{n+1} + c_{n+1}\sqrt{D}}{a_{n+1} + b_{n+1}\sqrt{D}} = \frac{d_n + c_n\sqrt{D}}{a_n + b_n\sqrt{D}},$$

hence by induction

$$\frac{d_n + c_n\sqrt{D}}{a_n + b_n\sqrt{D}} = \frac{b + \sqrt{D}}{2a},$$

and from this, Proposition 5.7.3 (1), Lemma 5.7.7 and its analog for $c$ instead of $a$, we obtain the identities $|a_{n+1}/a| = |d_n/c|$ and $|b_{n+1}/a| = |c_n/c|$, proving the proposition. □

### 5.7.3 Computation of the Regulator

We have already mentioned that the fundamental unit $\varepsilon$ itself can involve huge coefficients, and that what one usually needs is only the regulator to a reasonable degree of accuracy. Note first that for all $i \geq 1$, we have $a_i/b_i > 0$. This is an amusing exercise left to the reader (hint: consider separately the four cases $a > 0$ and $a < 0$, and $2|a| < \sqrt{D}$, $2|a| > \sqrt{D}$). Hence we have

$$R(D) = \ln\varepsilon = \ln\left(\frac{|a_k + b_k\sqrt{D}|}{|2a|}\right) = \sum_{i=0}^{k-1} \ln\left(\frac{|a_{i+1} + b_{i+1}\sqrt{D}|}{|a_i + b_i\sqrt{D}|}\right),$$

so by Proposition 5.7.3,

$$R(D) = \sum_{i=0}^{k-1} \ln\left(\frac{p_{i+1} + \sqrt{D}}{|q_i|}\right) = \sum_{i=1}^{k} \ln\left(\frac{p_i + \sqrt{D}}{|q_i|}\right),$$

since $q_k = q_0 = 2a$, and since the $p_i$ and $|q_i|$ are always small (less than $2\sqrt{D}$), this enables us to compute the regulator to any given accuracy without handling huge numbers. The computation of a logarithm is a time consuming operation however, and hence it is preferable to write

$$R(D) = \ln\left(\prod_{i=1}^{k} \frac{p_i + \sqrt{D}}{|q_i|}\right),$$

the product being computed to a given numerical accuracy. In most cases, this method will again not work, because the *exponents* of the floating point numbers become too large. The trick is to keep the exponent in a separate variable which is updated either at each multiplication, or as soon as there is the risk of having an exponent overflow in the multiplication. Note that we have the trivial inequality $(p_i + \sqrt{D})/|q_i| < \sqrt{D}$, hence exponent overflow can easily be checked. This leads to the following algorithm, analogous to Algorithm 5.7.1.

**Algorithm 5.7.10** (Regulator). Given a quadratic irrational $\tau = \dfrac{-b + \sqrt{D}}{2a}$ where $4a \mid (D - b^2)$ and $a > 0$, corresponding to a reduced form $(a, b, (b^2 -$

$D)/(4a)$), this algorithm computes the regulator $R(D)$ of $\mathbb{Q}(\sqrt{D})$ using the ordinary continued fraction expansion of $-\sigma(\tau)$.

1. [Initialize] Precompute $f \leftarrow \sqrt{D}$ to the desired accuracy, and set $d \leftarrow \lfloor f \rfloor$, $e \leftarrow 0$, $R \leftarrow 1$, $p \leftarrow b$, $q \leftarrow 2a$, and $q_1 \leftarrow (D - p^2)/q$. Finally, let $2^L$ be the highest power of 2 such that $2^L f$ does not give an exponent overflow.
2. [Euclidean step] Let $p + d = qA + r$ with $0 \leq r < |q|$ be the Euclidean division of $p + d$ by $q$, and set $p_1 \leftarrow p$, $p \leftarrow d - r$, $t \leftarrow q$, $q \leftarrow q_1 - A(p - p_1)$, $q_1 \leftarrow t$ and $R \leftarrow R(p + f)/q$. If $R \geq 2^L$, set $R \leftarrow R/2^L$, $e \leftarrow e + 1$.
3. [End of period?] If $q = 2a$ and $p \equiv b \pmod{2a}$, output $R(D) \leftarrow \ln R + eL \ln 2$ and terminate the algorithm. Otherwise, go to step 2.

In the case where we start with the unit form, we can use the symmetry of the period to obtain an algorithm similar to Algorithm 5.7.2. We leave this as an exercise for the reader (Exercise 23). We can also modify the algorithm so that it works for reduced forms with $a < 0$.

The running time of this algorithm is $O(D^{1/2+\epsilon})$ for all $\epsilon > 0$, but here this corresponds to the actual behavior since no multi-precision variables are being used. Although this is reasonable, we will now see that we can adapt Shanks's baby-step giant-step method to obtain a $O(D^{1/4+\epsilon})$ algorithm, bringing down the computation time to one similar to the case of imaginary quadratic fields.

**Remark.** If the regulator is computed to sufficient accuracy and is not too large, we can recover the fundamental unit by exponentiating. It is clear that it is impossible to find a sub-exponential algorithm for the fundamental unit in general, since, except when the regulator is very small, it already takes exponential time just to print it in the form $\epsilon = a + b\sqrt{D}$. It is possible however to write down explicitly the fundamental unit itself if we use a different representation, which H. Williams calls a *compact representation*. We will see in Section 5.8.3 how this is achieved.

## 5.8 The Infrastructure Method of Shanks

### 5.8.1 The Distance Function

The fundamental new idea introduced by Shanks in the theory of real quadratic fields is that one can introduce a distance function between quadratic forms or between ideals, and that this function will enable us to consider the principal cycle pretty much like a cyclic group. The initial theory is explained in [Sha3], and the refined theory which we will now explain can be found in [Len1].

**Definition 5.8.1.** *Let $\mathcal{O}$ be the quadratic order of discriminant $D$, and denote as usual by $\sigma$ real conjugation in $\mathcal{O}$. If $\mathfrak{a}$ and $\mathfrak{b}$ are fractional ideals of $\mathcal{O}$, we*

*define the* distance *of* $\mathfrak{a}$ *to* $\mathfrak{b}$ *as follows. If* $\mathfrak{a}$ *and* $\mathfrak{b}$ *are not equivalent (modulo principal ideals), the distance is not defined. Otherwise, write*

$$\mathfrak{b} = \gamma\mathfrak{a}$$

*for some* $\gamma \in K$. *We define the distance* $\delta(\mathfrak{a}, \mathfrak{b})$ *by the formula*

$$\delta(\mathfrak{a}, \mathfrak{b}) = \frac{1}{2} \ln \left| \frac{\gamma}{\sigma(\gamma)} \right|$$

*where* $\delta$ *is considered to be defined only modulo the regulator* $R$ *(i.e.* $\delta \in \mathbb{R}/R\mathbb{Z}$*).*

Note that this distance is well defined (modulo $R$) since if we take another $\gamma'$ such that $\mathfrak{b} = \gamma'\mathfrak{a}$, then $\gamma' = \epsilon\gamma$ where $\epsilon$ is a unit, hence the distance does not change modulo $R$. Note also that if $\mathfrak{a}$ is multiplied by a rational number, its distance to any other ideal does not change, hence in fact this distance carries over to the set $I$ of ideal classes defined in Section 5.2. This remark will be important later on.

In a similar manner, we can define the distance between two quadratic forms of positive discriminant $D$ as follows.

**Definition 5.8.2.** *Let* $f$ *and* $g$ *be two quadratic forms of discriminant* $D$, *and set* $(\mathfrak{a}, s) = \phi_{FI}(f)$, $(\mathfrak{b}, t) = \phi_{FI}(g)$ *as in Section 5.2, where* $s, t = \pm 1$. *If* $f$ *and* $g$ *are not equivalent modulo* $\mathrm{PSL}_2(\mathbb{Z})$, *the distance is not defined. If* $f$ *and* $g$ *are equivalent, then by Theorem 5.2.9 there exists* $\gamma \in K$ *such that*

$$\mathfrak{b} = \gamma\mathfrak{a} \quad \text{and} \quad t = s \cdot \mathrm{sign}(\mathcal{N}(\gamma)).$$

*We then define as above*

$$\delta(f, g) = \frac{1}{2} \ln \left| \frac{\gamma}{\sigma(\gamma)} \right|$$

*where* $\delta$ *is now considered to be defined modulo the regulator in the narrow sense* $R^+$, *i.e. the logarithm of the smallest unit greater than 1 which is of positive norm.*

Note once again that this distance is well defined, but this time modulo $R^+$, since if we take another $\gamma'$ we must have $\gamma' = \epsilon\gamma$ with $\epsilon$ a unit of positive norm. By abuse of notation, we will again denote by $\delta(f, g)$ the unique representative belonging to the interval $[0, R^+[$, and similarly for the distance between ideals.

Ideals are usually given by a $\mathbb{Z}$-basis, hence it is not easy to show that they are equivalent or not. Even if one knows for some reason that they are, it is still not easy to find a $\gamma \in K$ sending one into the other. In other words, it is not easy to compute the distance of two ideals (or of two quadratic forms) directly from the definition.

Luckily, we can bypass this problem in practice for the following reason. The quadratic forms which we will consider will almost always be obtained either by reduction of other quadratic forms (using the reduction step $\rho$ a number of times), or by composition of quadratic forms. Hence, it suffices to give transformation formulas for the distance $\delta$ under these two operations.

Composition is especially simple if one remembers that it corresponds to ideal multiplication. If, for $k = 1, 2$, we have $\mathfrak{b}_k = \gamma_k \mathfrak{a}_k$, then $\mathfrak{b}_1\mathfrak{b}_2 = \gamma_1\gamma_2\mathfrak{a}_1\mathfrak{a}_2$. This means that (*before any reduction step*), the distance function $\delta$ is exactly additive

$$\delta(\mathfrak{b}_1\mathfrak{b}_2, \mathfrak{a}_1\mathfrak{a}_2) = \delta(\mathfrak{b}_1, \mathfrak{a}_1) + \delta(\mathfrak{b}_2, \mathfrak{a}_2)$$

when all distances are defined. This is true for the distance function on ideals as well as for the distance function between quadratic forms *since $\delta$ does not change when one multiplies an ideal with a rational number.*

In the case of reduction, it is easier to work with quadratic forms. Let $f = (a, b, c)$ be a quadratic form of discriminant $D$. Then

$$\phi_{FI}(f) = \left(a\mathbb{Z} + \frac{-b + \sqrt{D}}{2}\mathbb{Z}, \operatorname{sign}(a)\right).$$

Furthermore, $\rho(f) = (c, b', a')$ where $b' \equiv -b \pmod{2c}$, hence

$$\phi_{FI}(\rho(f)) = \left(c\mathbb{Z} + \frac{b + \sqrt{D}}{2}\mathbb{Z}, \operatorname{sign}(c)\right),$$

since changing $b'$ modulo $2c$ does not change the ideal. Now clearly

$$c\mathbb{Z} + \frac{b + \sqrt{D}}{2}\mathbb{Z} = \gamma\left(a\mathbb{Z} + \frac{-b + \sqrt{D}}{2}\mathbb{Z}\right)$$

where

$$\gamma = \frac{b + \sqrt{D}}{2a}.$$

Hence we obtain

**Proposition 5.8.3.** *If $f = (a, b, c)$ is a quadratic form of discriminant $D$, then*

$$\delta(f, \rho(f)) = \frac{1}{2}\ln\left|\frac{b + \sqrt{D}}{b - \sqrt{D}}\right|.$$

Of course, the map $\phi_{IF}$ of Section 5.2 enables us also to compute distances between ideals.

If we have two quadratic forms $f$ and $g$ such that $g = \rho^n(f)$ for $n$ not too large, then by using the formula

$$\delta(f,g)=\sum_{i=1}^{n}\delta(\rho^{i-1}(f),\rho^{i}(f))$$

and this proposition, we can compute the distance of $f$ and $g$. When $n$ is large however, this formula, which takes time at least $O(n)$, becomes impractical. This is where we need to use composition.

For simplicity, we now assume that our forms are in the principal cycle, i.e. are equivalent to the unit form which we denote by $\mathbf{1}$. We then have the following proposition

**Proposition 5.8.4.** *Let $f_1$ and $f_2$ be two reduced forms in the principal cycle, and let $\mathbf{1}$ be the unit form. Then if we define $g=f_1\cdot f_2$ by the composition algorithm given in Section 5.4.2, $g$ may not be reduced, but let $f_3$ be a (non-unique) form obtained from $g$ by the reduction algorithm, i.e. by successive applications of $\rho$. Then we have*

$$\delta(\mathbf{1},f_3)=\delta(\mathbf{1},f_1)+\delta(\mathbf{1},f_2)+\delta(g,f_3),$$

*and furthermore*

$$|\delta(g,f_3)|<2\ln(D).$$

This proposition follows at once from the property that $\delta$ is exactly additive under composition (before any reductions are made). □

If we assume that we know $\delta(\mathbf{1},f_1)$ and $\delta(\mathbf{1},f_2)$, then it is easy to compute $\delta(\mathbf{1},f_3)$ since the number of reduction steps needed to go from $g$ to $f_3$ is very small. More precisely, it can be proved (see [Len1]) that $\delta(f,\rho^2(f))>\ln 2$, hence the number of reduction steps is at most $4\ln(D)/\ln 2$.

**Important Remark.** In the preceding section we have computed the regulator by adding $\ln((p_i+\sqrt{D})/|q_i|)$ over a cycle (or a half cycle). This corresponds to choosing a modified distance such that $\delta'(f,\rho(f))=\ln((b+\sqrt{D})/(2|a|))$, and this clearly corresponds to defining

$$\delta'(\mathfrak{a},\mathfrak{b})=\ln|\gamma|$$

instead of $\delta(\mathfrak{a},\mathfrak{b})=\frac{1}{2}\ln|\gamma/\sigma(\gamma)|$ if $\mathfrak{b}=\gamma\mathfrak{a}$. This distance, which was the initial one suggested by Shanks, can also be used for regulator computations since it is also additive. Note however that it is no longer defined on the set $I$ of ideals modulo the multiplicative action of $\mathbb{Q}^*$, but on the ideals themselves. In particular, with reference to Lemma 5.4.5, we must subtract $\ln(d)$ to the sum of the distances of $I_1$ and $I_2$ before starting the reduction of our composed quadratic form $(A,B,C)$. It also introduces extra factors when one computes the inverse of a form. For example, this would introduce many unnecessary complications in Buchmann's sub-exponential algorithm that we will study below (Section 5.9).

On the other hand, although Shanks's distance is less natural, it is computationally slightly better since it is simpler to multiply by $(b+\sqrt{D})/(2|a|)$ than by $|(b+\sqrt{D})/(b-\sqrt{D})|$. Note also that Proposition 5.8.4 is valid with $\delta$ replaced by $\delta'$, if we take care to subtract the $\ln(d)$ value after composition as we have just explained.

Hence, for simplicity, we will use the distance $\delta$ instead of Shanks's $\delta'$, except in the baby-step giant-step Algorithm 5.8.5 where the use of $\delta'$ gives a slightly more efficient algorithm.

### 5.8.2 Description of the Algorithm

We consider the set $S$ of pairs $(f,z)$, where $f$ is a reduced form of discriminant $D$ in the principal cycle, and $z=\delta(1,f)$. We can transfer the action of $\rho$ to $S$ by setting $\rho(f,z)=(\rho(f),z+\ln|(b+\sqrt{D})/(b-\sqrt{D})|/2)$ if $f=(a,b,c)$, using the above notations. Furthermore, we can transfer the composition operation by setting

$$(f_1,z_1)\cdot(f_2,z_2)=(f_3,z_1+z_2+\delta(g,f_3)),$$

using the notations of Proposition 5.8.4. Similar formulas are valid with $\delta$ replaced by $\delta'$. Recall that $f_3$ is not uniquely defined, but this does not matter for our purposes as long as we choose $f_3$ not too far away from the first reduced form that one meets after applying $\rho$ to $f_1\cdot f_2$.

Using these notations, we can apply Shanks's baby-step giant-step method to compute $R(D)$. Indeed, although the principal cycle is not a group, because of the set $S$ we can follow the value of $\delta$ through composition and reduction. This means that Shanks's method allows us to find the regulator in $O(D^{1/4+\epsilon})$ steps instead of the usual $O(D^{1/2+\epsilon})$. If, as for negative discriminants, we also use that the inverse of a form $(a,b,c)$ is a form equivalent to $(a,-b,c)$, i.e. $(a,r(-b,a),(r(-b,a)^2-D)/4a)$, we obtain the following algorithm, due in essence to Shanks, and modified by Williams. Note that we give the algorithm using Shanks's distance $\delta'$ instead of $\delta$ since it is slightly more efficient, and also we use the language of continued fractions as in Algorithm 5.7.10, in other words, instead of $(a,b,c)$ we use $(p,q)=(b,2|a|)$.

**Algorithm 5.8.5** (Regulator Using Infrastructure). Given a positive fundamental discriminant $D$, this algorithm computes $R(D)$. We assume that all the real numbers involved are computed with a finite and reasonably small accuracy. We make use of an auxiliary table $\mathcal{T}$ of quadruplets $(q,p,e,R)$ where $p$, $q$, $e$ are integers and $R$ is a real number.

1. [Initialize] Precompute $f\leftarrow\sqrt{D}$, and set $d\leftarrow\lfloor\sqrt{D}\rfloor$, $e\leftarrow 0$, $R\leftarrow 1$, $s\leftarrow\lceil 1.5\sqrt{d}\rceil$, $T\leftarrow s+\lceil\ln(4d)/(2\ln((1+\sqrt{5})/2))\rceil$ and $q\leftarrow 2$. If $d\equiv D\pmod 2$, set $p\leftarrow d$, otherwise set $p\leftarrow d-1$. Set $q_1=(D-p^2)/q$, $i\leftarrow 0$, and store the $(q,p,e,R)$ in $\mathcal{T}$. Finally, let $2^L$ be the highest power of 2 such that $2^Lf$ does not give an exponent overflow.
2. [Small steps] Set $i\leftarrow i+1$, and let $p+d=Aq+r$ with $0\le r<q$ be the Euclidean division of $p+d$ by $q$. Set $p_1\leftarrow p$, $p\leftarrow d-r$, $t\leftarrow q$,

$q \leftarrow q_1 - A(p - p_1)$, $q_1 \leftarrow t$, $R \leftarrow R(p+f)/q_1$. If $R \geq 2^L$, set $R \leftarrow R/2^L$, $e \leftarrow e+1$. If $q \leq d$, store $(q, p, e, R)$ in $\mathcal{T}$.

3. [Finished already?] If $p_1 = p$ and $i > 1$, then output

$$R(D) = 2(\ln(R) + eL\ln(2)) - \ln(q_1/2)$$

and terminate the algorithm. If $q_1 = q$ and $i > 1$, then output

$$R(D) = 2(\ln(R) + eL\ln(2)) - \ln((p+f)/2)$$

and terminate the algorithm. If $i = s$, then if $q \leq d$ set $(Q, P, E, R_1) \leftarrow (q, p, e, R)$ otherwise (still if $i = s$) set $s \leftarrow s+1$ and $T \leftarrow T+1$. Finally, if $i < T$ go to step 2.

4. [Initialize for giant steps] Sort table $\mathcal{T}$ lexicographically (or in any other way). Then using the composition Algorithm 5.8.6 given below, compute

$$(Q, P, E, R_1) \leftarrow (Q, P, E, R_1) \cdot (Q, P, E, R_1),$$

and set $R \leftarrow 1$, $e \leftarrow 0$, $j \leftarrow 1$, and $q \leftarrow Q$, $p \leftarrow P$.

5. [Match found?] If $(q, p) = (q_1, p_1)$ for some $(q_1, p_1, e_1, r_1) \in \mathcal{T}$, output

$$R(D) = j(\ln(R_1) + EL\ln(2)) + \ln(R) + eL\ln(2) - \ln(r_1) - e_1 L\ln(2)$$

and terminate the algorithm.

If $(q, r(-p, q)) = (q_1, p_1)$ for some $(q_1, p_1, e_1, r_1) \in \mathcal{T}$, output

$$R(D) = j(\ln(R_1)+EL\ln(2))+\ln(R)+eL\ln(2)+\ln(r_1)+e_1 L\ln(2)-\ln(q_1/2)$$

and terminate the algorithm.

6. [Giant steps] Using the composition Algorithm 5.8.6 below, compute

$$(q, p, e, R) \leftarrow (q, p, e, R) \cdot (Q, P, E, R_1),$$

set $j \leftarrow j+1$ and go to step 5.

We need to compose two quadratic forms of positive discriminant $D$, expressed as quadruplets $(q, p, e, R)$, where the pair $(e, R)$ keeps track of the distance from 1 (more precisely $\delta'(1, f) = eL\ln 2 + \ln R$), and the form itself is $(q, p, (p^2 - D)/q)$ or $(-q, p, (D - p^2)/q)$. The algorithm is identical to the positive definite case (Algorithm 5.4.7), except that the reduction in step 4 must be done using Algorithm 5.6.5 (i.e. powers of $\rho$) instead of Algorithm 5.4.2. We must also keep track of the distance function, and, since we use $\delta'$ instead of $\delta$, we must subtract a $\ln(d_1)$ (i.e. divide by $d_1$) where $d_1$ is the computed GCD.

This leads to the following algorithm.

**Algorithm 5.8.6** (Composition of Indefinite Forms with Distance Function). Given two quadruplets $(q_1, p_1, e_1, R_1)$ and $(q_2, p_2, e_2, R_2)$ as above (in particular with $q_i$ even and positive), this algorithm computes the composition

$$(q_3, p_3, e_3, R_3) = (q_1, p_1, e_1, R_1) \cdot (q_2, p_2, e_2, R_2).$$

We assume $f \leftarrow \sqrt{D}$ already computed to sufficient accuracy.

1. [Initialize] If $q_1 > q_2$, exchange the quadruplets. Then set $s \leftarrow \frac{1}{2}(p_1 + p_2)$, $n \leftarrow p_2 - s$.
2. [First Euclidean step] If $q_1 \mid q_2$, set $y_1 \leftarrow 0$ and $d \leftarrow q_1/2$. Otherwise, using Euclid's extended algorithm, compute $(u, v, d)$ such that $uq_2/2 + vq_1/2 = d = \gcd(q_2/2, q_1/2)$ and set $y_1 \leftarrow u$.
3. [Second Euclidean step] If $d \mid s$, set $y_2 \leftarrow -1$, $x_2 \leftarrow 0$ and $d_1 \leftarrow d$. Otherwise, using Euclid's extended algorithm, compute $(u, v, d_1)$ such that $us + vd = d_1 = \gcd(s, d)$, and set $x_2 \leftarrow u$, $y_2 \leftarrow -v$.
4. [Compose] Set $v_1 \leftarrow q_1/(2d_1)$, $v_2 \leftarrow q_2/(2d_1)$, $r \leftarrow ((y_1 y_2 n - x_2(p_2^2 - D)/(2q_2) \bmod v_1)$, $p_3 \leftarrow p_2 + 2v_2 r$, $q_3 \leftarrow 2v_1 v_2$.
5. [initialize reduction] Set $e_3 \leftarrow e_1 + e_2$ and $R_3 \leftarrow R_1 R_2/d_1$. If $R_3 \geq 2^L$, set $R_3 \leftarrow R_3/2^L$ and $e_3 \leftarrow e_3 + 1$.
6. [Reduced?] If $|f - q_3| < p_3$, then output $(q_3, p_3, e_3, R_3)$ and terminate the algorithm. Otherwise, set $p_3 \leftarrow r(-p_3, q_3/2)$, $R_3 \leftarrow R_3(p_3 + f)/q_3$, $q_3 \leftarrow (D - p_3^2)/q_3$, and if $R_3 \geq 2^L$ set $R_3 \leftarrow R_3/2^L$ and $e_3 \leftarrow e_3 + 1$. Finally, go to step 6.

Note that $r(-p_3, q_3/2)$ is easily computed by a suitable Euclidean division.

This algorithm performs very well, and one can compute regulators of real quadratic fields with discriminants with up to 20 digits in reasonable time. To go beyond this requires new ideas which are essentially the same as the ones used in McCurley's sub-exponential algorithm and will in fact give us simultaneously the regulator and the class group. We will study this in Section 5.9.

### 5.8.3 Compact Representation of the Fundamental Unit

The algorithms that we have seen above allow us to compute the regulator of a real quadratic field to any desired accuracy. If this accuracy is high, however, and in particular if we want infinite accuracy (i.e. the fundamental unit itself and not its logarithm), we must not apply the algorithms exactly as they are written. The reason for this is that by using the infrastructure ideas of Shanks (essentially the distance function), the knowledge of a crude approximation to the regulator $R(D)$ (say only its integer part) allows us to compute it very fast to any desired accuracy. Let us see how this is done.

Let $f$ be the form $\rho(\mathbf{1})$. It is the first form encountered in the principal cycle when we start at the unit form, and in particular has the smallest distance to

1. Assume that after applying one of the regulator algorithms we know that $R_1 < R(D) < R_2$ (this can be a very crude estimate, for example we could ask that $R_2 - R_1 < 1$). By using the same idea as in Exercise 4 of Chapter 1, it is easy to find in time $O(\ln(D))$ composition operations, an integer $n$ such that $\delta(\mathbf{1}, f^n) \leq R_1$ and $\delta(\mathbf{1}, f^{n+1}) > R_1$. This implies that $f^n$ is before the unit form in the principal cycle (counting in terms of increasing distances), but not much before since $R_2 - R_1$ is small. Hence, there exists a small $k \geq 0$ which one finds by simply trying $k = 0, 1, \ldots$ such that $\mathbf{1} = \rho^k(f^n)$. Note that this is checked on the exact components of the forms, not on the distance. Hence, we now assume that $k$ and $n$ have been found.

If we want the regulator very precisely, we recompute $f = \rho(\mathbf{1})$ to the desired accuracy, and then the distance component of $\rho^k(f^n)$ will give us the regulator to the accuracy that we want.

If we want the fundamental unit itself, note that by Proposition 5.8.4 the composition of two forms implies the addition of three distances, or equivalently the multiplication of three quadratic numbers. For the $\rho$ operator, only one such multiplication is required. Finally, note that $k$ will be $O(\ln(D))$ and $n$ will be $O(\sqrt{D})$ hence only $O(\ln(D))$ composition or reduction steps are required to compute $\rho^k(f^n)$. This implies that we can express the fundamental unit as a product of at most $O(\ln(D))$ terms of the form $(b + \sqrt{D})/(2|a|)$ (or $|(b + \sqrt{D})/(b - \sqrt{D})|$ if we use the distance $\delta$ instead of $\delta'$) and this is a compact way of keeping the fundamental unit even when $D$ is very large.

Let us give a numerical example. Take $D = 10209$. A rough computation using one of the regulator algorithms shows that $R(D) \approx 67.7$. Furthermore, one computes that $f = \rho(\mathbf{1}) = (-2, 99, 51)$. The binary algorithm gives $f^{14} = (1, 101, -2) = \mathbf{1}$ with $\delta'(\mathbf{1}, f^{14}) \approx 67.7$. Note that this exponent 14 is not at all canonical and depends on the number of reduction steps performed at each composition, and on the order in which the compositions steps are made. Here, we assume that we stop applying $\rho$ as soon as the form is reduced, and that $f^n$ is computed using the right-left binary powering Algorithm 1.2.1.

We now start again recomputing $f$ and $f^{14}$, keeping the quantities $(b + \sqrt{D})/(2|a|)|$ that are multiplied, along with their exponents. If $\epsilon$ is the fundamental unit, we obtain

$$\epsilon = \left(\frac{101 + \sqrt{D}}{2}\right)^{14} \left(\frac{111 + \sqrt{D}}{32}\right)^{3} \frac{1}{3} \frac{219 + \sqrt{D}}{242} \frac{351 + \sqrt{D}}{264} \frac{77 + \sqrt{D}}{428} \frac{93 + \sqrt{D}}{780}.$$

The lonely 1/3 in the middle is due to the use of the imperfect distance function $\delta'$ which as we have already mentioned introduces extra quantities $-\ln d$ in the compositions.

If we instead use the distance $\delta$, we obtain $\epsilon^2 = \tau/\overline{\tau}$ with

$$\tau = (101 + \sqrt{D})^{14}(111 + \sqrt{D})^3(219 + \sqrt{D})(197 + \sqrt{D})(103 + \sqrt{D}).$$

Hence, to represent $\epsilon$, we could simply keep the pairs $(101,14)$, $(111,3)$, $(219,1)$, $(197,1)$ and $(103,1)$. It is a matter of taste which of the two representations above is preferable. Note that in fact

$$\epsilon = 1309694962454302631594431787 75 + 1296219513663218157975941956\sqrt{D}$$

which does not really take more space, but for larger discriminants this kind of explicit representation becomes impossible, while the compact one survives without any problem since there are only $O(\ln(D))$ terms of size $O(\ln(D))$ to be kept.

In [Buc-Thi-Wil], the authors have given a slightly more elegant compact representation of the fundamental unit, but the basic principle is the same. This idea can be generalized to the representation of algebraic numbers (and not only units), and to any number field.

### 5.8.4 Other Application and Generalization of the Distance Function

An important aspect of the distance function should be stressed at this point. Not only does it give us a fundamental hold on the fine structure of units, but it also allows us to solve the *principal ideal* problem which is the following. Assume that $\mathfrak{a}$ is an integral ideal of $\mathbb{Z}_K$ which is known to be a principal ideal (for example because $\mathfrak{a} = \mathfrak{b}^h$ for some ideal $\mathfrak{b}$, where $h$ is the class number of $K$). Assume that we know the distance function $\delta(\mathbf{1},\mathfrak{a})$. Then it is easy to find an element $\gamma$ such that $\mathfrak{a} = \gamma\mathbb{Z}_K$ using the formulas

$$\gamma = \pm\sqrt{\mathcal{N}(\mathfrak{a})}e^{\delta(\mathbf{1},\mathfrak{a})}, \quad \sigma(\gamma) = \pm\sqrt{\mathcal{N}(\mathfrak{a})}e^{-\delta(\mathbf{1},\mathfrak{a})}.$$

This leaves only 2 possibilities for $\pm\gamma$, and usually only one will belong to $K$. Note that since $\delta$ is defined only in $\mathbb{R}/R\mathbb{Z}$, $\gamma$ will be defined up to multiplication by a unit.

Similarly, if the distance function $\delta'(\mathbf{1},\mathfrak{a})$ is known, we use the formulas

$$\gamma = \pm e^{\delta'(\mathbf{1},\mathfrak{a})}, \quad \sigma(\gamma) = \pm\mathcal{N}(\mathfrak{a})e^{-\delta'(\mathbf{1},\mathfrak{a})}.$$

The distance function $\delta$ can be naturally generalized to arbitrary number fields $K$ as follows. Let

$$L(x) = (\ln|\sigma_1(x)|, \ldots, \ln|\sigma_{r_1}(x)|, 2\ln|\sigma_{r_1+1}(x)|, \ldots, 2\ln|\sigma_{r_1+r_2}(x)|)$$

be the logarithmic embedding of $K^*$ into $\mathbb{R}^{r_1+r_2}$ seen in Definition 4.9.6, where $(r_1,r_2)$ is the signature of $K$. If $n = r_1 + 2r_2$ is the degree of $K$, we will set

$$\Delta(\mathfrak{a},\gamma\mathfrak{a}) = L(\gamma/|\mathcal{N}_{K/\mathbb{Q}}(\gamma)|^{1/n}),$$

where it is understood that the $\sigma_i$ act trivially on the $n$-th roots of the norms.

Then $\Delta$ belongs to the hyperplane $\sum_{1\le i\le r_1+r_2} x_i = 0$ of $\mathbb{R}^{r_1+r_2}$ and is defined modulo the lattice which is the image of the group of units $U(K)$ under the embedding $L(x)$.

In the case where $K$ is a real quadratic field, then clearly $\Delta = (\delta, -\delta)$, so this is a reasonable generalization of $\delta$. If $K$ is an imaginary quadratic field, we have $\Delta = 0$.

The principal ideal problem can, of course, be asked in general number fields and it is clear that $\Delta$ cannot help us to solve it in general since it cannot do so even for imaginary quadratic fields. For this specific application, the logarithmic embedding $L$ should be replaced by the ordinary embedding

$$(\sigma_1(x), \dots, \sigma_{r_1}(x), \sigma_{r_1+1}(x), \dots, \sigma_{r_1+r_2}(x))$$

of $K$ into $\mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$.

The components of this embedding are in general too large to be represented exactly, hence we will preferably choose the complex logarithmic embedding

$$L_C(x) = (\ln \sigma_1(x), \dots, \ln \sigma_{r_1}(x), 2\ln \sigma_{r_1+1}(x), \dots, 2\ln \sigma_{r_1+r_2}(x)),$$

where the logarithms are defined up to addition of an integer multiple of $2i\pi$. Note that this requires only twice as much storage space as the embedding $L$, and also that the first $r_1$ components have an imaginary part which is a multiple of $\pi$. Let $V = (n_i)_{1\le i\le r_1+r_2}$ be the vector such that $n_i = 1$ for $i \le r_1$ and $n_i = 2$ otherwise. We can then define

$$\Delta_C(\mathfrak{a}, \gamma\mathfrak{a}) = L_C(\gamma) - \frac{\ln(\mathcal{N}(\gamma))}{n} V,$$

and it is clear that the sum of the $r_1 + r_2$ components of $\Delta_C$ is an integral multiple of $2i\pi$. We will see the use of this function in Section 6.5.

## 5.9 Buchmann's Sub-exponential Algorithm

We will now describe a fast algorithm for computing the class group and the regulator of a real quadratic field, which uses essentially the same ideas as Algorithm 5.5.2.

Although the main ideas are in McCurley and Shanks, I have seen this algorithm explained only in manuscripts of J. Buchmann whom I heartily thank for the many conversations which we have had together. The first implementation of this algorithm is due to Cohen, Diaz y Diaz and Olivier (see [CohDiOl]).

### 5.9.1 Outline of the Algorithm

We will follow very closely Algorithm 5.5.2, and use the distance function $\delta$ and *not* Shanks's distance $\delta'$ which we used in Algorithm 5.8.5.

As we have already explained, in the quadratic case it is simpler to work with forms instead of directly with ideals. Note however that because of Theorem 5.2.9, we will be computing the *narrow* ideal class group and the regulator in the narrow sense, since this is the natural correspondence with quadratic forms. If, on the other hand, we want the ideal class group and the regulator in the ordinary sense, then, according to Proposition 5.6.1, we will have to identify the form $(a, b, c)$ with the form $(-a, b, -c)$. (This is implicitly what we did in Algorithm 5.8.5.) Although it is very easy to combine both procedures into a single algorithm, note that the computations are independent. More precisely, to the best of my knowledge it does not seem to be easy, given the ideal class group and regulator in one sense (narrow or ordinary) to deduce the ideal class group and regulator in the other sense, although of course only a factor of 2 is involved. We will describe the algorithm for the class group and regulator in the ordinary sense, leaving to the reader the simple modifications that must be made to obtain the class group and regulator in the narrow sense (see Exercise 26).

We now describe the outline of the algorithm. As in Algorithm 5.8.5, we keep track of the distance function as a pair $(e, R)$, but this time we will keep all three coefficients of the quadratic form. Also, we are going to use the distance $\delta$ instead of $\delta'$, and since there is a factor $1/2$ in the definition of $\delta$, we will use the correspondence $\delta(f_0, f) = (eL\ln 2 + \ln R)/2$ for some fixed form $f_0$ equivalent to $f$.

In other words, in this section a quadratic form of positive discriminant will be a quintuplet $f = (a, b, c, e, R)$ where $a$, $b$, $c$ and $e$ are integers and $R$ is a real number such that $1 \leq R < 2^L$.

We can compose two such forms by using the following algorithm, which is a trivial modification of Algorithm 5.8.6.

**Algorithm 5.9.1** (Composition of Indefinite Forms with Distance Function). Given two primitive quadratic forms $(a_1, b_1, c_1, e_1, R_1)$ and $(a_2, b_2, c_2, e_2, R_2)$ as above, this algorithm computes the composition

$$(a_3, b_3, c_3, e_3, R_3) = (a_1, b_1, c_1, e_1, R_1) \cdot (a_2, b_2, c_2, e_2, R_2).$$

We assume $f \leftarrow \sqrt{D}$ already computed to sufficient accuracy.

1. [Initialize] If $|a_1| > |a_2|$ exchange the quintuplets. Then set $s \leftarrow \frac{1}{2}(b_1 + b_2)$, $n \leftarrow b_2 - s$.
2. [First Euclidean step] If $a_1 \mid a_2$, set $y_1 \leftarrow 0$ and $d \leftarrow |a_1|$. Otherwise, using Euclid's extended algorithm, compute $u$, $v$ and $d$ such that $ua_2 + va_1 = d = \gcd(a_2, a_1)$ and set $y_1 \leftarrow u$.
3. [Second Euclidean step] If $d \mid s$, set $y_2 \leftarrow -1$, $x_2 \leftarrow 0$ and $d_1 \leftarrow d$. Otherwise, again using Euclid's extended algorithm, compute $(u, v, d_1)$ such that $us + vd = d_1 = \gcd(s, d)$, and set $x_2 \leftarrow u$ and $y_2 \leftarrow -v$.

4. [Compose] Set $v_1 \leftarrow a_1/d_1$, $v_2 \leftarrow a_2/d_1$, $r \leftarrow (y_1y_2n - x_2c_2 \bmod v_1)$, $b_3 \leftarrow b_2 + 2v_2r$ and $a_3 \leftarrow v_1v_2$.

5. [Initialize reduction] Set $e_3 \leftarrow e_1 + e_2$, $R_3 \leftarrow R_1R_2$. If $R_3 \geq 2^L$, set $R_3 \leftarrow R_3/2^L$ and $e_3 \leftarrow e_3 + 1$.

6. [Reduced?] If $|f - 2|a_3|| < b_3 < f$, then output $(a_3, b_3, c_3, e_3, R_3)$ and terminate the algorithm.

7. [Apply $\rho$] Set $R_3 \leftarrow R_3|(b_3 + f)/(b_3 - f)|$ and if $R_3 \geq 2^L$, set $R_3 \leftarrow R_3/2^L$ and $e_3 \leftarrow e_3 + 1$. Then set $a_3 \leftarrow c_3$, $b_3 \leftarrow r(-b_3, c_3)$, $c_3 \leftarrow (b_3^2 - D)/a_3$ and go to step 6.

Note that, apart from some absolute value signs, steps 1 to 4 are identical to the corresponding steps in Algorithm 5.4.7, but the reduction operation is quite different since it involves iterating the function $\rho$ in step 7 of the algorithm and the bookkeeping necessary for the distance function.

Returning to Buchmann's algorithm, what we will do is essentially, instead of keeping track only of $f_p = (p, b_p, (b_p^2 - D)/(4p))$, we also keep track of the distance function. Hence, in step 3 of Algorithm 5.5.2, we compute the product $\prod_{p \leq P} f_p^{e_p}$, doing the reduction at each product (of course the reduction being non-unique), and keeping track of the distance function thanks to Theorem 5.8.4. In this way we obtain a reduced form $f = (a, b, c)$ equivalent to the above product, and also the value of $\delta(\prod_{p \leq P} f_p^{e_p}, f)$. Since we identify $(a, b, c)$ with $(-a, b, -c)$, we will replace $(a, b, c)$ by $(|a|, b, -|c|)$.

If $a$ does not factor easily, in step 5 we have the option of doing more reduction steps instead of going back to step 4 in the hope of getting an easily factorable $a$. Since this is much faster than recomputing a new product, we will use this method as much as possible. Note that, although we have extra computations to make because of the distance function, the basic computational steps will be *faster* than in the imaginary quadratic case, hence this algorithm will be faster than the corresponding one for imaginary quadratic fields.

This behavior is to be expected since on heuristic and experimental grounds class numbers of real quadratic fields are much smaller than those of imaginary quadratic fields.

Finally, if $a$ factors easily, in step 5 we compute not only $a_{i,k}$ for $1 \leq i \leq n$, but also $a_{n+1,k} \leftarrow \delta(\mathbf{1}, fg^{-1})$ where $g = \prod_{p \leq P} f_p^{\epsilon_p v_p}$ and $\delta(\mathbf{1}, fg^{-1})$ is computed as usual at the same time as the product is done, using Theorem 5.8.4.

We thus obtain a matrix $A = (a_{i,j})$ with $n+1$ rows and $k$ columns, whose entries in the first $n$ rows are integers and the entries in the last row are real numbers. Note that by definition, for every $j \leq k$ we have

$$\delta\left(\mathbf{1}, \prod_{1 \leq i \leq n} f_{p_i}^{-a_{i,j}}\right) \equiv a_{n+1,j} \pmod{R(D)}.$$

Since the distance function that we have chosen is exactly additive, it follows that when performing column operations on the complete matrix $A$, this relation between the $n+1$-st component and the others is preserved.

Hence we apply Hermite reduction to the matrix formed by the first $n$ rows, but performing the corresponding column operations also the entries of the last row. The first $k-n$ columns of the resulting matrix will thus have only zero entries, except perhaps for the entry in the $n+1$-st row. By the remark made above, for $1 \leq j \leq k-n$ we will thus have

$$a_{n+1,j} \equiv \delta(\mathbf{1},\mathbf{1}) \equiv 0 \pmod{R(D)},$$

in other words $a_{n+1,j}$ is equal to a multiple of the regulator $R(D)$ for $1 \leq j \leq k-n$.

If $k$ is large enough, it follows that in a certain sense the GCD of the $a_{n+1,j}$ for $1 \leq j \leq k-n$ should be exactly equal to $R(D)$. We must be careful in the computation of this "GCD" since we are dealing with inexact real numbers. For this purpose, we can either use the LLL algorithm which will give us a small linear combination of the $a_{n+1,j}$ for $1 \leq j \leq k-n$ with integral coefficients, which should be the regulator $R(D)$, or use the "real GCD" Algorithm 5.9.3 as described below.

The rest of the algorithm will compute the class group structure in essentially the same way, except of course that in step 1 one must use the analytic class number formula for positive discriminants (Proposition 5.6.9).

### 5.9.2 Detailed Description of Buchmann's Sub-exponential Algorithm

A practical implementation of this algorithm should take into account at least two remarks. First, note that most of the time is spent in looking for relations. Hence, it is a waste of time to compute with the distance function during the search for relations: we do the search only with the components $(a,b,c)$ of the quadratic forms, and only in the rare cases where a relation is obtained do we recompute the relation with the distance function. The slight loss of time due to the recomputation of each relation is more than compensated by the gain obtained by not computing the distance function during the search for relations.

The second remark is that, as in McCurley's sub-exponential algorithm, the Hermite reduction of the first $n$ rows must be performed modulo a multiple of the determinant, which can be computed before starting the reduction. In other words, we will use Algorithm 2.4.8. The reduction of the last row is however another problem, and in the implementation due to the author, Diaz y Diaz and Olivier, the best method found was to compute the integer kernel of the integer matrix formed by the first $n$ rows using Algorithm 2.7.2, and multiply the $n+1$-st row of distances by this kernel, thus obtaining a vector whose components are (approximately) small multiples of the regulator, and

we find the regulator itself using one of the methods explained above, for example the LLL algorithm.

These remarks lead to the following algorithm.

**Algorithm 5.9.2** (Sub-Exponential Real Class Group and Regulator). If $D > 0$ is a non-square discriminant, this algorithm computes the class number $h(D)$, the class group $Cl(D)$ and the regulator $R(D)$. As before, in practice we work with binary quadratic forms. We also choose at will a positive real constant $b$.

1. [Compute primes and Euler product] Set $m \leftarrow b\ln^2 D$, $M \leftarrow L(D)^{1/\sqrt{8}}$, $P \leftarrow \lfloor \max(m, M) \rfloor$

$$\mathcal{P} \leftarrow \left\{ p \leq P, \left(\frac{D}{p}\right) \neq -1 \text{ and } p \text{ good} \right\}$$

and compute the product

$$B \leftarrow \frac{\sqrt{D}}{2} \prod_{p \leq P} \left(1 - \frac{\left(\frac{D}{p}\right)}{p}\right)^{-1}.$$

2. [Compute prime forms] Let $\mathcal{P}_0$ be the set made up of the smallest primes of $\mathcal{P}$ not dividing $D$ such that $\prod_{p \in \mathcal{P}_0} p > \sqrt{D}$. For the primes $p \in \mathcal{P}$ do the following. Compute $b_p$ such that $b_p^2 \equiv D \pmod{4p}$ using Algorithm 1.5.1 (and modifying the result to get the correct parity). If $b_p > p$, set $b_p \leftarrow 2p - b_p$. Set $f_p \leftarrow (p, b_p, (b_p^2 - D)/(4p))$ and $g_p \leftarrow (p, b_p, (b_p^2 - D)/(4p), 0, 1.0)$ Finally, let $n$ be the number of primes $p \in \mathcal{P}$.

3. [Compute powers] For each $p \in \mathcal{P}_0$ and each integer $e$ such that $1 \leq e \leq 20$ compute and store a reduced form equivalent to $f_p^e$. Set $k \leftarrow 0$.

4. [Generate random relations] Let $f_q$ be the primeform number $k + 1 \bmod n$ in the factor base. Choose random $e_p$ between 1 and 20, and compute a reduced form $(a, b, c)$ equivalent to

$$f_q \prod_{p \in \mathcal{P}_0} f_p^{e_p}$$

by using the composition algorithm for positive binary quadratic forms, replacing the final reduction step by a sufficient number of applications of the $\rho$ operator (note that $f_p^{e_p}$ has already been computed in step 3). Set $e_p \leftarrow 0$ if $p \notin \mathcal{P}_0$ then $e_q \leftarrow e_q + 1$. Set $(a_0, b_0, c_0) \leftarrow (a, b, c)$, $r \leftarrow 0$ and go to step 6.

5. [Apply $\rho$] Set $(a, b, c) \leftarrow \rho(a, b, c)$ and $r \leftarrow r + 1$. If $|a| = |a_0|$ and $r$ is odd, or if $b = b_0$ and $r$ is even, go to step 4.

6. [Factor $|a|$] Factor $|a|$ using trial division. If a prime factor of $|a|$ is larger than $P$, do not continue the factorization and go to step 5. Otherwise, if $|a| = \prod_{p \leq P} p^{v_p}$, set $k \leftarrow k + 1$, and for $i \leq n$ set

$$a_{i,k} \leftarrow e_{p_i} - \epsilon_{p_i} v_{p_i}$$

where $\epsilon_{p_i} = +1$ if $(b \bmod 2p_i) \leq p_i$, $\epsilon_{p_i} = -1$ otherwise.

7. [Recompute relation with distance] Compute

$$(a_0, b_0, c_0, e_0, R_0) \leftarrow g_q \prod_{p \in \mathcal{P}_0} g_p^{e_p}$$

by mimicking the order of squarings, compositions and reductions done to compute $(a_0, b_0, c_0)$, but this time using Algorithm 5.9.1 for composition. Then compute $(a, b, c, e, R) \leftarrow \rho^r(a_0, b_0, c_0, e_0, R_0)$ by applying the formulas of step 7 of Algorithm 5.9.1 to our forms. Finally, set $a_{n+1,k} \leftarrow (eL \ln 2 + \ln R)/2$.

8. [Enough relations?] If $k < n + 10$ go to step 4.

9. [Be honest] For each prime $q$ such that $P < q \leq 6 \ln^2 D$ do the following. Choose random $e_p$ between 1 and 20 (say), compute the primeform $f_q$ corresponding to $q$ and some reduced form $(a, b, c)$ equivalent to $f_q \prod_{p \in \mathcal{P}_0} f_p^{e_p}$. If $a$ does not factor into primes less than $q$, choose other exponents $e_p$ and continue until $a$ factors into such primes (or apply the $\rho$ operator as in step 5). Then go on to the next prime $q$ until the list is exhausted.

10. [Simple HNF] Perform a preliminary simple Hermite reduction on the $(n + 1) \times k$ matrix $A = (a_{i,j})$ as described in the remarks following Algorithm 5.5.2. In this reduction, only the first $n$ rows should be examined, but column operations should of course be done also with the $n+1$-st row. Let $A_1$ be the matrix thus obtained without its last row, and let $V$ be the last row (whose components are linear combinations of distances).

11. [Compute regulator] Using Algorithm 2.7.2, compute the LLL-reduced integral kernel $M$ of $A_1$ as a rectangular matrix, and set $V \leftarrow VM$. Let $s$ be the number of elements of $V$. Set $R \leftarrow |V_1|$, and for $i = 2, \ldots, s$ set $R \leftarrow RGCD(R, |V_i|)$ where RGCD is the real GCD algorithm described below. (Now $R$ is probably the regulator.)

12. [Compute determinant] Using standard Gaussian elimination techniques, compute the determinant of the lattice generated by the columns of the matrix $A_1$ modulo small primes $p$. Then compute the determinant $d$ exactly using the Chinese remainder theorem and Hadamard's inequality (see also Exercise 13).

13. [HNF reduction] Using Algorithm 2.4.8 compute the Hermite normal form $H = (h_{i,j})$ of the matrix $A_1$ using modulo $d$ techniques. Then for every $i$ such that $h_{i,i} = 1$, suppress row and column $i$. Let $W$ be the resulting matrix.

14. [Finished?] Let $h \leftarrow \det(W)$ (i.e. the product of the diagonal elements). If $hR \geq B\sqrt{2}$, get 5 more relations (in steps 4, 5 and 6) and go to step 10. (It will not be necessary to recompute the whole HNF, only that which takes into account the last 5 columns.) Otherwise, output $h$ as the class number and $R$ as the regulator.

15. [Class group] Compute the Smith normal form of $W$ using Algorithm 2.4.14. Output those among the diagonal elements $d_i$ which are greater than 1 as the invariants of the class group (i.e. $Cl(D) = \bigoplus \mathbb{Z}/d_i\mathbb{Z}$) and terminate the algorithm.

The real GCD algorithm is copied on the ordinary Euclidean algorithm, as follows. We use in an essential way that the regulator is bounded from below (by 1 for real quadratic fields of discriminant greater than 8) so as to have a reasonable stopping criterion. Since we will also use it for general number fields, we use 0.2 as a lower bound of the regulators of all number fields (see [Zim1], [Fri]).

**Algorithm 5.9.3** (Real GCD). Given two non-negative real numbers $a$ and $b$ which are known to be approximate integer multiples of some positive real number $R > 0.2$, this algorithm outputs the real GCD (RGCD) of $a$ and $b$, i.e. a non-negative real number $d$ which is an approximate integer multiple of $R$ and divisor of $a$ and $b$, and is the largest with this property. The algorithm also outputs an estimate on the absolute error for $d$.

1. [Finished?] If $b < 0.2$, then output $a$ as the RGCD, and $b$ as the absolute error and terminate the algorithm.
2. [Euclidean step] Let $r \leftarrow a - b\lfloor a/b \rfloor$, $a \leftarrow b$, $b \leftarrow r$ and go to step 1.

**Remarks.**

(1) It should be noted that not only does Algorithm 5.9.2 compute the class number and class group in sub-exponential time, but it is the only algorithm which is able to compute the regulator in sub-exponential time, even if we are not interested in the class number. In fact, in all the preceding algorithms, we first had to compute the regulator (for example using the infrastructure Algorithm 5.8.5), and combining this with the analytic class number formula giving the product $h(D)R(D)$, we could then embark on the computation of $h(D)$ and $Cl(D)$. The present algorithm goes the other way: we can in fact compute a small multiple of the class number alone, without using distances at all, and then compute the distances and the regulator, and at that point use the analytic class number formula to check that we have the correct regulator and class number, and not multiples.

(2) In an actual implementation of this algorithm, one should keep track of the absolute error of each real number. First, in the distance computation in step 7, the precision with which the computations are done gives a bound on the absolute error. Then, during steps 10 and 11, $\mathbb{Z}$-linear combinations of distances will be computed, and the errors updated accordingly (with suitable absolute value signs everywhere). Finally, in the last part of step 11 where real GCD's are computed, one should use the errors output by Algorithm 5.9.3.

(3) Essentially all the implementation details given for Algorithm 5.5.2 apply also here.

## 5.10 The Cohen-Lenstra Heuristics

The purpose of this section is to explain a number of observations which have been made on tables of class groups and regulators of quadratic fields. As already mentioned very few theorems exist (in fact essentially only the theorem of Brauer-Siegel and the theorem of Goldfeld-Gross-Zagier) so most of the explanations will be conjectural. These conjectures are however based on solid heuristic grounds so they may well turn out to be correct. As usual, we first start with imaginary quadratic fields.

### 5.10.1 Results and Heuristics for Imaginary Quadratic Fields

In this subsection $K$ will denote the unique imaginary quadratic field of discriminant $D < 0$. As we have seen, the only problem here is the behavior of the class group $Cl(D)$ and hence of the class number $h(D)$, all other basic problems being trivial to solve.

Here the Brauer-Siegel theorem says that $\ln(h(D)) \sim \ln(\sqrt{|D|})$ as $D \to -\infty$, which shows that $h(D)$ tends to infinity at least as fast as $|D|^{1/2-\varepsilon}$ and at most as fast as $|D|^{1/2+\varepsilon}$ for every $\varepsilon > 0$. The main problem is that this is not effective in a very strong sense, and this is why one has had to wait for the Gross-Zagier result to get any kind of effective result, and a very weak one at that since using their methods one can show only that

$$h(D) > \frac{1}{K} \ln(|D|) \prod_{p|D}^{*} \left(1 - \frac{2\sqrt{p}}{p+1}\right),$$

where $K = 55$ if $(D, 5077) = 1$ and $K = 7000$ otherwise, and the star indicates that the product is taken over all prime divisors $p$ of $D$ with the exception of the largest prime divisor (see [Oes]). This is of course much weaker than the Brauer-Siegel theorem.

Results in the other direction are much easier. For example, one can show that for all $D < -4$, we have

$$h(D) < \frac{1}{\pi} \sqrt{|D|} \ln(|D|)$$

(see Exercise 27). Similarly, it is very easy to obtain *average* results, which were known since Gauss. The result is as follows (see [Ayo]).

$$\sum_{|D| \le x} h(D) \sim \frac{x^{3/2}}{3\pi} C$$

where the sum runs over fundamental discriminants and

$$C = \prod_{p} \left(1 - \frac{1}{p^2(p+1)}\right) \approx 0.881538397.$$

Since by Exercise 1 the number of fundamental discriminants up to $x$ is asymptotic to $(3/\pi^2)x$, this shows that on average, $h(D)$ behaves as $C\pi/6\sqrt{|D|} \approx 0.461559\sqrt{|D|}$, and shows that the upper bound given for $h(D)$ is at most off by a factor $O(\ln(D))$.

All the above results deal with the size of $h(D)$. If we consider problems concerning its arithmetic properties (for example divisibility by small primes) or properties of the class group $Cl(D)$ itself, very little is known. If we make however the heuristic assumption that class groups behave as random groups except that they must be weighted by the inverse of the number of their automorphisms (this is a very common weighting factor in mathematics), then it is possible to make precise quantitative predictions about class numbers and class groups. This was done by H. W. Lenstra and the author in [Coh-Len1]. We summarize here some of the conjectures which are obtained in this way and which are well supported by numerical evidence.

It is quite clear that the prime 2 behaves in a special way, so we exclude it from the class group. More precisely, we will denote by $Cl_o(D)$ the odd part of the class group, i.e. the subgroup of elements of odd order. We then have the following conjectures.

**Conjecture 5.10.1** (Cohen-Lenstra). *For any odd prime $p$ and any integer $r$ including $r = \infty$ set $(p)_r = \prod_{1 \le k \le r}(1 - p^{-k})$, and let $A = \prod_{k \ge 2} \zeta(k) \approx 2.29486$, where $\zeta(s)$ is the ordinary Riemann zeta function.*

(1) *The probability that $Cl_o(D)$ is cyclic is equal to*

$$\zeta(2)\zeta(3)/(3(2)_\infty A\zeta(6)) \approx 0.977575.$$

(2) *If $p$ is an odd prime, the probability that $p \mid h(D)$ is equal to*

$$f(p) = 1 - (p)_\infty = \frac{1}{p} + \frac{1}{p^2} - \frac{1}{p^5} - \cdots$$

*For example, $f(3) \approx 0.43987$, $f(5) \approx 0.23967$, $f(7) \approx 0.16320$.*

(3) *If $p$ is an odd prime, the probability that the $p$-Sylow subgroup of $Cl(D)$ is isomorphic to a given finite Abelian $p$-group $G$ is equal to $(p)_\infty/|\operatorname{Aut}(G)|$, where $\operatorname{Aut}(G)$ denotes the group of automorphisms of $G$.*

(4) *If $p$ is an odd prime, the probability that the $p$-Sylow subgroup of $Cl(D)$ has rank $r$ (i.e. is isomorphic to a product of $r$ cyclic groups) is equal to $p^{-r^2}(p)_\infty/((p)_r)^2$.*

These conjectures explain the following qualitative observations which were made by studying the tables.

(1) The odd part of the class group is quite rarely non-cyclic. In fact, it was only in the sixties that the first examples of class groups with 3-rank greater or equal to 3 were discovered.

(2) Higher ranks are even more difficult to find, and the present record for $p = 3$, due to Quer (see [Llo-Quer] and [Quer]) is 3-rank equal to 6. Note that there is a very interesting connection with elliptic curves of high rank over $\mathbb{Q}$ (see Chapter 7), and Quer's construction indeed gives curves of rank 12.

(3) If $p$ is a small odd prime, the probability that $p \mid h(D)$ is substantially higher than the expected naïve value $1/p$. Indeed, it should be very close to $1/p + 1/p^2$.

### 5.10.2 Results and Heuristics for Real Quadratic Fields

Because of the presence of non-trivial units, the situation in this case is completely different and even less understood than the imaginary quadratic case. Here the Brauer-Siegel theorem tells us that $\ln(R(D)h(D)) \sim \ln(\sqrt{D})$ as $D \to \infty$, where $R(D)$ is the regulator. Unfortunately, we have little control on $R(D)$, and this is the main source of our ignorance about real quadratic fields. It is conjectured that $R(D)$ is "usually" of the order of $\sqrt{D}$, hence that $h(D)$ is usually very small, and this is what the tables show. For example, there should exist an infinite number of $D$ such that $h(D) = 1$, but this is not known to be true and is a famous conjecture. In fact, it is not even known whether there exists an infinite number of non-isomorphic number fields $K$ (all degrees taken together) with class number equal to one.

As in the imaginary case however, we can give an upper bound $h(D) < \sqrt{D}$ when $D > 0$, and the following average for $R(D)h(D)$:

$$\sum_{D \le x} R(D)h(D) \sim \frac{x^{3/2}}{6} C$$

where the sum runs over fundamental discriminants and the constant $C$ is as before.

It is possible to generalize the heuristic method used in the imaginary case. In fact, we could reinterpret Shanks's infrastructure idea as saying that the class group of a real quadratic field is equal to the quotient of the "group" of reduced forms by the "cyclic subgroup" formed by the principal cycle. This of course does not make any direct sense since the reduced forms form a group only in an approximate sense, and similarly for the principal cycle. It suggests however that we could consider the (odd part) of the class group of a real quadratic field as the quotient of a random finite Abelian group of odd order (weighted as before) by a random cyclic subgroup. This indeed works out very well and leads to the following conjectures.

**Conjecture 5.10.2** (Cohen-Lenstra). *Let $D$ be a positive fundamental discriminant.*

(1) *If $p$ is an odd prime, the probability that $p \mid h(D)$ is equal to*

$$1-\frac{(p)_\infty}{1-1/p}=\frac{1}{p^2}+\frac{1}{p^3}+\frac{1}{p^4}-\cdots$$

(2) *The probability that $Cl_o(D)$ is isomorphic to a given finite Abelian group $G$ of odd order $g$ is equal to $m(G)=1/(2g(2)_\infty A|\operatorname{Aut}(G)|)$. For example $m(\{0\})\approx 0.75446$, $m(\mathbb{Z}/3\mathbb{Z})\approx 0.12574$, $m(\mathbb{Z}/5\mathbb{Z})\approx 0.03772$.*

(3) *If $p$ is an odd prime, the probability that the $p$-Sylow subgroup of $Cl(D)$ has rank $r$ is equal to $p^{-r(r+1)}(p)_\infty/((p)_r(p)_{r+1})$.*

(4) *We have*

$$\sum_{p\le x} h(p)\sim\frac{x}{8},$$

*where the sum runs over primes congruent to 1 modulo 4.*

These conjectures explain in particular the experimental observation that most quadratic fields of prime discriminant $p$ (in fact more than three fourths) have class number one.

These heuristic conjectures have been generalized to arbitrary number fields by J. Martinet and the author (see [Coh-Mar1], [Coh-Mar2]). Note that contrary to what was claimed in these papers, apparently all the primes dividing the degree of the Galois closure should be considered as non-random (see [Coh-Mar3]), hence the numerical values given in [Coh-Mar1] should be corrected accordingly (e.g. by removing the 2-part for non-cyclic cubic fields or the 3-part for quartic fields of type $A_4$ or $S_4$).

## 5.11 Exercises for Chapter 5

1. Show that the number of imaginary quadratic fields with discriminant $D$ such that $|D|\le x$ is asymptotic to $3x/\pi^2$, and similarly for real quadratic fields.
2. Compute the probability that the discriminant of a quadratic field is divisible by a given prime number $p$ (beware: the result is not what you may expect).
3. Complete Theorem 5.2.9 by giving explicitly the correspondences between ideal classes, classes of quadratic forms and classes of quadratic numbers, at the level of $\mathrm{PSL}_2(\mathbb{Z})$.
4. Let $K$ be a quadratic field and $p$ a prime. Generalizing Theorem 1.4.1, find the structure of the multiplicative group $(\mathbb{Z}_K/p\mathbb{Z}_K)^*$, and in particular compute its cardinality.
5. (H.W. Lenstra and D. Knuth) Let $D$ denote the discriminant of an imaginary quadratic field. If $x\ge 0$, let $f(x,D)$ be the probability that a quadratic form $(a,b,c)$ with $-a<b\le a$ and $a<x\sqrt{|D|}$ is reduced. From Lemma 5.3.4, we know that $f(x,D)=1$ if $x\le 1/2$ and $f(x,D)=0$ if $x\ge 1/\sqrt{3}$. Show that $f(x,D)$ has a limit $f(x)$ as $|D|\to\infty$, and give a closed formula for $f(x)$, assuming that a quadratic number behaves like a random irrational number. Note that this exercise is difficult, and the complete result without the randomness assumption has only recently been proved by Duke (see [Duk]).

6. If $D_0$ is a fundamental negative discriminant and $D = D_0 f^2$, show directly from the formula given in the text that $h(D_0) \mid h(D)$.

7. Let $p$ be a prime number such that $p \equiv 3 \pmod 4$. Using Dirichlet's class number formula (Corollary 5.3.13) express $h(-p)$ as a function of

$$\sum_{1 \le n \le (p-1)/2} \left\lfloor \frac{n^2}{p} \right\rfloor .$$

Is this algorithmically better than Dirichlet's formula?

8. Carry out in detail the GCD computations of the proof of Lemma 5.4.5.

9. Show that the composite of two primitive forms is primitive, and also that primitivity is preserved under reduction (both for complex quadratic fields and real ones). Prove these results first using the interpretation in terms of ideals, then directly on the formulas.

10. Show that, in order to generalize Algorithm 5.4.7 to imprimitive forms, we can replace the assignment $v_1 \leftarrow a_1/d_1$ of Step 4 by $v_1 \leftarrow \gcd(d_1, c_1, c_2, n)a_1/d_1$.

11. Let $A$, $B$ and $C$ be integers, and assume that at most one of them is equal to zero. Show that the general integral solution to the equation

$$uA + vB + wC = 0$$

is given by

$$u = \frac{B}{(A,B)}\nu - \frac{C}{(A,C)}\mu, \quad v = \frac{C}{(B,C)}\lambda - \frac{A}{(A,B)}\nu, \quad w = \frac{A}{(A,C)}\mu - \frac{B}{(B,C)}\lambda$$

where $\lambda$, $\mu$ and $\nu$ are arbitrary integers.

12. Using the preceding exercise, show that as claimed after Definition 5.4.6 the class of $(a_3, b_3, c_3)$ modulo $\Gamma_\infty$ is well defined.

13. In step 9 of Algorithm 5.5.2, it is suggested to compute the determinant of the lattice generated by the columns of a rectangular matrix $A_1$ of full rank by computing this determinant modulo $p$ and using the Chinese remainder theorem together with Hadamard's inequality. Show that it is possible to modify the Gauss-Bareiss Algorithm 2.2.6 so as to compute this determinant directly, and compare the efficiency of the two methods, in theory as well as in practice (in the author's experience, the direct method is usually superior). Hint: use flags $c_k$ and/or $d_k$ as in Algorithm 2.3.1.

14. Implement the large prime variation explained after Algorithm 5.5.2 in the following manner. Choose some integer $k$ (say $k = 500$) and use $k$ lists of quadratic forms as follows. Each time that some $p_a$ is encountered, we store $p_a$ and the corresponding quadratic form in the $n$-th list, where $n = p_a \bmod k$. If $p_a$ is already in the list, we have a relation, otherwise we do nothing else. Study the efficiency of this method and the choice of $k$. (Note: this method is a special case of a well known method used in computer science called *hashing*, see [Knu3].)

15. Implement Atkin's variant of McCurley's algorithm assuming that the discriminant $D$ is a prime number and that the order of $f$ is larger than the bound given by the Euler product.

16. Let $\mathfrak{a}$ be an integral ideal in a number field $K$, $\ell(\mathfrak{a})$ the smallest positive rational integer belonging to $\mathfrak{a}$, and $\sigma_i$ the embeddings of $K$ into $\mathbb{C}$. We will say that $\mathfrak{a}$ is *reduced* if $\mathfrak{a}$ is primitive and if the conditions $\alpha \in \mathfrak{a}$ and for all $i$, $|\sigma_i(\alpha)| < \ell(\mathfrak{a})$ imply that $\alpha = 0$.

    a) If $(\mathfrak{a}, s) = \phi_{FI}(a, b, c)$, show that $\mathfrak{a}$ is reduced if and only if there exists a (unique) quadratic form in the $\Gamma_\infty$-class of $(a, b, c)$ which is reduced. (Since the cases $K$ real and imaginary must be treated separately, this is in fact two exercises in one.)

    b) In the case where $K = \mathbb{Q}(\sqrt{D})$ is a real quadratic field, show that $\mathfrak{a}$ is reduced if and only if there exists integers $a_1$ and $a_2$ such that $a_1 \equiv a_2 \equiv b \pmod{2a}$, $0 < a_1 < \sqrt{D}$ and $-\sqrt{D} < a_2 < 0$.

    c) Let $\mathfrak{a}$ be an ideal in the number field $K$. Show that there exists an $\alpha \in \mathfrak{a}$ such that $|\sigma_i(\beta)| < |\sigma_i(\alpha)|$ for all $i$ implies that $\beta = 0$. By considering the ideal $(d/\alpha)\mathfrak{a}$ for a suitable integer $d$, deduce from this that, as in the quadratic case, every ideal is equivalent to a (not necessarily unique) reduced ideal.

17. Show that in any cycle of reduced quadratic forms of discriminant $D > 0$, there exists a form $(a, b, c)$ with $|a| \leq \sqrt{D/5}$. In other words, show that in any ideal class there exists an ideal $\mathfrak{a}$ such that $\mathcal{N}(\mathfrak{a}) \leq \sqrt{D/5}$. (Hint: use Theorem 454 in [H-W].)

18. Prove Proposition 5.6.1.

19. Using Definition 4.9.11 and Proposition 5.1.4, show that if $K$ is a (real or imaginary) quadratic field of discriminant $D$ we have $\zeta_K(s) = \zeta(s)L_D(s)$, and hence that Propositions 5.3.12 and 5.6.9 are special cases of Dedekind's Theorem 4.9.12.

20. Modify Algorithm 5.7.1 so that it is still valid for $a < 0$.

21. Prove the following precise form of Lemma 5.7.6. If $p$ and $q$ are coprime integers, denote by $p'$ the inverse of $p$ modulo $q$ such that $1 \leq p' \leq q$. Let $\alpha$ be a real number. Then $p/q$ is a convergent in the continued fraction expansion of $\alpha$ if and only if
$$-\frac{1}{q(q+p')} < \alpha - \frac{p}{q} < \frac{1}{q(2q-p')}.$$

22. Show that the period of the continued fraction expansion of the quadratic irrational corresponding to the inverse of a reduced quadratic form $f$ of positive discriminant is the reverse of the period of the quadratic number corresponding to $f$. Conclude that for ambiguous forms, the period is symmetric.

23. Write an algorithm corresponding to Algorithm 5.7.2 as Algorithm 5.7.10 corresponds to Algorithm 5.7.1 for computing the regulator of a real quadratic field using the symmetry of the period when we start with the unit form instead of any reduced form.

24. Assume that one has computed the regulator of a real quadratic field using the method explained in Section 5.9 to a given precision which need not be very high. Show that one can then compute the regulator to any desired accuracy in a small extra amount of time (hint: using the distance function, we now know where to look in the cycle).

25. Similarly to the preceding exercise, show that one can also compute the $p$-adic regulator to any desired accuracy in a small extra amount of time.

26. Let $D$ be a fundamental discriminant.

a) Show that $h^+(D)R^+(D) = 2h(D)R(D)$ and that $R^+(D) = 2R(D)$ if and only if the fundamental unit is of norm equal to $-1$.

b) What modifications can be made to Algorithm 5.9.2 so that it computes the regulator and the class number in the narrow sense?

27. Let $D < -4$ be a fundamental discriminant, and set $f = |D|$.

a) Set $s(x) = \sum_{1 \le n \le x} \left(\frac{D}{n}\right)$. Show that $|s(x)| \le f/2$ and by Abel summation that $|\sum_{n>f} \left(\frac{D}{n}\right)/n| < 1/2$.

b) Show that $h(D) < \dfrac{1}{\pi}\sqrt{f}\ln f$.

c) Using the Polya-Vinogradov inequality (see Exercise 8 of Chapter 9), give a better explicit upper bound for $h(D)$, asymptotic to $\frac{1}{2\pi}\sqrt{f}\ln f$.

28. (S. Louboutin) Using again the function $s(x)$ defined in Exercise 27 and Abel summation, show that we can avoid the computation of the function $\operatorname{erfc}(x)$ in Proposition 5.3.14 using the fact that $h(D)$ is an integer whose parity can be computed in advance ($h(D)$ is odd if and only if $D = -4$, $D = -8$ or $D = -p$ where $p$ is a prime congruent to 3 modulo 4). Apply a similar method in Proposition 5.6.11.

# Chapter 6

# Algorithms for Algebraic Number Theory II

We now leave the realm of quadratic fields where the main computational tasks of algebraic number theory mentioned at the end of Chapter 4 were relatively simple (although as we have seen many conjectures remain), and move on to general number fields.

We first discuss practical algorithms for computing an integral basis and for the decomposition of primes in a number field $K$, essentially following a paper of Buchmann and Lenstra [Buc-Len], except that we avoid the explicit use of Artinian rings. We then discuss algorithms for computing Galois groups (up to degree 7, but see also Exercise 15). As examples of number fields of higher degree we then treat cyclic and pure cubic fields. Finally, in the last section of this chapter, we give a complete algorithm for class group and regulator computation which is sufficient for dealing with fields having discriminants of reasonable size. This algorithm also gives a system of fundamental units if desired.

## 6.1 Computing the Maximal Order

Let $K = \mathbb{Q}[\theta]$ be a number field, where $\theta$ is a root of a monic polynomial $T(X) \in \mathbb{Z}[X]$. Recall that $\mathbb{Z}_K$ has been defined as the set of algebraic integers belonging to $K$, and that it is called the maximal order since it is an order in $K$ containing every order of $K$. We will build it up by starting from a known order (in fact from $\mathbb{Z}[\theta]$) and by successively enlarging it.

### 6.1.1 The Pohst-Zassenhaus Theorem

The main tool that we will use for enlarging an order is the Pohst-Zassenhaus Theorem 6.1.3 below. We first need a few basic results and definitions.

**Definition 6.1.1.** *Let $\mathcal{O}$ be an order in a number field $K$ and let $p$ be a prime number.*

(1) *We will say that $\mathcal{O}$ is* $p$-maximal *if $[\mathbb{Z}_K : \mathcal{O}]$ is not divisible by $p$.*
(2) *We define the* $p$-radical $I_p$ *as follows.*

$$I_p = \{x \in \mathcal{O} \mid \exists m \geq 1 \text{ such that } x^m \in p\mathcal{O}\}$$

**Proposition 6.1.2.** *Let $\mathcal{O}$ be an order in a number field $K$ and let $p$ be a prime number.*

(1) *The $p$-radical $I_p$ is an ideal of $\mathcal{O}$.*
(2) *We have*

$$I_p = \prod_{1 \le i \le g} \mathfrak{p}_i$$

*the product being over all distinct prime ideals $\mathfrak{p}_i$ of $\mathcal{O}$ which lie above $p$.*
(3) *There exists an integer $m$ such that $I_p^m \subset p\mathcal{O}$.*

*Proof.* For (1), the only thing which is not completely trivial is that $I_p$ is stable under addition. If $x^m \in p\mathcal{O}$ and $y^n \in p\mathcal{O}$, then clearly $(x+y)^{n+m} \in p\mathcal{O}$ as we see by using the binomial theorem.

For (2) note that since $\mathfrak{p}_i$ lies above $p$ then $p\mathcal{O} \subset \mathfrak{p}_i$. So, if $x \in I_p$ there exists an $m$ such that $x^m \in p\mathcal{O} \subset \mathfrak{p}_i$, and hence $x \in \mathfrak{p}_i$ by definition of a prime ideal. By Proposition 4.6.4 this shows that $x \in \bigcap_{1 \le i \le g} \mathfrak{p}_i = \prod_{1 \le i \le g} \mathfrak{p}_i$ since the distinct maximal ideals $\mathfrak{p}_i$ are pairwise coprime.

Conversely, assume that $x \in \prod_{1 \le i \le g} \mathfrak{p}_i$. By definition, the set of ideals of $\mathcal{O}$ containing $p\mathcal{O}$ is in canonical one-to-one correspondence with the ideals of the finite quotient ring $R = \mathcal{O}/p\mathcal{O}$. We will use this at length later. For now, note that it implies that this set is finite, and in particular the ideals $\alpha^n R$ are finite in number, where $\alpha$ is the class of $x$ in $R$. In particular, there exists an $n$ such that $\alpha^n R = \alpha^{n+1} R$, i.e. $\alpha^n(1 - \alpha\beta) = 0$ for some $\beta \in R$. By assumption, $\alpha$ belongs to all the maximal ideals $\overline{\mathfrak{p}_i}$ of $R$ hence $(1 - \alpha\beta)$ cannot belong to any of them, otherwise 1 would also, which is impossible. It follows that the ideal $(1 - \alpha\beta)R$, not being contained in any maximal ideal, must be equal to $R$, i.e. $1 - \alpha\beta$ is invertible $R$. The equality $\alpha^n(1 - \alpha\beta) = 0$ thus implies that $\alpha^n = 0$ in $R$, i.e. that $x^n \in p\mathcal{O}$ or again that $x \in I_p$ as was to be proved.

Finally, for (3) note that since $I_p$ is an ideal of an order in a number field it has a finite $\mathbb{Z}$-basis $x_i$ for $1 \le i \le n$. For each $x_i$ there exists an $m_i$ such that $x_i^{m_i} \in p\mathcal{O}$, and if we set $m = \sum_{1 < i < n} m_i$ it is clear that $I_p^m \subset p\mathcal{O}$, using this time the multinomial theorem instead of the binomial theorem. □

The procedure that we will use to obtain the maximal order is to start with $\mathcal{O} = \mathbb{Z}[\theta]$ and enlarge it for successive primes so as to get an order which is $p$-maximal for every $p$, hence which will be the maximal order. The enlarging procedure which we will use, due to Pohst and Zassenhaus, is based on the following theorem.

**Theorem 6.1.3.** *Let $\mathcal{O}$ be an order in a number field $K$ and let $p$ be a prime number. Set*

$$O' = \{x \in K | xI_p \subset I_p\}.$$

*Then either $\mathcal{O}' = \mathcal{O}$, in which case $\mathcal{O}$ is $p$-maximal, or $\mathcal{O}' \supsetneq \mathcal{O}$ and $p \mid [\mathcal{O}' : \mathcal{O}] \mid p^n$.*

*Proof.* Since $I_p$ is an ideal, it is clear that $\mathcal{O}'$ is a ring containing $\mathcal{O}$. Furthermore, since $p \in I_p$, $x \in \mathcal{O}'$ implies that $xp \in I_p \subset \mathcal{O}$ and hence $\mathcal{O} \subset \mathcal{O}' \subset \frac{1}{p}\mathcal{O}$. This shows that $\mathcal{O}'$ has maximal rank, i.e. is an order in $K$, and it also shows that $[\mathcal{O}' : \mathcal{O}] | p^n$.

We now assume that $\mathcal{O}' = \mathcal{O}$. Define

$$\mathcal{O}_p = \{x \in \mathbb{Z}_K | \exists j \geq 1,\ p^j x \in \mathcal{O}\}.$$

It is clear that $\mathcal{O} \subset \mathcal{O}_p$ and that $\mathcal{O}_p$ is an order. Furthermore, $\mathcal{O}_p$ is $p$-maximal. Indeed, if $p$ divides the index $[\mathbb{Z}_K : \mathcal{O}_p]$, then there exists $x \in \mathbb{Z}_K$ such that $x \notin \mathcal{O}_p$ but $px \in \mathcal{O}_p$. The definition of $\mathcal{O}_p$ shows that this is impossible.

We are now going to show that $\mathcal{O}_p = \mathcal{O}$. Since $\mathcal{O}_p$ is an order, it is finitely generated over $\mathbb{Z}$. Hence there exists an $r \geq 1$ such that $p^r \mathcal{O}_p \subset \mathcal{O}$ (take $r$ to be the maximum of the $j$ such that $p^j x_i \in \mathcal{O}$ for a finite generating set $(x_i)$ of $\mathcal{O}_p$). Since $I_p^m \subset p\mathcal{O}$ it follows that $\mathcal{O}_p I_p^{mr} \subset \mathcal{O}$. Assume by contradiction that $\mathcal{O}_p \neq \mathcal{O}$, hence $\mathcal{O}_p \not\subset \mathcal{O}$. Let $n$ be the largest index such that $\mathcal{O}_p I_p^n \not\subset \mathcal{O}$ (hence $n$ exists and $0 \leq n < mr$). We thus also have $\mathcal{O}_p I_p^{n+1} \subset \mathcal{O}$. Choose any $x \in \mathcal{O}_p I_p^n \setminus \mathcal{O}$. Then $xI_p \subset \mathcal{O}$. Since $\mathcal{O}_p I_p^{n+m+1} \subset I_p^m \subset p\mathcal{O}$ it follows that if $y \in I_p$, then $(xy)^{n+m+1} \in p\mathcal{O}$ hence that $xy \in I_p$, so $xI_p \subset I_p$ thus showing that $x \in \mathcal{O}'$. This is a contradiction since $x \notin \mathcal{O}$ and we have assumed that $\mathcal{O}' = \mathcal{O}$. This finishes the proof of Theorem 6.1.3. □

(I thank D. Bernardi for the final part of the proof.)

### 6.1.2 The Dedekind Criterion

From the Pohst-Zassenhaus theorem, starting from a number field $K = \mathbb{Q}(\theta)$ defined by a monic polynomial $T \in \mathbb{Z}[X]$, we will enlarge the order $\mathbb{Z}[\theta]$ for every prime $p$ such that $p^2$ divides the discriminant of $T$ until we obtain an order which is $p$-maximal for every $p$, i.e. the maximal order. In practice however, even when the discriminant has square factors, $\mathbb{Z}[\theta]$ is quite often $p$-maximal for a number of primes $p$, and it is time consuming to have to compute $\mathcal{O}'$ as in Theorem 6.1.3 just to notice that $\mathcal{O}' = \mathbb{Z}[\theta]$, i.e. that $\mathbb{Z}[\theta]$ is $p$-maximal. Fortunately, there is a simple and important criterion due to Dedekind which allows us to decide, without the more complicated computations explained in the next section, whether $\mathbb{Z}[\theta]$ is $p$-maximal or not for prime numbers $p$, and if it is not, it will give us a larger order, which of course may still not be $p$-maximal.

It must be emphasized that this will work *only* for $\mathbb{Z}[\theta]$, or for any order $\mathcal{O}$ containing $\mathbb{Z}[\theta]$ with $[\mathcal{O} : \mathbb{Z}[\theta]]$ prime to $p$, but not for an order which has already been enlarged for the prime $p$ itself.

This being said the basic theorem that we will prove, of which Dedekind's criterion is a special case, is as follows.

**Theorem 6.1.4** (Dedekind). *Let $K = \mathbb{Q}(\theta)$ be a number field, $T \in \mathbb{Z}[X]$ the monic minimal polynomial of $\theta$ and let $p$ be a prime number. Denote by $^-$ reduction modulo $p$ (in $\mathbb{Z}$, $\mathbb{Z}[X]$ or $\mathbb{Z}[\theta]$). Let*

$$\overline{T}(X) = \prod_{i=1}^{k} \overline{t}_i(X)^{e_i}$$

*be the factorization of* $T(X)$ *modulo* $p$ *in* $\mathbb{F}_p[X]$, *and set*

$$g(X) = \prod_{i=1}^{k} t_i(X)$$

*where the* $t_i \in \mathbb{Z}[X]$ *are arbitrary monic lifts of* $\overline{t}_i$. *Then*

(1) *The* $p$*-radical* $I_p$ *of* $\mathbb{Z}[\theta]$ *at* $p$ *is given by*

$$I_p = p\mathbb{Z}[\theta] + g(\theta)\mathbb{Z}[\theta].$$

*In other words,* $x = A(\theta) \in I_p$ *if and only if* $\overline{g} \mid \overline{A}$.

(2) *Let* $h(X) \in \mathbb{Z}[X]$ *be a monic lift of* $\overline{T}(X)/\overline{g}(X)$ *and set*

$$f(X) = (g(X)h(X) - T(X))/p \in \mathbb{Z}[X].$$

*Then* $\mathbb{Z}[\theta]$ *is* $p$*-maximal if and only if*

$$(\overline{f}, \overline{g}, \overline{h}) = 1 \quad \text{in } \mathbb{F}_p[X].$$

(3) *More generally, let* $\mathcal{O}'$ *be the order given by Theorem 6.1.3 when we start with* $\mathcal{O} = \mathbb{Z}[\theta]$. *Then, if* $U$ *is a monic lift of* $\overline{T}/(\overline{f}, \overline{g}, \overline{h})$ *to* $\mathbb{Z}[X]$ *we have*

$$\mathcal{O}' = \mathbb{Z}[\theta] + \frac{1}{p}U(\theta)\mathbb{Z}[\theta]$$

*and if* $m = \deg(\overline{f}, \overline{g}, \overline{h})$, *then* $[\mathcal{O}' : \mathbb{Z}[\theta]] = p^m$, *hence* $\operatorname{disc}(O') = \operatorname{disc}(T)/p^{2m}$.

*Proof of (1).* $p \in I_p$ trivially, and since the exponents $e_i$ are at most equal to $n = [K : \mathbb{Q}] = \deg(T)$, we have $\overline{T} \mid \overline{g}^n$ hence $g^n(\theta) \equiv 0 \pmod{p\mathbb{Z}[\theta]}$ so $g(\theta) \in I_p$, thus proving that $I_p \supset p\mathbb{Z}[\theta] + g(\theta)\mathbb{Z}[\theta]$.

Now the minimal polynomial over $\mathbb{F}_p$ of $\theta$ in $\mathbb{Z}[\theta]/p\mathbb{Z}[\theta]$ (which is not a field in general) is clearly the polynomial $\overline{T}$. Indeed, it clearly divides $\overline{T}$, but it is of degree at least $n$ since $1, \theta, \ldots, \theta^{n-1}$ are $\mathbb{F}_p$-linearly independent.

Conversely let $x \in I_p$. Then $x = A(\theta)$ for $A \in \mathbb{Z}[X]$, and so there exists an integer $m$ such that $x^m \equiv 0 \pmod{p\mathbb{Z}[\theta]}$, in other words $\overline{A}^m(\theta) = 0$ in $\mathbb{Z}[\theta]/p\mathbb{Z}[\theta]$. Hence $\overline{T} \mid \overline{A}^m$. Since $e_i \geq 1$ for all $i$, this implies that $\overline{t}_i \mid \overline{A}^m$ hence $\overline{t}_i \mid \overline{A}$ since $t_i$ is irreducible in $\mathbb{F}_p[X]$, and since the $\overline{t}_i$ are pairwise coprime, we get $\overline{g} \mid \overline{A}$ which means that $x \in p\mathbb{Z}[\theta] + g(\theta)\mathbb{Z}[\theta]$ thus proving (1).

Since $\overline{T}$ is the minimal polynomial of $\theta$ in $\mathbb{Z}[\theta]/p\mathbb{Z}[\theta]$, it is clear that (2) follows from (3).

Let us now prove (3). Recall that $\mathcal{O}' = \{x \in K | xI_p \subset I_p\}$. From (1) we have that $x \in \mathcal{O}'$ if and only if $xp \in I_p$ and $xg(\theta) \in I_p$. Since $I_p \subset \mathbb{Z}[\theta]$, $xp \in I_p$ implies that

$$x = A_1(\theta)/p$$

where $A_1 \in \mathbb{Z}[X]$. Part (3) of the theorem will immediately follow from the following lemma.

**Lemma 6.1.5.** *Let $x = A_1(\theta)/p$ with $A_1 \in \mathbb{Z}[X]$. Then*

(1) *$xp \in I_p$ if and only if*

$$\overline{g} \mid \overline{A}_1 .$$

(2) *Let $\overline{k} = \overline{g}/(\overline{f}, \overline{g})$, where (here as elsewhere in this section) $k$ is implicitly considered to be a monic lift of $\overline{k}$ to $\mathbb{Z}[X]$. Then $xg(\theta) \in I_p$ if and only if*

$$\overline{hk} \mid \overline{A}_1 .$$

*Proof of the Lemma.* Part (1) of the lemma is an immediate consequence of part (1) of the theorem. Let us prove part (2).

From part (1) of the theorem, $xg(\theta) \in I_p$ if and only if there exist polynomials $A_2$ and $A_3$ in $\mathbb{Z}[X]$ such that

$$A_1(\theta)g(\theta) = p(pA_2(\theta) + g(\theta)A_3(\theta)),$$

and since $T$ is the minimal polynomial of $\theta$, this is true if and only if there exists $A_4 \in \mathbb{Z}[X]$ such that

$$A_1(X)g(X) = p^2A_2(X) + pg(X)A_3(X) + A_4(X)T(X).$$

For the rest of this proof, we will work only with polynomials (in $\mathbb{Z}[X]$ or $\mathbb{F}_p[X]$), and not any more in $K$.

Reducing modulo $p$, the above equation implies that $\overline{A}_1 = \overline{A_4 h}$. Hence write

$$A_1 = hA_4 + pA_5$$

with $A_5 \in \mathbb{Z}[X]$. We have that $xg(\theta) \in I_p$ if and only if there exist polynomials $A_i \in \mathbb{Z}[X]$ such that

$$(gh - T)A_4 = p^2A_2 + pg(A_3 - A_5),$$

hence if and only if there exist $A_i$ such that

$$fA_4 = pA_2 + gA_6 .$$

This last condition is equivalent to $\overline{g} \mid \overline{fA_4}$ so to $\overline{k} \mid \overline{A}_4$ where $\overline{k} = \overline{g}/(\overline{f}, \overline{g})$, and this is equivalent to the existence of $A_7$ and $A_8$ in $\mathbb{Z}[X]$ such that $A_4 = kA_7 + pA_8$.

To sum up, we see that if $x = A_1(\theta)/p$, then $xg(\theta) \in I_p$ if and only if there exist polynomials $A_5$, $A_7$ and $A_8$ in $\mathbb{Z}[X]$ such that

$$A_1 = hkA_7 + p(hA_8 + A_5),$$

and this is true if and only if there exist $A_9 \in \mathbb{Z}[X]$ such that $A_1 = hkA_7 + pA_9$ or equivalently $\overline{hk} \mid \overline{A_1}$, thus proving the lemma. □

We can now prove part (3) of the theorem. From the lemma, we have that $x = A_1(\theta)/p \in \mathcal{O}'$ if and only if both $\overline{g}$ and $\overline{hk}$ divide $\overline{A_1}$ in the PID $\mathbb{F}_p[X]$, hence if and only if the least common multiple of $\overline{g}$ and $\overline{hk}$ divides $A_1$. Since in any PID, $\operatorname{lcm}(x, y) = xy/(x, y)$ and $\operatorname{lcm}(zx, zy) = z \operatorname{lcm}(x, y)$, we have

$$\operatorname{lcm}(\overline{g}, \overline{hk}) = \overline{k} \operatorname{lcm}(\gcd(\overline{f}, \overline{g}), \overline{h}) = \frac{\overline{g}}{(\overline{f}, \overline{g})} \frac{\overline{h}(\overline{f}, \overline{g})}{(\overline{f}, \overline{g}, \overline{h})} = \frac{\overline{T}}{(\overline{f}, \overline{g}, \overline{h})} = \overline{U}$$

thus proving that $\mathcal{O}' = \mathbb{Z}[\theta] + (U(\theta)/p)\mathbb{Z}[\theta]$. Now it is clear that a system of representatives of $\mathcal{O}'$ modulo $\mathbb{Z}[\theta]$ is given by $A(\theta)U(\theta)/p$ where $A$ runs over uniquely chosen representatives in $\mathbb{Z}[X]$ of polynomials in $\mathbb{F}_p[X]$ such that $\deg(A) < \deg(T) - \deg(U) = m$, thus finishing the proof of the theorem. □

An important remark is that the proof of this theorem is *local* at $p$, in other words we can copy it essentially verbatim if we everywhere replace $\mathbb{Z}[\theta]$ by any overorder $\mathcal{O}$ of $\mathbb{Z}[\theta]$ such that $[\mathcal{O} : \mathbb{Z}[\theta]]$ is coprime to $p$. The final result is then that the new order enlarged at $p$ is

$$\mathcal{O} + \frac{U(\theta)}{p}\mathcal{O},$$

and $[\mathcal{O}' : \mathcal{O}] = p^m$.

### 6.1.3 Outline of the Round 2 Algorithm

From the Pohst-Zassenhaus theorem it is easy to obtain an algorithm for computing the maximal order. We will of course use the Dedekind criterion to simplify the first steps for every prime $p$.

Let $K = \mathbb{Q}(\theta)$ be a number field, where $\theta$ is an algebraic integer. Let $T$ be the minimal polynomial of $\theta$. We can write $\operatorname{disc}(T) = df^2$, where $d$ is either 1 or a fundamental discriminant. If $\mathbb{Z}_K$ is the maximal order which we are looking for, then the index $[\mathbb{Z}_K : \mathbb{Z}[\theta]]$ has only primes dividing $f$ as prime divisors because of Proposition 4.4.4. We are going to compute $\mathbb{Z}_K$ by successive enlargements from $\mathcal{O} = \mathbb{Z}[\theta]$, one prime dividing $f$ at a time. For every $p$ dividing $f$ we proceed as follows. By using Dedekind's criterion, we check whether $\mathcal{O}$ is $p$-maximal and if it is not we enlarge it once using Theorem 6.1.4 (3) applied to $\mathcal{O}$. If the new discriminant is not divisible by $p^2$, then we

are done, otherwise we compute $\mathcal{O}'$ as described in Theorem 6.1.3. If $\mathcal{O}' = \mathcal{O}$, then $\mathcal{O}$ is $p$-maximal and we are finished with the prime $p$, so we move on to the next prime, if any. (Here again we can start using Dedekind's criterion.) Otherwise, replace $\mathcal{O}$ by $\mathcal{O}'$, and use the method of Theorem 6.1.3 again. It is clear that this algorithm is valid and will lead quite rapidly to the maximal order. This algorithm was the second one invented by Zassenhaus for maximal order computations, and so it has become known as the round 2 algorithm (the latest and most efficient is round 4).

What remains is to explain how to carry out explicitly the different steps of the algorithm, when we apply Theorem 6.1.3.

First, $\theta$ is fixed, and all ideals and orders will be represented by their upper triangular HNF as explained in Section 4.7.2. We must explain how to compute the HNF of $I_p$ and of $\mathcal{O}'$ in terms of the HNF of $\mathcal{O}$. It is simpler to compute in $R = \mathcal{O}/p\mathcal{O}$. To compute the radical of $R$, we note the following lemma:

**Lemma 6.1.6.** *If $n = [K : \mathbb{Q}]$ and if $j \geq 1$ is such that $p^j \geq n$, then the radical of $R$ is equal to the kernel of the map $x \mapsto x^{p^j}$, which is the $j^{\text{th}}$ power of the Frobenius homomorphism.*

*Proof.* It is clear that the map in question is the $j^{\text{th}}$ power of the Frobenius homomorphism, hence talking about its kernel makes sense. By definition of the radical, it is clear that this kernel is contained in the radical. Conversely, let $x$ be in the radical. Then $x$ induces a nilpotent map defined by multiplication by $x$ from $R$ to $R$, and considering $R$ as an $\mathbb{F}_p$-vector space, this means that the eigenvalues of this map in $\overline{\mathbb{F}_p}$ are all equal to 0. Hence, its characteristic polynomial must be $X^n$ (since $n = \dim_{\mathbb{F}_p} R$), and by the Cayley-Hamilton theorem this shows that $x^n = 0$, and hence that $x^{p^j} = 0$, proving the lemma. □

Let $\omega_1, \ldots, \omega_n$ be the HNF basis of $\mathcal{O}$. Then it is clear that $\overline{\omega}_1, \ldots, \overline{\omega}_n$ is an $\mathbb{F}_p$-basis of $R$. For $k = 1, \ldots, n$, we compute $\overline{a}_{i,k}$ such that

$$\overline{\omega}_k^{p^j} = \sum_{i=1}^{n} \overline{a}_{i,k}\overline{\omega}_i,$$

the left hand side being computed as a polynomial in $\theta$ by the standard representation algorithms, and the coefficients $\overline{a}_{i,k}$ being easily found inductively since an HNF matrix is triangular. Hence, if $\overline{A}$ is the matrix of the $\overline{a}_{i,k}$, the radical is simply the kernel of this matrix.

Hence, if we apply Algorithm 2.3.1, we will obtain a basis of $\overline{I}_p$, the radical of $R$, in terms of the standard representation. Since $I_p$ is generated by pull-backs of a basis of $\overline{I}_p$ and $p\omega_1, \ldots, p\omega_n$, to obtain the HNF of $I_p$ we apply the HNF reduction algorithm to the matrix whose columns are the standard representations of these elements.

Now that we have $I_p$, we must compute $\mathcal{O}'$. For this, we use the following lemma:

**Lemma 6.1.7.** *With the notations of Theorem 6.1.3, if $U$ is the kernel of the map*

$$\alpha \longmapsto (\overline{\beta} \mapsto \overline{\alpha\beta})$$

*from $\mathcal{O}$ to $\mathrm{End}(I_p/pI_p)$, then $\mathcal{O}' = \frac{1}{p}U$.*

*Proof.* Trivial and left to the reader. Note that $\mathrm{End}(I_p/pI_p)$ is considered as a $\mathbb{Z}$-module. □

Hence, we first need to find a basis of $I_p/pI_p$. There are two methods to do this. From the HNF reduction above, we know a basis of $I_p$, and it is clear that the image of this basis in $I_p/pI_p$ is a basis of $I_p/pI_p$. The other method is as follows. We use only the $\mathbb{F}_p$-basis $\overline{\beta}_1, \ldots, \overline{\beta}_l$ of $\overline{I}_p$ found above. Using Algorithm 2.3.6, we can supplement this basis into a basis $\overline{\beta}_1, \ldots, \overline{\beta}_l, \overline{\beta}_{l+1}, \ldots, \overline{\beta}_n$ of $\mathcal{O}/p\mathcal{O}$, and then $\tilde{\beta}_1, \ldots, \tilde{\beta}_l, p\tilde{\beta}_{l+1}, \ldots, p\tilde{\beta}_n$ will be an $\mathbb{F}_p$-basis of $I_p/pI_p$, where $\tilde{}$ denotes reduction modulo $pI_p$, and $\beta_i$ denotes any pull-back of $\overline{\beta}_i$ in $\mathcal{O}$. (Note that the basis which one obtains depends on the pull-backs used.)

This method for finding a basis of $I_p/pI_p$ has the advantage of staying at the mod $p$ level, hence avoids the time consuming Hermite reduction, so it is preferable.

Now that we have a basis of $I_p/pI_p$, the elementary matrices give us a basis of $\mathrm{End}(I_p/pI_p)$. Hence, we obtain explicitly the matrix of the map whose kernel is $U$, and it is a $n^2 \times n$ matrix. Algorithm 2.3.1 makes sense only over a field, so we must first compute the kernel $\overline{U}$ of the map from $\mathcal{O}/p\mathcal{O}$ into $\mathrm{End}(I_p/pI_p)$ which can be done using Algorithm 2.3.1. If $\overline{v}_1, \ldots, \overline{v}_k$ is the basis of this kernel, to obtain $U$, we apply Hermite reduction to the matrix whose column vectors are $v_1, \ldots, v_k, p\omega_1, \ldots, p\omega_n$. In fact, we can apply Hermite reduction modulo the prime $p$, i.e. take $D = p$ in Algorithm 2.4.8.

Finally, note that to obtain the $n^2 \times n$ matrix above, if the $\overline{\gamma}_i$ form a basis of $I_p/pI_p$ one computes

$$\omega_k \overline{\gamma}_i = \sum_{1 \le j \le n} a_{k,i,j} \overline{\gamma}_j,$$

and $k$ is the column number, while $(i, j)$ is the row index. Unfortunately, in the round 2 algorithm, it seems unavoidable to use such large matrices. Note that to obtain the $a_{k,i,j}$, the work is much simpler if the matrix of the $\overline{\gamma}_j$ is triangular, and this is not the case in general if we complete the basis as explained above. On the other hand, this would be the case if we used the first method consisting of applying Hermite reduction to get the HNF of $I_p$ itself. Tests must be made to see which method is preferable in practice.

### 6.1.4 Detailed Description of the Round 2 Algorithm

Using what we have explained, we can now give in complete detail the round 2 algorithm.

**Algorithm 6.1.8** (Zassenhaus's Round 2). Let $K = \mathbb{Q}(\theta)$ be a number field given by an algebraic integer $\theta$ as root of its minimal monic polynomial $T$ of degree $n$. This algorithm computes an integral basis $\omega_1 = 1, \omega_2, \ldots, \omega_n$ of the maximal order $\mathbb{Z}_K$ (as polynomials in $\theta$) and the discriminant of the field. All the computations in $K$ are implicitly assumed to be done using the standard representation of numbers as polynomials in $\theta$.

1. [Factor discriminant of polynomial] Using Algorithm 3.3.7, compute $D \leftarrow \mathrm{disc}(T)$. Then using a factoring algorithm (see Chapters 8 to 10) factor $D$ in the form $D = D_0 F^2$ where $D_0$ is either equal to 1 or to a fundamental discriminant.

2. [Initialize] For $i = 1, \ldots, n$ set $\omega_i \leftarrow \theta^{i-1}$.

3. [Loop on factors of $F$] If $F = 1$, output the integral basis $\omega_i$ (which will be in HNF with respect to $\theta$), compute the product $G$ of the diagonal elements of the matrix of the $\omega_i$ (which will be the inverse of an integer by Corollary 4.7.6), set $d \leftarrow D \cdot G^2$, output the field discriminant $d$ and terminate the algorithm. Otherwise, let $p$ be the smallest prime factor of $F$.

4. [Factor modulo $p$] Using the mod $p$ factoring algorithms of Section 3.4, factor $T$ modulo $p$ as $\overline{T} = \prod \overline{t_i}^{e_i}$ where the $\overline{t_i}$ are distinct irreducible polynomials in $\mathbb{F}_p[X]$ and $e_i > 0$ for all $i$. Set $\overline{g} \leftarrow \prod \overline{t_i}$, $\overline{h} \leftarrow \overline{T}/\overline{g}$, $f \leftarrow (gh - T)/p$, $\overline{Z} \leftarrow (\overline{f}, \overline{g}, \overline{h})$, $\overline{U} \leftarrow \overline{T}/\overline{Z}$ and $m \leftarrow \deg(Z)$.

5. [Apply Dedekind] If $m = 0$, then $\mathcal{O}$ is $p$-maximal so while $p \mid F$ set $F \leftarrow F/p$, then go to step 3. Otherwise, for $1 \le i \le m$, let $v_i$ be the column vector of the components of $\omega_i U(\theta)$ on the standard basis $1, \theta, \ldots, \theta^{n-1}$ and set $v_{m+j} = p\omega_j$ for $1 \le j \le n$.

   Apply the Hermite reduction Algorithm 2.4.8 to the $n \times (n+m)$ matrix whose column vectors are the $v_i$. (Note that the determinant of the final matrix is known to divide $D$.) If $H$ is the $n \times n$ HNF reduced matrix which we obtain, set for $1 \le i \le n$, $\omega_i \leftarrow H_i/p$ where $H_i$ is the $i$-th column of $H$.

6. [Is the new order $p$-maximal?] If $p^{m+1} \nmid F$, then the new order is $p$-maximal so while $p \mid F$ set $F \leftarrow F/p$, then go to step 3.

7. [Compute radical] Set $q \leftarrow p$, and while $q < n$ set $q \leftarrow qp$. Then compute the $n \times n$ matrix $A = (a_{i,j})$ over $\mathbb{F}_p$ such that $\omega_j^q \equiv \sum_{1 \le i \le n} a_{i,j}\omega_i$. Note that the matrix of the $\omega_i$ will stay triangular, so the $a_{i,j}$ are easy to compute.

   Finally, using Algorithm 2.3.1, compute a basis $\overline{\beta_1}, \ldots, \overline{\beta_l}$ of the kernel of the matrix $A$ over $\mathbb{F}_p$ (this will be a basis of $I_p/p\mathcal{O}$).

8. [Compute new basis mod $p$] Using the known basis $\overline{\omega_1}, \ldots, \overline{\omega_n}$ of $\mathcal{O}/p\mathcal{O}$, supplement the linearly independent vectors $\overline{\beta_1}, \ldots, \overline{\beta_l}$ to a basis $\overline{\beta_1}, \ldots, \overline{\beta_n}$ of $\mathcal{O}/p\mathcal{O}$ using Algorithm 2.3.6.

9. [Compute big matrix] Set $\alpha_i \leftarrow \beta_i$ for $1 \leq i \leq l$, $\alpha_i \leftarrow p\beta_i$ for $l < i \leq n$, where $\beta_i$ is a lift to $\mathcal{O}$ of $\overline{\beta_i}$. Compute coefficients $c_{i,j,k} \in \mathbb{F}_p$ such that $\omega_k \alpha_j \equiv \sum_{1 \leq i \leq n} c_{i,j,k} \alpha_i \pmod{p}$. Let $C$ be the $n^2 \times n$ matrix over $\mathbb{F}_p$ such that $C_{(i,j),k} = c_{i,j,k}$.

10. [Compute new order] Using Algorithm 2.3.1, compute a basis $\gamma_1, \ldots \gamma_m$ for the kernel of $C$ (these are vectors in $\mathbb{F}_p^n$, and $m$ can be as large as $n^2$). For $1 \leq i \leq m$ let $v_i$ be a lift of $\gamma_i$ to $\mathbb{Z}^n$, and set $v_{m+j} = p\omega_j$ for $1 \leq j \leq n$. Apply the Hermite reduction Algorithm 2.4.8 to the $n \times (n+m)$ matrix whose column vectors are the $v_i$. (Note again that the determinant of the final matrix is known to divide $D$.) If $H$ is the $n \times n$ HNF reduced matrix which we obtain, set for $1 \leq i \leq n$, $\omega_i' \leftarrow H_i/p$ where $H_i$ is the $i$-th column of $H$.

11. [Finished with $p$?] If there exists an $i$ such that $\omega_i' \neq \omega_i$, then for every $i$ such that $1 \leq i \leq n$ set $\omega_i \leftarrow \omega_i'$ and go to step 7. Otherwise, $\mathcal{O}$ is $p$-maximal, so while $p \mid F$ set $F \leftarrow F/p$, and go to step 3.

This finishes our description of the round 2 algorithm. This algorithm seems complicated at first. Although it has been superseded by the round 4 algorithm, it is much simpler to implement and it performs very well. The major bottleneck is perhaps not where the reader expects it to be, i.e. in the handling of large matrices. It is, in fact, in the very first step which consists in factoring $\mathrm{disc}(T)$ in the form $D_0 F^2$. Indeed, as we will see in Chapter 10, factoring an 80 digit number takes a considerable amount of time, and factoring a 50 digit one is already not that easy. One can refine the methods given above to the case where one does not suppose $p$ to be necessarily prime (see [Buc-Len] and [Buc-Len2]), but unfortunately this does *not* avoid finding the largest square dividing $\mathrm{disc}(T)$, which is apparently almost as difficult as factoring it completely.

## 6.2 Decomposition of Prime Numbers II

As we shall see, the general problem of decomposing prime numbers in an algebraic number field is closely related to the problem of computing the maximal order. Consequently, we have already given most of the theory and auxiliary algorithms that we will need. As we have already seen, the problem is as follows. Given a prime $p$ and a $p$-maximal order $\mathcal{O}$, for example the maximal order $\mathbb{Z}_K$ itself, determine the maximal ideals $\mathfrak{p}_i$ and the exponents $e_i$ such that

$$p\mathcal{O} = \prod_{i=1}^{g} \mathfrak{p}_i^{e_i}.$$

As usual $\mathcal{O}$ will be given by its HNF on a power basis $1, \theta, \ldots, \theta^{n-1}$, and we want the HNF basis of the $\mathfrak{p}_i$. The determinant of the corresponding matrix is equal to $\mathcal{N}(\mathfrak{p}_i) = p^{f_i}$ in the traditional notation. For practical applications,

it will also be useful to have a two-element representation of the ideals $\mathfrak{p}_i$ (see Proposition 4.7.7).

In Theorem 4.8.13 we saw how to obtain this decomposition when $p$ does not divide the index $[\mathcal{O} : \mathbb{Z}[\theta]]$. Hence we will concentrate on the case where $p$ divides the index.

### 6.2.1 Newton Polygons

Historically the first method to deal with this problem is the so-called *Newton polygon method*. When it applies, it is very easy to use, but it must be stressed that it is not a general method. We will give a completely general method in the next section.

I am grateful to F. Diaz y Diaz and M. Olivier for the presentation of Newton polygons given here, which follows [Ore] and [Mon-Nar]. Essentially no proofs are given.

We may assume without loss of generality that the minimal polynomial $T(X)$ of $\theta$ is in $\mathbb{Z}[X]$ and is monic.

The first result tells us what survives of Theorem 4.8.13 in the case where $p$ divides the index.

**Proposition 6.2.1.** *Let*

$$T(X) \equiv \prod_{i=1}^{g} \overline{T_i(X)}^{e_i} \pmod{p}$$

*be the decomposition of $T$ into irreducible factors in $\mathbb{F}_p[X]$, where the $T_i$ are taken to be arbitrary monic lifts of $\overline{T_i(X)}$ in $\mathbb{Z}[X]$. Then*

$$p\mathbb{Z}_K = \prod_{i=1}^{g} \mathfrak{a}_i,$$

*where*

$$\mathfrak{a}_i = (p, T_i^{e_i}(\theta)) = p\mathbb{Z}_K + T_i^{e_i}(\theta)\mathbb{Z}_K$$

*and the $\mathfrak{a}_i$ are pairwise coprime (i.e. $\mathfrak{a}_i + \mathfrak{a}_j = \mathbb{Z}_K$ for $i \neq j$). Furthermore, if $n_i$ is the degree of $T_i$ we have $\mathcal{N}(\mathfrak{a}_i) = p^{e_i n_i}$, and all prime ideals dividing $\mathfrak{a}_i$ are of residual degree divisible by $n_i$.*

*Proof.* The proof follows essentially the same lines as that of Theorem 4.8.13. It is useful to also prove that the inverse of $\mathfrak{a}_i$ is given explicitly as

$$\mathfrak{a}_i^{-1} = (1, \prod_{j \neq i} T_j^{e_j}(\theta)/p)$$

(see Exercise 5). □

The problem is that the ideals $\mathfrak{a}_i$ are not necessarily of the form $\mathfrak{p}_i^{e_i}$ as in Theorem 4.8.13 (the reader can also check via examples that it would not do any good to set $\mathfrak{p}_i = (p, T_i(\theta))$). We must therefore try to split the ideals $\mathfrak{a}_i$ some more. For this we can proceed as follows. By successive Euclidean divisions of $T$ by $T_i$, we can write $T$ in a unique way in the form

$$T(X) = \sum_{j=0}^{\lfloor n/n_i \rfloor} Q_{i,j} T_i^j$$

with $\deg(Q_{i,j}) < n_i$. We will call this the *$T_i$-expansion* of $T$. We will write $d_i = \lfloor n/n_i \rfloor$.

If $Q = \sum_{0 \le k \le m} a_k X^k \in \mathbb{Z}[X]$, we will set

$$v_p(Q) = \min_k (v_p(a_k)),$$

where we set $v_p(0) = +\infty$ (or in other words we ignore coefficients equal to zero). The basic definition is as follows.

**Definition 6.2.2.** *With the above notations, for a fixed $i$, the convex hull of the set of points $(j, v_p(Q_{i,d_i-j}))$ for each $j$ such that $Q_{i,d_i-j} \neq 0$, is called the* Newton polygon *of $T$ relative to $T_i$ and the prime number $p$ (since $p$ is always fixed, we will in fact simply say "relative to $T_i$").*

Note that $Q_{i,j} = 0$ for $j < 0$ or $j > d_i$, hence the Newton polygon is bounded laterally by two infinite vertical half lines. Furthermore, since $T$ and the $T_i$ are monic, so is $Q_{i,d_i}$ hence $v_p(Q_{i,d_i}) = 0$. It follows that the first vertex of the Newton polygon is the origin $(0,0)$. Let $a$ be the largest real number (which is of course an integer) such that $(a,0)$ is still on the Newton polygon (we may have $a = 0$ or $a = d_i$). The part of the Newton polygon from the origin to $(a,0)$ is either empty (if $a = 0$) or is a horizontal segment. The rest of the Newton polygon, i.e. the points whose abscissa is greater than or equal to $a$, is called the *principal part* of the Newton polygon, and $(a,0)$ is its first vertex.

We assume now that $i$ is fixed.

Let $V_j$ for $0 \le j \le r$ be the vertices of the principal part of the Newton polygon of $T$ relative to $T_i$ (in the strict sense: if a point on the convex hull lies on the segment joining two other points, it is not a vertex), and set $V_j = (x_j, y_j)$. The *sides* of the polygon are the segments joining two consecutive vertices (not counting the infinite vertical lines), and the *slopes* are the slopes of these sides, i.e. the positive rational numbers $(y_j - y_{j-1})/(x_j - x_{j-1})$ for $1 \le j \le r$ (note that they cannot be equal to zero since we are in the principal part).

The second result gives us a more precise decomposition of $p\mathbb{Z}_K$ than the one given by Proposition 6.2.1 above, whose notations we keep. We refer to [Ore] for a proof.

**Proposition 6.2.3.** *Let $i$ be fixed.*

(1) *To each side $[V_{j-1}, V_j]$ of the principal part of the Newton polygon of $T$ relative to $T_i$ we can associate an ideal $\mathfrak{q}_{i,j}$ such that the $\mathfrak{q}_{i,j}$ are pairwise coprime and*

$$\mathfrak{a}_i = \prod_{j=1}^{r} \mathfrak{q}_{i,j}.$$

(2) *Set $h_j = y_j - y_{j-1}$ and $k_j = x_j - x_{j-1}$. If $h_j$ and $k_j$ are coprime for some $j$, then the corresponding ideal $\mathfrak{q}_{i,j}$ is of the form $\mathfrak{q}_{i,j} = \mathfrak{p}^{k_j}$ where $\mathfrak{p}$ is a prime ideal of degree $n_i$.*

(3) *In the special case when the principal part of the Newton polygon has a single side and $h_1 = y_1 - y_0 = y_1$ is equal to 1, then $\mathfrak{a}_i = \mathfrak{p}^{e_i}$ where $\mathfrak{p} = (p, T_i(\theta))$ is a prime ideal of degree $n_i$.*

**Corollary 6.2.4.** *Let $T \in \mathbb{Z}[X]$ be an* Eisenstein polynomial *with respect to a prime number $p$, i.e. a monic polynomial $T(X) = \sum_{i=0}^{n} a_i X^i$ with $p \mid a_i$ for all $i < n$ and $p^2 \nmid a_0$ (see Exercise 11 of Chapter 3). In the number field $K = \mathbb{Q}[\theta]$ defined by $T$ the prime $p$ is totally ramified, and more precisely $p\mathbb{Z}_K = \mathfrak{p}^n$ with $\mathfrak{p} = (p, \theta)$.*

*Proof.* In this case we have $T \equiv X^n \pmod{p}$, hence $T_1(X) = X$, $Q_{i,j} = a_j$, and since $p \mid a_i$ for all $i < n$, the principal part of the Newton polygon is the whole polygon, and since $p^2 \nmid a_0$ we are in the special case (3) of the proposition, so the corollary follows. □

Although Proposition 6.2.3 gives results in a number of cases, and can be generalized further (see [Ore] and [Mon-Nar]), it is far from being satisfactory from an algorithmic point of view.

### 6.2.2 Theoretical Description of the Buchmann-Lenstra Method

The second method for decomposing primes in number fields, which is completely general, is due to Buchmann and Lenstra ([Buc-Len]). We proceed as follows. (The reader should compare this to the method used for factoring polynomials modulo $p$ given in Chapter 3.) Write $I_p$ for the $p$-radical of $\mathcal{O}$. We know that $I_p = \prod_{i=1}^{g} \mathfrak{p}_i$. Set for any $j \geq 0$:

$$K_j = I_p^j + p\mathcal{O}.$$

It is clear that the valuation at $\mathfrak{p}_i$ of $K_j$ is equal to $\min(e_i, j)$, hence

$$K_j = \prod_{i=1}^{g} \mathfrak{p}_i^{\min(e_i, j)}.$$

It is also clear that $K_j \subset K_{j-1}$. Hence, if we set

$$J_j = K_j(K_{j-1})^{-1},$$

then $J_j$ is an integral ideal, and in fact $J_j = \prod_{e_i \geq j} \mathfrak{p}_i$ so in particular $J_j \subset J_{j+1}$. Finally, if we define

$$H_j = J_j(J_{j+1})^{-1},$$

we have

$$H_j = \prod_{e_i = j} \mathfrak{p}_i .$$

This exactly corresponds to the squarefree decomposition procedure of Section 3.4.2, the $H_i$ playing the role of the $A_i$, and without the inseparability problems. In other words, if we set $e = \max_i(e_i)$, we have

$$p\mathcal{O} = \prod_{j=1}^{e} H_j^j ,$$

and the $H_j$ are pairwise coprime and are products of distinct maximal ideals. To find the splitting of $p\mathcal{O}$, it is of course sufficient to find the splitting of each $H_j$.

Now, since $H_j$ is a product of distinct maximal ideals, i.e. is squarefree, the $\mathbb{F}_p$-algebra $\mathcal{O}/H_j$ is separable. Therefore, by the primitive element theorem there exists $\overline{\alpha}_j \in \mathcal{O}/H_j$ such that $\mathcal{O}/H_j = \mathbb{F}_p[\overline{\alpha}_j]$. Let $\overline{h}_j$ be the characteristic polynomial of $\overline{\alpha}_j$ over $\mathbb{F}_p$, and $h_j$ be any pullback in $\mathbb{Z}[X]$. Then exactly the same proof as in Section 4.8.2 shows that, if

$$h_j(X) \equiv \prod_{i=1}^{g_j} q_{i,j}(X) \pmod{p}$$

is the decomposition modulo $p$ of the polynomial $h_j$, then the ideals

$$\mathfrak{q}_{i,j} = H_j + q_{i,j}(\alpha_j)\mathcal{O}$$

are maximal and that

$$H_j = \prod_{i=1}^{g_j} \mathfrak{q}_{i,j}$$

is the desired decomposition of $H_j$ into a product of prime ideals.

We must now give algorithms for all the steps described above. Essentially, the two new things that we need are operations on ideals in our special case, and splitting of a separable algebra over $\mathbb{F}_p$.

### 6.2.3 Multiplying and Dividing Ideals Modulo $p$

Although the most delicate step in the decomposition of $p\mathbb{Z}_K$ is the final splitting of the ideals $H_j$, experiment (and complexity analysis) shows that this is paradoxically the fastest part. The conceptually easier steps of multiplying and dividing ideals take, in fact, most of the time and so must be speeded up as much as possible.

Looking at what is needed, it is clear that we use only the reductions modulo $p\mathcal{O}$ of the ideals involved. Hence, although for ease of presentation we have implicitly assumed that the ideals are represented by their HNF, we will in fact consider only ideals $I/p\mathcal{O}$ of $\mathcal{O}/p\mathcal{O}$ which will be represented by an $\mathbb{F}_p$-basis. All the difficulties of HNF (Euclidean algorithm, coefficient explosion) disappear and are replaced by simple linear algebra algorithms. Moreover, we are working with coefficients in a field which is usually of small cardinality. (Recall that $p$ divides the index, otherwise the much simpler algorithm of Section 4.8.2 can be used.)

If $I$ is given by its HNF with respect to $\theta$ (this will not happen in our case since we start working directly modulo $p$), then, since $I \supset p\mathcal{O} \supset p\mathbb{Z}[\theta]$, the diagonal elements of the HNF will be equal to 1 or $p$. Therefore, to find a basis of $\overline{I}$, we simply take the basis elements corresponding to the columns whose diagonal element is equal to 1.

The algorithm for multiplication is straightforward.

**Algorithm 6.2.5** (Ideal Multiplication Modulo $p\mathcal{O}$). Given two ideals $I/p\mathcal{O}$ and $J/p\mathcal{O}$ by $\mathbb{F}_p$-bases $(\alpha_i)_{1\le i\le r}$ and $(\beta_j)_{1\le j\le m}$ respectively, where the $\alpha_i$ and $\beta_j$ are expressed as $\mathbb{F}_p$-linear combinations of a fixed integral basis $\omega_1, \ldots, \omega_n$ of $\mathcal{O}$, this algorithm computes an $\mathbb{F}_p$-basis of the ideal $IJ/p\mathcal{O}$.

1. [Compute matrix] Using the multiplication table of the $\omega_i$, let $M$ be the $n \times rm$ matrix $M$ with coefficients in $\mathbb{F}_p$ whose columns express the products $\alpha_i\beta_j$ on the integral basis.
2. [Compute image] Using Algorithm 2.3.2 compute a matrix $M_1$ whose columns form an $\mathbb{F}_p$-basis of the image of $M$. Output the columns of $M_1$ and terminate the algorithm.

Ideal division modulo $p\mathcal{O}$ is slightly more difficult. We first need a lemma.

**Lemma 6.2.6.** *Denote by $^-$ reduction mod $p$. Let $I$ and $J$ two integral ideals of $\mathcal{O}$ containing $p\mathcal{O}$ and assume that $I \subset J$. Then, as a $\mathbb{Z}/p\mathbb{Z}$-vector space, $\overline{IJ^{-1}}$ is equal to the kernel of the map $\phi$ from $\mathcal{O}/p\mathcal{O}$ to* $\mathrm{End}(J/I)$ *given by*

$$\phi(\overline{\beta}) = (\overline{\alpha} \longmapsto \overline{\alpha\beta}) \ .$$

Indeed, $\phi(\overline{\beta})$ is equal to 0 if and only if $\alpha\beta \in I$ for every $\alpha \in J$, i.e. if $\beta J \subset I$, or in other words if $\beta \in IJ^{-1}$, proving the lemma. □

This leads to the following algorithm.

**Algorithm 6.2.7** (Ideal Division Modulo $p\mathcal{O}$). Given two ideals $I/p\mathcal{O}$ and $J/p\mathcal{O}$ by $\mathbb{F}_p$ bases $(\alpha_i)_{1\le i\le r}$ and $(\beta_j)_{1\le j\le m}$ respectively, where the $\alpha_i$ and $\beta_j$ are expressed as $\mathbb{F}_p$-linear combinations of a fixed integral basis $\omega_1, \ldots, \omega_n$ of $\mathcal{O}$, this algorithm computes an $\mathbb{F}_p$-basis of the ideal $IJ^{-1}/p\mathcal{O}$ assuming that $I \subset J$.

1. [Find basis of $J/I$] Apply Algorithm 2.3.7 to the subspaces $I/p\mathcal{O}$ and $J/p\mathcal{O}$ of $\mathbb{F}_p^n$, thus obtaining a basis $(\gamma_j)_{1\le j\le m-r}$ of a supplement of $I/p\mathcal{O}$ in $J/p\mathcal{O}$.
2. [Setup ideal division] By using the multiplication table of the $\omega_i$ and Algorithm 2.3.5, compute elements $a_{i,j,k}$ and $b_{i,j,k}$ in $\mathbb{F}_p$ such that

$$\overline{\omega_k}\gamma_i = \sum_j a_{i,j,k}\gamma_j + \sum_j b_{i,j,k}\alpha_j,$$

   and let $M$ be the $(m-r)^2 \times n$ matrix formed by the $a_{i,j,k}$ for $1 \le i, j \le m-r$ and $1 \le k \le n$ (we can forget the $b_{i,j,k}$).
3. [Compute $IJ^{-1}/p\mathcal{O}$] Using Algorithm 2.3.1, compute a matrix $M_1$ whose columns form an $\mathbb{F}_p$-basis of the kernel of $M$, output $M_1$ and terminate the Algorithm.

Indeed, $M$ is clearly equal to the matrix of $\phi$ in the standard basis of $\mathrm{End}(J/I)$. □

### 6.2.4 Splitting of Separable Algebras over $\mathbb{F}_p$

To avoid unnecessary indices, we set simply $H = H_j$. Using the above algorithms, it is straightforward to compute an $\mathbb{F}_p$-basis $\overline{\beta}_1, \ldots, \overline{\beta}_m$ of $\overline{H} = H/p\mathcal{O}$. Using Algorithm 2.3.6, we can supplement this basis to a basis $\overline{\beta}_1, \ldots, \overline{\beta}_n$ of $\mathcal{O}/p\mathcal{O}$ It is then clear that the images of $\beta_{m+1}, \ldots, \beta_n$ in $\mathcal{O}/H$ form an $\mathbb{F}_p$-basis of $\mathcal{O}/H$.

In order to finish the decomposition, there remains the problem of splitting the separable algebra $A = \mathcal{O}/H$ given by this $\mathbb{F}_p$-basis. As explained above, one method is to start by finding a primitive element $\overline{\alpha}$. Finding a primitive element is not, however, a completely trivial task. Perhaps the best way is to choose at random an element $x \in A \setminus \mathbb{F}_p$ (note that $\mathbb{F}_p$ can be considered naturally embedded in $A$), compute its *minimal* polynomial $P(X)$ over $\mathbb{F}_p$ (which need not be irreducible), and check whether $\deg(P) = \dim(A)$. Although practical, this method has the disadvantage of being completely non-deterministic, although it is easy to give estimates for the number of trials that one has to perform before succeeding in finding a suitable $x$, see Exercise 6.

We give another method which does not have this disadvantage. It is based on the following proposition.

**Proposition 6.2.8.** *Let $A$ be a finite separable algebra over $\mathbb{F}_p$. There exists an efficient probabilistic algorithm which either shows that $A$ is a field, or finds a non-trivial idempotent in $A$, i.e. an element $\varepsilon \in A$ such that $\varepsilon^2 = \varepsilon$ with $\varepsilon \neq 0$ and $\varepsilon \neq 1$.*

*Proof.* Since $A$ is a finite separable algebra, $A$ is isomorphic to a finite product of fields, say $A \simeq A_1 \times \cdots \times A_k$. Write any element $\alpha$ of $A$ as $(\alpha_1, \ldots, \alpha_k)$ where $\alpha_i \in A_i$. Consider the map $\phi$ from $A$ to $A$ defined by $\phi(x) = x^p - x$. It is clear that $\mathbb{F}_p$, considered as embedded in $A$, is in the kernel $V$ of $\phi$. By Algorithm 2.3.1, we can easily compute a basis for $V$, and, in particular, its dimension. Note that $\alpha = (\alpha_1, \ldots, \alpha_k) \in V$ if and only if for all $i$ such that $1 \leq i \leq k$, $\alpha_i \in \mathbb{F}_p$ where $\mathbb{F}_p$ is considered embedded in $A_i$. It follows that $\dim(V) = k$, and hence $\dim(V) = 1$ if and only if $A$ is a field.

Therefore assume that $\dim(V) > 1$, and let $\alpha \in V \setminus \mathbb{F}_p$. By computing successive powers of $\alpha$, we can find the minimal polynomial $m_\alpha(X)$ of $\alpha$ in $A$. If $\alpha = (\alpha_1, \ldots, \alpha_k)$, it is clear that $m_\alpha(X)$ is the least common multiple of the $m_{\alpha_i}(X)$, and since $\alpha \in V$, the polynomials $m_{\alpha_i}(X)$ are polynomials of degree 1. It follows that $m_\alpha(X)$ is a squarefree polynomial equal to a product of at least two linear factors (since $\alpha \notin \mathbb{F}_p$). Write $m_\alpha(X) = m_1(X)m_2(X)$ where $m_1$ and $m_2$ are non-constant polynomials in $\mathbb{F}_p[X]$. Since $m_\alpha$ is squarefree, $m_1$ and $m_2$ are coprime, so we can find polynomials $U(X)$ and $V(X)$ in $\mathbb{F}_p[X]$ such that $U(X)m_1(X) + V(X)m_2(X) = 1$. We now choose $\varepsilon = Um_1(\alpha)$. Since $m_1 m_2(\alpha) = 0$, $\varepsilon$ is an idempotent. In addition, it is clear that $(U, m_2) = (V, m_1) = 1$ and $m_1$, $m_2$ non-constant imply that $\varepsilon \neq 0$ and $\varepsilon \neq 1$. □

**Remark.** Note that it is not necessary to compute the complete basis of the kernel of $\phi$ in order to obtain the result. We need only, either show that the kernel $V$ is of dimension 1 (proving that $A$ is a field), or give an element of $V$ which is not in the one-dimensional subspace $\mathbb{F}_p$. Hence, we can stop algorithm 2.3.1 as soon as such an element is found.

Using this proposition, it is easy to finish the splitting of our ideals $H = H_j$. Set $A = \mathcal{O}/H$ as before. Using the above proposition, either we have shown that $A$ is a field (hence $H$ is a prime ideal, so we have shown that the splitting is trivial), or we have found a non-trivial idempotent $\varepsilon$. Set $H_1 = H + e\mathcal{O}$, $H_2 = H + (1 - e)\mathcal{O}$ where $e$ is any lift to $\mathcal{O}$ of $\varepsilon$. I claim that $H = H_1 \cdot H_2$. Indeed, since $e(1 - e) \in H$, it is clear that $H_1 \cdot H_2 \subset H$. Conversely, if $x \in H$ we can write $x = ex + (1 - e)x$, and $ex \in e\mathcal{O} \cdot H$, $(1 - e)x \in (1 - e)\mathcal{O} \cdot H$ so $x \in H_1 \cdot H_2$ as claimed.

Hence, we have split $H$ non-trivially (since $e$ is a non-trivial idempotent) and we can continue working on $H_1$ and $H_2$ separately. This process terminates in at most $k$ steps, where $k$ is the number of prime factors of $H$.

A more efficient method would be to use the complete splitting of $m_\alpha(X)$ (in the notation of the proof of Proposition 6.2.8) which gives a corresponding splitting of $H$ as a product of more than two ideals. This will be done in the algorithm given below.

**Remark.** For some applications, such as computing the values of zeta and $L$-functions, it is not necessary to obtain the explicit decomposition of $p\mathcal{O}$, but only the ramification indices and residual degrees $e_i$ and $f_i$. Once the $H_j$ above have been computed, this can be done without much further work, as explained in Exercise 8 (this remark is due to H. W. Lenstra).

Once $H$ has been shown to be a maximal ideal by successive splittings, what remains is the problem of representing $H$. Since we will have computed an $\mathbb{F}_p$-basis $(\alpha_i)_{1\le i\le m}$ of $H/p\mathcal{O}$, to obtain the HNF of $H$ we arbitrarily lift the $\alpha_i$ to $a_i \in \mathcal{O}$, and then do an HNF reduction of the matrix whose first $m$ columns are the components of the $a_i$ on the $\omega_j$, and whose last $n$ columns form $p$ times the $n \times n$ identity matrix. It is obviously possible to do this HNF reduction modulo $p$ (Algorithm 2.4.8), so no coefficient explosion can take place.

Even after finding the HNF of $H$ we should still not be satisfied, because in practice, it is much more efficient to represent prime ideals by a two-element representation. To obtain this, we apply Algorithm 4.7.10. Note that we know the degree of $H$ (the number $f$ in the notation of Algorithm 4.7.10), which is simply equal to $n - m$ (since $p^n = [\mathcal{O} : p\mathcal{O}] = [\mathcal{O} : H][H : p\mathcal{O}] = p^f p^m$). Also we do not need to compute the HNF of $H$ at all to apply Algorithm 4.7.10 since (together with $p$) the $a_i$ clearly form a $\mathbb{Z}_K$-generating set.

### 6.2.5 Detailed Description of the Algorithm for Prime Decomposition

We can summarize the preceding discussions in the following algorithm

**Algorithm 6.2.9** (Prime Decomposition). Let $K = \mathbb{Q}(\theta)$ be a number field given by an algebraic integer $\theta$ as root of its minimal monic polynomial $T$ of degree $n$. We assume that we have already computed an integral basis $\omega_1 = 1, \ldots, \omega_n$ and the discriminant $d(K)$ of $K$, for example, by using the round 2 Algorithm 6.1.8.

Given a prime number $p$, this algorithm outputs the decomposition $p\mathbb{Z}_K = \prod_{1\le i\le g} \mathfrak{p}_i^{e_i}$ by giving for each $i$ the values of $e_i$, $f_i = \deg(\mathfrak{p}_i)$ and a two-element representation $\mathfrak{p}_i = (p, \alpha_i)$. All the ideals $I$ which we will use (except for the final $\mathfrak{p}_i$) will be represented by $\mathbb{F}_p$ bases of $I/p\mathcal{O}$.

1. [Check if easy] If $p \nmid \operatorname{disc}(T)/d(K)$, then by applying the algorithms of Section 3.4 factor the polynomial $T(X)$ modulo $p$, output the decomposition of $p\mathbb{Z}_k$ given by Theorem 4.8.13 and terminate the algorithm.

2. [Compute radical] Set $q \leftarrow p$, and while $q < n$ set $q \leftarrow qp$. Now compute the $n \times n$ matrix $A = (a_{i,j})$ over $\mathbb{F}_p$ such that $\omega_j^q \equiv \sum_{1\le i\le n} a_{i,j}\omega_i$. Note that the matrix of the $\omega_i$ will stay triangular, so the $a_{i,j}$ are easy to compute.

   Finally, using Algorithm 2.3.1, compute a basis $\overline{\beta_1}, \ldots, \overline{\beta_l}$ of the kernel of the matrix $A$ over $\mathbb{F}_p$ (this will be a basis of $I_p/p\mathcal{O}$). (Note that this step

has already been performed as step 7 of the round 2 algorithm, so if the result has been kept it is not necessary to recompute this again.)

3. [Compute $\overline{K_i}$] Set $\overline{K_1} \leftarrow I_p/p\mathcal{O}$ (computed in step 2), $i \leftarrow 1$ and while $\overline{K_i} \neq \{0\}$ set $i \leftarrow i+1$ and $\overline{K_i} \leftarrow \overline{K_1 K_{i-1}}$ computed using Algorithm 6.2.5.

4. [Compute $\overline{J_j}$] Set $\overline{J_1} \leftarrow \overline{K_1}$ and for $j = 2, \ldots, i$ set $\overline{J_j} \leftarrow \overline{K_j K_{j-1}^{-1}}$ using Algorithm 6.2.7.

5. [Compute $\overline{H_j}$] For $j = 1, \ldots, i-1$ set $\overline{H_j} \leftarrow \overline{J_j J_{j+1}^{-1}}$ using Algorithm 6.2.7, and set $\overline{H_i} \leftarrow \overline{J_i}$.

6. [Initialize loop] Set $j \leftarrow 0$, $c \leftarrow 0$.

7. [Finished?] If $c = 0$ do the following: if $j = i$ terminate the algorithm, otherwise set $j \leftarrow j+1$ and if $\dim_{\mathbb{F}_p}(\overline{H_j}) < n$ set $\mathcal{L} \leftarrow \{\overline{H_j}\}$ and $c \leftarrow 1$, else go to step 7 ($\mathcal{L}$ will be a list of $c$ ideals of $\mathcal{O}/p\mathcal{O}$).

8. [Compute separable algebra $A$] Let $\overline{H}$ be an element of $\mathcal{L}$. Compute an $\mathbb{F}_p$-basis of $A = \mathcal{O}/H = (\mathcal{O}/p\mathcal{O})/(H/p\mathcal{O})$ in the following way. If $\beta_1, \ldots, \beta_r$ is the given $\mathbb{F}_p$-basis of $\overline{H}$, set $\beta_{r+1} \leftarrow (1, 0, \ldots, 0)^t$ (which will be linearly independent of the $\beta_i$ for $i \leq r$ since $1 \notin H$), supplement this family of vectors using Algorithm 2.3.6 to a basis $\beta_1, \ldots, \beta_n$ of $\mathcal{O}/p\mathcal{O}$. Then, as an $\mathbb{F}_p$-basis of $A$, take $\beta_{r+1}, \ldots, \beta_n$. (This insures that the first vector of our basis of $A$ is always $(1, 0, \ldots, 0)^t$, which would not be the case if we applied Algorithm 2.3.6 directly.)

9. [Compute multiplication table] Denote by $\gamma_1, \ldots, \gamma_f$ the $\mathbb{F}_p$-basis of $A$ just obtained (hence $\gamma_i = \beta_{r+i}$ and $f = n - r$). By using the multiplication table of the $\omega_i$ and Algorithm 2.3.5, compute elements $a_{i,j,k}$ and $b_{i,j,k}$ in $\mathbb{F}_p$ such that

$$\gamma_i \gamma_j = \sum_{1 \leq j \leq f} a_{i,j,k} \gamma_j + \sum_{1 \leq j \leq r} b_{i,j,k} \beta_j .$$

The multiplication table of the $\gamma_i$ (which will be used implicitly from now on) is given by the $a_{i,j,k}$ (we can forget the $b_{i,j,k}$).

10. [Compute $V = \ker(\phi)$] Let $M$ be the matrix of the map $\alpha \mapsto \alpha^p - \alpha$ from $A$ to $A$ on the $\mathbb{F}_p$ basis that we have found. Compute a basis $M_1$ of the kernel of $M$ using Algorithm 2.3.1. Note that if some other algorithm is used to find the kernel, we should nonetheless insure that the first column of $M_1$ is equal to $(1, 0, \ldots, 0)^t$.

11. [Do we have a field?] If $M_1$ has at least two columns (i.e. if the kernel of $M$ is not one-dimensional), go to step 12. Otherwise, set $f \leftarrow \dim_{\mathbb{F}_p}(A)$, let $(p, \alpha)$ be the two-element representation of $H$ obtained by applying Algorithm 4.7.10 to $\overline{H}$. Output $j$ as ramification index, $f$ as residual degree of $H$, and the prime ideal $(p, \alpha)$. Then remove $H$ from the list $\mathcal{L}$, set $c \leftarrow c-1$ and go to step 7.

12. [Find $m(X)$] Let $\alpha \in A$ correspond to a column of $M_1$ which is not proportional to $(1, 0, \ldots, 0)^t$. By computing the successive powers of $\alpha$ in $A$, let $m(X) \in \mathbb{F}_p[X]$ be the minimal monic polynomial of $\alpha$ in $A$.

13. [Factor $m(X)$] (We know that $m(X)$ is a squarefree product of linear polynomials.) By using one of the final splitting methods described in Section 3.4, or simply by trial and error if $p$ is small, factor $m(X)$ into linear factors as $m(X) = m_1(X) \cdots m_k(X)$.

14. [Split $H$] Let $r = \dim_{\mathbb{F}_p}(\overline{H})$. For $s = 1, \ldots, k$ do as follows. Set $\beta_s \leftarrow m_s(\alpha)$, let $M_s$ be the $n \times (r+n)$ matrix over $\mathbb{F}_p$ whose first $r$ columns give the basis of $\overline{H}$ and the last $n$ express $\omega_i \beta_s$ on the integral basis. Finally, let $\overline{H_s}$ be the image of $M_s$ computed using Algorithm 2.3.2.

15. [Update list] Remove $\overline{H}$ and add $\overline{H_1}, \ldots, \overline{H_k}$ to the list $\mathcal{L}$, set $c \leftarrow c + k - 1$ and go to step 8.

The dimension condition in step 7 was added so as to avoid considering values of $j$ such that there are no prime ideals over $p$ whose ramification index is equal to $j$.

The validity of steps 14 and 15 of the algorithm is left as an exercise for the reader (Exercise 27).

**Remark.** If we want to avoid writing routines for ideal multiplication and division, we can also proceed as follows. After step 2 of the above algorithm set $\mathcal{L} \leftarrow \{\overline{I_p}\}$ and go directly to step 8 to compute the decomposition of the separable algebra $A = \mathcal{O}/I_p$. In step 11, we must compute the ramification index $j$ of each prime ideal found, and this is easily done by using Algorithm 4.8.17. We leave the details of these modifications to the reader (Exercise 11). This method is in practice much faster than the method using ideal multiplication and division.

## 6.3 Computing Galois Groups

### 6.3.1 The Resolvent Method

I am indebted to Y. Eichenlaub for help in writing this section.

Let $K = \mathbb{Q}(\theta)$ be a number field of degree $n$, where $\theta$ is an algebraic integer whose minimal monic polynomial is denoted $T(X)$. An important algebraic question is to compute the *Galois group* $\mathrm{Gal}(T)$ of the polynomial $T$, in other words the Galois group of the splitting field of $T$, or equivalently of the Galois closure of $K$ in $\overline{\mathbb{Q}}$. Since by definition elements of $\mathrm{Gal}(T)$ act as permutations on the roots of $T$, once an ordering of the roots is given, $\mathrm{Gal}(T)$ can naturally be considered as a subgroup of $S_n$, the symmetric group on $n$ letters. Changing the ordering of the roots clearly transforms $\mathrm{Gal}(T)$ into a conjugate group, and since the ordering is not canonical, the natural objects to consider are subgroups of $S_n$ up to conjugacy. It will be important in what follows to remember that we have chosen a specific, but arbitrary ordering, since it will sometimes be necessary to change it.

Furthermore, since the polynomial $T$ is irreducible, the group $\mathrm{Gal}(T)$ is a *transitive* subgroup of $S_n$, i.e. there is a single orbit for the action of $\mathrm{Gal}(T)$ on the roots $\theta_i$ of $T$ (each orbit corresponding to an irreducible factor of $T$). Hence, the first task is to classify transitive subgroups of $S_n$ up to conjugacy. This is a non-trivial (but purely) group-theoretical question. It has been solved up to $n = 32$ (see [But-McKay] and [Hül]), but the number of groups becomes unwieldy for higher degrees. We will give the classification for $n \leq 7$.

Note that since the cardinality of an orbit divides the order of $\mathrm{Gal}(T)$, the cardinality of a transitive subgroup of $S_n$ is divisible by $n$.

Once the transitive groups are classified, we must still determine which corresponds to our Galois group $\mathrm{Gal}(T)$. We first note the following simple, but important proposition.

**Proposition 6.3.1.** *Let $A_n$ be the alternating group on $n$ letters corresponding to the even permutations. Then $\mathrm{Gal}(T) \subset A_n$ if and only if $\mathrm{disc}(T)$ is a square.*

*Proof.* Let $\theta_i$ be the roots of $T$. By Proposition 3.3.5, we know that

$$\mathrm{disc}(T) = f^2, \quad \text{where} \quad f = \prod_{1 \leq i < j \leq n} (\theta_j - \theta_i).$$

Clearly $f$ is an algebraic integer, and for any $\sigma \in \mathrm{Gal}(T)$ we have

$$\sigma(f) = \epsilon(\sigma) f,$$

where $\epsilon(\sigma)$ denotes the signature of $\sigma$. Hence, if $\mathrm{Gal}(T) \subset A_n$, all permutations of $\mathrm{Gal}(T)$ are even, so $f$ is invariant under $\mathrm{Gal}(T)$. Thus by Galois theory, $f \in \mathbb{Z}$. Conversely, if $f \in \mathbb{Z}$, we have $f \neq 0$ since the roots of $T$ are distinct. Therefore $\epsilon(\sigma) = 1$ for all $\sigma \in \mathrm{Gal}(T)$, so $\mathrm{Gal}(T) \subset A_n$. Note that since $A_n$ is a normal subgroup, that a group is a subgroup of $A_n$ depends only on its conjugacy class, and not on the precise conjugate. □

We now need to introduce a definition which will be basic to our work.

**Definition 6.3.2.** *Let $G$ be a subgroup of $S_n$ containing $\mathrm{Gal}(T)$ (not up to conjugacy, but for the given numbering of the roots), and let $F(X_1, X_2, \ldots, X_n)$ be a polynomial in $n$ variables with coefficients in $\mathbb{Z}$. If $H$ is the stabilizer of $F$ in $G$, i.e.*

$$H = \{\sigma \in G, F\left(X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(n)}\right) = F(X_1, X_2, \ldots, X_n)\},$$

*we define the* resolvent polynomial $R_G(F, T)$ *with respect to $G$, $F$ and the polynomial $T$ by*

$$R_G(F, T)(X) = \prod_{\sigma \in G/H} \left(X - F\left(\theta_{\sigma(1)}, \theta_{\sigma(2)}, \ldots, \theta_{\sigma(n)}\right)\right),$$

*where $G/H$ denotes any set of left coset representatives of $G$ modulo $H$.*

When $G = S_n$, we will omit the subscript in the notation.

It is clear from elementary Galois theory that $R_G(F,T) \in \mathbb{Z}[X]$. The main theorem which we will use concerning resolvent polynomials is as follows.

**Theorem 6.3.3.** *With the notation of the preceding definition, set $m = [G : H] = \deg(R_G(F,T))$. Then, if $R_G(F,T)$ is squarefree, its Galois group (as a subgroup of $S_m$) is equal to $\phi(\mathrm{Gal}(T))$, where $\phi$ is the natural group homomorphism from $G$ to $S_m$ given by the natural left action of $G$ on $G/H$. In particular, the list of degrees of the irreducible factors of $R_G(F,T)$ in $\mathbb{Z}[X]$ is the same as the list of the length of the orbits of the action of $\phi(\mathrm{Gal}(T))$ on $[1,\dots,m]$. For example, $R_G(F,T)$ has a root in $\mathbb{Z}$ if and only if $\mathrm{Gal}(T)$ is conjugate under $G$ to a subgroup of $H$.*

For the proof, see [Soi].

Note that it is important to specify that $\mathrm{Gal}(T)$ is conjugate under $G$, since this is a stronger condition than being conjugate under $S_n$.

Now it will often happen that $R_G(F,T)$ is not squarefree. In that case, to be able to apply the theorem, we use the following algorithm.

**Algorithm 6.3.4** (Tschirnhausen Transformation). Given a monic irreducible polynomial $T$ defining a number field $K = \mathbb{Q}(\theta)$, we find another such polynomial $U$ defining the same number field.

1. [Choose random polynomial] Let $n \leftarrow \deg(T)$. Choose at random a polynomial $A \in \mathbb{Z}[X]$ of degree less than or equal to $n-1$.
2. [Compute characteristic polynomial] Using the method explained in Section 4.3, compute the characteristic polynomial $U$ of $\alpha = A(\theta)$. In other words, using the sub-resultant Algorithm 3.3.7, set $U \leftarrow R_Y(T(Y), X - A(Y))$.
3. [Check degree] Using Euclid's algorithm, compute $V \leftarrow \gcd(U, U')$. If $V$ is constant, then output $U$ and terminate the algorithm, otherwise go to step 1.

The validity of this algorithm is clear.

Modifying $T$ if necessary by using such a Tschirnhausen transformation, it is always easy to reduce to the case where $R_G(F,T)$ is squarefree.

Finally, we need some notation. The elements of the set $G/H$ will be given as products of disjoint cycles, with $I$ denoting the identity permutation. Usually, apart from $I$, $G/H$ will contain only transpositions.

We denote by $C_n$ the cyclic group $\mathbb{Z}/n\mathbb{Z}$, and by $D_n$ the dihedral group of order $2n$, isomorphic to the isometries of a regular $n$-gon. As before, $A_n$ and $S_n$ denote the alternating group and symmetric group on $n$ letters respectively. Finally, $A \rtimes B$ denotes the semi-direct product of the groups $A$ and $B$, where the action of $B$ on $A$ is understood.

When we compute a group, we will output not only the isomorphism class of the group, but also a sign expressing whether the group is contained in $A_n$ (+ sign) or not (− sign). This will help resolve a number of ambiguities since isomorphic groups are not always conjugate in $S_n$.

Let us now examine in turn each degree up to degree 7. The particular choices of resolvents that we give are in no way canonical, although we have tried to give the ones which are the most efficient. The reader can find many other choices in the literature ([Stau], [Gir], [Soi] and [Soi-McKay], [Eic1]). The validity of the algorithms given can be checked using Theorem 6.3.3.

In degrees 1 and 2 there is of course nothing to say since the only possible group is $S_n$ in these cases, so we always output $(S_n, -)$.

### 6.3.2 Degree 3

In degree 3, it is obvious that the only transitive subgroups of $S_3$ are $C_3 \simeq A_3$ and $S_3 \simeq D_3$ which may be separated by the discriminant. In other words:

**Proposition 6.3.5.** *If $n = 3$, we have either $\mathrm{Gal}(T) \simeq C_3$ or $\mathrm{Gal}(T) \simeq S_3$ depending on whether $\mathrm{disc}(T)$ is a square or not.*

Thus we output $(C_3, +)$ or $(S_3, -)$ depending on $\mathrm{disc}(T)$.

### 6.3.3 Degree 4

In degree 4, there are (up to conjugacy) five transitive subgroups of $S_4$. These are $C_4$ (the cyclic group), $V_4 = C_2^2$ (the Klein 4-group), $D_4$ (the dihedral group of order 8, group of isometries of the square), $A_4$ and $S_4$.

Some inclusions are $V_4 \subset D_4 \cap A_4$, and $C_4 \subset D_4$.

Important remark: note that although we consider the groups only up to conjugacy, the notion of inclusion for two groups $G_1$ and $G_2$ can reasonably be defined by saying that $G_1 \subset G_2$ only when $G_1$ is a subgroup of some conjugate of $G_2$. On the other hand, when we consider *abstract* groups such as $V_4$, $D_4$, etc ... , the notion of inclusion is much more delicate since some subgroups of $S_n$ can be isomorphic as abstract groups but not conjugate in $S_n$. In this case, we write $G_1 \subset G_2$ only if this is valid for all conjugacy classes isomorphic to $G_1$ and $G_2$ respectively.

A simple algorithm is as follows.

**Algorithm 6.3.6** (Galois Group for Degree 4). Given an irreducible monic polynomial $T \in \mathbb{Z}[X]$ of degree 4, this algorithm computes its Galois group.

1. [Compute resolvent] Using Algorithm 3.6.6, compute the roots $\theta_i$ of $T$ in $\mathbb{C}$. Let
$$F \leftarrow X_1X_2^2 + X_2X_3^2 + X_3X_4^2 + X_4X_1^2$$

and let $R \leftarrow R(F,T)$, where a system of representatives of $G/H$ is given by

$$G/H = \{I, (12), (13), (14), (23), (34)\}.$$

Then round the coefficients of $R$ to the nearest integer (note that the roots $\theta_i$ must be computed to a sufficient accuracy for this rounding to be correct, and the needed accuracy is easily determined, see Exercise 13).

2. [Squarefree?] Compute $V \leftarrow (R, R')$ using the Euclidean algorithm. If $V$ is non-constant, replace $T$ by the polynomial obtained by applying a Tschirnhausen transformation using Algorithm 6.3.4 and go to step 1.
3. [Factor resolvent] Using Algorithm 3.5.7, factor $R$ over $\mathbb{Z}$. Let $L$ be the list of the degrees of the irreducible factors sorted in increasing order.
4. [Conclude] If $R$ is irreducible, i.e. if $L = (6)$, then output $(A_4, +)$ or $(S_4, -)$ depending on whether $\mathrm{disc}(T)$ is a perfect square or not. Otherwise, output $(C_4, -)$, $(V_4, +)$ or $(D_4, -)$ depending on whether $L = (1,1,4)$, $L = (2,2,2)$ or $L = (2,4)$ respectively. Terminate the algorithm.

Note that with this choice of resolvent, we have $H = C_4 =< (1234) >$, the group of cyclic permutations, but this fact is needed in checking the correctness of the algorithm, not in the algorithm itself, where only $G/H$ is used.

Another algorithm which is computationally slightly simpler is as follows. We give it also to illustrate the importance of the root ordering.

**Algorithm 6.3.7** (Galois Group for Degree 4). Given an irreducible monic polynomial $T \in \mathbb{Z}[X]$ of degree 4, this algorithm computes its Galois group.

1. [Compute resolvent] Using Algorithm 3.6.6, compute the roots $\theta_i$ of $T$ in $\mathbb{C}$. Let
$$F \leftarrow X_1X_3 + X_2X_4$$
and let $R \leftarrow R(F,T)$, where a system of representatives of $G/H$ is given by
$$G/H = \{I, (12), (14)\}.$$
Round the coefficients of $R$ to the nearest integer.
2. [Squarefree?] Compute $V \leftarrow (R, R')$ using the Euclidean algorithm. If $V$ is non-constant, replace $T$ by the polynomial obtained by applying a Tschirnhausen transformation using Algorithm 6.3.4 and go to step 1.
3. [Integral root?] Check whether $R$ has an integral root by explicitly computing them in terms of the $\theta_i$. (This is usually much faster than using the general factoring procedure 3.5.7.)
4. [Can one conclude?] If $R$ does not have an integral root (so $R$ is irreducible), then output $(A_4, +)$ or $(S_4, -)$ depending on whether $\mathrm{disc}(T)$ is a perfect square or not and terminate the algorithm. Otherwise, if $\mathrm{disc}(T)$ is a square, output $(V_4, +)$ and terminate the algorithm.

5. [Renumber] (Here $R$ has an integral root and $\mathrm{disc}(T)$ is not a square. The Galois group must be isomorphic either to $C_4$ or to $D_4$.) Let $\sigma$ be the element of $S_4$ corresponding to the integral root of $R$, and set $(t_i) \leftarrow (t_{\sigma(i)})$ (i.e. we renumber the roots of $T$ according to $\sigma$).

6. [Use new resolvent] Set

$$d \leftarrow ((\theta_1 - \theta_3)(\theta_2 - \theta_4)(\theta_1 + \theta_3 - \theta_2 - \theta_4))^2$$

rounded to the nearest integer (with the same remarks as before about the accuracy needed for the $\theta_i$). If $d \neq 0$, output $(C_4, -)$ or $(D_4, -)$ depending on whether $d$ is a perfect square or not and terminate the algorithm.

7. [Replace] (Here $d = 0$.) Replace $T$ by the polynomial obtained by applying a Tschirnhausen transformation $A$ using Algorithm 6.3.4. Set $\theta_i \leftarrow A(\theta_i)$ (which will be the roots of the new $T$). Reorder the $\theta_i$ so that $\theta_1\theta_3 + \theta_2\theta_4 \in \mathbb{Z}$, (only the 3 elements of $G/H$ given in step 1 need to be tried), then go to step 6.

In principle, this algorithm involves factoring polynomials of degree 3, hence is computationally simpler than the preceding algorithm, although its structure is more complicated due to the implicit use of two different resolvents. The first resolvent corresponds to $G = S_4$ and $H = D_4 = <(1234), (13)>$. The second resolvent corresponds to $F = X_1X_2^2 + X_2X_3^2 + X_3X_4^2 + X_4X_1^2$, $G = D_4$, $H = C_4$ and $G/H = \{I, (13)\}$, hence the polynomial of degree 2 need not be explicitly computed in order to find its arithmetic structure.

**Remark.** (This remark is valid in any degree.) As can be seen from the preceding algorithm, it is not really necessary to compute the resolvent polynomial $R$ explicitly, but only a sufficiently close approximation to its roots (which are known explicitly by definition). To check whether $R$ is squarefree or not can also be done by simply checking that $R$ does not have any multiple root (to sufficient accuracy). In fact, we have the following slight strengthening of Theorem 6.3.3 which can be proved in the same way.

**Proposition 6.3.8.** *We keep the notations of Theorem 6.3.3, but we do not necessarily assume that $R_G(F,T)$ is squarefree. If $R_G(F,T)$ has a* simple *root in $\mathbb{Z}$, then* $\mathrm{Gal}(T)$ *is conjugate under $G$ to a subgroup of $H$.*

This proposition shows that it is not necessary to assume $R_G(F,T)$ squarefree in order to apply the above algorithms, as well as any other which depend only on the existence of an integral root and not more generally on the degrees of the irreducible factors of $R_G(F,T)$. (This is the case for the algorithms that we give in degree 4 and 5.) This remark should of course be used when implementing these algorithms.

### 6.3.4 Degree 5

In degree 5 there are also (up to conjugacy) five transitive subgroups of $S_5$. These are $C_5$ (the cyclic group), $D_5$ (the dihedral group of order 10), $M_{20}$ (the metacyclic group of degree 5), $A_5$ and $S_5$.

Some inclusions are

$$C_5 \subset D_5 \subset A_5 \cap M_{20}.$$

The algorithm that we suggest is as follows.

**Algorithm 6.3.9** (Galois Group for Degree 5). Given an irreducible monic polynomial $T \in \mathbb{Z}[X]$ of degree 5, this algorithm computes its Galois group.

1. [Compute resolvent] Using Algorithm 3.6.6, compute the roots $\theta_i$ of $T$ in $\mathbb{C}$. Let

$$F \leftarrow X_1^2(X_2X_5 + X_3X_4) + X_2^2(X_1X_3 + X_4X_5) + X_3^2(X_1X_5 + X_2X_4) + X_4^2(X_1X_2 + X_3X_5) + X_5^2(X_1X_4 + X_2X_3)$$

   and let $R \leftarrow R(F,T)$, where a system of representatives of $G/H$ is given by

$$G/H = \{I, (12), (13), (14), (15), (25)\}.$$

   Round the coefficients of $R$ to the nearest integer.

2. [Squarefree?] Compute $V \leftarrow (R, R')$ using the Euclidean algorithm. If $V$ is non-constant, replace $T$ by the polynomial obtained by applying a Tschirnhausen transformation using Algorithm 6.3.4 and go to step 1.

3. [Factor resolvent] Factor $R$ using Algorithm 3.5.7. (Note that one can show that either $R$ is irreducible or $R$ has an integral root. So, as in the algorithm for degree 4, it may be better to compute the roots of $R$ which are known explicitly.)

4. [Can one conclude?] If $R$ is irreducible, then output $(A_5, +)$ or $(S_5, -)$ depending on whether $\mathrm{disc}(T)$ is a perfect square or not, and terminate the algorithm. Otherwise, if $\mathrm{disc}(T)$ is not a perfect square, output $(M_{20}, -)$ and terminate the algorithm.

5. [Renumber] (Here $R$ has an integral root and $\mathrm{disc}(T)$ is a square. The Galois group must be isomorphic either to $C_5$ or to $D_5$.) Let $\sigma$ be the element of $S_5$ corresponding to the integral root of $R$, and set $(t_i) \leftarrow (t_{\sigma(i)})$ (i.e. we renumber the roots of $T$ according to $\sigma$).

6. [Compute discriminant of new resolvent] Set

$$d \leftarrow (\theta_1\theta_2(\theta_2 - \theta_1) + \theta_2\theta_3(\theta_3 - \theta_2) + \theta_3\theta_4(\theta_4 - \theta_3) + \theta_4\theta_5(\theta_5 - \theta_4) + \theta_5\theta_1(\theta_1 - \theta_5))^2$$

rounded to the nearest integer (with the same remarks as before about the accuracy needed for the $\theta_i$). If $d \neq 0$, output $(C_5,+)$ or $(D_5,+)$ depending on whether $d$ is a perfect square or not, and terminate the algorithm.

7. [Replace] (Here $d = 0$.) Replace $T$ by the polynomial obtained by applying a Tschirnhausen transformation $A$ using Algorithm 6.3.4. Set $\theta_i \leftarrow A(\theta_i)$ (which will be the roots of the new $T$). Reorder the $\theta_i$ so that $F(\theta_1, \theta_1, \theta_3, \theta_4, \theta_5) \in \mathbb{Z}$ where $F$ is as in step 1, (only the 6 elements of $G/H$ given in step 1 need to be tried), then go to step 6.

The first resolvent corresponds to $G = S_5$ and

$$H = M_{20} = <(12345), (2354)>.$$

Step 6 corresponds implicitly to the use of the second degree resolvent obtained with $F = X_1X_2^2 + X_2X_3^2 + X_3X_4^2 + X_4X_5^2 + X_5X_1^2$, $G = D_5$, $H = C_5$ and $G/H = \{I, (12)(35)\}$.

### 6.3.5 Degree 6

In degree 6 there are up to conjugation, 16 transitive subgroups of $S_6$. The inclusion diagram is complicated, and the number of resolvent polynomials is high. The best way to study this degree is to work using *relative extensions*, that is study the number field $K$ as a quadratic or cubic extension of a cubic or quadratic subfield respectively, if they exist. This is done in [Oli2] and [BeMaOl].

In this book we have not considered relative extensions. Furthermore, when a sextic field is given by a sixth degree polynomial over $\mathbb{Q}$, it is not immediately obvious, even if it is theoretically possible, how to express it as a relative extension, although the POLRED Algorithm 4.4.11 often gives such information. Hence, we again turn to the heavier machinery of resolvent polynomials.

It is traditional to use the notation $G_k$ to denote a group of cardinality $k$. Also, special care must be taken when considering abstract groups. For example, the group $S_4$ occurs as two different conjugacy classes of $S_6$, one which is in $A_6$, the other which is not (the traditional notation would then be $S_4^+$ and $S_4^-$ respectively).

We will describe the groups as we go along the algorithm. There are many possible resolvents which can be used. The algorithm that we suggest has the advantage of needing a single resolvent, except in one case, similarly to degrees 4 and 5.

**Algorithm 6.3.10** (Galois Group for Degree 6). Given an irreducible monic polynomial $T \in \mathbb{Z}[X]$ of degree 6, this algorithm computes its Galois group.

1. [Compute resolvent] Using Algorithm 3.6.6, compute the roots $\theta_i$ of $T$ in $\mathbb{C}$. Let

$$F \leftarrow X_1^2X_5^2(X_2X_4+X_3X_6)+X_2^2X_4^2(X_1X_5+X_3X_6)+X_3^2X_6^2(X_1X_5+X_2X_4)$$
$$+X_1^2X_6^2(X_2X_5+X_3X_4)+X_2^2X_5^2(X_1X_6+X_3X_4)+X_3^2X_4^2(X_1X_6+X_2X_5)$$
$$+X_1^2X_3^2(X_2X_6+X_4X_5)+X_2^2X_6^2(X_1X_3+X_4X_5)+X_4^2X_5^2(X_1X_3+X_2X_6)$$
$$+X_1^2X_4^2(X_2X_3+X_5X_6)+X_2^2X_3^2(X_1X_4+X_5X_6)+X_5^2X_6^2(X_1X_4+X_2X_3)$$
$$+X_1^2X_2^2(X_3X_5+X_4X_6)+X_3^2X_5^2(X_1X_2+X_4X_6)+X_4^2X_6^2(X_1X_2+X_3X_5)$$

and let $R \leftarrow R(F,T)$, where a system of representatives of $G/H$ is given by

$$G/H = \{I, (12), (13), (14), (15), (16)\}.$$

Round the coefficients of $R$ to the nearest integer.

2. [Squarefree?] Compute $V \leftarrow (R, R')$ using the Euclidean algorithm. If $V$ is non-constant, replace $T$ by the polynomial obtained by applying a Tschirnhausen transformation using Algorithm 6.3.4 and go to step 1.

3. [Factor resolvent] Factor $R$ using Algorithm 3.5.7. If $R$ is irreducible, then go to step 5, otherwise let $L$ be the list of the degrees of the irreducible factors sorted in increasing order.

4. [Conclude]

a) If $L = (1,2,3)$, let $f_1$ be the irreducible factor of $R$ of degree equal to 3. Output $(C_6, -)$ or $(D_6, -)$ depending on whether $\mathrm{disc}(f_1)$ is a square or not.

b) If $L = (3,3)$, let $f_1$ and $f_2$ be the irreducible factors of $R$. If both $\mathrm{disc}(f_1)$ and $\mathrm{disc}(f_2)$ are not squares output $(G_{36}, -)$, otherwise output $(G_{18}, -)$. Note that $G_{36}^- = C_3^2 \rtimes C_2^2 \simeq D_3 \times D_3$, and $G_{18} = C_3^2 \rtimes C_2 \simeq C_3 \times D_3$.

c) If $L = (2,4)$ and $\mathrm{disc}(T)$ is a square, output $(S_4, +)$. Otherwise, if $L = (2,4)$ and $\mathrm{disc}(T)$ is not a square, let $f_1$ be the irreducible factor of degree 4 of $R$. Then output $(A_4 \times C_2, -)$ or $(S_4 \times C_2, -)$ depending on whether $\mathrm{disc}(f_1)$ is a square or not.

d) If $L = (1,1,4)$ then output $(A_4, +)$ or $(S_4, -)$ depending on whether $\mathrm{disc}(T)$ is a square or not.

e) If $L = (1,5)$, then output $(\mathrm{PSL}_2(\mathbb{F}_5), +)$ or $(\mathrm{PGL}_2(\mathbb{F}_5), -)$ depending on whether $\mathrm{disc}(T)$ is a square or not. Note that $\mathrm{PSL}_2(\mathbb{F}_5) \simeq A_5$ and that $\mathrm{PGL}_2(\mathbb{F}_5) \simeq S_5$.

f) Finally, if $L = (1,1,1,3)$, output $(S_3, -)$.

Then terminate the algorithm.

5. [Compute new resolvent] (Here our preceding resolvent was irreducible. Note that we do *not* have to reorder the roots.) Let

$$F \leftarrow X_1X_2X_3 + X_4X_5X_6$$

and let $R \leftarrow R(F,T)$, where a system of representatives of $G/H$ is now given by

$$G/H = \{I, (14), (15), (16), (24), (25), (26), (34), (35), (36)\}.$$

Round the coefficients of $R$ to the nearest integer.

6. [Squarefree?] Compute $V \leftarrow (R, R')$ using the Euclidean algorithm. If $V$ is non-constant, replace $T$ by the polynomial obtained by applying a Tschirnhausen transformation using Algorithm 6.3.4 and go to step 5.
7. [Factor resolvent] Factor $R$ using Algorithm 3.5.7 (Note that in this case either $R$ is irreducible, or it has an integral root, so again it is probably better to compute these 10 roots directly from the roots of $T$ and check whether they are integral.)
8. [Conclude] If $R$ is irreducible (or has no integral root), then output $(A_6, +)$ or $(S_6, -)$ depending on whether $\mathrm{disc}(T)$ is a square or not. Otherwise, output $(G_{36}, +)$ or $(G_{72}, -)$ depending on whether $\mathrm{disc}(T)$ is a square or not. Then terminate the algorithm. Note that $G_{36}^{+} = C_3^2 \rtimes C_4$ and $G_{72} = C_3^2 \rtimes D_4$.

The first resolvent corresponds to $G = S_6$ and

$$H = \mathrm{PGL}_2(\mathbb{F}_5) = <(12345), (16)(23)(45)>.$$

The second resolvent, used in step 5, corresponds to $G = S_6$ and

$$H = G_{72} = <(123), (14)(25)(36), (1524)(36)>.$$

**Remark.** It can be shown that a sextic field has a quadratic subfield if and only if its Galois group is isomorphic to a (transitive) subgroup of $G_{72}$. This corresponds to the groups $(C_6, -)$, $(S_3, -)$, $(D_6, -)$, $(G_{18}, -)$, $(G_{36}, -)$, $(G_{36}, +)$ and $(G_{72}, -)$.

Similarly, it has a cubic subfield if and only if its Galois group is isomorphic to a (transitive) subgroup of $S_4 \times C_2$. This corresponds to the groups $(C_6, -)$, $(S_3, -)$, $(D_6, -)$, $(A_4, +)$, $(S_4, +)$, $(S_4, -)$, $(A_4 \times C_2, -)$ and $(S_4 \times C_2, -)$.

Hence, it has both a quadratic and a cubic subfield if and only if its Galois group is isomorphic to $(C_6, -)$, $(S_3, -)$ or $(D_6, -)$.

If the field is primitive, i.e. does not have quadratic or cubic subfields, this implies that its Galois group can only be $\mathrm{PSL}_2(\mathbb{F}_5) \simeq A_5$, $\mathrm{PGL}_2(\mathbb{F}_5) \simeq S_5$, $A_6$ or $S_6$.

### 6.3.6 Degree 7

In degree 7, there are seven transitive subgroups of $S_7$ which are $C_7$, $D_7$, $M_{21}$, $M_{42}$, $\mathrm{PSL}_2(\mathbb{F}_7) \simeq \mathrm{PSL}_3(\mathbb{F}_2)$, $A_7$ and $S_7$.

Some inclusions are

$$C_7 \subset D_7 \subset M_{42} \ , \quad C_7 \subset M_{21} \subset \mathrm{PSL}_2(\mathbb{F}_7) \subset A_7 \quad \text{and} \quad M_{21} \subset M_{42}.$$

In this case there exists a remarkably simple algorithm.

**Algorithm 6.3.11** (Galois Group for Degree 7). Given an irreducible monic polynomial $T \in \mathbb{Z}[X]$ of degree 7, this algorithm computes its Galois group.

1. [Compute resolvent] Using Algorithm 3.6.6, compute the roots $\theta_i$ of $T$ in $\mathbb{C}$. Let
$$R \leftarrow \prod_{1 \le i < j < k \le 7} (X - (\theta_i + \theta_j + \theta_k))$$
which is a polynomial of degree 35, and round the coefficients of $R$ to the nearest integer.
2. [Squarefree?] Compute $V \leftarrow (R, R')$ using the Euclidean algorithm. If $V$ is non-constant, replace $T$ by the polynomial obtained by applying a Tschirnhausen transformation using Algorithm 6.3.4 and go to step 1.
3. [Factor resolvent and conclude] Factor $R$ using Algorithm 3.5.7. If $R$ is irreducible, then output $(A_7, +)$ or $(S_7, -)$ depending on whether $\mathrm{disc}(T)$ is a square or not. Otherwise, let $L$ be the list of the degrees of the irreducible factors sorted in increasing order. Output $(\mathrm{PSL}_2(\mathbb{F}_7), +)$, $(M_{42}, -)$, $(M_{21}, +)$, $(D_7, -)$ or $(C_7, +)$ depending on whether $L = (7, 28)$, $L = (14, 21)$, $L = (7, 7, 21)$, $L = (7, 7, 7, 14)$ or $L = (7, 7, 7, 7, 7)$ respectively. Then terminate the algorithm.

Note that this algorithm does not exactly correspond to the framework based on Theorem 6.3.3 but it has the advantage of being very simple, and computationally not too inefficient. It does involves factoring a polynomial of degree 35 over $\mathbb{Z}$ however, and this can be quite slow. (To give some idea of the speed: on a modern workstation the algorithms take a few seconds for degrees less than or equal to 6, while for degree 7, a few minutes may be required using this algorithm.)

Several methods can be used to improve this basic algorithm in practice. First of all, one expects that the overwhelming majority of polynomials will have $S_7$ as their Galois group, and hence that our resolvent will be irreducible. We can test for irreducibility, without actually factoring the polynomial, by testing this modulo $p$ for small primes $p$. If it is already irreducible modulo $p$ for some $p$, then there is no need to go any further. Of course, this is done automatically if we use Algorithm 3.5.7, but that algorithm will start by doing the distinct degree factorization 3.4.3, when it is simpler here to use Proposition 3.4.4.

Even if one expects that the resolvent will factor, we can use the divisibility by 7 of the degrees of its irreducible factors in almost every stage of the factoring Algorithm 3.5.7.

Another idea is to use the resolvent method as explained at the beginning of this chapter. Instead of factoring polynomials having large degrees, we simply find the list of all cosets $\sigma$ of $G$ modulo $H$ such that
$$F\left(\theta_{\sigma(1)}, \theta_{\sigma(2)}, \ldots, \theta_{\sigma(n)}\right) \in \mathbb{Z}.$$
If there is more than one coset, this means that the resolvent is not squarefree, hence we must apply a Tschirnhausen transformation. If there is exactly one, then the Galois group is isomorphic to a subgroup of $H$, and the coset gives

the permutation of the roots which must be applied to go further down the tree of subgroups. If there are none, the Galois group is not isomorphic to a subgroup of $H$. Of course, all this applies to any degree, not only to degree 7.

As the reader can see, I do not give explicitly the resolvents and cosets for degree 7. The resolvents themselves are as simple as the ones that we have given in lower degrees. On the other hand, the list of cosets is long. For example for the pair $(S_7, M_{42})$ we need 120 elements. This is cumbersome to write down. It should be noted however that the resulting algorithm is much more efficient than the preceding one (again at most a few seconds on a modern workstation). These cosets and resolvents in degree 7, 8, 9, 10 and 11 may be obtained in electronic form upon request from M. Olivier (same address as the author).

### 6.3.7 A List of Test Polynomials

As a first check of the correctness of an implementation of the above algorithms, we give a polynomial for each of the possible Galois groups occurring in degree less than or equal to 7. This list is taken from [Soi-McKay]. Note that for many of the given polynomials, it will be necessary to apply a Tschirnhausen transformation. We list first the group as it is output by the algorithm, then a polynomial having this as Galois group.

$(S_1, -)$: $X$
$(S_2, -)$: $X^2 + X + 1$
$(C_3, +)$: $X^3 + X^2 - 2X - 1$
$(S_3, -)$: $X^3 + 2$
$(C_4, -)$: $X^4 + X^3 + X^2 + X + 1$
$(V_4, +)$: $X^4 + 1$
$(D_4, -)$: $X^4 - 2$
$(A_4, +)$: $X^4 + 8X + 12$
$(S_4, -)$: $X^4 + X + 1$
$(C_5, +)$: $X^5 + X^4 - 4X^3 - 3X^2 + 3X + 1$
$(D_5, +)$: $X^5 - 5X + 12$
$(M_{20}, -)$: $X^5 + 2$
$(A_5, +)$: $X^5 + 20X + 16$
$(S_5, -)$: $X^5 - X + 1$
$(C_6, -)$: $X^6 + X^5 + X^4 + X^3 + X^2 + X + 1$
$(S_3, -)$: $X^6 + 108$
$(D_6, -)$: $X^6 + 2$
$(A_4, +)$: $X^6 - 3X^2 - 1$
$(G_{18}, -)$: $X^6 + 3X^3 + 3$
$(A_4 \times C_2, -)$: $X^6 - 3X^2 + 1$
$(S_4, +)$: $X^6 - 4X^2 - 1$

$(S_4,-)$: $X^6-3X^5+6X^4-7X^3+2X^2+X-4$
$(G_{36},-)$: $X^6+2X^3-2$
$(G_{36},+)$: $X^6+6X^4+2X^3+9X^2+6X-4$
$(S_4\times C_2,-)$: $X^6+2X^2+2$
$(\mathrm{PSL}_2(\mathbb{F}_5),+)\simeq(A_5,+)$: $X^6-2X^5-5X^2-2X-1$
$(G_{72},-)$: $X^6+2X^4+2X^3+X^2+2X+2$
$(\mathrm{PGL}_2(\mathbb{F}_5),-)\simeq(S_5,-)$: $X^6-X^5-10X^4+30X^3-31X^2+7X+9$
$(A_6,+)$: $X^6+24X-20$
$(S_6,-)$: $X^6+X+1$
$(C_7,+)$: $X^7+X^6-12X^5-7X^4+28X^3+14X^2-9X+1$
$(D_7,-)$: $X^7+7X^3+7X^2+7X-1$
$(M_{21},+)$: $X^7-14X^5+56X^3-56X+22$
$(M_{42},-)$: $X^7+2$
$(\mathrm{PSL}_2(\mathbb{F}_7),+)\simeq(\mathrm{PSL}_3(\mathbb{F}_2),+)$: $X^7-7X^3+14X^2-7X+1$
$(A_7,+)$: $X^7+7X^4+14X+3$
$(S_7,-)$: $X^7+X+1$

## 6.4 Examples of Families of Number Fields

### 6.4.1 Making Tables of Number Fields

It is important to try to describe the family of all number fields (say of a given degree, Galois group of the Galois closure and signature) up to isomorphism. Unfortunately, this is a hopeless task except for some special classes of fields such as quadratic fields, cyclic cubic fields, cyclotomic fields, etc. We could, however, ask for a list of such fields whose discriminant is in absolute value bounded by a given constant, i.e. ask for *tables* of number fields. We first explain briefly how this can be done, referring to [Mart] and [Poh1] for complete details.

We need two theorems. The first is an easy result of the geometry of numbers (which we already used in Section 2.6 to show that the LLL algorithm terminates) which we formulate as follows.

**Proposition 6.4.1.** *There exists a positive constant $\gamma_n$ having the following property. In any lattice $(L,q)$ of $\mathbb{R}^n$, there exists a non-zero vector $x$ such that $q(x)\le\gamma_n D^{2/n}$ where $D=\det(L)=\det(Q)^{1/2}$ is the determinant of the lattice (here $Q$ is the matrix of $q$ in some $\mathbb{Z}$-basis of $L$, see Section 2.5).*

See for example [Knu2] (Section 3.3.4, Exercise 9) for a proof.

The best possible constant $\gamma_n$ is called Hermite's constant, and is known only for $n\le 8$:

$$\gamma_1=1,\ \gamma_2^2=\frac{4}{3},\ \gamma_3^3=2,\ \gamma_4^4=4,\ \gamma_5^5=8,\ \gamma_6^6=\frac{64}{3},\ \gamma_7^7=64,\ \gamma_8^8=256.$$

For larger values of $n$, the recursive upper bound

$$\gamma_n^n \leq \gamma_{n-1}^{(n-1)n/(n-2)}$$

gives useful results. The best known bounds are given for $n \leq 24$ in [Con-Slo], Table 1.2 and Formula (47).

The basic theorem, due to Hunter (see [Hun] and Exercise 26), is as follows.

**Theorem 6.4.2** (Hunter). *Let $K$ be a number field of degree $n$ over $\mathbb{Q}$. There exists $\theta \in \mathbb{Z}_K \setminus \mathbb{Z}$ having the following property. Call $\theta_i$ the conjugates of $\theta$ in $K$. Then*

$$\sum_{i=1}^{n} |\theta_i|^2 \leq \frac{1}{n}\operatorname{Tr}(\theta)^2 + \gamma_{n-1}\left(\frac{|d(K)|}{n}\right)^{1/(n-1)},$$

*where $d(K)$ is the discriminant of $K$ and $\operatorname{Tr}(\theta) = \sum_{i=1}^{n} \theta_i$ is the trace of $\theta$ over $\mathbb{Q}$. In addition, we may assume that $0 \leq \operatorname{Tr}(\theta) \leq n/2$.*

This theorem is used as follows. Assume that we want to make a table of number fields of degree $n$ and having a given signature, with discriminant $d(K)$ satisfying $|d(K)| \leq M$ for a given bound $M$. Then replacing $d(K)$ by $M$ in Hunter's theorem gives an upper bound for the $|\theta_i|$ and hence for the coefficients of the characteristic polynomial of $\theta$ in $K$.

If $K$ is primitive, i.e. if the only subfields of $K$ are $\mathbb{Q}$ and $K$ itself, then since $\theta \notin \mathbb{Z}$ we know that $K = \mathbb{Q}(\theta)$, and thus we obtain a finite (although usually large) collection of polynomials to consider. Most of these polynomials can be discarded because their roots will not satisfy Hunter's inequality. Others can be discarded because they are reducible, or because they do not have the correct signature. Note that a given signature will give several inequalities between the coefficients of acceptable polynomials, and these should be checked before using Sturm's Algorithm 4.1.11 which is somewhat longer. (We are talking of millions if not billions of candidate polynomials here, depending on the degree and, of course, the size of $M$.)

Finally, using Algorithm 6.1.8 compute the discriminant of the number fields corresponding to each of the remaining polynomials. This is the most time-consuming part. After discarding the polynomials which give a field discriminant which is larger than $M$ in absolute value, we have a list of polynomials which define all the number fields that we are interested in. Many polynomials may give the same number field, so this is the next thing to check. Since we have computed an integral basis for each polynomial during the computation of the discriminant of the corresponding number field, we can use the POLRED algorithm (or more precisely Algorithm 4.4.12) to give a pseudo-canonical polynomial for each number field. This will eliminate practically all the coincidences.

When two distinct polynomials give the same field discriminant, we must now check whether or not the corresponding number fields are isomorphic,

and this is done by using one of the algorithms given in Section 4.5.4. Note that this will now occur very rarely (since most cases have been dealt with using Algorithm 4.4.12).

If the field $K$ is not primitive, we must use a relative version of Hunter's theorem due to Martinet (see [Mart]), and make a separate table of imprimitive fields.

In the rest of this chapter we will give some examples of families of number fields.

The simplest of all number fields (apart from $\mathbb{Q}$ itself) are quadratic fields. This case has been studied in detail in Chapter 5, and we have also seen that there exist methods for computing regulators and class groups which do not immediately generalize to higher degree fields. Note also that higher degree fields are not necessarily Galois.

The next simplest case is probably that of cyclic cubic fields, which we now consider.

### 6.4.2 Cyclic Cubic Fields

Let $K$ be a number field of degree 3 over $\mathbb{Q}$, i.e. a cubic field. If $K$ is Galois over $\mathbb{Q}$, its Galois group must be isomorphic to the cyclic group $\mathbb{Z}/3\mathbb{Z}$, hence we say that $K$ is a cyclic cubic field. The Galois group has, apart from its identity element, two other elements which are inverses. We denote them by $\sigma$ and $\sigma^{-1} = \sigma^2$. The first proposition to note is as follows.

**Proposition 6.4.3.** *Let $K = \mathbb{Q}(\theta)$ be a cubic field, where $\theta$ is an algebraic integer whose minimal monic polynomial will be denoted $P(X)$. Then $K$ is a cyclic cubic field if and only if the discriminant of $P$ is a square.*

*Proof.* This is a restatement of Proposition 6.3.5. □

This proposition clearly gives a trivial algorithm to check whether a cubic field is Galois or not.

In the rest of this (sub)section, we assume that $K$ is a cyclic cubic field. Our first task is to determine a general equation for such fields. Let $\theta$ be an algebraic integer such that $K = \mathbb{Q}(\theta)$, and let $P(X) = X^3 - SX^2 + TX - N$ be the minimal monic polynomial of $\theta$, with integer coefficients $S$, $T$ and $N$.

Note first that since any cubic field has at least one real embedding (as does any odd degree field) and since $K$ is Galois, all the roots of $P$ must be in $K$ hence they must all be real, so a cyclic cubic field must be totally real (i.e. $r_1 = 3$ real embeddings, and $r_2 = 0$ complex ones). Of course, this also follows because the discriminant of $P$ is a square.

In what follows, we set $\zeta = e^{2i\pi/3}$, i.e. a primitive cube root of unity. Since $K$ is totally real, $\zeta \notin K$, hence the extension field $K(\zeta)$ is a sixth degree field over $\mathbb{Q}$. It is easily checked that it is still Galois, with Galois group generated

by commuting elements $\sigma$ and $\tau$, where $\sigma$ acts on $K$ as above and trivially on $\zeta$, and $\tau$ denotes complex conjugation.

The first result that we need is as follows.

**Lemma 6.4.4.** *Set $\gamma = \theta + \zeta^2\sigma(\theta) + \zeta\sigma^2(\theta) \in K(\zeta)$, and $\beta = \gamma^2/\tau(\gamma)$. Then $\beta \in \mathbb{Q}(\zeta)$ and we have*

$$P(X) = X^3 - SX^2 + \frac{S^2 - e}{3}X - \frac{S^3 - 3Se + eu}{27},$$

*where we have set $e = \beta\tau(\beta)$ and $u = \beta + \tau(\beta)$ (i.e. $e$ and $u$ are the norm and trace of $\beta$ considered as an element of $\mathbb{Q}(\zeta)$).*

*Proof.* We have $\tau(\gamma) = \theta + \zeta\sigma(\theta) + \zeta^2\sigma^2(\theta)$. One sees immediately that $\sigma(\gamma) = \zeta\gamma$ and $\sigma(\tau(\gamma)) = \zeta^2\tau(\gamma)$ hence $\beta$ is invariant under the action of $\sigma$, so by Galois theory $\beta$ must belong to the quadratic subfield $\mathbb{Q}(\zeta)$ of $K(\zeta)$. In particular, $e$ and $u$ as defined above are in $\mathbb{Q}$. Now we have the matrix equation

$$\begin{pmatrix} S \\ \gamma \\ \tau(\gamma) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \zeta^2 & \zeta \\ 1 & \zeta & \zeta^2 \end{pmatrix} \begin{pmatrix} \theta \\ \sigma(\theta) \\ \sigma^2(\theta) \end{pmatrix},$$

so it follows by inverting the matrix that

$$\begin{pmatrix} \theta \\ \sigma(\theta) \\ \sigma^2(\theta) \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \zeta & \zeta^2 \\ 1 & \zeta^2 & \zeta \end{pmatrix} \begin{pmatrix} S \\ \gamma \\ \tau(\gamma) \end{pmatrix}.$$

From the formulas $T = \theta\sigma(\theta) + \theta\sigma^2(\theta) + \sigma(\theta)\sigma^2(\theta)$ and $N = \theta\sigma(\theta)\sigma^2(\theta)$, a little computation gives the result of the lemma. □

We will now modify $\theta$ (hence its minimal polynomial $P(X)$) so as to obtain a unique equation for each cyclic cubic field. First note that replacing $\gamma$ by $(b + c\zeta)\gamma$ is equivalent to changing $\theta$ into $b\theta + c\sigma(\theta)$, and $\beta$ is changed into

$$\beta\frac{(b + c\zeta)^2}{b + c\zeta^2}.$$

Let $p_k$ be the primes which split in $\mathbb{Q}(\zeta)$ (as $p_k = \pi_k\overline{\pi_k}$), i.e. such that $p_k \equiv 1 \pmod 3$, let $q_k$ be the inert primes, i.e. such that $q_k \equiv 2 \pmod 3$, and let $\rho = 1 + 2\zeta = \sqrt{-3}$ be a ramified element (i.e. a prime element above the prime 3). We can write

$$b + c\zeta = (-\zeta)^g\rho^f \prod \pi_k^{e_k} \prod \overline{\pi_k}^{f_k} \prod q_k^{g_k}.$$

Hence, since $b + c\zeta^2 = \overline{b + c\zeta}$, we have

$$\frac{(b+c\zeta)^2}{b+c\zeta^2} = (-1)^{g+f}\rho^f \prod \pi_k^{2e_k-f_k} \prod \overline{\pi_k}^{2f_k-e_k} \prod q_k^{g_k}.$$

If the decomposition of $\beta$ (which is in $\mathbb{Q}(\zeta)$ but perhaps not in $\mathbb{Z}[\zeta]$) is

$$\beta = (-\zeta)^n \rho^m \prod \pi_k^{l_k} \prod \overline{\pi_k}^{m_k} \prod q_k^{n_k}$$

then we can choose $g_k = -n_k$ and $f = -m$. Furthermore, for each $k$ consider the quantity $m_k + 2l_k$. If it is congruent to 0 or 1 modulo 3, we will choose $e_k = \lfloor(-m_k - 2l_k + 1)/3\rfloor$ and $f_k = l_k + 2e_k$. If it is congruent to 2 modulo 3, then $l_k + 2m_k \equiv 1 \pmod 3$ and we choose $f_k = \lfloor(-l_k - 2m_k + 1)/3\rfloor$ and $e_k = m_k + 2f_k$.

It is easy to check that, with this choice of exponents, the new value of $\beta$ is an element of $\mathbb{Z}[\zeta]$ (and not only of $\mathbb{Q}(\zeta)$), is not divisible by any inert or ramified prime, and is divisible by split primes only to the first power. Also, at most one of $\pi_k$ or $\overline{\pi_k}$ divides $\beta$. In other words, if $e = \beta\tau(\beta)$ is the new value of the norm of $\beta$, then $e$ is equal to a product of distinct primes congruent to 1 modulo 3.

Finally, since $1 + \zeta + \zeta^2 = 0$, if we change $\theta$ into $a + \theta$ with $a \in \mathbb{Q}$, then $\gamma$ does not change and so neither do $\beta$ or $e$. Taking $a = S/3$, we obtain a new value of $\theta$ whose trace is equal to 0. Putting all this together we have almost proved the following lemma.

**Lemma 6.4.5.** *For any cyclic cubic field $K$, there exists a unique pair of integers $e$ and $u$ such that $e$ is equal to a product of distinct primes congruent to 1 modulo 3, $u \equiv 2 \pmod 3$ and such that $K = \mathbb{Q}(\theta')$ where $\theta'$ is a root of the polynomial*

$$Q(X) = X^3 - \frac{e}{3}X - \frac{eu}{27},$$

*or equivalently $K = \mathbb{Q}(\theta)$ where $\theta$ is a root of*

$$P(X) = 27Q(X/3) = X^3 - 3eX - eu.$$

*Proof.* Since $\beta = (u + v\sqrt{-3})/2$, $u$ cannot be divisible by 3 since $\beta$ is not divisible by the ramified prime. Hence, by suitably choosing the exponent $g$ above (which amounts to changing $\beta$ into $-\beta$ if necessary), we may assume $u \equiv 2 \pmod 3$.

For the uniqueness statement, note that all the possible choices of generators of $K$ are of the form $a + b\theta + c\sigma(\theta)$, and since we want a trace equal to 0, this gives us the value of $a$ as a function of $b$ and $c$, where these last values are determined because we want $e$ to be equal to a product of primes congruent to 1 modulo 3, hence $\beta$ is unique. The last statement is trivial. □

We can now state the main theorem of this section.

**Theorem 6.4.6.** *All cyclic cubic fields $K$ are given exactly once (up to isomorphism) in the following way.*

(1) *If the prime 3 is ramified in $K$, then $K = \mathbb{Q}(\theta)$ where $\theta$ is a root of the equation with coefficients in $\mathbb{Z}$*

$$P(X) = X^3 - \frac{e}{3}X - \frac{eu}{27}, \quad \text{where}$$

$$e = \frac{u^2 + 27v^2}{4},\ u \equiv 6 \pmod 9,\ 3 \nmid v,\ u \equiv v \pmod 2,\ v > 0$$

*and $e/9$ is equal to the product of distinct primes congruent to 1 modulo 3.*

(2) *If the prime 3 is unramified in $K$, then $K = \mathbb{Q}(\theta)$ where $\theta$ is a root of the equation with coefficients in $\mathbb{Z}$*

$$P(X) = X^3 - X^2 + \frac{1-e}{3}X - \frac{1-3e+eu}{27}, \quad \text{where}$$

$$e = \frac{u^2 + 27v^2}{4},\ u \equiv 2 \pmod 3,\ u \equiv v \pmod 2,\ v > 0$$

*and $e$ is equal to the product of distinct primes congruent to 1 modulo 3.*

*In both cases, the discriminant of $P$ is equal to $e^2v^2$ and the discriminant of the number field $K$ is equal to $e^2$.*

(3) *Conversely, if $e$ is equal to 9 times the product of $t-1$ distinct primes congruent to 1 modulo 3, (resp. is equal to the product of $t$ distinct primes congruent to 1 modulo 3), then there exists up to isomorphism exactly $2^{t-1}$ cyclic cubic fields of discriminant $e^2$ defined by the polynomials $P(X)$ given in (1) (resp. (2)).*

To prove this theorem, we will need in particular to compute explicitly integral bases and discriminants of cyclic cubic fields. Although there are other (essentially equivalent) methods, we will apply the round 2 algorithm to do this.

So, let $K$ be a cyclic cubic field. By Lemma 6.4.5, we have $K = \mathbb{Q}(\theta)$ where $\theta$ is a root of the equation

$$P(X) = X^3 - 3eX - eu, \quad \text{where} \quad e = \frac{u^2 + 3v^2}{4}, \quad u \equiv 2 \pmod 3$$

and $e$ is equal to a product of distinct primes congruent to 1 modulo 3.

We first prove a few lemmas.

**Lemma 6.4.7.** *Let $p \mid e$. Then the order $\mathbb{Z}[\theta]$ is $p$-maximal.*

*Proof.* We apply Dedekind's criterion. Since $p \mid e$, $\overline{P}(X) = X^3$, therefore with the notations of Theorem 6.1.4, $t_1(X) = X$, $g(X) = X$, $h(X) = X^2$

and $f(X) = (3e/p)X + eu/p$. Since $p \mid e$ we cannot have $p \mid u$, otherwise $p \mid v$, hence $p^2 \mid e$ which was assumed not to be true. Therefore, $p \nmid eu/p$ so $(\overline{f}, \overline{g}, \overline{h}) = 1$, showing that $\mathbb{Z}[\theta]$ is $p$-maximal. □

**Corollary 6.4.8.** *The discriminant of $P(X)$ is equal to $81e^2v^2$. The discriminant of the number field $K$ is divisible by $e^2$.*

*Proof.* The discriminant of $X^3 + aX + b$ is equal to $-(4a^3 + 27b^2)$ (see Exercise 7 of Chapter 3), hence the discriminant of $P$ is equal to

$$-(4(-3e)^3 + 27e^2u^2) = -27e^2(u^2 - 4e) = 81e^2v^2$$

thus proving the first formula. For the second, we know that the discriminant of the field $K$ is a square divisor of $81e^2v^2$. By the preceding lemma, $\mathbb{Z}[\theta]$ is $p$-maximal for all primes dividing $e$, and since $e$ is coprime to $81v^2$, the primes for which $\mathbb{Z}[\theta]$ may not be $p$-maximal are divisors of $81v^2$, hence the discriminant of $K$ is divisible by $e^2$. □

Since, as we will see, the prime divisors of $v$ other than 3 are irrelevant, what remains is to look at the behavior of the prime 3.

**Lemma 6.4.9.** *Assume that $3 \nmid v$. Then $\mathbb{Z}[\theta]$ is 3-maximal.*

*Proof.* Again we use Dedekind's criterion. Since $eu \equiv 2 \pmod 3$, we have $\overline{P} = (X+1)^3$ in $\mathbb{F}_3[X]$ hence $t_1(X) = X+1$, $g(X) = X+1$, $h(X) = (X+1)^2$ and $f(X) = X^2 + (e+1)X + (1+eu)/3 = (X+1)(X+e) + (eu+1-3e)/3$ hence

$$(\overline{f}, \overline{g}, \overline{h}) = (X+1, \overline{f}) = (X+1, \overline{(eu+1-3e)/3}).$$

Now we check that

$$r = \frac{eu+1-3e}{3} = \frac{(u^2+3v^2)(u-3)+4}{12} = \frac{(u-2)^2(u+1)+3v^2(u-3)}{12}$$

hence, since $u \equiv 2 \pmod 3$, $4r \equiv v^2(u-3) \pmod 9$ and, in particular, since $3 \nmid v$, $r \equiv 1 \pmod 3$ so $(\overline{f}, \overline{g}, \overline{h}) = 1$, which proves the lemma. □

**Lemma 6.4.10.** *With the above notation, let $\theta$ be a root of $P(X) = X^3 - 3eX - eu$, where $e = (u^2+3v^2)/4$ and $u \equiv 2 \pmod 3$. The conjugates of $\theta$ are given by the formulas*

$$\sigma(\theta) = \frac{-2e}{v} - \frac{u+v}{2v}\theta + \frac{1}{v}\theta^2,$$

$$\sigma^2(\theta) = \frac{2e}{v} + \frac{u-v}{2v}\theta - \frac{1}{v}\theta^2.$$

*Proof.* From the proof of Proposition 6.4.3, we have $f = (\theta-\theta_2)(\theta_2-\theta_3)(\theta_3-\theta) = \pm 9ev$ (since the discriminant is equal to $81e^2v^2$). If necessary, by exchanging $\theta_2$ and $\theta_3$, we may assume that $\theta_2 - \theta_3 = 9ev/(\theta-\theta_2)(\theta-\theta_3) = 9ev/P'(\theta) = 9ev/(3\theta^2 - 3e)$. Using the extended Euclidean algorithm with $A(X) = X^3 - 3eX - eu$ and $B(X) = X^2 - e$, one finds immediately that the inverse of $B$ modulo $A$ is equal to $(2X^2 - uX - 4e)/(3v^2e)$ hence

$$\theta_2 - \theta_3 = \frac{1}{v}(2\theta^2 - u\theta - 4e).$$

On the other hand, since the trace of $\theta$ is equal to 0, we have $\theta_2 + \theta_3 = -\theta$, and the formulas for $\theta_2 = \sigma(\theta)$ and $\theta_3 = \sigma^2(\theta)$ follow immediately.

It would of course have been simple, but less natural, to check directly with the given formulas that $(X-\theta)(X-\sigma(\theta))(X-\sigma^2(\theta)) = X^3 - 3eX - eu$. □

We can now prove a theorem which immediately implies the first two statements of Theorem 6.4.6.

**Theorem 6.4.11.** *Let $K = \mathbb{Q}(\theta)$ be a cyclic cubic field where $\theta$ is a root of $X^3-3eX-eu = 0$ and where, as above, $e = (u^2+3v^2)/4$ is equal to a product of distinct primes congruent to 1 modulo 3.*

(1) *Assume that $3 \nmid v$. Then $(1, \theta, \sigma(\theta))$ (where $\sigma(\theta)$ is given by the above formula) is an integral basis of $K$ and the discriminant of $K$ is equal to $(9e)^2$.*
(2) *Assume now that $3 \mid v$. Then, if $\theta' = (\theta+1)/3$, $(1, \theta', \sigma(\theta'))$ is an integral basis of $K$ and the discriminant of $K$ is equal to $e^2$.*

*Proof.* 1) Since $\theta^2 = v\sigma(\theta) + ((u+v)/2)\theta + 2e$, the $\mathbb{Z}$-module $\mathcal{O}$ generated by $(1, \theta, \sigma(\theta))$ contains $\mathbb{Z}[\theta]$. One computes immediately (in fact simply from the formula that we have just given for $\theta^2$) that $\mathbb{Z}[\theta]$ is of index $v$ in $\mathcal{O}$. Hence, the discriminant of $\mathcal{O}$ is equal to $81e^2$. Since we know that $\mathbb{Z}[\theta]$, and a fortiori that $\mathcal{O}$ is 3-maximal and $p$-maximal for every prime dividing $e$, it follows that $\mathcal{O}$ is the maximal order, thus proving the first part of the theorem.

2) We now consider the case where $3 \mid v$. The field $K$ can then be defined by the polynomial

$$Q(X) = P(3X-1)/27 = X^3 - X^2 + \frac{1-e}{3}X - \frac{1-3e+eu}{27}.$$

Since $e \equiv 1 \pmod 3$, $u \equiv 2 \pmod 3$ and $3 \mid v$, a simple calculation shows that $Q \in \mathbb{Z}[X]$. Furthermore, from Proposition 3.3.5 the discriminant of $Q$ is equal to the discriminant of $P$ divided by $3^6$, i.e. to $e^2(v/3)^2$. Set $\theta' = (\theta+1)/3$, which is a root of $Q$, and let $\mathcal{O}$ be the $\mathbb{Z}$-module generated by $(1, \theta', \sigma(\theta'))$. We compute that

$$\sigma(\theta') = \frac{2+u+3v-4e}{6v} - \frac{4+u+v}{2v}\theta' + \frac{3}{v}\theta'^2$$

and so, as in the proof of the first part, one checks that $\mathcal{O} \supset \mathbb{Z}[\theta']$ and $[\mathcal{O} : \mathbb{Z}[\theta']] = v/3$. Therefore the discriminant of $\mathcal{O}$ is equal to $e^2$. By Corollary 6.4.8 the discriminant of $K$ must also be divisible by $e^2$, and so the theorem follows. □

*Proof of Theorem 6.4.6.* First, we note that the polynomials given in Theorem 6.4.6 are irreducible in $\mathbb{Q}[X]$ (see Exercise 17).

From Theorem 6.4.11, one sees immediately that 3 is ramified in $K$ (i.e. 3 divides the discriminant of $K$) if and only if $3 \nmid v$. Hence, Lemma 6.4.5 tells us that $K$ is given by an equation $P(X) = X^3 - 3eX - eu$ (with several conditions on $e$ and $u$). If we set $u_1 = 3u$, $v_1 = v$ and $e_1 = 9e$, we have $e_1 = (u_1^2 + 27v_1^2)/4$, $u_1 \equiv 6 \pmod 9$, $3 \nmid v_1$, and $P(X) = X^3 - (e_1/3)X - (e_1u_1)/27$ as claimed in Theorem 6.4.6 (1).

Assume now that 3 is not ramified, i.e. that $3 \mid v$. From the proof of the second part of Theorem 6.4.11, we know that $K$ can be defined by the polynomial $X^3 - X^2 + ((1-e)/3)X - (1-3e+eu)/27 \in \mathbb{Z}[X]$ and this time setting $e_1 = e$, $v_1 = v/3$ and $u_1 = u$, it is clear that the second statement of Theorem 6.4.6 follows.

We still need to prove that any two fields defined by different polynomials $P(X)$ given in (1) or (2) are not isomorphic, i.e. that the pair $(e, u)$ determines the isomorphism class. This follows immediately from the uniqueness statement of Lemma 6.4.5. (Note that the $e$ and $u$ in Lemma 6.4.5 are either equal to the $e$ and $u$ of the theorem (in case (2)), or to $e/9$ and $u/3$ (in case (1)).)

Let us prove (3). Assume that $e$ is equal to a product of $t$ distinct primes congruent to 1 modulo 3 (the case where $e$ is equal to 9 times the product of $t-1$ distinct primes congruent to 1 modulo 3 is dealt with similarly, see Exercise 18). Let $A = \mathbb{Z}[(1+\sqrt{-3})/2]$ be the ring of algebraic integers of $\mathbb{Q}(\sqrt{-3})$. It is trivial to check (and in fact we have already implicitly used this in the proof of (2)) that if $\alpha \in A$ with $3 \nmid \mathcal{N}(\alpha)$, there exists a unique $\alpha'$ associate to $\alpha$ (i.e. generating the same principal ideal) such that

$$\alpha' = (u + 3v\sqrt{-3})/2\,, \qquad u \equiv 2 \pmod 3.$$

Furthermore, since $A$ is a Euclidean domain and in particular a PID, Proposition 5.1.4 shows that if $p_i$ is a prime congruent to 1 modulo 3, then $p_i = \alpha_i\overline{\alpha}_i$ for a unique $\alpha_i = (u_i + 3v_i\sqrt{-3})/2$ with $u_i \equiv 2 \pmod 3$ and $v_i > 0$.

Hence, if $e = \prod_{1 \le i \le t} p_i$, then $e = (u^2 + 27v^2)/4 = \mathcal{N}(u + 3v\sqrt{-3})/2$ if and only if

$$(u + 3v\sqrt{-3})/2 = \prod_{1 \le i \le t} \beta_i$$

where $\beta_i = \alpha_i$ or $\beta_i = \overline{\alpha}_i$, and this gives $2^t$ solutions to the equation $e = (u^2 + 27v^2)/4$. (Note that using associates of $\beta_i$ do not give any new solutions.)

But, we have seen above that the isomorphism class of a cyclic cubic field is determined uniquely by the pair $(e,u)$ satisfying appropriate conditions. Since $e=(u^2+27(-v)^2)/4$ gives the same field as $e=(u^2+27v^2)/4$, this shows, as claimed, that there exist exactly $2^{t-1}$ distinct values of $u$, hence $2^{t-1}$ non-isomorphic fields of discriminant $e^2$. This finishes the proof of Theorem 6.4.6. □

**Corollary 6.4.12.** *With the notation of Theorem 6.4.6 (i.e.* not *those of Theorem 6.4.11), the conjugates of $\theta$ are given by the formula*

$$\sigma^{\pm1}(\theta)=\mp\frac{2e}{9v}+\frac{-3v\mp u}{6v}\theta\pm\frac{1}{v}\theta^2$$

*when 3 is ramified in $K$ (i.e. in case (1)), and by the formula*

$$\sigma^{\pm1}(\theta)=\frac{9v\pm(u+2-4e)}{18v}+\frac{-3v\mp(u+4)}{6v}\theta\pm\frac{1}{v}\theta^2$$

*when 3 is not ramified in $K$ (i.e. in case (2)).*

*In addition, in all cases the discriminant of the polynomial $P$ is equal to $e^2v^2$, the discriminant of the field $K$ is equal to $e^2$ and $(1,\theta,\sigma(\theta))$ is an integral basis of $K$.*

The proof of this corollary follows immediately from Lemma 6.4.10 and the proof of Theorems 6.4.11 and 6.4.6. □

For another way to describe cyclic cubic fields parametrically see Exercise 21.

### 6.4.3 Pure Cubic Fields

Another class of fields which is easy to describe is the class of pure cubic fields, i.e. fields $K=\mathbb{Q}(\sqrt[3]{m})$ where $m$ is an integer which we may assume not to be divisible by a cube other than $\pm1$.

The defining polynomial is $P(X)=X^3-m$ whose discriminant is equal to $-27m^2$. Let $\theta$ be the root of this polynomial which is in $K$.

As in the case of cyclic cubic fields, we must compute the maximal order of $K$. This is very easy to do using Dedekind's criterion (see Exercise 2). I would like to show however how the Pohst-Zassenhaus Theorem 6.1.3 is really used in the round 2 algorithm, so I will deliberately skip the steps of Algorithm 6.1.8 which use the Dedekind criterion. This will of course make the computations longer, but will illustrate the full use of the round 2 algorithm.

Let $p$ be a prime dividing $m$ and not equal to 3. Then $p^2$ divides the discriminant of $P$. Let $r$ be 1 if $p\equiv1\pmod3$, $r=2$ if not. Then, clearly $\theta^p=m^{(p-r)/3}\theta^r$. Hence, in the basis $1$, $\theta$, $\theta^2$ the matrix of the Frobenius at $p$ (or of its square if $p=2$) is clearly equal to

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

This implies that a basis of the $p$-radical is given by $(\theta, \theta^2)$. Hence, in step 9 we take $\alpha_1 = \theta$, $\alpha_2 = \theta^2$ and $\alpha_3 = p$.

The 9 by 3 matrix $C$ is obtained by stacking the following three matrices:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & m/p \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & m/p & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

It follows from the first three equations that, if $p^2 \nmid m$, the kernel of $C$ is trivial, hence that $\mathbb{Z}[\theta]$ is $p$-maximal. Therefore, we will write

$$m = ab^2, \quad a \text{ and } b \text{ squarefree}, \ (a, b) = 1.$$

Indeed, $a$ is chosen squarefree, but since $m$ is cubefree the other conditions follow.

With these notations, we have just shown that if $p \mid a$ then $\mathbb{Z}[\theta]$ is $p$-maximal. Take now $p \mid b$ (still with $p \neq 3$). The kernel of the matrix $C$ is now clearly generated over $\mathbb{F}_p$ by the column vector $(0, 0, 1)$ corresponding to $\theta^2$, hence in step 10 we will compute the Hermite normal form of the matrix

$$\begin{pmatrix} 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 1 & 0 & 0 & p \end{pmatrix}.$$

This is clearly equal to the matrix

$$\begin{pmatrix} p & 0 & 0 \\ 0 & p & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

thus enlarging the order $\mathbb{Z}[\theta]$ to the order whose $\mathbb{Z}$-basis is $(1, \theta, \theta^2/p)$. If we apply the round 2 algorithm again to this new order, one checks immediately that the new matrix $C$ will be the same as the one above with $m/p$ replaced by $m/p^2$. Since $m$ is cubefree, this is not divisible by $p$ which shows that the kernel is trivial and so the new order is $p$-maximal.

Putting together all the local pieces, we can enlarge our order to $(1, \theta, \theta^2/b')$ where $b' = b$ if $3 \nmid b$, $b' = b/3$ if $3 \mid b$. This order will then be $p$-maximal for every prime $p$ except perhaps the prime 3, which we now consider.

We start from the order $(1, \theta, \theta^2/b')$ and consider separately the cases where $3 \mid m$ and $3 \nmid m$.

Assume first that $3 \mid m$. The matrix of the Frobenius with respect to the basis $(1, \theta, \theta^2/b')$ is equal to

$$\begin{pmatrix} 1 & m & a^2b^4/b'^3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and modulo 3 both $m$ and $a^2b^4/b'^3$ are equal to 0. Hence, as in the case $p \neq 3$, the kernel of the Frobenius is generated by $(\theta, \theta^2/b')$. Therefore, in step 9 we take $\alpha_1 = \theta$, $\alpha_2 = \theta^2/b'$ and $\alpha_3 = 3$. The matrix $C$ is then obtained by stacking the following three matrices:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & b' & 0 \\ 0 & 0 & m/(3b') \end{pmatrix}, \begin{pmatrix} 0 & 0 & m/b'^2 \\ 1 & 0 & 0 \\ 0 & m/(3b') & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Since $3 \nmid b'$ but $3 \mid m$, we have $m/b'^2 \equiv 0 \pmod 3$. On the other hand, $m/(3b')$ is equal to 0 modulo 3 if and only if $3^2 \mid m$, i.e. $3 \mid b$. Hence, we consider two sub-cases.

If $3 \nmid b$, the first three relations show that the kernel of $C$ is equal to 0 and so our order is 3-maximal. Thus, in that case $b' = b$ so an integral basis is $(1, \theta, \theta^2/b)$ and the discriminant of the field $K$ is equal to $-27a^2b^2$.

If $3 \mid b$, the kernel of $C$ is generated by $(0, 0, 1)$ corresponding to $\theta^2/b'$. The Hermite normal form obtained in step 10 is, as for $p \neq 3$, equal to the matrix

$$\begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

giving the larger order $(1, \theta, \theta^2/b'/3) = (1, \theta, \theta^2/b)$.

Since the discriminant of this order is still divisible by 9, we must start again. A similar computation shows that the matrix $C$ is obtained by stacking the following 3 matrices:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & ab/3 \end{pmatrix}, \begin{pmatrix} 0 & 0 & a \\ 1 & 0 & 0 \\ 0 & ab/3 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

and since $3 \nmid ab/3$, the first, third and sixth relation show that the kernel of $C$ is trivial, hence that our order is now 3-maximal. So if $3 \mid b$, an integral basis is $(1, \theta, \theta^2/b)$ and the discriminant of $K$ is equal to $-27a^2b^2$, giving exactly the same result as when $3 \nmid b$.

We now assume that $3 \nmid m$, and so in particular we have $b' = b$. The matrix of the Frobenius is equal to

$$\begin{pmatrix} 1 & ab^2 & a^2b \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Since $a^2 \equiv b^2 \equiv 1 \pmod 3$, this shows that the kernel of the Frobenius is equal to the set of elements $x + y\theta + z\theta^2/b$ such that $x + ay + bz \equiv 0 \pmod 3$. Hence modulo 3 it is, for example, generated by $(\theta - a, \theta^2/b - b)$. This means that in step 9 we can take $\alpha_1 = \theta - a$, $\alpha_2 = \theta^2/b - b$ and $\alpha_3 = 3$. The matrix $C$ is obtained by stacking the following three matrices:

$$\begin{pmatrix} 1 & -a & 0 \\ 0 & b & -a \\ 0 & (b^2-a^2)/3 & 0 \end{pmatrix}, \begin{pmatrix} 0 & -b & a \\ 1 & 0 & -b \\ 0 & 0 & (a^2-b^2)/3 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & a & b \end{pmatrix}.$$

We consider two subcases. First assume that $a^2 \not\equiv b^2 \pmod 9$. Then from the first, third and sixth relation we see that the kernel of $C$ is trivial, hence that our order is 3-maximal. This means, as in the case $3 \mid m$, that $(1, \theta, \theta^2/b)$ is an integral basis and the discriminant of $K$ is equal to $-27a^2b^2$.

Assume now that $a^2 \equiv b^2 \pmod 9$. In this case, one sees easily that the kernel of $C$ is generated by $(b, ab, 1)$ corresponding to $\theta^2/b + ab\theta + b$, and the computation of the Hermite normal form of the matrix

$$\begin{pmatrix} b & 3 & 0 & 0 \\ ab & 0 & 3 & 0 \\ 1 & 0 & 0 & 3 \end{pmatrix}$$

leads to the matrix

$$\begin{pmatrix} 3 & 0 & b \\ 0 & 3 & ab \\ 0 & 0 & 1 \end{pmatrix},$$

thus giving a larger order generated by $(1, \theta, (\theta^2 + ab^2\theta + b^2)/(3b))$, and the discriminant of this order being equal to $-3a^2b^2$, hence not divisible by $3^2$, this enlarged order is 3-maximal.

We summarize what we have proved in the following theorem.

**Theorem 6.4.13.** *Let $K = \mathbb{Q}(\sqrt[3]{m})$ be a pure cubic field, where $m$ is cubefree and not equal to $\pm 1$. Write $m = ab^2$ with $a$ and $b$ squarefree and coprime. Let $\theta$ be the cube root of $m$ belonging to $K$. Then*

(1) *If $a^2 \not\equiv b^2 \pmod 9$ then*

$$\left(1, \theta, \frac{\theta^2}{b}\right)$$

*is an integral basis of $K$ and the discriminant of $K$ is equal to $-27a^2b^2$.*

(2) *If $a^2 \equiv b^2 \pmod 9$ then*

$$\left(1, \theta, \frac{\theta^2 + ab^2\theta + b^2}{3b}\right)$$

*is an integral basis of $K$ and the discriminant of $K$ is equal to $-3a^2b^2$.*

*Proof.* Simply note that since $a$ and $b$ are coprime, when $3 \mid m$ we cannot have $a^2 \equiv b^2 \pmod 9$. □

**Remark.** The condition $a^2 \equiv b^2 \pmod 9$ is clearly equivalent to the condition $m \equiv \pm 1 \pmod 9$.

### 6.4.4 Decomposition of Primes in Pure Cubic Fields

As examples of applications of Algorithm 6.2.9, we will give explicitly the decomposition of primes in pure cubic fields. We could also treat the case of cyclic cubic fields, but the results would be a little more complicated.

Let $\theta$ be the real root of the polynomial $X^3 - m$, and let $K = \mathbb{Q}(\theta)$. First consider the case of "good" prime numbers $p$, i.e. such that $p$ does not divide the index $[\mathbb{Z}_K : \mathbb{Z}[\theta]]$ (which, by Theorem 6.4.13 is equal to $3b$ or $b$ depending on whether $a^2 \equiv b^2 \pmod 9$ or not). In this case we can directly apply Theorem 4.8.13. In other words the decomposition of $p\mathbb{Z}_K$ mimics that of the polynomial $T(X) = X^3 - m$ modulo $p$.

Now this decomposition is obtained as follows (compare with Section 1.4.2 where the Legendre symbol is defined).

**Proposition 6.4.14.** *Let $p$ be a prime number not dividing $m$. The decomposition of $X^3 - m$ modulo $p$ is of the following type.*

(1) *If $p \equiv 2 \pmod 3$, then $X^3 - m \equiv (X-u)(X^2 - vX + w) \pmod p$ (where it is of course implicitly understood that the polynomial $X^2 - vX + w$ is irreducible in $\mathbb{F}_p[X]$).*
(2) *If $p \equiv 1 \pmod 3$ and $m^{(p-1)/3} \equiv 1 \pmod p$ then $X^3 - m \equiv (X-u_1)(X-u_2)(X-u_3) \pmod p$, where $u_1$, $u_2$ and $u_3$ are distinct elements of $\mathbb{F}_p$.*
(3) *If $p \equiv 1 \pmod 3$ and $m^{(p-1)/3} \not\equiv 1 \pmod p$, then $X^3 - m$ is irreducible in $\mathbb{F}_p[X]$.*
(4) *If $p = 3$, then $X^3 - m \equiv (X-a)^3 \pmod p$.*

*Proof.* Consider the group homomorphism $\phi$ such that $\phi(x) = x^3$ from $\mathbb{F}_p^*$ into itself. It is clear that if $\phi(x) = 1$, then $(x-1)(x^2+x+1) = 0$ (in $\mathbb{F}_p$) hence

$$(x-1)((2x+1)^2 + 3) = 0.$$

If $p \equiv 2 \pmod 3$ the quadratic reciprocity law 1.4.7 shows that $\left(\frac{-3}{p}\right) = -1$, hence $-3$ is not equal to a square in $\mathbb{F}_p$. This shows that $(2x+1)^2 + 3 = 0$ is impossible, hence that the function $\phi$ is injective, hence bijective. In particular, there exists a unique $u \in \mathbb{F}_p^*$ such that $\phi(u) = m$, hence a unique root of $X^3 - m$ in $\mathbb{F}_p$, proving (1).

For (2) and (3), by quadratic reciprocity we have $\left(\frac{-3}{p}\right) = 1$, hence there exists $z \in \mathbb{F}_p^*$ such that $z^2 = -3$. This immediately implies that the kernel of $\phi$ has exactly 3 elements, and hence that the image of $\phi$ has $(p-1)/3$ elements.

Furthermore, if $g$ is a primitive root modulo $p$, then clearly the image of $\phi$ is the set of elements $x$ of the form $g^{3k}$ for $0 \le k < (p-1)/3$, and these are exactly those elements such that $x^{(p-1)/3} = 1$ in $\mathbb{F}_p$, proving (2) and (3).

Finally, (4) is trivial. □

When $p \mid m$ we trivially have $X^3 - m \equiv X^3 \pmod{p}$, so we immediately obtain the following corollary in the "easy" cases where $p$ does not divide the index.

**Corollary 6.4.15.** *As above let $K = \mathbb{Q}(\sqrt[3]{m})$ and recall that we have set $m = ab^2$. Assume that $p \nmid b$ and that if $a^2 \equiv b^2 \pmod{9}$, then also $p \neq 3$. Then the decomposition of $p\mathbb{Z}_K$ is given as follows.*

(1) *If $p \mid a$, then $p\mathbb{Z}_K = \mathfrak{p}^3$ where $\mathfrak{p} = p\mathbb{Z}_K + \theta\mathbb{Z}_K$.*

(2) *If $p \nmid a$ and $p \equiv 2 \pmod{3}$, then $p\mathbb{Z}_K = \mathfrak{p}_1\mathfrak{p}_2$ where $\mathfrak{p}_1 = p\mathbb{Z}_K + (\theta - u)\mathbb{Z}_K$ is an ideal of degree 1 and $\mathfrak{p}_2 = p\mathbb{Z}_K + (\theta^2 - v\theta + w)\mathbb{Z}_K$ is an ideal of degree 2.*

(3) *If $p \nmid a$, $p \equiv 1 \pmod{3}$ and $m^{(p-1)/3} \equiv 1 \pmod{p}$, then $p\mathbb{Z}_K = \mathfrak{p}_1\mathfrak{p}_2\mathfrak{p}_3$ where $\mathfrak{p}_i = p\mathbb{Z}_K + (\theta - u_i)\mathbb{Z}_K$ are three distinct ideals of degree 1.*

(4) *If $p \nmid a$, $p \equiv 1 \pmod{3}$ and $m^{(p-1)/3} \not\equiv 1 \pmod{p}$, then the ideal $p\mathbb{Z}_K$ is inert.*

(5) *If $p = 3$ and $p \nmid a$, then $p\mathbb{Z}_K = \mathfrak{p}^3$, where $\mathfrak{p} = p\mathbb{Z}_K + (\theta - a)\mathbb{Z}_K$ is an ideal of degree 1.*

We must now consider the more difficult cases where $p$ divides the index. Here we will follow the Algorithm 6.2.9 more closely, and we will skip the detailed computations of products and quotients of ideals, which are easy but tedious.

Assume first that $a^2 \not\equiv b^2 \pmod{9}$. Then Theorem 6.4.13 tells us that $1$, $\theta$, $\theta^2/b$ is an integral basis, and according to the algorithm described in Section 6.2 we start by computing the $p$-radical of $\mathbb{Z}_K$, assuming that $p \mid b$. It is easily seen that the matrix of the Frobenius at $p$ (or its square for $p = 2$) is always equal to the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

in $\mathbb{F}_p$. Therefore $(\theta, \theta^2/b)$ is an $\mathbb{F}_p$-basis of $\overline{I_p}$. From this, using Algorithm 6.2.5 we obtain the following $\mathbb{F}_p$ bases.

$\overline{K_1} = (\theta, \theta^2/b)$, $\overline{K_2} = (\theta)$ and $\overline{K_j} = \{0\}$ for $j \ge 3$.

As a consequence, using Algorithm 6.2.7 we obtain

$\overline{J_1} = \overline{J_2} = \overline{J_3} = (\theta, \theta^2/b)$, and $\overline{J_j} = (1, \theta, \theta^2/b)$ for $j \ge 4$.

From this, it is clear that we have $H_1 = H_2 = \mathbb{Z}_K$, $H_3 = K_1$ and $H_j = \mathbb{Z}_K$ for $j \ge 4$, from which it follows that

$$p\mathbb{Z}_K = K_1^3.$$

Since $K$ is a field of degree equal to 3, this implies that $K_1$ is a prime ideal (which of course can be checked directly since it is of norm $p$). This shows that $p$ is totally ramified, and the unique prime ideal $\mathfrak{p}$ above $p$ is generated over $\mathbb{Z}$ by $(p, \theta, \theta^2/b)$.

Note that most of these computations can be avoided. Indeed, once we know a $\mathbb{Z}$-basis of $I_p$, a trivial determinant computation shows that $I_p$ is of norm $p$, hence is a prime ideal of degree 1. Using the notations of Section 6.2, it follows that $g = 1$ and that $p\mathbb{Z}_K = I_p^{e_1}$ and since we are in a field of degree 3, the relation $\sum e_i f_i = 3$ tells us that $e_1 = 3$, thus showing that $p$ is totally ramified.

We have kept the computations however, so that the reader can check his implementation of ideal multiplication and division.

Assume now that $a^2 \equiv b^2 \pmod 9$. Recall that in this case we have $3 \nmid b$. Then Theorem 6.4.13 tells us that $1$, $\theta$, $(\theta^2 + ab^2\theta + b^2)/(3b)$ is an integral basis, and we must first compute the $p$-radical of $\mathbb{Z}_K$, assuming that $p \mid 3b$.

Consider first the case where $p \neq 3$, i.e. $p \mid b$. It is easily seen that the matrix of the Frobenius at $p$ (or its square for $p = 2$) is still equal to the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

in $\mathbb{F}_p$ hence we obtain that a $\mathbb{F}_p$-basis of $\overline{I_p}$ is $(\theta, (\theta^2 + ab^2\theta + b^2)/(3b))$. As in the preceding case, one checks trivially that $I_p$ has norm equal to $p$ so is a prime ideal of degree 1, so as before $p$ is totally ramified and $p\mathbb{Z}_K = I_p^3$. For the sake of completeness (or again as exercises), we give the computations as they would have been carried out without noticing this.

By Algorithm 6.2.5 we obtain the following $\mathbb{F}_p$-bases.

$\overline{K_1} = (\theta, (\theta^2 + ab^2\theta + b^2)/(3b))$, $\overline{K_2} = (\theta)$ and $\overline{K_j} = \{0\}$ for $j \geq 3$.

As a consequence, using Algorithm 6.2.7, we obtain

$\overline{J_1} = \overline{J_2} = \overline{J_3} = (\theta, (\theta^2 + ab^2\theta + b^2)/(3b))$, and $\overline{J_j} = (1, \theta, (\theta^2 + ab^2\theta + b^2)/(3b))$ for $j \geq 4$.

From this, as before, we have $H_1 = H_2 = \mathbb{Z}_K$, $H_3 = K_1$ and $H_j = \mathbb{Z}_K$ for $j \geq 4$, from which it follows that

$$p\mathbb{Z}_K = K_1^3.$$

Therefore $p$ is totally ramified, and the unique prime ideal $\mathfrak{p}$ above $p$ is generated over $\mathbb{Z}$ by $(p, \theta, (\theta^2 + ab^2\theta + b^2)/(3b))$.

Finally, still assuming $a^2 \equiv b^2 \pmod 9$, consider the case $p = 3$. The matrix of the Frobenius at 3 is now equal to the matrix

$$\begin{pmatrix} 1 & ab^2 & \frac{b(a^2-2b^2+3a^2b^2-3a^2b^4+a^4b^4)}{27} \\ 0 & 0 & ab\frac{1-a^2b^4}{9} \\ 0 & 0 & b^2\frac{1+a^2+a^2b^2}{3} \end{pmatrix}$$

with coefficients in $\mathbb{F}_3$. Since $a^2 \equiv b^2 \pmod 9$ and $3 \nmid ab$, we have

$$1+a^2+a^2b^2 \equiv 1+a^2+a^4 \equiv (1-a^2)(1+2a^2)+3a^4 \equiv 3a^4 \pmod 9,$$

hence $3 \nmid (1+a^2+a^2b^2)/3$. This shows that $(3, \theta-a, (\theta^2+ab^2\theta+b^2)/b)$ is a $\mathbb{Z}$-basis of $I_p$, and hence $(\theta-a)$ is an $\mathbb{F}_p$-basis of $\overline{I}_p$. Here the norm of $I_p$ is equal to 9, so we cannot obtain the decomposition of $3\mathbb{Z}_K$ directly, and it is really necessary to do the computations of Algorithm 6.2.9

By Algorithm 6.2.5, we obtain the following $\mathbb{F}_3$-bases.

$\overline{K}_1 = (\theta-a)$ and $\overline{K}_j = \{0\}$ for $j \geq 2$.

As a consequence, using Algorithm 6.2.7, we obtain

$\overline{J}_1 = (\theta-a)$, $\overline{J}_2 = (\theta-a, (\theta^2+ab^2\theta+a^2b^4)/(3b))$ and $\overline{J}_j = (1, \theta, (\theta^2+ab^2\theta+b^2)/(3b))$ for $j \geq 3$.

From this we obtain (after lifting to $\mathcal{O}$) that $H_1 = (3, \theta-a, (\theta^2+ab^2\theta-b^2(1+a^2))/(3b))$, $H_2 = J_2 = (3, \theta-a, (\theta^2+ab^2\theta+a^2b^4)/(3b))$ and $H_j = \mathbb{Z}_K$ for $j \geq 3$. It is immediately checked (for example using the determinant of the matrix of $H_j$) that $H_1$ and $H_2$ are of norm equal to 3, hence are prime ideals. Thus, we obtain that the prime ideal decomposition of $3\mathbb{Z}_K$ is given by

$$3\mathbb{Z}_K = H_1H_2^2$$

where $H_1$ and $H_2$ are distinct prime ideals with $\mathbb{Z}$-basis given above. Hence, 3 is ramified (as it must be since the discriminant of the field is divisible by 3), but not totally ramified as in the case $a^2 \not\equiv b^2 \pmod 9$.

We summarize the above in the following theorem.

**Theorem 6.4.16.** *Let $(1, \theta, \omega)$ be the integral basis of $\mathbb{Z}_K$ given by Theorem 6.4.13 (hence $\omega = \theta^2/b$ if $a^2 \not\equiv b^2 \pmod 9$, $\omega = (\theta^2+ab^2\theta+b^2)/(3b)$ if $a^2 \equiv b^2 \pmod 9$). Then*

(1) *If $p \mid b$, then $p$ is totally ramified, and we have $p\mathbb{Z}_K = \mathfrak{p}^3$, where $\mathfrak{p}$ is a prime ideal of degree 1 given by*

$$\mathfrak{p} = p\mathbb{Z}+\theta\mathbb{Z}+\omega\mathbb{Z} = p\mathbb{Z}_K+\omega\mathbb{Z}_K.$$

(2) *If $p = 3$ and $a^2 \equiv b^2 \pmod 9$, then 3 is partially ramified and we have $3\mathbb{Z}_K = \mathfrak{p}_1\mathfrak{p}_2^2$ where $\mathfrak{p}_1$ and $\mathfrak{p}_2$ are prime ideals of degree 1 given by*

$$\mathfrak{p}_1 = 3\mathbb{Z}+(\theta-a)\mathbb{Z}+(\omega-b(2+a^2)/3)\mathbb{Z} = 3\mathbb{Z}_K+(\omega-b(a^2+2)/3)\mathbb{Z}_K$$

*and*

$$\mathfrak{p}_2 = 3\mathbb{Z}+(\theta-a)\mathbb{Z}+(\omega-b(a^2-1)/3)\mathbb{Z} = 3\mathbb{Z}_K+\alpha\mathbb{Z}_K$$

*where*

$$\alpha = \omega-b(a^2-1)/3 \quad \text{if} \quad a^2b^4 \not\equiv 1 \pmod{27},$$
$$\alpha = \omega+\theta-a-b(a^2-1)/3 \quad \text{if} \quad a^2b^4 \equiv 1 \pmod{27}.$$

*Proof.* We have shown everything except the generating systems over $\mathbb{Z}_K$. If $p \mid b$, a simple HNF computation shows that one has $p\mathbb{Z}_K + \omega\mathbb{Z}_K = (p, \theta, \omega)$.

If $p = 3$ and $a^2 \equiv b^2 \pmod 9$, we could also check the result via a HNF computation. Another method is to notice that $3\mathbb{Z}_K = \mathfrak{p}_1\mathfrak{p}_2^2$ and that if we set $\alpha_1 = \omega - b(a^2+2)/3$, then $\alpha_1 \in \mathfrak{p}_1$, but $\alpha_1 \notin \mathfrak{p}_2$ otherwise $\mathfrak{p}_1 \subset p_2$ which is absurd, so that $\alpha_1 = \mathfrak{p}_1^e\mathfrak{q}$ with $\mathfrak{q}$ prime to 3, so $3\mathbb{Z}_K + \alpha_1\mathbb{Z}_K = \mathfrak{p}_1$.

For $\mathfrak{p}_2$, if we set $\alpha_2 = \omega - b(a^2-1)/3$, then again $\alpha_2 \in \mathfrak{p}_2$ and $\alpha_2 \notin \mathfrak{p}_1$. Hence $\alpha_2 = \mathfrak{p}_2^e\mathfrak{q}$ with $\mathfrak{q}$ prime to 3. This implies that $3\mathbb{Z}_K + \alpha_2\mathbb{Z}_K = \mathfrak{p}_2^{\min(e,2)}$ hence this can be equal to $\mathfrak{p}_2$ or to its square. To distinguish the two cases, we must compute the norm of $\alpha_2$, whose 3-adic valuation will be equal to $e$. As it happens, it is simpler to work with the norm of $\alpha_2' = \alpha_2 + b(a^2b^2 + a^2 - 2)/3$ (note that $a^2b^2 + a^2 - 2 \equiv (a^2-1)(a^2+2) \pmod 9$ hence $3\mathbb{Z}_K + \alpha_2\mathbb{Z}_K = 3\mathbb{Z}_K + \alpha_2'\mathbb{Z}_K$).

One computes that $n = \mathcal{N}(\alpha_2') = a^2b(1 - a^2b^4)^2/27$. Hence, if $a^2b^4 \not\equiv 1 \pmod{27}$, the 3-adic valuation of $n$ is equal to 1, therefore $3\mathbb{Z}_K + \alpha_2\mathbb{Z}_K = \mathfrak{p}_2$.

If $a^2b^4 \equiv 1 \,(\text{mod } 2)7$, a similar computation shows that the 3-adic valuation of $\mathcal{N}(\alpha_2' + \theta - a)$ is equal to 1, thus proving the theorem. □

### 6.4.5 General Cubic Fields

In this section, we give without proof a few results concerning the decomposition of primes in general cubic extensions of $\mathbb{Q}$.

Let $K$ be a cubic field. The discriminant $d(K)$ of the number field $K$ can (as any discriminant) be written in a unique way in the form $d(K) = df^2$ where $d$ is either a fundamental discriminant or is equal to 1. The field $k = \mathbb{Q}(\sqrt{d})$ is either $\mathbb{Q}$ if $d = 1$, or is a quadratic field, and is the unique subfield of index 3 of the Galois closure of $K$.

In particular, cyclic cubic fields correspond to $d = 1$, i.e. $k = \mathbb{Q}$, and pure cubic fields correspond to $d = -3$, i.e. $k = \mathbb{Q}(\sqrt{-3})$ the cyclotomic field of third roots of unity.

Let $p$ be a prime number. If $p \nmid d(K)$, then $p$ is unramified. Therefore by Proposition 4.8.10 we have the following cases.

(1) If $\left(\frac{d(K)}{p}\right) = -1$, then $g = 2$. Hence, we have a decomposition of $p$ in the form $p\mathbb{Z}_K = \mathfrak{p}_1\mathfrak{p}_2$ where $\mathfrak{p}_1$ is a prime ideal of degree 1 and $\mathfrak{p}_2$ is a prime ideal of degree 2.

(2) If $\left(\frac{d(K)}{p}\right) = 1$, then $g$ is odd. Hence, either $p$ is inert or $p\mathbb{Z}_K$ is equal to the product of three prime ideals of degree 1.

If $p$ does not divide the index $[\mathbb{Z}_K : \mathbb{Z}[\theta]]$ where $K = \mathbb{Q}(\theta)$, then the two cases are distinguished by the splitting modulo $p$ of the minimal polynomial $T(X)$ of $\theta$.

If $p$ divides the index, then $T$ has at least a double root modulo $p$. If $T$ has a double root, but not a triple root, then $T$ also has a simple root which corresponds to a prime ideal of degree 1. In this case $p\mathbb{Z}_K$ is the

product of three ideals of degree 1. Finally, if $T$ has a triple root modulo $p$, we must apply other techniques such as the ones in Section 6.2.

Assume now that $p \mid d(K) = df^2$, hence that $p$ is ramified. Then the result is as follows.

(1) If $p \mid f$, then $p$ is totally ramified. In other words, $p\mathbb{Z}_K = \mathfrak{p}^3$ where $\mathfrak{p}$ is a prime ideal of degree 1.

(2) If $p \mid d$ and $p \nmid f$, then $p$ is partially ramified. In other words, $p\mathbb{Z}_K = \mathfrak{p}_1\mathfrak{p}_2^2$, where $\mathfrak{p}_1$ and $\mathfrak{p}_2$ are distinct prime ideals of degree 1.

(3) Furthermore, if there exists a $p$ such that $p \mid (d, f)$, then we must have $p = 3$ (and we are in case (1), since $p \mid f$).

See for example [Has] for proofs of these results.

## 6.5 Computing the Class Group, Regulator and Fundamental Units

In this section, we shall give a practical generalization of Buchmann's sub-exponential Algorithm 5.9.2 to an arbitrary number field. This algorithm computes the class group, the regulator and also if desired a system of fundamental units, for a number field whose discriminant is not too large. Although based on essentially the same principles as Algorithm 5.9.2, we do not claim that its running time is sub-exponential, even assuming some reasonable conjectures. On the other hand it performs very well in practice. The algorithm originates in an unpublished paper of J. Buchmann, but the present formulation is due to F. Diaz y Diaz, M. Olivier and myself. As almost all other algorithms in this book, this algorithm has been fully implemented in the author's PARI package (see Appendix A). It is still in an experimental state, hence many refinements need to be made to achieve optimum performance.

We assume that our number field $K$ is given as usual as $K = \mathbb{Q}[\theta]$ where $\theta$ is an algebraic integer. Let $T(X)$ be the minimal monic polynomial of $\theta$. Let $n = [K : \mathbb{Q}] = r_1 + 2r_2$, denote by $\sigma_i$ the complex embeddings of $K$ ordered as usual, and finally let $\omega_1, \ldots, \omega_n$ be an integral basis of $\mathbb{Z}_K$, found using for example the round 2 Algorithm 6.1.8.

### 6.5.1 Ideal Reduction

The only notion that we have not yet introduced and that we will need in an essential way in our algorithm is that of ideal reduction.

**Definition 6.5.1.** *Let $I$ be a fractional ideal and $\alpha$ a non-zero element of $I$. We will say that $\alpha$ is a* minimum *in $I$ if, for all $\beta \in I$, we have*

$$(\forall i \quad |\sigma_i(\beta)| < |\sigma_i(\alpha)|) \implies \beta = 0.$$

*We will say that the ideal $I$ is* reduced *if $\ell(I)$ is a minimum in $I$, where $I \cap \mathbb{Q} = \ell(I)\mathbb{Z}$.*

The reader can check that this definition of reduction coincides with the definitions given for the imaginary and real quadratic case (see Exercise 16 of Chapter 5).

**Definition 6.5.2.** *Let $v = (v_i)_{1 \le i \le n}$ be a vector of real numbers such that $v_{r_2+i} = v_i$ for $r_1 < i \le r_1 + r_2$. We define the* $v$-norm $\|\alpha\|_v$ *of $\alpha$ by the formula*

$$\|\alpha\|_v^2 = \sum_{i=1}^{n} e^{v_i} |\sigma_i(\alpha)|^2.$$

If $\alpha_1, \ldots, \alpha_n$ is a $\mathbb{Z}$-basis for the ideal $I$, then $\| \sum_j x_j \alpha_j \|_v^2$ defines a positive definite quadratic form on $I$.

**Definition 6.5.3.** *We say that a $\mathbb{Z}$-basis $\alpha_1, \ldots, \alpha_n$ of an ideal $I$ is* LLL-reduced along the vector $v$ *if it is LLL-reduced for the quadratic form defined by $\|\alpha\|_v^2$.*

Thanks to the LLL algorithms seen in Section 2.6 we can efficiently LLL-reduce along $v$ any given basis.

The main point of these definitions is the following.

**Proposition 6.5.4.** *If $\alpha \in I$ is a (non-zero) minimum for the quadratic form $\|\alpha\|_v^2$, then $\alpha$ is a minimum of $I$ in the sense of Definition 6.5.1 above, and $I/\alpha$ is a reduced ideal.*

*Proof.* If $\beta \in I$ is such that for all $i$, $|\sigma_i(\beta)| < |\sigma_i(\alpha)|$, then clearly $\|\beta\|_v^2 < \|\alpha\|_v^2$. Hence, since $\alpha$ is a minimum non-zero value of the quadratic form, we must have $\beta = 0$ so $\alpha$ is a minimum in $I$. Let us show that $I/\alpha$ is a reduced ideal. First, I claim that $I/\alpha \cap \mathbb{Q} = \mathbb{Z}$. Indeed, if $r \in \mathbb{Q}^*$, $r \in I/\alpha$ is equivalent to $r\alpha \in I$ and since $\alpha$ is a minimum and $r$ is invariant under the $\sigma_i$, this implies that $|r| \ge 1$. Since $1 \in I/\alpha$, this proves my claim, hence $\ell(I/\alpha) = 1$. The proposition now follows since $\alpha$ minimum in $I$ is clearly equivalent to 1 minimum in $I/\alpha$. □

The LLL-algorithm allows us to find a small vector for our quadratic form, corresponding to an $\alpha \in I$. This $\alpha$ may not be a true minimum, but the inequalities proved in Chapter 2 show that it will in any case be a small vector. If we choose this $\alpha$ instead of a minimum, the ideal $I/\alpha$ will not be necessarily reduced, but it will be sufficient for our needs. For lack of a better term, we will say that $I/\alpha$ is *LLL-reduced in the direction $v$*.

To summarize, this gives the following algorithm for reduction.

**Algorithm 6.5.5** (LLL-Reduction of an Ideal Along a Direction $v$). Given a vector $v$ as above and an ideal $I$ by a $\mathbb{Z}$-basis $\alpha_1, \ldots, \alpha_n$, this algorithm computes $\alpha \in I$ and a new ideal $J = I/\alpha$ such that the $v$-norm of $\alpha$ is small.

1. [Set up quadratic form] Let

$$q_{i,j} = \sum_{k=1}^{n} e^{v_k} \overline{\sigma_k(\alpha_i)}\, \sigma_k(\alpha_j)$$

   (note that these are all real numbers), and let $Q$ be the quadratic form on $\mathbb{R}^n$ whose matrix is $(q_{i,j})$.

2. [Apply LLL] Using the LLL Algorithm 2.6.3, compute an LLL-reduced basis $\beta_1, \ldots, \beta_n$ of $I$ corresponding to this quadratic form, and let $\alpha \leftarrow \beta_1$.

3. [Compute $J$] Output $\alpha$ and the $\mathbb{Z}$-basis $\beta_i/\alpha$ of the ideal $J = I/\alpha$ and terminate the algorithm.

**Remarks.**

(1) The ideal $J$ is a fractional ideal. If desired, we can multiply it by a suitable rational number to make it integral and primitive.

(2) In practice the basis elements $\alpha_i$ are given in terms of a fixed basis $\mathcal{B}$ of $K$ (for example either a power basis or an integral basis of $\mathbb{Z}_K$). If we compute once and for all the quadratic form $Q_{\mathcal{B}}$ attached to $\mathcal{B}$, it is then easier to compute the quadratic form attached to the ideal $I$. Note however that this argument is only valid for a fixed choice of the vector $v$.

### 6.5.2 Computing the Relation Matrix

As in the quadratic case we choose a suitable integer $L$ such that non-inert prime ideals of norm less than or equal to $L$ generate the class group. The GRH implies that we can take $L = 12\ln^2|D|$ where $D$ is the discriminant of $K$ (see [Bach]). This is only twice the special value used for quadratic fields. However, if we allow ourselves to be not completely rigorous, we could choose a lower value.

To obtain relations, we will compute random products $I$ of powers of prime ideals. Let $J = I/\alpha$ be an LLL-reduced ideal along a certain direction $v$, obtained using Algorithm 6.5.5. If $J$ factors on a given factor base, as in the quadratic case we will obtain a relation of the type $\prod_i \mathfrak{p}_i^{x_i} = \alpha\mathbb{Z}_K$. This relation will be stored in two parts. The non-Archimedean information $(x_i)$ will be stored as a column of an integral relation matrix $M$. The Archimedean information $\alpha$ will be stored as an $r_1 + r_2$-component column vector, by using the complex logarithmic embedding $L_C(\alpha) - \frac{\ln(\mathcal{N}(\alpha))}{n} V$ defined in Section 5.8.4.

Note that, by definition, the sum of the $r_1 + r_2$ components of this vector is an integral multiple of $2i\pi$.

We now give the algorithm which computes the factor bases and the relation matrix.

**Algorithm 6.5.6** (Computation of the Relation Matrix). Given a number field $K$ as above, this algorithm computes integers $k$ and $k_2$ with $k_2 > k$, a $k \times k_2$ integral relation matrix $M$, an $(r_1 + r_2) \times k_2$ complex logarithm matrix $M_C$ and an Euler product $z$. These objects will be needed in the class group and unit Algorithm 6.5.9 below. We set $r_u \leftarrow r_1 + r_2$ (this is equal to the unit rank plus one). We choose at will a positive real number $B_1$ and we set $B_2 \leftarrow 12$.

1. [Compute integral basis and limits] Using Algorithm 6.1.8 compute the field discriminant $D = D(K)$ and an integral basis $\omega_1 = 1, \ldots, \omega_n$. Set $L_1 \leftarrow B_1 \ln^2 |D|$, $L_2 \leftarrow B_2 \ln^2 |D|$ and $L_s \leftarrow (4/\pi)^{r_2} n!/n^n \sqrt{|D|}$.

2. [Compute small factor base] Set $u \leftarrow 1$, $S \leftarrow \emptyset$ and for each prime $p$ such that $p \nmid D$ (i.e. $p$ unramified) do the following until $u > L_s$. Let $p\mathbb{Z}_K = \prod_{1 \le i \le g} \mathfrak{p}_i$ be the prime ideal decomposition of $p\mathbb{Z}_K$ obtained using Algorithm 6.2.9. For each $i \le g - 1$ such that $\mathcal{N}(\mathfrak{p}_i) \le L_2$, set $S \leftarrow S \cup \{\mathfrak{p}_i\}$ and $u \leftarrow u\mathcal{N}(\mathfrak{p}_i)$. Then $S$ will be a set of prime ideals which we call the small factor base. Let $s$ be its cardinality.

3. [Compute and store powers] For each $\mathfrak{p} \in S$ and each integer $e$ such that $0 \le e \le 20$, compute and store an LLL-reduced ideal equivalent to $\mathfrak{p}^e$, where the reduction is done using Algorithm 6.5.5 with $v$ equal to the zero vector. Note that the Archimedean information must also be stored, using the function $L_C$ as explained above.

4. [Compute factor bases and Euler product] For all primes $p \le L_2$ compute the prime ideal decomposition of $p\mathbb{Z}_K$ using Algorithm 6.2.9, and let the large factor base LFB be the list of all non-inert prime ideals of norm less than or equal to $L_2$ (where if necessary we also add the elements of $S$), and let the factor base FB be the subset of LFB containing only those primes of norm less than or equal to $L_1$ as well as the elements of $S$. Set $k$ equal to the cardinality of FB, and set $k_2 \leftarrow k + r_u + 10$. Finally, using the prime ideal decompositions, compute the Euler product

$$z \leftarrow \prod_{p \le L_2} \frac{1 - 1/p}{\prod_{\mathfrak{p} \mid p} (1 - 1/\mathcal{N}(\mathfrak{p}))}.$$

5. [Store trivial relations] Set $m \leftarrow 0$. For each $p \le L_1$ such that all the prime ideals above $\mathfrak{p}$ are in FB, set $m \leftarrow m + 1$ and store the relation $p\mathbb{Z}_K = \prod_{1 \le i \le g} \mathfrak{p}_i^{e_i}$ found in step 4 as the $m$-th column of the matrices $M$ and $M_C$ as explained above.

6. [Generate random power products] Call $S_i$ the elements of the small factor base $S$. Let $\mathfrak{q}$ be the ideal number $m + 1$ mod $k$ in $FB$. Choose random nonnegative integers $v_i \le 20$ for $i \le s + r_u$, set $v_{i+r_2} \leftarrow v_i$ for $s < i \le s + r_u$,

compute the ideal $I \leftarrow \mathfrak{q}\prod_{1\le i\le s} S_i^{v_i}$ and let $J = I/\alpha$ be the ideal obtained by LLL-reducing $I$ along the direction determined by the $v_i$ for $s < i \le s+n$ using Algorithm 6.5.5. Note that the $S_i^{v_i}$ have been precomputed in step 4.

7. [Relation found?] Using Algorithm 4.8.17, try to factor $\alpha$ (or equivalently the ideal $J$) on the factor base FB. If it factors, set $m \leftarrow m+1$ and store the relation $IJ^{-1} = \alpha\mathbb{Z}_K$ as the $m$-th column of the matrices $M$ and $M_C$ as explained above.

8. [Enough relations?] If $m \le k_2$ go to step 6.

9. [Be honest] For all prime ideals $\mathfrak{q}$ in the large factor base LFB and not belonging to FB, do as follows. Choose randomly integers $v_i$ as in step 6, compute $I \leftarrow \mathfrak{q}\prod_{1\le i\le s} S_i^{v_i}$ and let $J = I/\alpha$ be the ideal obtained by LLL-reducing $I$ along the direction determined by the $v_i$ for $s \le s+n$. If all the prime ideals dividing $J$ belong to FB or have been already checked in this test, then $\mathfrak{q}$ is OK, otherwise choose other random integers $v_i$ until $\mathfrak{q}$ passes this test.

10. [Eliminate spurious factors] For each ramified prime ideal $\mathfrak{q}$ which belongs to the factor base FB, check whether the GCD of the coefficients occurring in the matrix $M$ in the row corresponding to $\mathfrak{q}$ is equal to 1 (this is always true if $\mathfrak{q}$ is unramified). If not, as in step 9, choose random $v_i$, compute $I \leftarrow \mathfrak{q}\prod_{1\le i\le s} S_i^{v_i}$, LLL-reduce along the $v_i$ for $i > s$ and see if the resulting ideal factors on FB. If it does, add the relation to the matrices $M$ and $M_C$, set $k_2 \leftarrow k_2+1$, and continue doing this until the GCD of the coefficients occurring in the row corresponding to $\mathfrak{q}$ is equal to 1.

**Remarks.**

(1) The constant $B_1$ is usually chosen between 0.1 and 0.8, and controls the execution speed of the general algorithm, as in the quadratic case. On the other hand, the constant $B_2$ must be taken equal to 12 according to Bach's result. It can be taken equal to $B_1$ for maximum speed, but in this case, the result may not be correct even under the GRH. This is useful for long searches.

(2) As in the quadratic case, the constants 10 and 20 used in this algorithm are quite arbitrary but usually work.

(3) Step 10 of this algorithm was added only after the implementation was finished since it was noticed that for number fields of small discriminant, the class number was usually a multiple of the correct value due to the presence of ramified primes.

(4) The Euler product that is computed is closely linked to $h(K)R(K)$ since

$$\frac{h(K)R(K)}{w(K)} = 2^{-r_1}(2\pi)^{-r_2}\sqrt{|d(K)|}\prod_p \frac{1-1/p}{\prod_{\mathfrak{p}|p}(1-1/\mathcal{N}(\mathfrak{p}))},$$

where the outer product runs over all primes $p$ and the innermost product runs over the prime ideals above $p$ (see Exercise 23).

### 6.5.3 Computing the Regulator and a System of Fundamental Units

Before giving the complete algorithm, we need to explain how to extract from the Archimedean information that we have computed, both the regulator and a system of fundamental units of $K$.

After suitable column operations on the matrices $M$ and $M_C$ as explained below in Algorithm 6.5.9, we will obtain a complex matrix $C$ whose columns correspond to the Archimedean information associated to zero exponents, i.e. to a relation of the form $\mathbb{Z}_K = \alpha\mathbb{Z}_K$. In other words, the columns are complex logarithmic embeddings of units. As in the real quadratic case, we can obtain the regulator of the subgroup spanned by these units (which hopefully is equal to the field regulator) by computing a real GCD of $(r_u - 1) \times (r_u - 1)$ sub-determinants as follows.

**Algorithm 6.5.7** (Computation of the Regulator and Fundamental Unit Matrix). Given a $r_u \times r$ complex matrix $C$ whose columns are the complex logarithmic embeddings of units, this algorithm computes the regulator $R$ of the subgroup spanned by these units as well as an $r_u \times (r_u - 1)$ complex matrix $F$ whose columns give a basis of the lattice spanned by the columns of $C$. As usual we denote by $C_j$ the columns of the matrix $C$ and we assume that the real part of $C$ is of rank equal to $r_u - 1$.

1. [Initialize] Let $R \leftarrow 0$ and $j \leftarrow r_u - 2$.
2. [Loop] Set $j \leftarrow j + 1$. If $j > r$, let $F$ be the matrix formed by the last $r_u - 1$ columns of $C$, output $R$ and $F$ and terminate the algorithm.
3. [Compute determinant] Let $A$ be the $(r_u - 1) \times (r_u - 1)$ matrix obtained by extracting from $C$ any $r_u - 1$ rows, columns $j - r_u + 2$ to $j$, and taking the real part. Let $R_1 \leftarrow \det(A)$. Using the real GCD Algorithm 5.9.3, compute the RGCD $d$ of $R$ and $R_1$ as well as integers $u$ and $v$ such that $uR + vR_1 = d$ (note that Algorithm 5.9.3 does not give $u$ and $v$, but it can be easily extended to do so, as in Algorithm 1.3.6).
4. [Replace] Set $R \leftarrow d$, $C_j \leftarrow vC_j + (-1)^{r_u} uC_{j-r_u+1}$ (where $C_0$ is to be understood as the zero column) and go to step 2.

The proof of the validity of this algorithm is immediate once we notice that the GCD and replacement operations in steps 3 and 4 correspond to computing the sum of two sub-lattices of the unit lattice given by two $\mathbb{Z}$-bases differing by a single element. The sign $(-1)^{r_u}$ is the signature of the cyclic permutation that is performed. Note also that the real GCD Algorithm 5.9.3 may be applied since by [Zim1] and [Fri] we know that regulators of number fields are uniformly bounded from below by 0.2. □

To compute the regulator, we have only used the real part of the matrix $C$. We now explain how the use of the imaginary part, and more precisely of

the matrix $F$ output by this algorithm, allows us in principle to compute a system of fundamental units. Note that, by construction, the columns of $F$ are the complex logarithmic embeddings of a system of fundamental units of $\mathbb{Z}_K$. However this may be a very badly skewed basis of units, hence the first thing is to compute a nice basis using the LLL algorithm. This leads to the following algorithm.

**Algorithm 6.5.8** (Computation of a System of Fundamental Units). Given the regulator $R$ and the $r_u \times (r_u - 1)$ matrix $F$ output by Algorithm 6.5.7, this algorithm computes a system of fundamental units, expressing them on an integral basis $\omega_i$. We let $f_{i,j}$ be the coefficients of $F$.

1. [Build matrix] Set $r \leftarrow r_u - 1$. For $j = 1, \ldots, j = r$ set $b_{i,j} \leftarrow f_{i,j}$ if $i \leq r_1$, $b_{i,j} \leftarrow f_{i,j}/2$ if $r_1 < i \leq r_u$ and $b_{i,j} \leftarrow \overline{f_{i-r_2,j}}/2$ if $r_u < i \leq n$. Let $B$ be the $n \times r$ matrix with coefficients $b_{i,j}$.
2. [LLL reduce] Using the LLL Algorithm 2.6.3 on the real part of the matrix $B$, compute a $r \times r$ unimodular matrix $U$ such that the real part of $BU$ is LLL-reduced. Let $E = (e_{i,j})$ be the $n \times r$ matrix such that $e_{i,j} = \exp(b'_{i,j})$, where $BU = (b'_{i,j})$. (Note that the exponential taken here may overflow the possibilities of the implementation, in which case the algorithm must be aborted.)
3. [Solve linear system] Let $\Omega = (w_{i,j})$ be the $n \times n$ matrix such that $w_{i,j} = \sigma_j(\omega_i)$ (where, as before, $(\omega_i)$ is an integral basis of $\mathbb{Z}_K$). Set $F_u \leftarrow \Omega^{-1}E$.
4. [Round] The coefficients of $F_u$ should be close to rational integers. If this is not the case, then either the precision used to make the computations was insufficient or the units are too large, and the algorithm fails. Otherwise, round all the coefficients of $F_u$ to the nearest integer.
5. [Check] Check that the columns of $F_u$ correspond to units and that the usual regulator determinant constructed using the columns of $F_u$ is equal to $R$. If this is the case, output the matrix $F_u$ and terminate the algorithm (the columns of this matrix gives the coefficients of a system of fundamental units expressed on the integral basis $\omega_i$). Otherwise, output an error message saying that the accuracy is insufficient to compute the fundamental units.

### 6.5.4 The General Class Group and Unit Algorithm

We are now ready to give a general algorithm for class group, regulator and fundamental unit computation.

**Algorithm 6.5.9** (Class Group, Regulator and Units for General Number Fields). Let $K = \mathbb{Q}[\theta]$ be a number field of degree $n$ given by a primitive algebraic number $\theta$, let $T$ be the minimal monic polynomial of $\theta$. We assume that we have already computed the signature $(r_1, r_2)$ of $K$ using Algorithm 4.1.11. This algorithm computes the class number $h(K)$, the class group $Cl(K)$, the

order of the subgroup of roots of unity $w(K)$, the regulator $R(K)$ and a system of fundamental units of $\mathbb{Z}_K$.

1. [Compute relation matrices and Euler product] Using Algorithm 6.5.6, compute the discriminant $D(K)$, a $k \times k_2$ integral relation matrix $M$, a $r_u \times k_2$ complex logarithm matrix $M_C$ and an Euler product $z$.

2. [Compute roots of unity] Using Algorithm 4.9.9 compute the order $w(K)$ of the group of roots of unity in $K$. Output $w(K)$ and set

$$z \leftarrow 2^{-r_1}(2\pi)^{-r_2} w(K)\sqrt{|D(K)|} \cdot z$$

   (now $z$ should be close to $h(K)R(K)$).

3. [Simple HNF] Perform a preliminary simple Hermite reduction on the matrix $M$ as described in the remarks after Algorithm 5.5.2. All column operations done on the matrix $M$ should also be done on the corresponding columns of the matrix $M_C$. Denote by $M'$ and $M'_C$ the matrices obtained in this way.

4. [Compute probable regulator and units] Using Algorithm 2.7.2, compute the LLL-reduced integral kernel $A$ of $M'$ as a rectangular matrix, and set $C \leftarrow M'_C A$. By applying Algorithm 6.5.7 and if desired also Algorithm 6.5.8, compute a probable value for the regulator $R$ and the corresponding system of units which will be fundamental if $R$ is correct.

5. [HNF reduction] Using Algorithm 2.4.8, compute the Hermite normal form $H = (h_{i,j})$ of the matrix $M'$ using modulo $d$ techniques, where $d$ can be computed using standard Gaussian elimination (or simply use Algorithm 2.4.5). If one of the matrices $H$ or $C$ is not of maximal rank, get 10 more relations as in steps 6 and 7 of Algorithm 6.5.6 and go to step 3. (It will not be necessary to recompute the whole HNF.)

6. [Simplify $H$] For every $i$ such that $h_{i,i} = 1$, suppress row and column $i$, and let $W$ be the resulting matrix.

7. [Finished?] Let $h \leftarrow \det(W)$ (i.e. the product of the diagonal elements). If $hR \geq z\sqrt{2}$, get 10 more relations in steps 6 and 7 of Algorithm 6.5.6 and go to step 3 (same remark as above). Otherwise, output $h$ as the class number, $R$ as the regulator, and the system of fundamental units if it has been computed.

8. [Class group] Compute the Smith normal form of $W$ using Algorithm 2.4.14. Output those among the diagonal elements $d_i$ which are greater than 1 as the invariants of the class group (i.e. $Cl(K) = \bigoplus \mathbb{Z}/d_i\mathbb{Z}$) and terminate the algorithm.

**Remarks.**

(1) Most implementation remarks given after Algorithm 5.5.2 also apply here. In particular the correctness of the results given by this algorithm depends on the validity of GRH and the constant $B_2 = 12$ chosen in Algorithm 6.5.6. To speed up this algorithm, one can take $B_2$ to be a much lower value, and practice shows that this works well, but the results are not

anymore guaranteed to be correct even under GRH until someone improves Bach's bounds.

(2) The randomization of the direction of ideal reduction performed in step 6 of Algorithm 6.5.6 is absolutely essential for the correct performance of the algorithm. Intuitively the first $s$ values of $v_i$ correspond to randomization of the non-Archimedean components, while the last $r_u$ values randomize the Archimedean components. If the reduction was always done using the zero vector for instance, we would almost never obtain a relation matrix giving us the correct class number and regulator.

(3) An important speedup can be obtained by generating some relations in a completely different way. Assume that we can generate many elements $\alpha \in \mathbb{Z}_K$ of reasonably small norm. Then it is reasonable to expect that $\alpha\mathbb{Z}_K$ will factor on the factor base FB, thus giving us a relation. To obtain elements of small norm we can use the Fincke-Pohst Algorithm 2.7.7 on the quadratic form $\|\alpha\|_0^2$ defined on the lattice $\mathbb{Z}_K$, where 0 denotes the zero vector. If $\|\alpha\|_0^2 \leq nB^{2/n}$ then the inequality between arithmetic and geometric mean easily shows that $|\mathcal{N}(\alpha)| \leq B$, hence this indeed allows us to find elements of small norm. The reader is warned however that the relations that may be obtained in this way will in general not be random and may generate sub-lattices of the correct lattice.

(4) It is often useful, not only to compute the class group as an abstract group $Cl(K) = \bigoplus \mathbb{Z}/d_i\mathbb{Z}$, but to compute explicitly a generating set of ideal classes $g_i$ such that $g_i$ is of order $d_i$. This can easily be done by keeping track of the Smith reduction matrices in the above algorithm.

### 6.5.5 The Principal Ideal Problem

As in the real quadratic case, we can now solve the principal ideal problem for general number fields. In other words, given an ideal $I$ of $\mathbb{Z}_K$, determine whether $I$ is a principal ideal, and if this is the case, find an $\alpha \in K$ such that $I = \alpha\mathbb{Z}_K$.

To do this, we need to keep some information that was discarded in Algorithm 6.5.9. More precisely, we must keep better track of the Hermite reduction which is performed, including the simple Hermite reduction stage. If we do so, we will have kept a matrix $M''$ of relations which will be of the form

$$M'' = \begin{pmatrix} 0 & W & B \\ 0 & 0 & I \end{pmatrix},$$

where 0 denotes the zero matrix, $I$ is some identity matrix and $W$ is the square matrix in Hermite normal form computed in Step 6 of Algorithm 6.5.9. Together with this matrix, we must also compute the corresponding complex matrix $M''_C$, so that each column of $M''$ and $M''_C$ still corresponds to a relation. Finally, in Step 8 of Algorithm 6.5.9, we also keep the unimodular matrix $U$ such that $D = UWV$ is in Smith normal form (it is not necessary to keep the unimodular matrix $V$).

Now given an ideal $I$ we can first compute the norm of $I$. If it is small, then $I$ will factor on the factor base FB chosen in Algorithm 6.5.6. Otherwise, as in Algorithm 6.5.6, we choose random exponents $v_i$ and compute $I\prod_{1\le i\le s} S_i^{v_i}$ and reduce this ideal (along the direction 0 for instance, here it does not matter). Since this reduced ideal has a reasonably small norm, we may hope to factor it on our factor base, thus expressing $I$ in the form $I=\alpha\prod_{1\le i\le k}\mathfrak{p}_i^{x_i}$, where we denote by $\mathfrak{p}_i$ the elements of FB.

Once such an equality is obtained, we proceed as follows. Since the columns of $M''$ generate the lattice of relations among the $\mathfrak{p}_i$ in the class group, it is clear that $I$ is a principal ideal if and only if the column vector of the $x_i$ is in the image of $M''$. Let $r$ (resp. $c$) be the number of rows (resp. columns) of the matrix $B$ occuring in $M''$ as described above, and let $c_1$ be the number of initial columns of zeros in $M''$. Then if $X$ (resp. $Y$) is the column vector of the $x_i$ for $1\le i\le r$ (resp. $r<i\le k$), then $I$ is a principal ideal if and only if there exists an integral column vector $Z$ such that $WZ+BY=X$. This is equivalent to $U^{-1}DV^{-1}Z=X-BY$, and since $V$ is unimodular this is equivalent to the existence of an integral column vector $Z_1$ such that

$$DZ_1=U(X-BY).$$

Since $D$ is a diagonal matrix, this means that the $j$-th element of $U(X-BY)$ must be divisible by the $j$-th diagonal element of $D$.

If $I$ is found in this way to be a principal ideal, the use of the complex matrix $M''_C$ allows us to find $\alpha$ such that $I=\alpha\mathbb{Z}_K$.

This gives the following algorithm.

**Algorithm 6.5.10** (Principal Ideal Testing). Given an ideal $I$ of $\mathbb{Z}_K$, this algorithm tests whether $I$ is a principal ideal, and if it is, computes an $\alpha\in K$ such that $I=\alpha\mathbb{Z}_K$. We assume computed the matrices $M''$ and $M''_C$ (and hence the matrices $W$ and $B$), as well as the unimodular matrices $U$ and $V$ and the diagonal matrix $D$ such that $UWV=D$ is in Smith normal form, as explained above. We keep the notations of Algorithm 6.5.6.

1. [Reduce to primitive] If $I$ is not a primitive integral ideal, compute a rational number $a$ such that $I/a$ is primitive integral, and set $I\leftarrow I/a$.
2. [Small norm?] If $\mathcal{N}(I)$ is divisible only by prime numbers below the prime ideals in the factor base FB (i.e. less than or equal to $L_1$) set $v_i\leftarrow 0$ for $i\le s$, $\beta\leftarrow a$ and go to step 4.
3. [Generate random relations] Choose random nonnegative integers $v_i\le 20$ for $i\le s$, compute the ideal $I_1\leftarrow I\prod_{1\le i\le s}S_i^{v_i}$, and let $J=I_1/\gamma$ be the ideal obtained by LLL-reducing $I_1$ along the direction of the zero vector. If $\mathcal{N}(J)$ is divisible only by the prime numbers less than equal to $L_1$, set $I\leftarrow J$, $\beta\leftarrow a\gamma$ and go to step 4. Otherwise, go to step 3.
4. [Factor $I$] Using Algorithm 4.8.17, factor $I$ on the factor base FB. Let $I=\prod_{1\le i\le k}\mathfrak{p}_i^{x_i}$. Let $X$ (resp. $Y$) be the column vector of the $x_i-v_i$ for $i\le r$

(resp. $i > r$), where $r$ is the number of rows of the matrix $B$, as above, and where we set $v_i = 0$ for $i > s$.

5. [Check if principal] Let $Z \leftarrow D^{-1}U(X - BY)$ (since $D$ is a diagonal matrix, no matrix inverse must be computed here). If some entry of $Z$ is not integral, output a message saying that the ideal $I$ is not a principal ideal and terminate the algorithm.

6. [Use Archimedean information] Let $A$ be the $(c_1+k)$-column vector whose first $c_1$ elements are zero, whose next $r$ elements are the elements of $Z$, and whose last $k-r$ elements are the elements of $Y$. Let $A_C = (a_i)_{1\le i\le r_u} \leftarrow M''_C A$.

7. [Restore correct information] Set $s \leftarrow (\ln \mathcal{N}(I))/n$, and let $A' = (a'_i)_{1\le i\le n}$ be defined by $a'_i \leftarrow \exp(s+a_i)$ if $i \le r_1$, $a'_i \leftarrow \exp(s+(a_i/2))$ if $r_1 < i \le r_u$ and $a'_i \leftarrow \exp(s+\overline{(a_{i-r_2}/2)})$ if $r_u < i \le n$. (As in Algorithm 6.5.8, the exponential which is computed here may overflow the possibilities of the implementation, in which case the algorithm must be aborted.)

8. [Round] Set $A'' \leftarrow \Omega^{-1}A'$, where $\Omega = \sigma_j(\omega_i)$ as in Algorithm 6.5.8. The coefficients of $A''$ must be close to rational integers. If this is not the case, then either the precision used to make the computation was insufficient or the desired $\alpha$ is too large. Otherwise, round the coefficients of $A''$ to the nearest integer.

9. [Terminate] Let $\alpha'$ be the element of $\mathbb{Z}_K$ whose coordinates in the integral basis are given by the vector $A''$. Set $\alpha \leftarrow \beta\alpha'$ (product computed in $K$). If $I \ne \alpha\mathbb{Z}_K$, output an error message stating that the accuracy is not sufficient to compute $\alpha$. Otherwise, output $\alpha$ and terminate the algorithm.

Note that, since we chose the complex logarithmic embedding $L_C(\alpha) - \frac{\ln(\mathcal{N}(\alpha))}{n}V$ as defined in Section 5.8.4, we must adjust the components by $s = (\ln \mathcal{N}(I))/n$ before computing the exponential in Step 7.

**Remark.** It is often useful in step 5 to give more information than just the negative information that $I$ is not a principal ideal. Indeed, if as suggested in Remark (4) after Algorithm 6.5.9, the explicit generators $\overline{g_i}$ of order $d_i$ of the class group $Cl(K)$ have been computed, we can easily compute $\alpha$ and $k_i$ such that $I = \alpha\prod_i g_i^{k_i}$ and $0 \le k_i < d_i$. The necessary modifications to the above algorithm are easy and left to the reader.

## 6.6 Exercises for Chapter 6

1. By Theorem 6.1.4, $\mathbb{Z}[\theta] + (U(\theta)/p)\mathbb{Z}[\theta]$ is an order, hence a ring. Clearly the only non-trivial fact to check about this is that $(U(\theta)/p)^2$ is still in this order. Using the notations of Theorem 6.1.4, show how to compute polynomials $A$ and $B$ in $\mathbb{Z}[X]$ such that

$$\frac{U(\theta)^2}{p^2} = A(\theta) + \frac{U(\theta)}{p}B(\theta).$$

2. Compute the maximal order of pure cubic fields using only Dedekind's criterion (Theorem 6.1.4) instead of the Pohst-Zassenhaus theorem.

3. (F. Diaz y Diaz.) With the notations of Theorem 6.1.4, show that a restatement of the Dedekind criterion is the following. Let $r_i(X)$ be the remainder of the Euclidean division of $T(X)$ by $t_i(X)$. We have evidently $r_i \in p\mathbb{Z}[X]$. Set $d_i = 1$ if $e_i \geq 2$ and $r_i \in p^2\mathbb{Z}[X]$, $d_i = 0$ otherwise. Then in (3) we can take $U(X) = \prod_{1 \leq i \leq k} t_i^{e_i - d_i}$. In particular, $\mathbb{Z}[\theta]$ is $p$-maximal if and only if $r_i \notin p^2\mathbb{Z}[X]$ for every $i$ such that $e_i \geq 2$.

4. Let $\mathcal{O}$ be an order in a number field $K$ and let $p$ be a prime number. Show that $\mathcal{O}$ is $p$-maximal if and only if every ideal $\mathfrak{p}_i$ of $\mathcal{O}$ which lies above $p$ is invertible in $\mathcal{O}$.

5. Prove Proposition 6.2.1 by first proving the formula for $\mathfrak{a}_i^{-1}$ given in the text.

6. Given a finite separable algebra $A$ over $F_p$ isomorphic to a product of $k$ fields $A_i$, compute the probability that a random element $x$ of $A$ is a generator of $A$ in terms of the dimensions $d_i$ of the $A_i$ (hint: use Exercise 13 of Chapter 3).

7. Let $m$ and $n$ be distinct squarefree (positive or negative) integers different from 1. Compute an integral basis for the quartic field $K = \mathbb{Q}(\sqrt{n}, \sqrt{m})$. Find also the explicit decomposition of prime numbers in $K$.

8. (H. W. Lenstra)

   a) Let $A$ be a separable algebra of degree $n$ over $\mathbb{F}_p$ (for example $A = \mathcal{O}/H_j$ in the notation of Section 6.2). Then $A$ is isomorphic to a product of fields $K$, and let $\chi_m$ be the number of such fields which are of degree $m$ over $\mathbb{F}_p$ (if $A = \mathcal{O}/H_j$, then $\chi_m$ is the number of prime ideals of $\mathcal{O}$ of degree $m$ dividing $H_j$). Show that for all $d$ such that $1 \leq d \leq n$ one has

$$\sum_{1 \leq m \leq n} \gcd(d, m)\chi_m = \dim_{\mathbb{F}_p}(\ker(\sigma^d - 1)),$$

   where $\sigma$ denotes the Frobenius homomorphism $x \mapsto x^p$ from $A$ to $A$.

   b) Compute explicitly the inverse of the matrix $M_n = (\gcd(i, j))_{1 \leq i,j \leq n}$ and give an algorithm which computes the local Euler factor

$$L_p = \prod_{\mathfrak{p} \mid p} (1 - \mathcal{N}(\mathfrak{p})^{-s})^{-1}$$

   without splitting explicitly the $H_j$ of Section 6.2.

9. Using the ideas used in decomposing prime numbers into a product of prime ideals, write a general algorithm for factoring polynomials over $\mathbb{Q}_p$. You may assume that the coefficients are known to any necessary accuracy (for example that they are in $\mathbb{Q}$), and that the required $p$-adic precision for the result is sufficiently high. (Hint: If $K = \mathbb{Q}[\theta]$ with $T(\theta) = 0$ and if $p\mathbb{Z}_K = \prod_i \mathfrak{p}_i^{e_i}$, consider the characteristic polynomial of the map multiplication by $\theta$ in the $\mathbb{Z}/p^k\mathbb{Z}$-module $\mathbb{Z}_K/\mathfrak{p}_i^{ke_i}$.)

10. (Dedekind) Let $K = \mathbb{Q}(\theta)$ be the cubic field defined by the polynomial $P(X) = X^3 + X^2 - 2X + 8$.

    a) Compute the discriminant of $P(X)$.

b) Show that $(1, \theta, (\theta + \theta^2)/2)$ is an integral basis of $\mathbb{Z}_K$ and that the discriminant of $K$ is equal to $-503$.

c) Using Algorithm 6.2.9 show that the prime 2 is totally split in $K$.

d) Conclude from Theorem 4.8.13 that 2 is an inessential discriminantal divisor, i.e. that it divides the index $[\mathbb{Z}_K : \mathbb{Z}[\alpha]]$ for *any* $\alpha \in \mathbb{Z}_K$.

11. So as to avoid ideal multiplication and division, implement the idea given in the remark after Algorithm 6.2.9, and compare the efficiency of this modified algorithm with Algorithm 6.2.9.

12. Compute the Galois group of the fields generated by the polynomials $X^3 - 2$, $X^3 - X^2 - 2X + 1$ and $X^4 - 10X^2 + 1$.

13. Compute the accuracy needed for the roots of $T$ so that the rounding procedures used in computing the resolvents in all the Galois group finding algorithms given in the text be correct.

14. Implement the Galois group algorithms and check your implementation with the list of 37 polynomials given at the end of Section 6.3.

15. a) Using Proposition 4.5.3, give an algorithm which determines whether or not a number field $K$ is Galois over $\mathbb{Q}$ (without explicitly computing its Galois group).

b) Using the methods of Section 4.5 write an algorithm which finds explicitly the conjugates of an element of a number field $K$ belonging to $K$. The correctness of the results given by your algorithm should *not* depend on approximations, that is once a tentative formula has been found it must be checked exactly. Note that this algorithm may allow to compute the Galois group of $K$ if $K$ is Galois over $\mathbb{Q}$, even when the degree of $K$ is larger than 7.

16. Determine the decomposition of prime numbers dividing the index in cyclic cubic fields by using the method of Algorithm 6.2.9. (Note: if the reader wants to find also the explicit decomposition of prime numbers not dividing the index, which is given by Theorem 4.8.13, he will first need to solve Exercise 28 of Chapter 1.)

17. Show that the polynomials $P(X)$ given in Theorem 6.4.6 (1) and (2) are irreducible in $\mathbb{Q}[X]$.

18. Complete the proof of Theorem 6.4.6 (3) in the case where $e$ is equal to 9 times a product of $t - 1$ primes congruent to 1 modulo 3.

19. Check that the fields defined in Theorem 6.4.6 (2) are not isomorphic for distinct pairs $(e, u)$ (the proof was given explicitly in the text only for case (1)).

20. Generalize the formulas and results of Section 6.4.2 to cyclic quartic fields, replacing $\mathbb{Q}(\zeta)$ by $\mathbb{Q}(i)$. (Hint: start by showing that such a field has a unique quadratic subfield, which is real.)

21. Using the notations of Theorem 6.4.6, find the minimal equation of $\alpha = (\sigma(\theta) - \theta)/3$, and deduce from this another complete parametrization of cyclic cubic fields.

22. Let $K$ be a cubic field.

a) Show that there exists a $\theta \in \mathbb{Z}_K$ and $a$, $b$ and $c$ in $\mathbb{Z}$ such $(1, \theta, (\theta^2 + a\theta + b)/c)$ is an integral basis, and give an algorithm for finding $\theta$, $a$, $b$ and $c$.

b) Such a $\theta$ being found, show that there exists a $k \in \mathbb{Z}$ such that if we set $\omega = \theta + k$, then $(1, \omega, (\omega^2 + a_2\omega)/a_3)$ is an integral basis of $\mathbb{Z}_K$ for some integers $a_2$ and $a_3$.

c) Deduce from this that for any cubic field $K$ there exists $\alpha \in K$ which is not necessarily an algebraic integer such that $\mathbb{Z}_K = \mathbb{Z}[\alpha]$ in the sense of Exercise 15 of Chapter 4.

d) Generalize this result to the case of an arbitrary order in a cubic field $K$ by allowing the polynomial used in Exercise 15 of Chapter 4 to have a content larger than 1.

23. Prove that, as claimed in the text, Theorem 4.9.12 (4) implies the formula

$$\frac{h(K)R(K)}{w(K)} = 2^{-r_1}(2\pi)^{-r_2}\sqrt{|d(K)|}\prod_p \frac{1-1/p}{\prod_{\mathfrak{p}|p}(1-1/\mathcal{N}(\mathfrak{p}))}.$$

24. Using Algorithm 6.5.9 compute the class group, the regulator and a system of fundamental units for the number fields defined by the polynomials $T(X) = X^4 + 6$, $T(X) = X^4 - 3X + 5$ and $T(X) = X^4 - 3X - 5$.

25. Compute the different of pure cubic fields and of cyclic cubic fields using Proposition 4.8.19 and Algorithm 4.8.21.

26. Let $(\omega_i)_{1\le i\le n}$ be an integral basis for a number field $K$ of degree $n$ such that $\omega_n = 1$, and set $t_i = \mathrm{Tr}_{K/\mathbb{Q}}(\omega_i)/n$. Consider the lattice $\mathbb{Z}^{n-1}$ together with the quadratic form

$$q(\mathbf{x}) = \sum_{k=1}^{n}\left|\sigma_k\left(\sum_{1\le i\le n-1} x_i(\omega_i - t_i)\right)\right|^2.$$

a) Show that the determinant of this lattice is equal to $\sqrt{|d(K)|/n}$.

b) Setting $\theta = \sum_{i=1}^{n-1} x_i\omega_i - \left\lfloor\sum_{i=1}^{n-1} x_i t_i\right\rceil$ prove Hunter's Theorem 6.4.2.

27. Let $m(X) = m_1(X)\cdots m_k(X)$ be the decomposition of $m(X)$ obtained in step 13 of Algorithm 6.2.9. For $1 \le r \le k$, let $e_r$ be a lift to $\mathcal{O}$ of $m_r(\alpha)$, and set $H_r = H + e_r\mathcal{O}$. Show that $H = H_1\cdots H_r$, and hence that steps 14 and 15 of Algorithm 6.2.9 are valid. (Note: the $e_r$ are *not* orthogonal idempotents.)