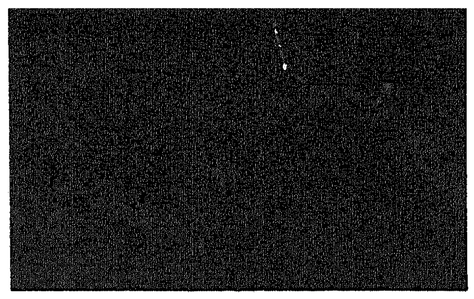
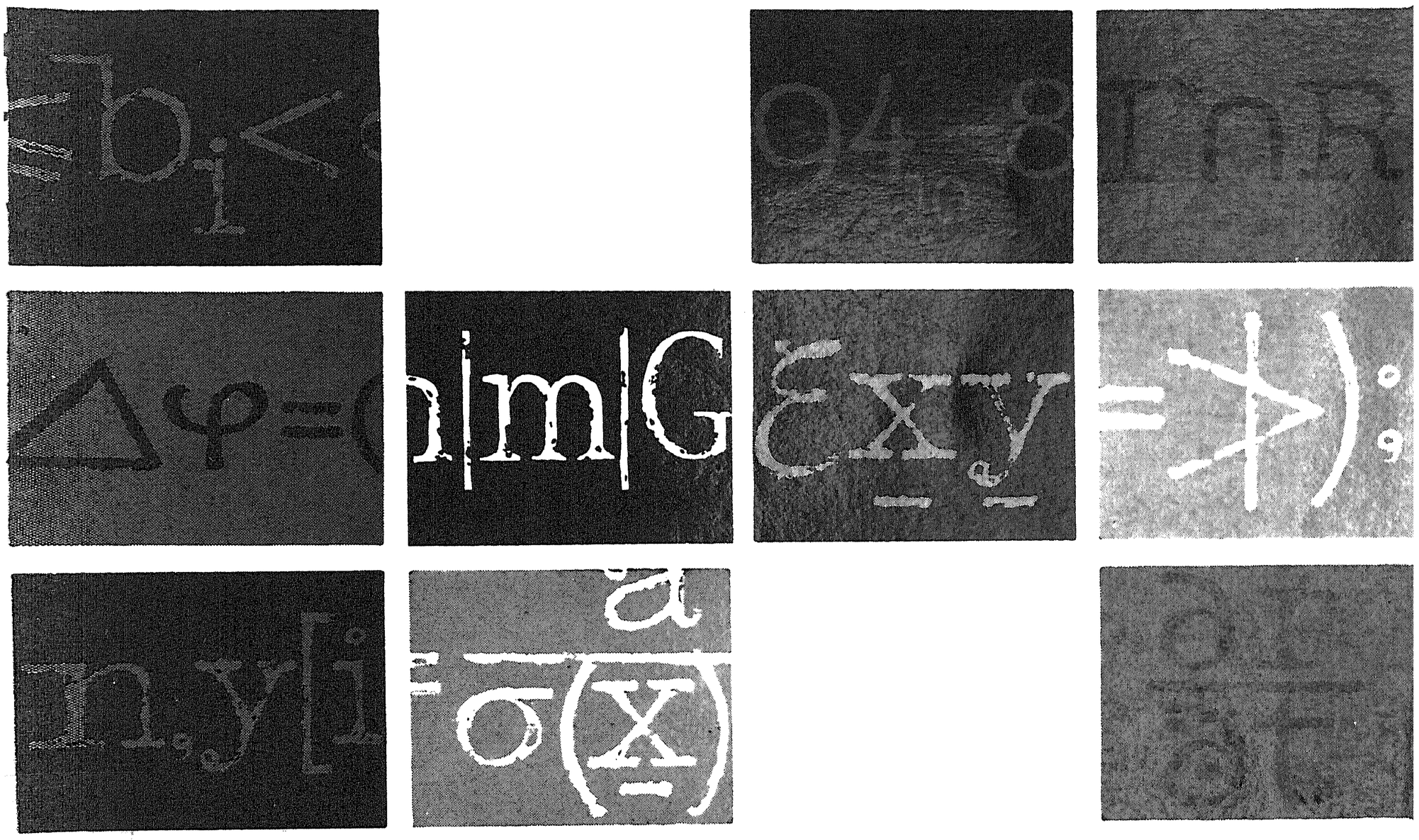


COMPUTATIONAL METHODS IN NUMBER THEORY

PART II

edited by H.W. LENSTRA, JR.
R. TIJDEMAN



MATHEMATICAL CENTRE TRACTS 155

MATHEMATICAL CENTRE TRACTS 155

COMPUTATIONAL METHODS IN NUMBER THEORY

PART II

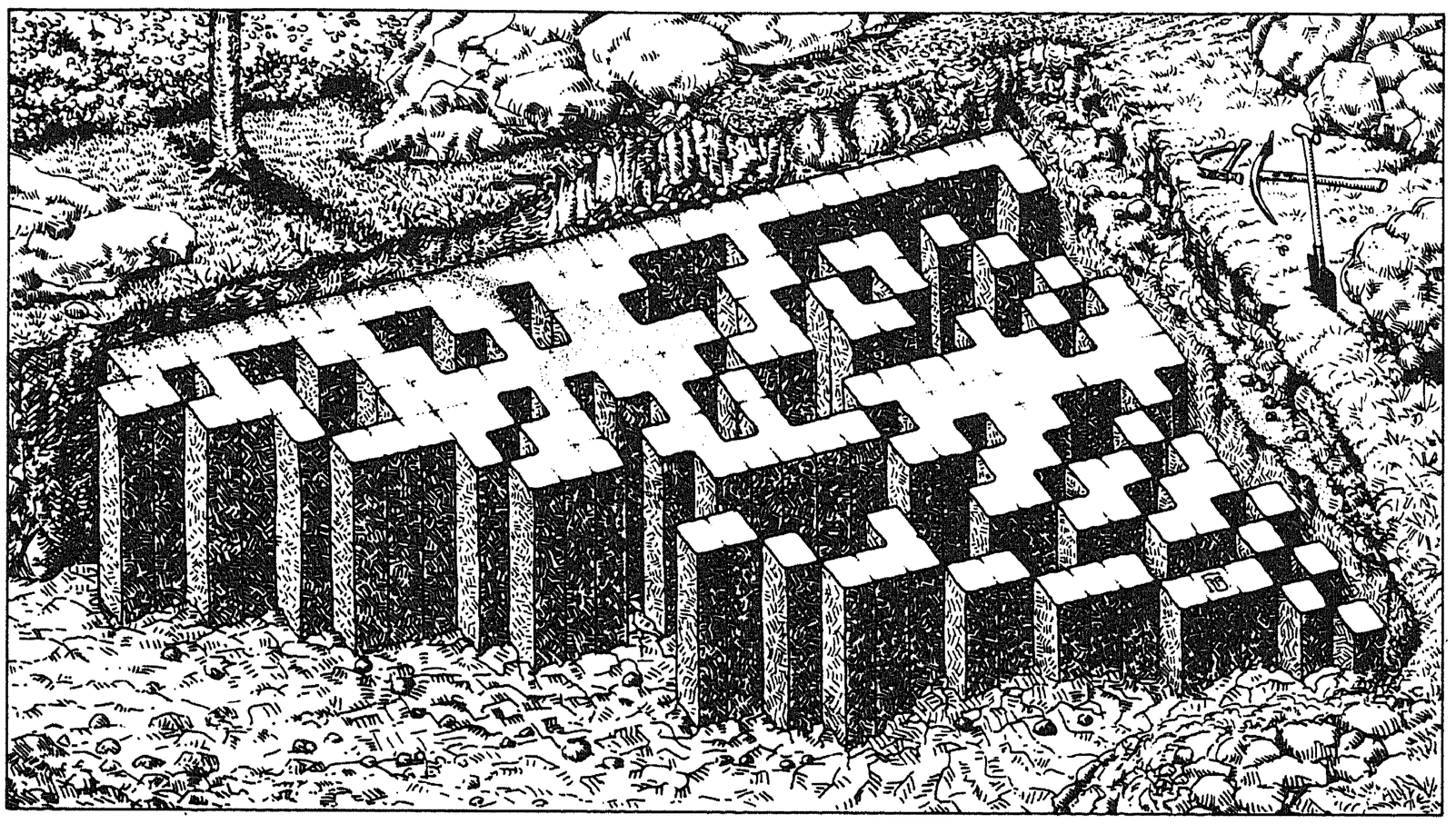
edited by

H.W. LENSTRA, JR.

R. TIJDEMAN

MATHEMATISCH CENTRUM AMSTERDAM 1982

ISBN 90 6196 249 8



CONTENTS PART II

CONTENTS PART I

PREFACE

A preliminary version of this tract appeared in 1980 under the title "Studieweek getaltheorie en computers". It contained the written versions of the lectures presented during the study week "Number theory and computers" that was held at the Mathematical Centre, September 1-5, 1980. The contents have been thoroughly revised for the present edition. We are happy to include Carl Pomerance's paper "Analysis and comparison of some integer factoring algorithms", which does not correspond to a lecture during the study week.

The editors are grateful to all those at the Mathematical Centre who have contributed to the technical realization of the tract.

H.W. Lenstra, Jr.
R. Tijdeman

ADDRESSES OF AUTHORS

R.P. BRENT : The Australian National University
Department of Computer Science
P.O. Box 4
Canberra ACT 2600
Australia

A.J. BRENTJES : Papiermolen 4
Molenwijk
2317 SV Leiden
The Netherlands

P.L. CIJSOUW : Technische Hogeschool Eindhoven
Onderafdeling der Wiskunde en Informatica
Postbus 513
5600 MB Eindhoven
The Netherlands

P. VAN EMDE BOAS : Universiteit van Amsterdam
Instituut voor Interdisciplinaire Wiskunde
Roetersstraat 15
1018 WB Amsterdam
The Netherlands

P.J. HOOGENDOORN : Stichting Mathematisch Centrum
Kruislaan 413
1098 SJ Amsterdam
The Netherlands

A. KORLAAR : Technische Hogeschool Eindhoven
Onderafdeling der Wiskunde en Informatica
Postbus 513
5600 MB Eindhoven
The Netherlands

A.K. LENSTRA : Stichting Mathematisch Centrum
Kruislaan 413
1098 SJ Amsterdam
The Netherlands

H.W. LENSTRA, JR. : Universiteit van Amsterdam
Mathematisch Instituut
Roetersstraat 15
1018 WB Amsterdam
The Netherlands

F.J. VAN DER LINDEN : Universiteit van Amsterdam
Mathematisch Instituut
Roetersstraat 15
1018 WB Amsterdam
The Netherlands

J. VAN DE LUNE : Stichting Mathematisch Centrum
Kruislaan 413
1098 SJ Amsterdam
The Netherlands

C. POMERANCE : Department of Mathematics
University of Georgia
Athens, GA 30602
U.S.A.

H.J.J. TE RIELE : Stichting Mathematisch Centrum
Kruislaan 413
1098 SJ Amsterdam
The Netherlands

R.J. SCHOOF : Rijksuniversiteit Leiden
Mathematisch Instituut
Postbus 9512
2300 RA Leiden
The Netherlands

R.J. STROEKER : Erasmus Universiteit Rotterdam
Econometrisch Instituut
Burgemeester Oudlaan 50
3062 PA Rotterdam
The Netherlands

R. TIJDEMAN : Rijksuniversiteit Leiden
Mathematisch Instituut
Postbus 9512
2300 RA Leiden
The Netherlands

J.W.M. TURK : Erasmus Universiteit Rotterdam
Econometrisch Instituut
Burgemeester Oudlaan 50
3062 PA Rotterdam
The Netherlands

M. VOORHOEVE : Philips Data Systems
Postbus 245
7300 AE Apeldoorn
The Netherlands

D.T. WINTER : Stichting Mathematisch Centrum
Kruislaan 413
1098 SJ Amsterdam
The Netherlands

H. ZANTEMA : Universiteit van Amsterdam
Mathematisch Instituut
Roetersstraat 15
1018 WB Amsterdam
The Netherlands

THE COMPUTATION OF GALOIS GROUPS

by

F.J. VAN DER LINDEN

INTRODUCTION

Let f be a monic polynomial of degree n in $\mathbb{Z}[x]$. Assume f is *square free*, i.e. f has no double zeros. One of the fundamental invariants of f is its *Galois group*, which may be described as follows. Let $\alpha_1,\ldots,\alpha_n$ be the zeros of f, then $\mathbb{Q}(\alpha_1,\ldots,\alpha_n)$ is called the *splitting field* of f. This is a Galois extension of $\mathbb{Q}$. The Galois group G of this extension is called the Galois group of f. The elements of G permute the α_i so we can consider G as a subgroup of S_n, the symmetric group on n letters. We want to know which subgroup it is; it is determined up to conjugacy.

We will discuss the existing techniques to determine G with the help of an electronic computer. For simplicity we often restrict ourselves to the case that f is *irreducible*. For G this means that it is a *transitive* subgroup of S_n. We will give two major methods for the computation of G in Sections 1, 2. The first one does not compute G in all cases, but it leaves us sometimes with a choice between several subgroups of S_n. If we assume certain *generalized Riemann hypotheses* more subgroups can be eliminated, but even then we may be left with several possibilities. The method has the advantage that essentially the same program can be used for different values of n. The second method determines G always, but we must use multiprecision real and integral arithmetic, and for different values of n different programs have to be used.

In Section 3 we show how the advantages of both methods can be combined. Some methods of lesser importance are discussed in Section 4.

The cases n = 2, 3 are particularly easy. For n = 2 we always have $G = S_2$ when f is irreducible. For n = 3 we only have two transitive groups: S_3 and A_3. In Section 2 we show how to distinguish between them.

1. THE METHOD OF VAN DER WAERDEN

In this section we fix a square free monic polynomial $f \in \mathbb{Z}[x]$ of degree n. Let $G \subset S_n$ be its Galois group.

VAN DER WAERDEN gave in [19], §66, a method to compute G (see also [9]). ZASSENHAUS [24] and COCKAYNE [2] used it to determine G with the help of electronic computers. This method will be described in this section.

Let $d_1,\ldots,d_r$ be positive integers with $\Sigma_{i=1}^{r} d_i = n$. We say that $\sigma \in S_n$ has *cycle pattern* $(d_1,\ldots,d_r)$ if σ is the product of r disjoint cycles of lengths $d_1,\ldots,d_r$. Let p be a prime number. Suppose that $\bar{f} = (f \bmod p) \in \mathbb{F}_p[x]$ factorizes as

$$\bar{f} = \prod_{i=1}^{r} \bar{f}_i, \quad \text{with degree } (\bar{f}_i) = d_i,$$

where the $\bar{f}_i$ are distinct monic irreducible polynomials in $\mathbb{F}_p[x]$. In this situation we say that p *belongs* to the cycle pattern $(d_1,\ldots,d_r)$. We also want to speak of the cycle pattern belonging to the "prime at infinity". This can be defined by replacing in the above definition $\mathbb{F}_p[x]$ by $\mathbb{R}[x]$ and $(f \bmod p)$ by f. In this case all d_i are 1 or 2.

THEOREM 1. *Suppose that* p *is a, possibly infinite, prime, which belongs to the cycle pattern* $(d_1,\ldots,d_r)$. *Then there exists an element* σ *of* G *of cycle pattern* $(d_1,\ldots,d_r)$.

PROOF. For finite primes see [19], §66. For the infinite prime we embed $\mathbb{Q}(\alpha_1,\ldots,\alpha_n)$ in $\mathbb{C}$. Then σ = (complex conjugation) has cycle pattern $(d_1,\ldots,d_r)$. □

We can now use the methods of A.K. LENSTRA [7] to factorize f modulo several primes. For $n \leq 5$ it can be easier because we have the following theorem:

THEOREM 2 (STICKELBERGER). *Let* K *be a finite field of odd characteristic. Let* g *be a square free polynomial of degree* n *over* K. *Let* r *be the number of irreducible factors of* g *over* K. *Let* Δ *be the discriminant of* g *over* K. *Then* $r \equiv n \bmod 2$ *iff* Δ *is a square in* K.

PROOF. See [17], Cor. 1. □

For $n \leq 5$ there are at most two non-linear facotrs. So in this case we only have to find the linear factors of $f \bmod p$ to get the complete cycle pattern.

Now we can make a list of all subgroups of S_n which contain elements with the encountered cycle pattern. Such a list is already available for subgroups of S_n with $n \leq 20$, cf. [6]. By Theorem 1 we know that G is in this list. This does not suffice to determine G except if we find $G = S_n$ (or $G = A_n$, if we know that the discriminant of f is a square; see Section 2). In the other cases, it can be useful to know that for every cycle pattern occurring in G there is a prime number p belonging to it; and these p's occur with the expected frequency:

THEOREM 3 (FROBENIUS-TSCHEBOTAREFF). *Let* $(d_1,\ldots,d_r)$ *be a cycle pattern. Let* C *be the set of elements of* G *with cycle pattern* $(d_1,\ldots,d_r)$, *and let* P *be the set of primes belonging to* $(d_1,\ldots,d_r)$. *Then*

$$\lim_{x\to\infty} \frac{\#\{p\in x:\ p\in P\}}{\#\{p\leq x:\ p \text{ prime}\}} = \frac{\#C}{\#G}\ .$$

PROOF. See [3], [18], [5]. □

This theorem is not very useful for our purpose. We want to have an explicit error term, or at least an upper bound for the smallest $p \in P$. A few years ago some results in this direction have been found. Unfortunately, they are very weak.

THEOREM 4 (LAGRARIAS-ODLYZKO-MONTGOMERY). *Let* C *and* P *be as in Theorem* 3. *Suppose* $C \neq \emptyset$, *and that all prime divisors of the discriminant of* f *are divisors of the discriminant* $\mathcal{D}$ *of* $\mathbb{Q}(\alpha_1,\ldots,\alpha_n)$ *over* $\mathbb{Q}$. *Then there exists* $p \in P$ *with*

$$p \leq 2\cdot|\mathcal{D}|^A,$$

where A *is an absolute, effectively computable constant.*

PROOF. See [4], Theorem 1.1. □

The assumption of the divisors of the discriminant of f can probably be omitted by altering the definition of $\mathcal{D}$, but no such theorem has been published. The value of A in Theorem 4 is not given explicitly. OESTERLÉ has given a much better result assuming the generalized Riemann hypothesis:

THEOREM 5. (OESTERLÉ). *Let the assumptions and notations be as in Theorem* 3, *and assume moreover that the zeta-function of* $\mathbb{Q}(\alpha_1,\ldots,\alpha_n)$ *satisfies the generalized Riemann hypothesis. Then there exists a* $p \in P$ *with*

$$p \leq 70\cdot(\log |D|)^2.$$

PROOF. Promised in [10], Théorème 4; cf. [5], Corollary 1.2. □

Oesterlé also announced a completely explicit remainder term in Theorem 3, still assuming the generalized Rîemann hypothesis.

With these theorems we cannot get G for sure. We still know that G must belong to the same list, but in addition we have the moral certainty that G must belong to a much smaller list just by looking at the frequencies of the cycle patterns that are found. Often this list contains only one subgroup of the S_n. The smallest n for which S_n contains two non-conjugate transitive subgroups with the same frequencies of cycle patterns is n = 8, see [11].

Table 1, abstracted from [24], gives all transitive subgroups of S_n for n = 4, 5, together with their cycle patterns. The cases n = 2, 3 are trivial; see the Introduction and Section 2.

TABLE 1

Transitive subgroups of S_n for n = 4, 5.

n	#G	Cycle pattern	Frequency	G
4	24	1111	1	S_4
		211	6	
		22	3	
		31	8	
		4	6	
4	12	1111	1	A_4
		22	3	
		31	8	
4	8	1111	1	D_4
		211	2	
		22	3	
		4	2	
4	4	1111	1	V_4
		22	3	
4	4	1111	1	C_4
		22	1	
		4	2	

n	#G	Cycle pattern	Frequency	G
5	120	11111	1	S_5
		2111	10	
		221	15	
		311	20	
		32	20	
		41	30	
		5	24	
5	60	11111	1	A_5
		221	15	
		311	20	
		5	24	
5	20	11111	1	N_5
		221	5	
		41	10	
		5	4	
5	10	11111	1	D_5
		221	5	
		5	4	
5	5	11111	1	C_5
		5	4	

EXAMPLE. Let $f = x^4 - 4x^3 - 4x + 13$. This polynomial is irreducible over $\mathbb{Q}$. The zeros of f are approximately

$$\alpha_1 = 1.4159768...$$
$$\alpha_2 = 4.0481248...$$
$$\alpha_3 = -0.7320508... + i(1.3160740...)$$
$$\alpha_4 = -0.7320508... - i(1.3160740...)$$

The discriminant of f is $-2^8.3^3.13^2$. Factoring modulo primes gives the following cycle patterns:

Cycle pattern	Primes
211	∞,11,23
22	7,19
4	5,17

Using Table 1, we see that $G = D_4$ or $G = S_4$, and that we are morally sure that $G = D_4$.

2. THE METHOD OF STAUDUHAR

The method proposed by STAUDUHAR [16], cf. [8], to compute Galois groups, is based on the use of *Galois resolvents*. These are defined as follows: Write $\underline{u} = (u_1,\ldots,u_n)$, where the u_i are indeterminants over $\mathbb{Q}$. Consider the field $\mathbb{Q}(\underline{u})$. We have an action of S_n on $\mathbb{Q}(\underline{u})$ by permuting the u_i. For every subgroup H of S_n we denote by $\mathbb{Q}(\underline{u})^H$ the fixed field of H. By Galois theory we have $\mathrm{Gal}(\mathbb{Q}(\underline{u})/\mathbb{Q}(\underline{u})^H) = H$. Let H' be a subgroup of H; then we have $\mathbb{Q}(\underline{u})^{H'} = \mathbb{Q}(\underline{u})^H(F(\underline{u}))$ for some $F(\underline{u}) \in \mathbb{Q}(\underline{u})^{H'}$. We may choose $F(\underline{u}) \in \mathbb{Z}[\underline{u}]$. Let $\Phi_{H,H'}(z,\underline{u}) = \prod_{\sigma\in R}(z-\sigma F(\underline{u}))$, where R is a set of left coset representatives of H' in H, i.e. H is the disjoint union of $\sigma H'$ for $\sigma \in R$. We call $\Phi_{H,H'}$ the *Galois resolvent* of H' in H corresponding to $F(\underline{u})$.

THEOREM 6. *Let* $f \in \mathbb{Z}[x]$ *be monic and irreducible, and* $G \subset S_n$ *its Galois group. Let* $H \subset S_n$ *be a subgroup containing* G, *and* $H' \subset H$ *a subgroup. Let* $\sigma \in H$. *Write* $\underline{\alpha} = (\alpha_1,\ldots,\alpha_n)$. *Then*

a) $\Phi_{H,H'}(z,\underline{\alpha}) \in \mathbb{Z}[z]$;

b) *if* $G \subset \sigma H'\sigma^{-1}$ *then the zero* $\sigma F(\underline{\alpha})$ *of* $\Phi_{H,H'}(z,\underline{\alpha})$ *is in* $\mathbb{Z}$;

c) *conversely, if* $\sigma F(\underline{\alpha}) \in \mathbb{Z}$, *and* $\sigma F(\underline{\alpha})$ *is not a double zero of* $\Phi_{H,H'}(z,\underline{\alpha})$, *then* $G \subset \sigma H\sigma^{-1}$.

PROOF. See [16], Theorems 4, 5. For an important special case ($H = S_n$) see [8]. □

Stauduhar uses this theorem in the following way. Suppose one knows that $G \subset H$, where H is a transitive subgroup of S_n; e.g. one knows this for $H = S_n$. Using Galois resolvents and Theorem 6, we can determine whether or not $G \subset \sigma H'\sigma^{-1}$ for some maximal transitive subgroup $H' \subset H$ and some $\sigma \in H$. If this does not occur then $G = H$. If however $G \subset \sigma H'\sigma^{-1}$, we replace H by $\sigma H'\sigma^{-1}$ and repeat the procedure.

Some remarks are in order here.

1) We compute $\Phi_{H,H'}(z,\underline{\alpha})$ with the help of its zeros which we get from the zeros of f. Because we know that $\Phi_{H,H'}(z,\underline{\alpha}) \in \mathbb{Z}[z]$, we can calculate it exactly on an electronic computer using multiprecision arithmetic. If a zero of $\Phi_{H,H'}(z,\underline{\alpha})$ is "almost" an integer, we can round it to an integer. With the help of multiprecision integer arithmetic we can show that this integer is a zero of $\Phi_{H,H'}(z,\underline{\alpha})$. In some cases there is an alternative: for small

n we can express $\Phi_{H,H'}(z,\underline{\alpha})$ in the coefficients of f. For example, the resolvent Φ_{S_4,D_4}, which is also called the *cubic resolvent* of a quartic polynomial, will be given below. The resolvent $\Phi_{S_5,N_5}(z,\underline{\alpha}) = \Phi_{A_5,D_5}(z,\underline{\alpha})$ is given in [1], app.1. For Φ_{S_n,A_n} see below.

2) From every H-conjugacy class of maximal transitive subgroups $H' \subset H$ we only have to consider one subgroup.

3) Suppose we get $G \subset \sigma H'\sigma^{-1}$ for some $\sigma \in H$. Then we renumber the zeros of f to get $G \subset H'$.

4) Let H_1 and H_2 be subgroups of S_n. When we get $G \not\subset H_1$ but $G \subset H_2$, then we have of course not to look whether $G \subset H_1 \cap H_2$, when $H_1 \cap H_2$ is a maximal transitive subgroup of H_2.

5) When $\Phi_{H,H'}(z,\underline{\alpha})$ has a double integral zero one has to take another Galois resolvent of H' in H. For most f this does not occur and for some pairs $H' \subset H$ it never occurs. If $\Phi_{H,H'}(z,\underline{\alpha})$ has a double integral zero the α_i must satisfy a given algebraic relation, which happens with "probability" zero.

One special case of Galois resolvents is the resolvent of A_n in S_n. In this case we can take $F = \prod_{1\le i<j\le n}(u_i-u_j)$; then

$$\Phi_{S_n,A_n}(z,\underline{u}) = (z - \prod_{1\le i<j\le n}(u_i-u_j))(z + \prod_{1\le i<j\le n}(u_i-u_j))$$
$$= z^2 - \prod_{1\le i<j\le n}(u_i-u_j)^2.$$

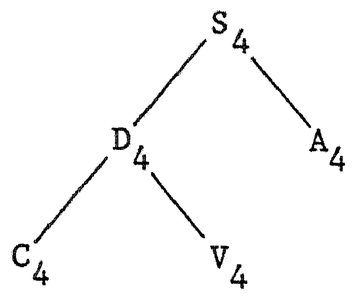
So $\Phi_{S_n,A_n}(z,\underline{\alpha}) = z^2 - \Delta$, where Δ is the discriminant of f. So we have $G \subset A_n$ iff Δ is a square in $\mathbb{Z}$. Because there are faster methods to compute the discriminant, see the talk of H. ZANTEMA [23], we do not use the method given above to look if $G \subset A_n$. We also see that Φ_{S_n,A_n} cannot have a double integral zero, because $\Delta \neq 0$.

Let $f = x^4 + a_1x^3 + a_2x^2 + a_3x + a_4$ be a quartic polynomial. For the resolvent Φ_{S_4,D_4} we can take $F(\underline{u}) = u_1u_3 + u_2u_4$. Then we get $\Phi_{S_4,D_4}(z,\underline{\alpha}) = z^3 - a_2z^2 + (a_1a_3 - 4a_4)z + 4a_2a_4 - a_1^2a_4 - a_3^2$, the cubic resolvent of f. It can be shown that its discriminant is equal to that of f. Moreover it has no double zero if f has none. Also VAN DER WAERDEN has given a cubic resolvent in [19], §64. He took $F(\underline{u}) = (u_1 + u_2)(u_3 + u_4)$, and he considered only the case that $a_1 = 0$. In this case, i.e. $a_1 = 0$, his resolvent is equal to $-\Phi_{S_4,D_4}(-z,\underline{\alpha})$.

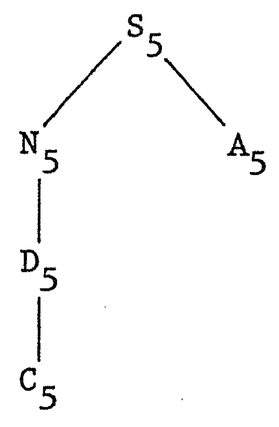
STAUDUHAR has given in [16] data for using this method for $4 \le n \le 7$. He does not consider n = 2, 3 because these are easy: For n = 2 we have only

S_2, for n = 3 we have only S_3 and A_3 as possibilities. One can distinguish S_3 and A_3 by the discriminant. Stauduhar has made search trees (of depth ≤ 5) of subgroups of S_n. He has given $F(\underline{u})$ and systems of representatives for the various pairs of subgroups appearing in these search trees. Below we give these data for n = 4, 5.

n = 4



n = 5



In Table 2 we give the Galois resolvents for pairs of subgroups $H' \subset H$ of S_n. Here Δ means: if $G \subset H$, then $G \subset H'$ iff Δ is a square in $\mathbb{Z}$.

Table 2

n	H	H'	Generators of H'	$F(\underline{u})$	Representatives of H' in H
4	S_4	D_4	(1234), (13)	$u_1u_3+u_2u_4$	(1), (23), (34)
4	S_4	A_4	(123), (134)	Δ	(1), (12)
4	D_4	C_4	(1234)	$u_1u_2^2+u_2u_3^2+u_3u_4^2+u_4u_1^2$	(1), (12)(34)
4	D_4	V_4	(12)(34), (13)(24)	Δ	(1), (13)
5	S_5	N_5	(12345), (2354)	$(u_1u_2+u_2u_3+u_3u_4+u_4u_5+$ $+u_5u_1-u_1u_3-u_3u_5-u_5u_2+$ $-u_2u_4-u_4u_1)^2$	(1), (12)(34), (12435), (15243), (12453), (12543)
5	S_5	A_5	(123), (134), (12)(35)	Δ	(1), (12)
5	N_5	D_5	(12345), (25)(34)	Δ	(1), (2354)
5	D_5	C_5	(12345)	$u_1u_2^2+u_2u_3^2+u_3u_4^2+u_4u_5^2+$ $+u_5u_1^2$	(1), (12)(35)

SOICHER [15] has written a thesis on the computation of Galois groups. He introduces linear resolvents, i.e. resolvents in which the function $F(\underline{u})$ is linear. He gives examples of computer programs for determining whether a zero of a resolvent is integral, and for determining the Galois group using the resolvents.

Recently Girstmair (unpublished) made an improvement on Stauduhar's method. He used resolvents to distinguish whether or not a Galois group is contained in some set of subgroups of S_n. Such a set does not necessarily consists only of conjugates of a given group, but it can contain more groups. Moreover he calculated the resolvents in terms of the coefficients of the polynomial, instead of the zeros, cf. Remark 1 above.

3. THE USE OF BOTH METHODS TOGETHER

We can use the methods of Sections 1 and 2 together in the following way. The method of Section 1 gives us a list of subgroups of S_n and it is known that one of its members has a conjugate contained in G. Now we can use the Galois resolvent of Section 2 to show that G is contained in one of the conjugates of one of the subgroups of the list. If this is the case we know G exactly. If not, G must be bigger than our first guess. BUHLER [1] has used this method to get many polynomials of which the Galois group is equal to A_5.

We can use this on the example of Section 1 where $f = x^4 - 4x^3 - 4x + 13$. We had the possibility $G = D_4$.

We have

$$\Phi_{S_4,D_4}(z,\underline{u}) = (z-u_1u_3-u_3u_4)(z-u_1u_2-u_3u_4)(z-u_1u_4-u_2u_3).$$

When we calculate the roots of $\Phi(z,\underline{\alpha})$ we get

$$\alpha_1\alpha_2 + \alpha_3\alpha_4 = 7.9999999...$$

and

$$\Phi(z,\underline{\alpha}) = z^3 - 36z - 224$$

of which 8 is the only integral zero. So we conclude that $G = D_4$.

4. OTHER METHODS

In this section we discuss some other methods to calculate Galois groups. First there are methods in which one calculates $\mathbb{Q}(\alpha_1,\ldots,\alpha_n)$ and in the course of this calculation we get the Galois group. This can be done in different ways. For examples, we can look at $\mathbb{Q}(\alpha_1)$ and factorize f over it. If f has an irreducible factor of degree ≥ 2 over $\mathbb{Q}(\alpha_1)$ we take a zero, α_2 say of this factor and do the same over $\mathbb{Q}(\alpha_1,\alpha_2)$. We repeat this until f factorizes as a product of linear polynomials. For methods for factoring polynomials over number fields see [21], [22] or the talk of A.K. LENSTRA [7].

Another way of computing $\mathbb{Q}(\alpha_1,\ldots,\alpha_n)$ is to use the methods of SMADJA [14]. He gives methods how to compute in number fields, how to determine the automorphism group of such a field and how one can show that an element with given conjugates is in the ring of integers of the field. So as above we look at $\mathbb{Q}(\alpha_1)$. If all α_i are contained in its ring of integers we are ready, we compute its automorphism group which is the Galois group of f. If α_2 is not contained in it we look at $\mathbb{Q}(\alpha_1,\alpha_2)$ and so on.

The disadvantage of these methods is that if the Galois group of f is S_n, which it is in most cases (see [20]), we have to do the greatest amount of work, contrary to the earlier methods, which are faster when the Galois group is S_n. But one can use these methods when the methods of Section 1 suggest that G is small.

One can also compute G with the help of the ramifying primes, in contrast with Section 1, where we use the primes which do not ramify. We can do this because we know that the inertia groups are subgroups of G which generate G. These inertia groups are cyclic when the ramification is tame. SCHUR [12], [13] has used this method to compute the Galois groups of some sequences of polynomials. These are the following sequences:

I $$L_n = \frac{e^x}{n!}\frac{d^n(x^n e^{-x})}{dx^n} = \sum_{i=0}^{n} \binom{n}{i}\frac{(-x)^i}{i!},$$ the *Laguerre* polynomials.

II $$E_n = \sum_{i=0}^{n}\frac{x^i}{i!},$$ the "truncated exponential series".

III $$J_n = \frac{1}{x}\int_0^x L_n(t)\,dt = \sum_{i=0}^{n}\binom{n}{i}\frac{(-x)^i}{(i+1)!}$$

IV $$K_n^{(0)} = \sum_{i=0}^{n} (-1)^i \binom{2n}{2i} \cdot 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2i-1) x^{n-i}.$$

V $$K_n^{(1)} = \sum_{i=0}^{n} (-1)^i \binom{2n+1}{2i} \cdot 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2i-1) x^{n-i}.$$

The polynomials $(-1)^n n! L_n$, $n! E_n$, $(-1)^n (n+1)! J_n$, $K_n^{(i)}$ are monic and belong to $\mathbb{Z}[x]$. Schur found the following result.

THEOREM 7. (SCHUR).

a) *if* $f = (-1)^n n! L_n$ *then* $G = S_n$;
b) *if* $f = n! E_n$ *then* $G = A_n$ *if* $4 | n$,
$G = S_n$ *if* $4 \nmid n$;
c) *if* $f = (-1)^n (n+1)! J_n$ *then* $G = A_n$ *if* $2 \nmid n$,
$G = A_n$ *if* $n+1$ *is a square*,
$G = S_n$ *in other cases;*
d) *if* $f = K_n^{(i)}$, $i = 0,1$, *then* $G = S_n$ *if* $n > 12$.

PROOF. See [12], [13]. □

REFERENCES

[1] BUHLER, J.P., *Icosahedral Galois representations*, Lecture Notes in Mathematics 654, Springer (1978).

[2] COCKAYNE, E.J., *Computation of Galois group elements of a polynomial equation*, Math. Comp. 23, 425-428 (1969).

[3] FROBENIUS, F.G., *Über Beziehungen zwischen den Primidealen eines algebraischen Körpers und den Substitutionen seiner Gruppe*, S.-ber Kön. Preuss. Akad. Wiss. Berlin, 689-703 (1896). Gesammelte Abhandlungen II 719-733, Springer (1968).

[4] LAGARIAS, J.C., H.L. MONTGOMERY & A.M. ODLYZKO, *A bound for the least prime ideal in the Chebotarev density theorem*, Invent. Math. 54 271-296 (1979).

[5] LAGARIAS, J.C. & A.M. ODLYZKO, *Effective versions of the Chebotarev density theorem*, Algebraic Number Fields, A. Fröhlich (ed.), Academic Press, 409-464 (1977).

[6] LAND, R., *Computation of Pólya polynomials of primitive permutation groups*, Math. Comp. 36 267-278 (1981).

[7] LENSTRA, A.K., *Factorization of polynomials*, This volume.

[8] LEFTON, P., *Galois resolvents of permutation groups*, Amer. Math. Monthly 84, 642-644 (1977).

[9] MAURER, W.D., *The uses of computers in Galois theory*, in: *Computational Problems in Abstract Algebra*, J. Leech (ed.), Pergamon Press 325-328 (1970).

[10] OESTERLÉ, J., *Versions effectives du théoremè de Chebotarev sous l'hypothèse de Riemann généralisée*, Astérisque 61, 165-167 (1979).

[11] PLAYTIS, A.S., S. SEHGAL & H. ZASSENHAUS, *Equidistributed permutation groups*, Comm. Algebra 6, 35-57 (1978).

[12] SCHUR, I., *Gleichungen ohne Affekt*, S.-ber. Preuss. Akad. Wiss. Phys.-Math. Kl. 443-449 (1930), Gesammelte Abhandlungen III, 191-197, Springer (1973).

[13] SCHUR, I., *Affektlose Gleichungen in der Theorie der Laguerreschen und Hermiteschen Polynome*, J. Reine Angew. Math. 165, 52-58 (1931), Gesammelte Abhandlungen III, 227-233.

[14] SMADJA, R., *Utilisation des ordinateurs dans les calculs sur les idéaux des corps de nombres algébriques*, Astérisque 41 - 42, 277-282 (1977).

[15] SOICHER, L., *The computation of Galois groups*, Thesis Concordia Univ., Montreal (1981).

[16] STAUDUHAR, R.P., *The determination of Galois groups*, Math. Comp. 27, 981-996 (1973).

[17] SWAN, R.G., *Factorization of polynomials over finite fields*, Pac. J. Math. 12, 1099-1106 (1962).

[18] TSCHEBOTAREFF, N., *Die Bestimmung der Dichtigkeit einer Menge von Primzahlen, welche zu einer gegebenen Substitutionsklasse gehören*, Math. Ann. 95, 191-228 (1925 - 1926).

[19] WAERDEN, B.L. VAN DER, *Algebra I*, Springer (1971)[12].

[20] WAERDEN, B.L. VAN DER, *Die Seltenheit der Gleichungen mit Affekt*, Math. Ann. 109, 13-16 (1934).

[21] WANG, P.S., *Factoring multivariate polynomials over algebraic number fields*, Math. Comp. 30, 324-336 (1976).

[22] WEINBERGER, P.J. & L.P. ROTHSCHILD, *Factoring polynomials over algebraic number fields*, ACM Trans. Math. Software 2, 335-350 (1976).

[23] ZANTEMA, H., *Class numbers and units*, This volume.

[24] ZASSENHAUS, H., *On the group of an equation*, in: *Computers in Algebra and Number Theory*, G. Birkhoff and M. Hall (eds), SIAM AMS Proc. IV, 69-88 (1971).

CLASS NUMBERS AND UNITS

by

H. ZANTEMA

1. INTRODUCTION

Every non-zero rational number q has a unique expression

$$q = \varepsilon \cdot \prod_{p \text{ prime}} p^{a(p)} \tag{1}$$

where $\varepsilon = \pm 1$ and the $a(p)$ are integers, almost all zero. This statement embodies two properties of the ring of integers $\mathbb{Z}$: first, that it has unique factorization into primes, and secondly, that it only has "trivial" units ± 1.

Let K be an algebraic number field, i.e. a finite extension of $\mathbb{Q}$, and define

$$\mathcal{O}(K) := \{x \in K \mid g(x) = 0 \text{ for some monic } g \in \mathbb{Z}[X]\}.$$

This is a subring of K, called the *ring of integers* of K. From Gauss' lemma we know that $\mathcal{O}(\mathbb{Q}) = \mathbb{Z}$. The ring $\mathcal{O}(K)$ may fail to have the two properties of $\mathbb{Z}$ mentioned above. To recover uniqueness of factorization we have to pass to ideals of $\mathcal{O}(K)$. Put

$$I(K) := \{\underline{a} \subset K \mid x\underline{a} \text{ is a nonzero ideal of } \mathcal{O}(K) \text{ for some } x \in K\}.$$

Elements of I(K) are called *ideals* of K; to avoid confusion ideals of $\mathcal{O}(K)$ will be called *integral ideals*. The set I(K) is an abelian group under multiplication:

$$\underline{a} \cdot \underline{b} := \{\sum_{i=1}^{t} a_i b_i \mid t \in \mathbb{Z},\ t > 0,\ a_i \in \underline{a},\ b_i \in \underline{b}\}.$$

Every ideal $\underline{a}$ has a unique decomposition

$$\underline{a} = \prod_{\underline{p}} \underline{p}^{a(\underline{p})} \tag{2}$$

where $\underline{p}$ ranges over the non-zero prime ideals of $\mathcal{O}(K)$, and $a(\underline{p}) \in \mathbb{Z}$, almost all equal to zero. An ideal of K is called a *principal ideal* if it can be written as $\alpha\mathcal{O}(K)$ for some $\alpha \in K$, $\alpha \neq 0$; the element α is called a *generator*. The set of principal ideals is a subgroup $P(K)$ of $I(K)$; the *class group* $Cl(K)$ is defined by

$$Cl(K) := I(K)/P(K).$$

For $\underline{a} \in I(K)$ we write

$$[\underline{a}] := (\underline{a} \bmod P(K)) \in Cl(K)$$

and call it the *ideal class* of $\underline{a}$. In 3.1 we shall see that $Cl(K)$ is finite; its order is called the *class number* $h(K)$ of K. Roughly speaking, the class number measures how far $\mathcal{O}(K)$ fails to have unique factorization. More precisely, we have $h(K) = 1$ if and only if $\mathcal{O}(K)$ is a principal ideal domain, and if and only if $\mathcal{O}(K)$ has unique factorization.

If an ideal is principal, its generator is determined up to a *unit* of $\mathcal{O}(K)$. The structure of the group of units $\mathcal{O}(K)^*$ of $\mathcal{O}(K)$ is given by Dirichlet's theorem, see 3.3. It is an infinite group, except for the cases $K = \mathbb{Q}$ and $K = \mathbb{Q}(\sqrt{\Delta})$, $\Delta < 0$.

Class numbers and units play an important role in algebraic number theory. For more details and proofs we refer to [SA]. In this paper we describe a computational technique for determining the class group and units for general algebraic number fields K. Throughout the paper we suppose that K is given as $K = \mathbb{Q}(\lambda)$ where λ is a zero of the polynomial

$$f(X) = X^n + a_{n-1}X^{n-1} + \ldots + a_0, \quad a_i \in \mathbb{Z}, \quad i = 0,\ldots,n-1,$$

and f is irreducible over $\mathbb{Z}$. It is well known that every K can be written in this way.

For specific fields K there exist faster methods, see [SCH] for quadratic fields and [A] and [B] for cubic fields; see also 3.5.

2. THE DISCRIMINANT

2.1. Write $f = \prod_{i=1}^{n} (X-\lambda_i)$ for $\lambda_i \in \mathbb{C}$, $i = 1,\ldots,n$; $\lambda = \lambda_1$. The *discriminant* $\Delta(f)$ of f is defined by

$$\Delta(f) := \prod_{i<j} (\lambda_i - \lambda_j)^2 \tag{3}$$

Clearly this doesn't depend on the chosen labeling of the λ_i, so $\Delta(f)$ can be expressed in the coefficients of f. For example, we have

$$\Delta(f) = a_1^2 - 4a_0, \quad \text{if } n = 2,$$

and

$$\Delta(f) = a_1^2 a_2^2 - 4a_1^3 - 4a_2^3 a_0 - 27a_0^2 + 18a_0 a_1 a_2, \quad \text{if } n = 3.$$

Similar expressions can be given for larger n, but they rapidly become unwieldy.

It is possible to compute $\Delta(f)$ by determining all zeros of f numerically and then substituting them in (3). A more efficient way makes use of the properties of the resultant of two polynomials, which we shall now discuss. Let $g,h \in \mathbb{C}[X]$ be two nonzero polynomials, and write

$$g = a \prod_{i=1}^{s} (X-\alpha_i), \quad h = b \prod_{j=1}^{t} (X-\beta_j)$$

with $a,\alpha_1,\ldots,\alpha_s,b,\beta_1,\ldots,\beta_t \in \mathbb{C}$, $a,b \neq 0$. The *resultant* $R(g,h)$ of g and h is defined by

$$R(g,h) = a^t b^s \prod_{i=1}^{s} \prod_{j=1}^{t} (\alpha_i-\beta_j).$$

Clearly

$$R(g,h) = a^t \prod_{i=1}^{s} h(\alpha_i) \tag{4}$$

and

$$R(g,h) = (-1)^{st} R(h,g). \tag{5}$$

If $h \equiv h_1 \bmod g$ and h_1 has degree t_1, one derives from (4) that

$$R(g,h_1) = a^{t_1-t} R(g,h). \tag{6}$$

Combining (5), (6) and the Euclidean algorithm for polynomials we obtain an efficient method for computing resultants. Since $f'(\lambda_i) = \prod_{j\neq i} (\lambda_i-\lambda_j)$ we have

$$\Delta(f) = (-1)^{n(n-1)/2} R(f,f')$$

which gives an efficient way for calculating the discriminant of a polynomial. In 3.2 we shall see that resultants also can be used for computing norms of elements.

2.2. Let K be a number field of degree n over $\mathbb{Q}$. There are exactly n embeddings $\sigma_1,\ldots,\sigma_n$ of K into $\mathbb{C}$. If σ is an embedding of K into $\mathbb{C}$, then so is $\bar{\sigma}$, hence we can label the σ_i's such that $\sigma_1,\ldots,\sigma_{r_1}$ are real and

$$\sigma_{r_1+i} = \bar{\sigma}_{r_1+r_2+i},$$

for $i = 1,\ldots,r_2$, where r_1 and r_2 satisfy $r_1 + 2r_2 = n$. Now K is embedded in $\mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$ by identifying $x \in K$ with

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

this identification makes $\mathcal{O}(K)$ into a lattice in

$$\mathbb{R}^{r_1} \times \mathbb{C}^{r_2} \cong \mathbb{R}^n.$$

If $\{a_1,\ldots,a_n\}$ is a basis of $\mathcal{O}(K)$ as a lattice, the *discriminant* $\Delta(K)$ of K is defined by

$$\Delta(K) := (\det(\sigma_i(a_j)))^2. \tag{7}$$

Using $\Delta(f) = (\det(\lambda_i^{j-1}))^2$ (Vandermonde) one shows

$$\Delta(f) = (\text{index}(\mathcal{O}(K)\colon \mathbb{Z}[\lambda]))^2 \cdot \Delta(K). \tag{8}$$

From (7) we see that $\Delta(K) \in \mathcal{O}(K)$, and (8) gives $\Delta(K) \in \mathbb{Q}$, hence $\Delta(K) \in \mathbb{Z}$.

It would be nice if we could choose λ such that the index of $\mathbb{Z}[\lambda]$ in $\mathcal{O}(K)$ equals 1, but this is not possible for each K. For example, if $K = \mathbb{Q}(\lambda)$ where λ is a zero of $X^3 + 2X^2 - 9X - 2$, the prime 2 splits completely and $\text{index}(\mathcal{O}(K): \mathbb{Z}[\lambda]) = 2$. If 2 would not divide this index for some other choice of λ in the same field, the method of 3.2 would give 3 distinct zeros of some polynomial modulo 2, which is impossible. For the theory about primes dividing the index we refer to [H]; by "the index" we mean $\text{index}(\mathcal{O}(K): \mathbb{Z}[\lambda])$.

The following theorem of DEDEKIND, see [D] and [U], is very useful to determine the prime divisors of the index. For a prime p decompose f mod p, i.e. choose $g_i \in \mathbb{Z}[X]$ monic and $e_i \geq 1$ such that each g_i is irreducible mod p, the g_i's are different mod p and $f \equiv \Pi_i\ g_i^{e_i} \bmod p$. Then

(9)
$$p \mid \text{index}(\mathcal{O}(K): \mathbb{Z}[\lambda]) \quad \text{if and only if}$$
$$(g_j \bmod p) \mid (p^{-1}(f - \Pi_i\ g_i^{e_i}) \bmod p)$$

as elements of $\mathbb{F}_p[X]$, for some j with $e_j \geq 2$. If (9) holds, then

$$p^{-1}(g_j(\lambda))^{-1} \prod_i g_i(\lambda)^{e_i}$$

is an element of $\mathcal{O}(K)$ which is not contained in $\mathbb{Z}[\lambda]$, and $p^{\deg(g_j)}$ divides the index of λ. In many cases, this suffices to determine $\Delta(K)$ using (8). In some cases it is difficult to determine to which power a prime divides the index. More information about K can be helpful in such cases. If p decomposes in K as

$$p\mathcal{O}(K) = \prod \underline{p}^{e_{\underline{p}}},$$

its contribution to $\Delta(K)$ is

$$\prod N_{K/\mathbb{Q}}(\underline{p}^{e_{\underline{p}}-1})$$

if p doesn't divide any $e_{\underline{p}}$. In particular, unramified primes, i.e. primes such that all $e_{\underline{p}} = 1$, will not occur in $\Delta(K)$. If p is wildly ramified in $K/\mathbb{Q}$, i.e. p divides some $e_{\underline{p}}$, one needs information about higher ramification groups to compute the contribution of p in $\Delta(K)$, see [SE].

2.3. We now discuss some attempts to make the index smaller or remove primes from it, i.e. to find $\eta \in \mathcal{O}(K)$ such that $K = \mathbb{Q}(\eta)$, and

$$\text{index}(\mathcal{O}(K): \mathbb{Z}[\eta]) < \text{index}(\mathcal{O}(K): \mathbb{Z}[\lambda])$$

or $\text{index}(\mathcal{O}(K): \mathbb{Z}[\eta])$ is divisible by a smaller power of some prime than $\text{index}(\mathcal{O}(K): \mathbb{Z}[\lambda])$.

If for some p we have $p^{n-k} | a_k$ for $k = 0,1,\ldots,n-1$, choose $\eta = \lambda/p$ and the index decreases by a factor $p^{n(n-1)/2}$. If $a_0 = \pm p^{k_0}$ for some $\frac{1}{2}n < k_0 \leq n$ and $p^{k_0-k} | a_k$ for $k < k_0$, choose $\eta = p/\lambda$ and the index decreases by a factor $p^{(2k_0-n)(n-1)/2}$.

Transformations like these can also be applied to minimum polynomials of $\lambda - a$ for $a \in \mathbb{Z}$ or of other simple expressions in λ. In practice they often succeed for small primes, in particular 2, that divide the index to a high power.

Another method is the following. Choose $\mu \in \mathcal{O}(K) \setminus \mathbb{Z}[\lambda]$, for example

$$\mu = p^{-1} g_j(\lambda)^{-1} \prod_i g_i(\lambda)^{e_i}$$

in the notation of (9). Try to find an element $\eta \neq 0$ of the lattice $\mathbb{Z}[\lambda,\mu]$ which is close to zero in the euclidean metric of $\mathbb{R}^{r1} \times \mathbb{C}^{r2}$. Then by (3) the minimum polynomial of η has a rather small discriminant. In practice this is a useful method, but we can never be sure that it works. In fact, the problem whether the index can be removed completely, i.e. $\mathcal{O}(K) = \mathbb{Z}[\eta]$ for some $\eta \in \mathcal{O}(K)$, can be formulated as a rather difficult diophantine equation. It can be shown, see [G], that for each K there exist at most finitely many $\eta \in \mathcal{O}(K)$ such that $\mathcal{O}(K) = \mathbb{Z}[\eta]$, up to translation by $\mathbb{Z}$. These can all be determined effectively.

If η is a polynomial expression in λ, the minimum polynomial of η can be constructed in the following way. Write $\lambda^i \eta$ as a linear combination of $1,\lambda,\lambda^2,\ldots,\lambda^{n-1}$ with rational coefficients for $i = 1,2,\ldots,n-1$. These expressions give rise to an $n \times n$-matrix R satisfying

$$(\eta I - R)\vec{v} = \vec{0},$$

where I denotes the $n \times n$ unit matrix and $\vec{v}$ the vector with coefficients $1,\lambda,\lambda^2,\ldots,\lambda^{n-1}$. Let

$g(X) = \det(XI - R)$;

we see that $g(\eta) = 0$. If $\mathbb{Q}(\lambda) = \mathbb{Q}(\eta)$ then g is the minimum polynomial of η, else g is the $[\mathbb{Q}(\lambda): \mathbb{Q}(\eta)]$-th power of the minimum polynomial of η. We see that $\eta \in \mathcal{O}(K)$ if and only if $g \in \mathbb{Z}[X]$; this gives a method to check if $\eta \in \mathcal{O}(K)$.

2.4. We shall construct a basis for $\mathcal{O}(K)$. Given $\lambda \in \mathcal{O}(K)$ there is a unique basis of the following form

$$\{1, \frac{h_1(\lambda)}{a_1}, \frac{h_2(\lambda)}{a_2}, \ldots, \frac{h_{n-1}(\lambda)}{a_{n-1}}\}$$

such that $h_i \in \mathbb{Z}[X]$, h_i is monic of degree i, all coefficients at degree $< i$ of h_i are in the interval $(-a_i/2a_{i-1}, a_i/2a_{i-1}]$ and a_i is a positive integer for $i = 1,2,\ldots,n-1$, while $a_0 = 1$. If the index of λ is one, this basis is simply $\{1,\lambda,\ldots,\lambda^{n-1}\}$. It is trivial that

$$\text{index}(\mathcal{O}(K): \mathbb{Z}[\lambda]) = \prod_{i=1}^{n-1} a_i$$

and

$$a_i \cdot a_j \mid a_{i+j} \quad \text{for } i+j \leq n-1.$$

In particular $a_i \mid a_{i+1}$ for $i \leq n-2$. Even if $\Delta(K)$ is not known, there is only a finite number of possibilities for $(a_1,a_2,\ldots,a_{n-1})$ satisfying (8) and the relations above. Under these restrictions a_1 is the maximal possible value such that $h_1(\lambda)/a_1 \in \mathcal{O}(K)$ for some choice of h_1, this can be found by trying all possibilities for h_1. The same can be done for a_2, a_3 and so on until the whole basis has been constructed. Then also $\Delta(K)$ is given by (8).

Although this method always works, it is not fast. For primes p such that

$$p \parallel \text{index}(\mathcal{O}(K): \mathbb{Z}[\lambda])$$

we can avoid this method by using the theorem of Dedekind discussed in 2.2. A good algorithm which works in general is described in [Z1] in a more general context of orders in a commutative $\mathbb{Q}$-algebra.

3. THE CLASS GROUP

3.1. As we saw in 2.2, $\mathcal{O}(K)$ can be embedded in $\mathbb{R}^n \cong \mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$ as a lattice. One easily derives from (7) that its determinant is $2^{-r_2}\cdot\sqrt{|\Delta(K)|}$. Each non-zero ideal $\underline{a}$ of $\mathcal{O}(K)$ is a sublattice of determinant

$$\det(\underline{a}) = N(\underline{a})\cdot 2^{-r_2}\sqrt{|\Delta(K)|},$$

where $N(\underline{a})$ denotes index$(\mathcal{O}(K): \underline{a})$, the *norm* of $\underline{a}$. Define the *norm* $N(x)$ of $x \in K$ by

$$N(x) := \prod_{i=1}^{n} \sigma_i(x);$$

one has $N(x) \in \mathbb{Q}$ for $x \in K$ and $N(x\mathcal{O}(K)) = |N(x)|$ for $x \in \mathcal{O}(K)$, $x \neq 0$. Define for $t \in \mathbb{R}$, $t > 0$:

$$B_t := \{(y_1,\ldots,y_{r_1},z_1,\ldots,z_{r_2}) \in \mathbb{R}^{r_1} \times \mathbb{C}^{r_2} \mid \sum_{i=1}^{r_1} |y_i| + 2\sum_{j=1}^{r_2} |z_j| \leq t\}.$$

The volume of B_t is

$$2^{r_1}\cdot(\pi/2)^{r_2}\cdot t^n/n!\ .$$

The inequality of arithmetic and geometric means gives

$$B_t \cap \mathcal{O}(K) \subset \{x \in \mathcal{O}(K) \mid |N(x)| \leq t^n/n^n\}. \tag{10}$$

Minkowski's theorem from the geometry of numbers states that for each lattice L in $\mathbb{R}^n$ and for each convex 0-symmetric closed set $S \subset \mathbb{R}^n$ satisfying $\mathrm{vol}(S) \geq 2^n\cdot\det(L)$, there exists a non-zero element of $S \cap L$. We apply this theorem to $L = \underline{a}$ and $S = B_t$, where t is chosen such that $\mathrm{vol}(B_t) = 2^n\cdot\det(\underline{a})$. Then one obtains from (10) that each ideal $\underline{a}$ of $\mathcal{O}(K)$ contains an element $x \neq 0$ such that

$$|N(x)| \leq N(\underline{a})\cdot(4/\pi)^{r_2}\cdot n!\cdot n^{-n}\sqrt{|\Delta(K)|}. \tag{11}$$

Now let g be an element of $C\ell(K)$. Choose an integral ideal $\underline{a}$ such that $[\underline{a}] = g^{-1}$, and an element $0 \neq x \in \underline{a}$ satisfying (11). Then $x\cdot\underline{a}^{-1}$ is an integral ideal of norm at most

$$(4/\pi)^{r_2} \cdot n! \cdot n^{-n} \cdot \sqrt{|\Delta(K)|},$$

while $[x \cdot a^{-1}] = g$. We have proved:

THEOREM. *Every ideal class of* K *contains an integral ideal* $\underline{b}$ *with the property*

$$N(\underline{b}) \leq (4/\pi)^{r_2} \cdot n! \cdot n^{-n} \cdot \sqrt{|\Delta(K)|}. \tag{12}$$

This theorem is very useful for computing class numbers: the class group is generated by the ideal classes of the prime ideals of norm not exceeding the right hand side of (12). Since at most finitely many integral ideals satisfy (12), the class group is finite.

The bound in (12) is not best possible. Define $M(n,r_2)$ to be the smallest value so that for each field K of degree n over $\mathbb{Q}$ with $2r_2$ non-real embeddings in $\mathbb{C}$, each ideal class contains an integral ideal $\underline{b}$ of norm at most $M(n,r_2)\sqrt{|\Delta(K)|}$. A reformulation of (12) is

$$M(n,r_2) \leq (4/\pi)^{r_2} \cdot n! \cdot n^{-n}.$$

The following values of $M(n,r_2)$ are known, see [C]:

n	r_2	$M(n,r_2)$	polynomial for which the bound is sharp
2	0	$5^{-1/2}$	$X^2 + X - 1$
2	1	$3^{-1/2}$	$X^2 + X + 1$
3	0	7^{-1}	$X^3 + X^2 - 2X - 1$
3	1	$23^{-1/2}$	$X^3 - X^2 + 1$

The values of $M(n,r_2)$ are improvements of (12). For $r_2 = 0$, they are isolated bounds, i.e. if the field for which the bound is sharp is excluded, the bounds can be improved again. It would be desirable to extend the table to higher values of n, since for n = 2 or 3 better techniques for computing class numbers are available. Though for $n \geq 4$ exact values of $M(n,r_2)$ are not known, ZIMMERT gave in [Z2] sharper upperbounds than the Minkowski bounds given by (12), as we see in the next table.

n	r_2	upperbound for $M(n,r_2)$ given by Minkowski	Zimmert
4	0	.9375	.06921
4	1	.1194	.1026
4	2	.1520	.1473
5	0	.03840	.01992
5	1	.04890	.03114
5	2	.06226	.04737
6	0	.01544	.005317
6	3	.03186	.02140
100	0	$9.333*10^{-43}$	$1.184*10^{-73}$
100	50	$1.643*10^{-37}$	$4.138*10^{-56}$

For larger n the difference between Minkowski's and Zimmert's results increases. In fact from Minkowski's result it follows that for n large enough

$$M(n,r_2) \leq (.38)^n(1.28)^{r_2}$$

while Zimmert improved this to

$$M(n,r_2) \leq (.141)^n(2.55)^{r_2}.$$

So far we only considered upper bounds for the smallest ideal in a class which are valid for *all* classes in $C\ell(K)$. But for our purpose it suffices that the bound is valid for a set of *generators* of $C\ell(K)$, which is a far weaker condition. If a certain generalized Riemann hypothesis is true, then the classes of integral ideals of norm less than $A\cdot(\log|\Delta(K)|)^2$ generate the whole class group, see [L.M.O.]. Here A is some absolute constant, for which no explicit value has been published.

3.2. Let B denote the right hand side of (12) or an improved version of it. We now construct all prime ideals with norm less than B.

If p is a prime not dividing the index, the prime decomposition of $p\mathcal{O}(K)$ is the same as the decomposition of f mod p in $\mathbb{F}_p[X]$ into irreducible polynomials. More precisely, if

$$f(X) \equiv \prod_i g_i(X)^{e_i} \bmod p,$$

with g_i monic and irreducible mod p, and $g_i \not\equiv g_j \bmod p$ for $i \neq j$, then

$$p\mathcal{O}(K) = \prod_i \underline{p}_i^{e_i}$$

for different prime ideals $\underline{p}_i = (p, g_i(\lambda))$ and

$$N(\underline{p}_i) = p^{\deg(g_i)}.$$

In particular, for primes p not dividing the index, the prime ideals of norm p are precisely the ideals of the form $(p, \lambda - k)$, where k is a zero of f mod p, i.e. $f(k) \equiv 0 \bmod p$. For factorizing f mod p we remark that for odd primes p not dividing $\Delta(f)$ the number of irreducible factors of f mod p has the same parity as n if and only if $\Delta(f)$ is a quadratic residue mod p, see [SW]. For $n \leq 5$ this determines the decomposition type of f mod p completely if the number of zeros of f mod p is known. For further information we refer to [LE].

If p is a prime dividing the index we have to factorize f in $\mathbb{Z}_p[X]$, where

$$\mathbb{Z}_p := \lim_{\leftarrow m} (\mathbb{Z}/p^m\mathbb{Z}).$$

This means that we have to consider f modulo a power of p instead of modulo p itself. It can be shown that, for $p^k \| \Delta(f)$, the decomposition type of f in $\mathbb{Z}_p[X]$ is the same as that of f mod p^{k+1}. Finding zeros of f in $\mathbb{Z}_p$ can be done by Newton's method, see [W]: let $a_0 \in \mathbb{Z}$, $p^s \| f(a_0)$, $p^t \| f'(a_0)$. Assume $s > 2t$, then the sequence $(a_i)_{i=0}^{\infty}$ defined by

$$a_{i+1} = a_i - (f(a_i)/f'(a_i)), \quad i \geq 0,$$

converges to a root $\alpha \in \mathbb{Z}_p$ of f satisfying

$$\alpha \equiv a_0 \bmod p^{s-2t}.$$

The condition $s > 2t$ can always be satisfied by choosing a_0 to be a zero of f modulo a sufficiently high power of p.

If $f = \prod_{i=1}^{u} g_i$ is the decomposition of f into irreducible polynomials in $\mathbb{Z}_p[X]$, then $p\mathcal{O}(K)$ decomposes as $\prod_{i=1}^{u} \underline{p}_i^{e_i}$, where $e_i f_i = \deg(g_i)$ for $i = 1,\ldots,u$. Here f_i denotes the *residue class degree* of $\underline{p}_i$, defined by

$$N(\underline{p}_i) = p^{f_i}.$$

If p is unramified we have $e_i = 1$ and $f_i = \deg(g_i)$ for $i = 1,\ldots,u$. If p is ramified or the ramification behaviour of p is not known, one has to study the ramification behaviour of the corresponding extensions of $\mathbb{Q}_p$ in order to determine the e_i and f_i.

If for all primes $p \leq B^{(1/i)}$ the prime ideals of norm p^i have been found for $i = 1,2,\ldots$, as described above, we know all prime ideals of norm less than B. The ideal classes of these primes generate the class group. We now have to find relations among these generators. These are obtained by factorizing principal ideals; if $\mu\mathcal{O}(K) = \Pi\underline{p}^{a(\underline{p})}$ for $\mu \in K$, one has the relation $\Pi[\underline{p}]^{a(\underline{p})} = 1$ in the class group.

For factorization of the principal ideal $\mu\mathcal{O}(K)$, $\mu \in \mathcal{O}(K)$, first remark that if $\mu\mathcal{O}(K) = \Pi\underline{p}^{a(\underline{p})}$ then $|N(\mu)| = \Pi(N(\underline{p}))^{a(\underline{p})}$. Let $\mu = h(\lambda)$, with $h \in \mathbb{Q}[X]$; then $N(\mu) = R(f,h)$ by (3), so $N(\mu)$ can be computed by the method of 2.1. In particular, we have

$$N(a-b\lambda) = b^n \cdot f(a/b)$$

for $a,b \in \mathbb{Z}$, $b \neq 0$. If $h \in \mathbb{Z}[X]$ and $p|N(\mu)$ for a prime p not dividing the index, then a prime ideal $(p,g(\lambda))$ divides $\mu\mathcal{O}(K)$ if and only if $(g \bmod p) \mid (h \bmod p)$. In particular, a prime $(p,\lambda-a)$ divides $\mu\mathcal{O}(K)$ if and only if $h(a) \equiv 0 \bmod p$. If several primes $\underline{p}$ above p divide $\mu\mathcal{O}(K)$, the exact power of $\underline{p}$ dividing $\mu\mathcal{O}(K)$ can usually be determined by remarking that $\underline{p}^k$ cannot divide $\mu\mathcal{O}(K)$ if $N(\underline{p})^k$ doesn't divide $N(\mu+\eta)$ for some $\eta \in \underline{p}^k$.

To generate many relations, one applies the above technique to $\mu = p$ for small primes p, and to $\mu = h(\lambda)$ for several $h \in \mathbb{Z}[X]$ with small degree and small coefficients. In particular $\pi = a - b\lambda$ may be used for small integers a and b. Also $\mu = a - b\lambda$ where a/b is close to a real zero of f may be a good choice; such a and b can be found with a continued fraction algorithm. If many relations have been generated, select a small set of prime ideals among which many relations have been found. In the cases that can be treated by hand with the help of a pocket calculator, usually no more than ten prime ideals and a few more relations are sufficient. Define G to be the free abelian group generated by these prime ideals, divided by the subgroup generated by the relations. This group is easily determined explicitly. There is a natural group homomorphism $\phi\colon G \to Cl(K)$. If G is finite, we may hope to prove that ϕ is an isomorphism.

To prove that ϕ is surjective we have to express all prime ideals of norm less than B in the generators of G, modulo principal ideals. If we do not yet have an expression for $(p,g(\lambda))$ among the relations already found, we can look for one by decomposing elements of $(p,g(\lambda)) = p\mathcal{O}(K) + g(\lambda)\mathcal{O}(K)$ of small norm. Theoretically a suitable expression can always be found if ϕ is surjective. However, if a suitable expression is hard to find it is better, in practice, to go on with other prime ideals. In this way the set of "expressed" prime ideals increases, and this makes it easier to deal with the difficult prime ideal. If such expressions have been found for all prime ideals of norm less than B then ϕ is surjective. If we do not succeed we have to change G by adding a generator.

3.3. To prove that ϕ is injective we do the following. If ϕ is not injective then there exists a prime p and $x \in \ker\phi$ such that the order of x is p. For a fixed p define

$$H := \{x \in G \mid x^p = e\};$$

if $|H| = p^t$ choose generators $g_1,\ldots,g_t$ for H and integral ideals $\underline{a}_1,\ldots,\underline{a}_t$ such that

$$\phi(g_i) = [\underline{a}_i], \quad i = 1,\ldots,t.$$

To prove that ϕ is injective we have to show that no $(\ell_1,\ldots,\ell_t)$ exists, $\ell_i \in \mathbb{Z}$ and not all ℓ_i are divisible by p, such that

$$\prod_{i=1}^{t} \underline{a}_i^{\ell_i}$$

is a principal ideal, and this for all p dividing $|G|$. Define b_i to be a generator of the principal ideal $\underline{a}_i^p$, $i = 1,\ldots,t$. Assume

$$\prod_{i=1}^{t} \underline{a}_i^{\ell_i} = a\mathcal{O}(K),$$

not all ℓ_i divisible by p. Then

$$\left(\prod_{i=1}^{t} b_i^{\ell_i}\right)\mathcal{O}(K) = a^p\mathcal{O}(K),$$

hence $\prod_{i=1}^{t} b_i^{\ell_i}$ is a unit times the p-th power of an element of K. By Dirichlet's unit theorem there exists a set $\{u_1,\ldots,u_{r_1+r_2-1}\}$ of units, which is called a fundamental set of units, such that each unit u of $\mathcal{O}(K)$ can uniquely be written as

$$u = \zeta \cdot \prod_{i=1}^{r_1+r_2-1} u_i^{c_i}$$

for $c_i \in \mathbb{Z}$ and ζ is a root of unity contained in K. The set of roots of unity in K is easy to determine; most times it is $\{\pm 1\}$. Constructing a fundamental set of units will be discussed in 4; constructing a generating set of units modulo p-th powers is far easier, and is done as follows. Observe that by Dirichlet's theorem, $\mathcal{O}(K)^*/\mathcal{O}(K)^{*p}$ is an s-dimensional vector space over $\mathbb{F}_p$, where $s = r_1 + r_2$ if K contains a primitive p-th root of unity and $s = r_1 + r_2 - 1$ if not. If, in the procedure of 3.2, the same relation among prime ideals is found twice, then we have found two generators of the same principal ideal. Its quotient is then a unit. In this way we can generate as many units as we wish. Continue doing this until s units $u_1,\ldots,u_s$ have been found whose images in $\mathcal{O}(K)^*/\mathcal{O}(K)^{*p}$ are linearly independent over $\mathbb{F}_p$; this is checked by the method described below, with t = 0. Then $u_1,\ldots,u_s$ generate $\mathcal{O}(K)^*$ modulo p-th powers.

To prove the injectivity of ϕ, we now have to derive a contradiction from the hypothesis that there exist integers $k_1,\ldots,k_s$, $\ell_1,\ldots,\ell_t$, not all divisible by p, such that

$$\prod_{j=1}^{s} u_j^{k_j} \cdot \prod_{i=1}^{t} b_i^{\ell_i} \quad \text{is a p-th power.} \tag{13}$$

We may regard $k_1,\ldots,k_s$, $\ell_1,\ldots,\ell_t$ as elements of $\mathbb{F}_p$. Let $\underline{q}$ be a prime ideal not containing any of the b_i, for which $N(\underline{q}) \equiv 1 \bmod p$. Then (13) taken modulo $\underline{q}$, gives rise to a linear relation among the k_j and ℓ_i over $\mathbb{F}_p$, since p divides the order of $(\mathcal{O}(K)/\underline{q})^*$. Similar relations can be found by reducing (13) modulo a power of prime ideal above p. Further, if p = 2, such relations can also be found by looking at signs in (13) at real embeddings of K. Continue finding linear relations among the k_j and ℓ_i until s + t of them are independent, then all k_j and ℓ_i are zero, which is the required contradiction.

If we do not succeed in finding s + t independent relations then probably ϕ is not injective and we have to find another relation among prime ideals and define G anew. It can be proved, that if ϕ is injective, then s + t

independent relations can always be found using prime ideals $\underline{q}$ as above.

3.4. We give a simple example of computing a class group using the technique described in 3.2 and 3.3. Let K be given as $K = \mathbb{Q}(\lambda)$ where λ is a zero of $f(X) = X^3 + 2X^2 - 8X + 1$. One has $\Delta(f) = 1957 = 19.103$; since $\Delta(f)$ is square free it follows from (8) that $\Delta(f) = \Delta(K)$. Since f has 3 real zeros we have $r_1 = 3$, $r_2 = 0$ and by (12) we may choose $B = (2/9)\sqrt{|\Delta(K)|} < 10$. Modulo 2 the Modulo 2 the polynomial f decomposes as $(X+1)(X^2 + X + 1)$, modulo 3 and modulo 7 it is irreducible because none of the values $f(0)$, $f(\pm 1)$, $f(\pm 2)$, $f(\pm 3)$ are divisible by 3 or by 7. Modulo 5 the decomposition of f is $(X+1)(X^2+X+1)$. Hence there are only 3 prime ideals of norm less than ten: $(2,\lambda+1)$, $(2,\lambda^2+\lambda+1)$ and $(5,\lambda+1)$. Writing the ideal group additively we get the following relations among these three generators of the class group (each row is a relation):

principal ideal	$(2,\lambda+1)$	$(2,\lambda^2+\lambda+1)$	$(5,\lambda+1)$
(2)	1	1	0
$(\lambda+1)$	1	0	1
$(\lambda-1)$	2	0	0

From this table we see that the ideal $(2,\lambda+1)$ is a generator of $Cl(K)$ and that its order in $Cl(K)$ is at most 2; hence $Cl(K) \cong \mathbb{Z}/2\mathbb{Z}$ or $Cl(K) = \{1\}$. Assume $Cl(K) = \{1\}$. Then $(2,\lambda+1) = \beta O(K)$ for some $\beta \in O(K)$ and since $(\lambda-1) = (2,\lambda+1)^2$ some unit u will exist such that $u(\lambda-1) = \beta^2$. Since $r_1 = 3$, and ± 1 are the only roots of unity in K, the set of units is as a group isomorphic to $(\mathbb{Z}/2\mathbb{Z}) \oplus \mathbb{Z}^2$, hence $O(K)^*/O(K)^{*2}$ has dimension 3 over $\mathbb{F}_2$. We immediately discover the units -1, λ and $\lambda + 4$, and we wonder if they generate all units modulo squares. In the next table we write 0 if an element is a square modulo a prime, or positive, and 1 if it is not. For example, $\lambda + 4$ is not a square $\bmod(5,\lambda-4)$, because $3 \equiv \lambda + 4 \bmod(5,\lambda-4)$ and 3 is not a square modulo 5.

	$(5,\lambda-4)$	$(11,\lambda-3)$	$\lambda = -4.0410$	$\lambda = 0.1295$	$\lambda = 1.9115$
-1	0	1	1	1	1
$\lambda+4$	1	1	1	0	0
λ	0	0	1	0	0
$\lambda-1$	1	1	1	1	0

Since the first three rows are linearly independent over $\mathbb{F}_2$ our three units generate all units modulo squares. Since all rows are independent over $\mathbb{F}_2$ there is no unit u such that $u(\lambda-1)$ is a square, hence $Cl(K) \neq \{1\}$, hence $Cl(K) \cong \mathbb{Z}/2\mathbb{Z}$.

3.5. In this section we spend a few words on other techniques helpful for the determination of the class number of a number field. ODLYZKO [OD] found a set of universal constants A, B and E such that for each number field K:

$$|\Delta(K)| > A^{r_1} \cdot B^{2r_2} \cdot e^{-E}. \tag{14}$$

From class field theory we know that each number field K has a maximal abelian totally unramified extension H(K), the *Hilbert class field* of K and that the Galois group of H(K)/K is isomorphic to $Cl(K)$. Applying (14) to H(K) gives

$$h(K) < E(r_1 \log A + 2r_2 \log B - \log\Delta(K))^{-1} \tag{15}$$

if the right hand side of (15) is positive; i.e. if $\Delta(K)$ is not too large, then we have an upperbound for h(K).

If we can construct an abelian extension M/K which is totally unramified, the field M has to be contained in H(K) and the Galois group of M/K has to be a factor group of $Cl(K)$ and we have a divisor of h(K).

Until now all methods can be used for each number field K, which in general will not be Galois over $\mathbb{Q}$ or over another non-trivial subfield. If K/K' is Galois for some non-trivial subfield K' of K, then its Galois group has a natural action on $Cl(K)$, from which a lot of restrictions on $Cl(K)$ can be obtained. Together with (15) and some more class field theory this leads to techniques of determination of class groups of abelian extensions of $\mathbb{Q}$, see [MA]. Finally, we mention the analytic class number formula (17), see 4.4, which relates the class number and the number of units to the value of some analytic function.

4. UNITS

4.1. In 3.3 we described a procedure to find a generating set of units of K modulo p-th powers for a given prime number p. Now we want to find a generating set of *all* units. As an element of $O(K)$ each unit u corresponds to

an element $(u_1,\ldots,u_{r_1},u_{r_1+1},\ldots,u_{r_1+r_2})$ of $\mathbb{R}^{r_1} \times C^{r_2}$. Taking the logarithm we let the unit u correspond to

$$(\log|u_1|,\log|u_2|,\ldots,\log|u_{r_1}|,2\log|u_{r_1+1}|,\ldots,2\log|u_{r_1+r_2}|) \in \mathbb{R}^{r_1+r_2}.$$

Since u has norm ± 1, the sum of all these coordinates is zero and we may omit one coordinate, say the last one, without loss of information. We will identify this element of $\mathbb{R}^{r_1+r_2-1}$ with the unit u modulo roots of unity. According to Dirichlet's unit theorem the set of all units modulo roots of unity is a lattice in $\mathbb{R}^{r_1+r_2-1}$. The *regulator* of K is defined to be the determinant of this lattice; a basis of this lattice is called a fundamental set of units. As mentioned in 3.3 one can find as many units as one wants; now find a set of units large enough to generate a lattice L in $\mathbb{R}^{r_1+r_2-1}$. If L is not the whole lattice of units modulo roots of unity, another unit has to be contained in some bounded set S in $\mathbb{R}^{r_1+r_2-1}$. Such a bounded set can be pulled back to a bounded set $\tilde{S}$ in $\mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$. Since $\mathcal{O}(K)$ is a lattice in $\mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$ only a finite number of elements of $\mathcal{O}(K)$ is contained in $\tilde{S}$. The units among this finite number of elements generate together with the units already found, the whole set of units.

4.2. Next a few words on the choice of S for which the technique of 4.1 works. Assume we know from the technique of 3.3 that, for some q, the lattice L contains all units modulo p-th powers for all primes $p < q$. Then the index of L in the lattice of all units is only divisible by primes larger than or equal to q. Choose a basis $\{\vec{a}_i \mid i = 1,\ldots,r_1+r_2-1\}$ of L; for computations it will be pleasant if $\|\vec{a}_i\|$ is not too large. If L is not equal to the lattice of units, then

$$S := \{ \sum_{i=1}^{r_1+r_2-1} \mu_i \vec{a}_i \mid \mu_i \in \mathbb{R},\ |\mu_1| \le 1/q;\ |\mu_i| \le 1/2,\ i = 2,\ldots,r_1+r_2-1\}$$

contains a unit not contained in L. Another choice of S with the same property is

$$S := \{ \sum_{i=1}^{r_1+r_2-1} \mu_i \vec{a}_i \mid |\mu_i| \le b_i\}$$

with $b_i > 0$,

$$\prod_{i=1}^{r_1+r_2-1} b_i = q^{-1}.$$

Although the volume of the latter set is larger, it has the advantage that for large q for a good choice of b_i *all* coordinates become small instead of just *one*. If u is a unit in S not contained in L, then so is u^{-1}, and of one of them the sum of the coordinates is positive. Hence for $\tilde{S}$ we may choose:

$$\tilde{S} := \{(u_1,\ldots,u_{r_1},u_{r_1+1},\ldots,u_{r_1+r_2}) \in \mathbb{R}^{r_1} \times C^{r_2} \mid |u_{r_1+r_2}| \le 1,$$

$$(\log|u_1|,\ldots,\log|u_{r_1}|,\ 2\log|u_{r_1+1}|,\ldots,2\log|u_{r_1+r_2-1}|) \in S\}.$$

4.3. As an example we compute all units of the same field as in 3.4, $K = \mathbb{Q}(\lambda)$ where λ is a zero of $f(X) = X^3 + 2X^2 - 8X + 1$. In 3.4 we saw that -1, λ and $\lambda+4$ generate all units modulo squares and now we examine if in fact they generate *all* units. Two numerical values of λ are -4.0410 and 0.1295; thus $\lambda+4$ corresponds to $(-3.1950,\ 1.4181)$ and λ to $(1.3965,\ -2.0444)$. We choose $\vec{a}_1 = (1.3965,\ -2.0444)$ corresponding to λ and $\vec{a}_2 = (1.7985,\ 0.6263)$ corresponding to $(\lambda+4)^{-1}\lambda^{-1}$; now $S = \{\mu_1\vec{a}_1 + \mu_2\vec{a}_2 \mid |\mu_1| \le \frac{1}{3},\ |\mu_2| \le \frac{1}{2}\}$. We get

$$\tilde{S} = \{(x,y,z) \in \mathbb{R}^3 \mid |0.6263\ \log|x| - 1.7985\ \log|y|| \le 1.5172,$$

$$|2.0444\ \log|x| + 1.3965\ \log|y|| \le 2.2757,$$

$$|z| \le 1\}.$$

Notice that $\tilde{S} \subset B := \{(x,y,z) \in \mathbb{R}^3 \mid |x| \le 3.9147,\ |y| \le 2.7037,\ |z| \le 1\}$. A basis of $\mathcal{O}(K)$ is $\{1,\lambda,\lambda^2\}$, or written in coordinates

$$\{(1,1,1),\ (-4.0410,0.1295,1.9115),\ (16.3294,0.0168,3.6538)\}.$$

The intersection of B and $\mathcal{O}(K)$ is $\{0,\pm1\}$; this doesn't give new units, hence -1, λ and $\lambda+4$ generate all units of K.

4.4. The procedure of 4.1 and 4.2 is not the only way to compute the units of K. For example, there exist universal lower bounds for the regulator, see [Z2]. Let the determinant of a lattice L of units be less than k times such a lower bound. Assume we know from the technique of 3.3 that for all primes $p < k$ the lattice L contains all units modulo p-th powers. Then L contains all units. By this approach choosing a bounded set S and determining all integral elements in $\tilde{S}$ are avoided. We conclude by suggesting an

alternative method for computing the units of K.

By geometrical arguments one shows that the number A_t of integral ideals in K of norm less than t satisfies

$$A_t = \rho_K t + O(t^{(n-1)/n}), \quad t \to \infty, \tag{16}$$

where

$$\rho_K := 2^{r_1+r_2} \cdot \pi^{r_2} \cdot R(K) \cdot h(K) \cdot w(K)^{-1} \cdot (|\Delta(K)|)^{-1/2},$$

in which $R = R(K)$ is the regulator, $h(K)$ the class number and $w(K)$ the number of roots of unity in K. Define for $s > 1$:

$$\zeta_K(s) := \sum_{0 \neq \underline{a} \subset \mathcal{O}(K)} (N\underline{a})^{-s} = \sum_{n=1}^{\infty} (A_{n+1} - A_n) n^{-s}.$$

One derives from (16): $\lim_{s \downarrow 1} (s-1)\zeta_K(s) = \rho_K$, hence

$$\lim_{s \downarrow 1} \zeta_K(s)/\zeta_{\mathbb{Q}}(s) = \rho_K/\rho_{\mathbb{Q}} = \rho_K. \tag{17}$$

By decomposing the ideals in the definition of ζ_K into prime ideals we obtain the Euler product

$$\zeta_K(s) = \prod_{p \subset \mathcal{O}(K) \text{ prime}} (1 - (N(\underline{p}))^{-s})^{-1}.$$

Hence:

$$\rho_K = \lim_{s \downarrow 1} \zeta_K(s)/\zeta_{\mathbb{Q}}(s) = \lim_{s \downarrow 1} \prod_{p \text{ prime}} ((1-p^{-s}) \prod_{\underline{p} \mid p} (1 - (N(\underline{p}))^{-s})^{-1})$$

It can be shown that $s = 1$ simply may be substituted in the right hand side, i.e.

$$\rho_K = \prod_{p \text{ prime}} a_p, \tag{18}$$

where

$$a_p := (1-p^{-1}) \prod_{\underline{p} \mid p} (1 - N(\underline{p}))^{-1})^{-1}.$$

The values of a_p follow immediately from the decomposition types of $p\mathcal{O}(K)$ which were discussed in 3.2. As in 4.2, let L denote the lattice of units containing all units modulo p-th powers for primes $p < q$. Denote R' to be the determinant of L, then we know that $R'/R \in \mathbb{Z}$, and if $R \neq R'$ then $R'/R \geq q$. Assume we know h(K), then

$$\rho_K' := 2^{r_1+r_2} \cdot \pi^{r_2} R' \cdot h(K) \cdot w(K)^{-1} \cdot |\Delta(K)|^{-1/2}$$

equals $\prod_{p \text{ prime}} a_p$ or is at least q times larger. The problem now is how fast $\prod_{p \text{ prime}} a_p$ converges; it doesn't converge absolutely. We would like to have

$$(19) \qquad \left| \sum_{p>x} \log a_p \right| < F(x)$$

for some explicit function F satisfying $\lim_{x\to\infty} F(x) = 0$. Denote G to be the Galois group of the normal closure of K over $\mathbb{Q}$; then G is a transitive subgroup of S_n and is discussed in [LI]. There is a theorem that states that for each $(b_1,\ldots,b_t)$, $t,b_1,\ldots,b_t \in \mathbb{Z}$, > 0, $\Sigma_{i=1}^{t} b_i = n$:

$$\lim_{x\to\infty} \frac{|\{p \text{ prime} < x \mid f \bmod p = \prod_{i=1}^{t} f_{i,p}, f_{i,p} \text{ irreducible mod } p, \deg f_{i,p} = b_i\}|}{|\{p \text{ prime} < x\}|}$$

$$= \frac{|\{g \in G \mid g \text{ splits into } t \text{ disjoint cycles of order } b_1,\ldots,b_t\}|}{|G|}$$

With a slightly weaker notion of density this theorem was already proved by FROBENIUS, see [F]. A stronger version of Frobenius' theorem, known as Chebotarev's density theorem, states an analogous result on conjugacy classes of Frobenius symbols instead of cycle types, see [LA]. Assuming certain generalized Riemann hypotheses one can prove an effective version of Chebotarev's theorem, for results see [OE]. From this effective theorem an explicit function F satisfying (19) can be derived, of order $x^{-1/2} \log x$. Having such a function F, for proving $R = R'$ it is sufficient to compute values of a_p until

$$\log \rho_K' - \sum_{p<x} \log a_p + F(x) < \log q.$$

It's likely that much sharper results can be obtained than those deduced from [OE]. Although this might possibly lead to an efficient way to determine the regulator, no results have been published yet. The convergence of $\prod_{p<x} a_p$ is illustrated by the next table for the field given by $f(X) = X^3 + 2X^2 - 8X + 1$, where p_i denotes the i-th prime number:

x	$\prod_{p<x} a_p$	x	$\prod_{p<x} a_p$
$p_5 = 11$	0.8267	$p_{25} = 97$	0.8268
$p_{10} = 29$	0.9557	$p_{30} = 113$	0.8507
$p_{15} = 47$	0.9021	$p_{35} = 149$	0.8403
$p_{20} = 71$	0.8582	$p_{40} = 173$	0.8241

The value of ρ_K is 0.8231.

REFERENCES

[A] ANGELL, I.O., *A table of complex cubic fields*, Bull. London Math. Soc., 5 (1973), 37-38; *A table of totally real cubic fields*, Mathematics of Computation, 30, no. 133 (1976), 184-187.

[B] BRENTJES, A.J., *Multi-dimensional continued fraction algorithms*, this volume.

[C] CASSELS, J.W.S., *An introduction to the geometry of numbers*, Springer (1959), Chapter II, Sections 3.1, 4.1, 5.1.

[D] DEDEKIND, R., *Gesammelte Mathematische Werke Band I*, (1930), Chap. XV, §3.

[F] FROBENIUS, G.F., *Über Beziehungen zwischen den Primidealen eines algebraischen Körpers und den Substitutionen seiner Gruppe*, (1896), in Gesammelte Abhandlungen, Springer (1968), Band II, Chapter 52.

[G] GYÖRY, K., *Sur les polynômes à coefficients entiers et de discriminant donné*, Part 1 in Acta Arith. 23 (1973), 419-426, Part 2 in Publicationes Mathematicae Debrecen 21 (1974), 125-144.

[H] HANCOCK, H., *Foundations of the theory of algebraic numbers*, Vol. II, MacMillan (1932), Chap. VI.

[LA] LANG, S., *Algebraic number theory*, Addison-Wesley (1970), Chap. VIII, §4.

[LE] LENSTRA, A.K. *Factorization of polynomials*, this volume.

[LI] LINDEN, F.J. VAN DER, *The computation of Galois groups*, this tract.

[L.M.O.] LAGARIAS, J.C., H.L. MONTGOMERY & A.M. ODLYZKO, *A bound for the least prime ideal in the Chebotarev density theorem*, Inventiones Mathematicae 54 (1979), 271-296, Corollary 1.3.

[MA] MASLEY, J.M., *Class numbers of real cyclic number fields with small conductor*, Compositio Mathematica, 37, Fasc. 3 (1978), 297-319.

[OD] ODLYZKO, A.M., *Lower bounds for discrimants of number fields*, Part 1, Acta Arithmetica 29 (1976) 275-297; Part 2, Tohoku Math. J. 29 (1977) 209-216.

[OE] OESTERLÉ, J., *Versions effectives du théorème de Chebotarev sous l'hypothèse de Riemann généralisée*, Astérisque 61 (1979), 165-167.

[P] POITOU, G., *Minorations de discriminants*, Séminaire Bourbaki, février 1976, no. 479.

[SA] SAMUEL, P., *Théorie algébrique des nombres*, Hermann Paris 1967.

[SCH] SCHOOF, R.J., *Quadratic number fields and factorization*, this volume.

[SE] SERRE, J.P., *Corps locaux*, Hermann Paris 1962, Chapter IV, §1.

[SW] SWAN, R.G., *Factorization of polynomials over finite fields*, Pacific Journal of Mathematics 12 (1962), 1099-1106, Corollary 1.

[U] UCHIDA, K., *When is* $\mathbb{Z}[\alpha]$ *the ring of integers?*, Osaka Journal of Mathematics 14 (1977), 155-157.

[W] WEISS, E., *Algebraic number theory*, McGraw-Hill (1963), Chapter 3.1.2.

[Z1] ZIMMER, H.G., *Computational Problems, Methods, and Results in Algebraic Number Theory*, Springer L.N. in M. 262 (1972), Chap. 6(b).

[Z2] ZIMMERT, R., *Ideale kleiner Norm in Idealklassen und eine Regulatorabschätzung*, Inventiones Mathematicae 62 (1981), 367-380.

QUADRATIC FIELDS AND FACTORIZATION

by

R.J. SCHOOF

1. INTRODUCTION

Let K be an algebraic number field and $\mathcal{O} = \mathcal{O}(K)$ its ring of integers. We recall a few basic definitions and facts concerning algebraic number fields.

By I(K) we denote the group of fractional $\mathcal{O}$-ideals and by P(K) its subgroup of principal fractional $\mathcal{O}$-ideals, which is a subgroup of I(K). The class group $C\ell(K)$ of K is defined by

$$C\ell(K) = I(K)/P(K).$$

The class group is a finite abelian group and its order is denoted by h(K), the class number of K. By $\mathcal{O}^\times$ we denote the multiplicative group of units of $\mathcal{O}$. The structure of $\mathcal{O}^\times$ as an abelian group, is given by Dirichlet's Unit Theorem:

THEOREM 1.1. *Let* K *be an algebraic number field and* $\mathcal{O}$ *its ring of integers, then*

$$\mathcal{O}^\times \simeq \mu(\mathcal{O}) \oplus \mathbb{Z}^{r_1+r_2-1}.$$

Here $\mu(\mathcal{O})$ denotes the finite group of roots of unity in K,

r_1 = number of embeddings $K \hookrightarrow \mathbb{R}$,

r_2 = half the number of embeddings $K \hookrightarrow \mathbb{C}$ with $\mathrm{im}(K) \not\subset \mathbb{R}$.

It holds that $r_1 + 2r_2 = n = [K:\mathbb{Q}]$, the (absolute) degree of K. For these and more definitions and facts from algebraic number theory see for instance [17].

In general, it is hard to determine the class group of a number field, which, for instance, is given by a generator; for this general problem see ZANTEMA's talk [13]. Here we shall concentrate on fields with small degrees; in this case, the rings of integers do not contain too many units and the computation of the class group is relatively easy. It appears to be possible to determine the class group of fields with small degrees, which have very large discriminants.

First we consider *complex quadratic number fields*. A field K is called complex quadratic if $[K:\mathbb{Q}] = 2$ and if $r_1 = 0$, $r_2 = 1$; it follows from Dirichlet's Unit Theorem that $\mathcal{O}(K)$ contains only finitely many units.

The study of the class groups of these fields is a very old one; it was initiated by Gauss, in the beginning of the 19th century [12]. Gauss studied the problem in the language of "binary quadratic forms" and he made extensive lists of class groups of complex quadratic fields. In Section 2 we shall discuss the complex quadratic fields in more detail; it appears that for our purposes, it is useful to formulate matters in the old-fashioned terms of binary quadratic forms again. An algorithm, due to D. SHANKS [31], to compute class groups of complex quadratic fields will be treated in Section 3.

Next we consider *real quadratic number fields* i.e. fields of degree 2 with $r_1 = 2$ and $r_2 = 0$. For a real quadratic field K, Dirichlet's Unit Theorem boils down to

$$\mathcal{O}(K)^\times \cong \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}.$$

The determination of the class group of a real quadratic field cannot go, it seems, without the determination of the group of units; the latter is equivalent to finding a unit ε in $\mathcal{O}^\times$ such that $\mathcal{O}^\times$ is generated by ε and -1, which in turn is easily seen to be equivalent to solving a so-called Pellian equation, a problem which dates back to Fermat. The study of class groups of real quadratic fields was also begun by Gauss, who studied the subject in terms of binary quadratic forms.

In Section 4 we will discuss the structure of the class groups and unit groups of real quadratic fields in more detail, here some new ideas of LENSTRA and SHANKS come in [18,31], which give rise to a new, fast algorithm to determine the class group and, in some sense, the size of the group of units of a real quadratic field. We will describe this algorithm in Section 5; it is closely related to Shanks', discussed in Section 3, but slightly more complicated.

Complex cubic fields are fields of degree 3 over $\mathbb{Q}$ with $r_1 = 1$ and $r_2 = 1$. Complex cubic fields are not Galois extensions of $\mathbb{Q}$; the structure of their groups of units is the same as for real quadratic fields: if K is a complex cubic fields we have that

$$\mathcal{O}(K)^\times \simeq \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}.$$

We do not discuss this class of fields; an algorithm to compute the class groups and the groups of units of these fields is being developed by WILLIAMS and SCHMID [40], their algorithm is along the same lines as the algorithm for real quadratic fields discussed in Section 5.

In Section 6 we point out how the algorithms, discussed in Sections 3 and 5 may be used to factor the discriminant of the number fields under consideration. We will in fact, describe two deterministic factorization algorithms which, on assumption of certain generalized Riemann hypotheses, factor an integer N in time, bounded by $N^{1/5+\varepsilon}$ for all $\varepsilon > 0$. For a discussion of related algorithms see [13,23].

The algorithms discussed are suitable to compute the class groups and units of quadratic fields that have *very* large discriminants. In fact, determining the class group and units of quadratic fields with discriminants of fewer than, say, 6 decimal digits, may be done faster by simpler and more direct methods. Therefore, in application of these algorithms, one should think of discriminants of 10 to 30 decimal digits.

Using these algorithms, one can practice a kind of experimental mathematics; it seems to be generally believed, that every finite abelian group occurs as a subgroup of the class group of some, say, complex quadratic field but theoretical results on this question are very scarce indeed. By means of these algorithms, however, one is able to compute class groups of quadratic fields with very large discriminants, and, guided by heuristic, one can search for explicit examples of quadratic fields that have unusual subgroups of their class groups. Only recently, some progress in this direction has been made. Some old and new results will be discussed in Section 7.

Finally, in Section 8, we will give a few details on the actual implementation of the algorithms on the SARA CDC-Cyber 170-750 computer.

2. CLASS GROUPS OF COMPLEX QUADRATIC NUMBER FIELDS

It is well known that the discriminant of complex quadratic number fields are negative integers, congruent to 0 or 1 (mod 4). Furthermore, complex quadratic fields are characterized by their discriminants, but it is not true that every negative integer $\equiv 0$ or 1 (mod 4) is the discriminant of some complex quadratic number field.

However, every $\Delta \in \mathbb{Z}_{<0}$, $\Delta \equiv 0$ or 1 (mod 4), can in one and only one way be written as $\Delta = f^2 D$, where D is the discriminant of a complex quadratic number field K, and $f \in \mathbb{Z}_{\geq 1}$. Now, $\Delta = \Delta(\mathcal{O})$ is the discriminant of the unique subring $\mathcal{O}$ of index f in $\mathcal{O}(K)$: the unique *quadratic order* of discriminant Δ.

So for every $\Delta \in \mathbb{Z}_{<0}$, $\Delta \equiv 0$ or 1 (mod 4), there exists a unique *complex quadratic order* $\mathcal{O} = \mathcal{O}(\Delta)$, with discriminant Δ, contained in the ring of integers of some complex quadratic number field. Rings of integers themselves are also called *maximal orders*. It is also possible to define the notion of class group for non-maximal orders:

Let $\mathcal{O}$ be a complex quadratic order, contained in a complex quadratic field K. By definition, a fractional $\mathcal{O}$-ideal M is a non-zero finitely generated $\mathcal{O}$-submodule of K, and M is called *invertible* if there is a fractional $\mathcal{O}$-ideal $N \subset K$ such that $MN = \mathcal{O}$. By $I(\mathcal{O})$ we denote the *group* of invertible fractional $\mathcal{O}$-ideals and by $P(\mathcal{O})$ the group of principal fractional ideals, a subgroup of $I(\mathcal{O})$. The class group of $\mathcal{O}$ is denoted by $Cl(\mathcal{O})$ and defined by $Cl(\mathcal{O}) = I(\mathcal{O})/P(\mathcal{O})$. The group $Cl(\mathcal{O})$ is finite abelian and its order will be denoted by $h(\mathcal{O})$, the class number of $\mathcal{O}$.

REMARK. If $\mathcal{O}$ is a *maximal* complex quadratic order, i.e. the ring of integers of some complex quadratic field, then $\mathcal{O}$ is a Dedekind ring and *all* fractional $\mathcal{O}$-ideals are invertible.

EXERCISE. Let $\mathcal{O}$ be a complex quadratic order, and M a fractional $\mathcal{O}$-ideal; then M is invertible iff $\{\alpha \in K \mid \alpha M \subset M\} = \mathcal{O}$.

The ring $\{\alpha \in K \mid \alpha M \subset M\}$ is called "the ring of coefficients of M", cf. [1].

For definitions, notations, terminology and facts on complex quadratic orders see [1]. Next we will discuss the correspondence between ideal classes of $\mathcal{O}$ and primitive positive definite binary quadratic forms of discriminant $\Delta(\mathcal{O})$.

DEFINITION 2.1. A polynomial $f = aX^2 + bXY + cY^2 \in \mathbb{Z}[X,Y]$ with $b^2 - 4ac = \Delta$ is called a *binary quadratic form* of discriminant Δ. A binary quadratic form $f = aX^2 + bXY + cY^2$ is called *positive definite* if $\Delta < 0$ and $a > 0$, and is called *primitive* if $\gcd(a,b,c) = 1$.

We will often denote a form $aX^2 + bXY + cY^2$ by (a,b,c), or even (a,b) since c is determined by $b^2 - 4ac = \Delta$.

DEFINITION 2.2. Let $f = aX^2 + bXY + cY^2$ and $g = a'X^2 + b'XY + c'Y^2$ be positive definite binary quadratic forms. We shall call f and g *equivalent* if there is a $\sigma = \begin{pmatrix} \alpha & \gamma \\ \beta & \delta \end{pmatrix} \in SL_2(\mathbb{Z})$ such that

$$a'U^2 + b'UV + c'V^2 = aX^2 + bXY + cY^2;$$

where $U = \alpha X + \gamma Y$ and $V = \beta X + \delta Y$.

Since $SL_2(\mathbb{Z})$ is a group, "equivalence" is indeed an equivalence relation.

THEOREM 2.3. *Let $\mathcal{O}$ be a complex quadratic order with discriminant Δ. There is a 1-1 correspondence between classes of invertible fractional $\mathcal{O}$-ideals and equivalence classes of primitive positive definite binary quadratic forms of discriminant Δ.*

PROOF. Let M be a primitive invertible fractional $\mathcal{O}$-ideal i.e. a non-zero $\mathcal{O}$-submodule of K with its ring of coefficients equal to $\mathcal{O}$, we shall attach a primitive positive definite form f to M.

The fractional ideal M is a free $\mathbb{Z}$-module of rank 2 in K, i.e. a two dimensional lattice in $K \hookrightarrow \mathbb{C}$. We can attach a quadratic form to M in the following way:

Let $\{\alpha,\beta\}$ be an oriented $\mathbb{Z}$-basis for M (i.e. $\mathrm{Im}(\beta/\alpha) > 0$) and take

$$f = \frac{N(\alpha X+\beta Y)}{N(M)} .$$

(Here N denotes the *norm*; for definitions and properties of the norm see [1].)

The choice of the basis $\{\alpha,\beta\}$ does not affect the $SL_2(\mathbb{Z})$ class of f; it can be proved that f is of discriminant Δ, and that f is primitive if M is invertible. However, for future purposes, we prefer to give another

construction of the form f.

Let M denote a fractional $\mathcal{O}$-ideal; the ideal class, represented by M contains the ideals βM, with $\beta \in K^{\times}$, so we can find an $\mathcal{O}$-submodule of K, equivalent to M and of the form $\mathbb{Z} + \mathbb{Z}\alpha$, $\alpha \in K$; This module will be denoted by M again.

We can always choose α in the upper half plane and under this condition α is unique up to $SL_2(\mathbb{Z})$-action.

Next, let's exploit the fact that M is an $\mathcal{O}$-module: assume Δ is even, then $\{1, \frac{1}{2}\sqrt{\Delta}\}$ is a $\mathbb{Z}$-basis for $\mathcal{O}$ and $\frac{1}{2}\sqrt{\Delta} \cdot M \subset M$:

$$\frac{1}{2}\sqrt{\Delta} \in M \to \frac{1}{2}\sqrt{\Delta} = -\frac{1}{2}b + \alpha \cdot a \qquad (-\tfrac{1}{2}b, a \in \mathbb{Z})$$

$$\to \alpha = \frac{b+\sqrt{\Delta}}{2a}$$

with $a > 0$ since α is in the upper half plane;

$$\frac{1}{2}\sqrt{\Delta}\cdot\alpha \in M \to \frac{1}{2}\sqrt{\Delta}\cdot\alpha = c + \alpha \cdot d \qquad (c,d \in \mathbb{Z})$$

which, combined with the fact, that

$$\alpha = \frac{b+\sqrt{\Delta}}{2a},$$

gives us that

$$\frac{\Delta - b^2}{4a} = c \in \mathbb{Z}.$$

We conclude that $M = \mathbb{Z} + \mathbb{Z}\,\frac{b+\sqrt{\Delta}}{2a}$ with $a > 0$ and $c \in \mathbb{Z}$ such that $b^2 - 4ac = \Delta$. If Δ is odd, $\{1, \frac{1}{2}(1+\sqrt{\Delta})\}$ is a $\mathbb{Z}$-basis for $\mathcal{O}$, and a completely analogous proof gives exactly the same result.

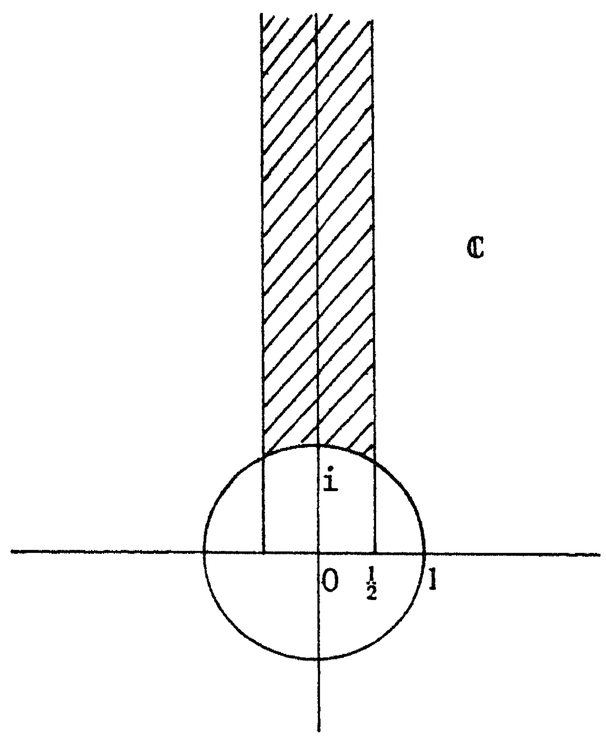
To the ideal class represented by M we associate the positive definite quadratic form $f = aX^2 + bXY + cY^2$. It remains to check that this association respects equivalence and that f is primitive if M is invertible; this is straightforward and left to the reader, see [1]. Next let $f = aX^2 + bXY + cY^2$ be a primitive binary quadratic form of discriminant Δ with $a > 0$. To f we associate the ideal class represented by $M = \mathbb{Z} + \frac{b+\sqrt{\Delta}}{2a}\mathbb{Z}$; since f is primitive, M is invertible and the association is correctly defined with respect to equivalence. This completes the proof of Theorem 2.3. □

Let $\mathcal{O}$ be a complex quadratic order with discriminant Δ, and let k be an $\mathcal{O}$-ideal class; then k consists of fractional ideals $(\mathbb{Z} + b+\sqrt{\Delta}/2a\ \mathbb{Z}) \cdot \beta$, $\beta \in K^{\times}$, and to k is associated the quadratic form $aX^2 + bXY + cY^2$. The integers a and b are unique up to $SL_2(\mathbb{Z})$-action on $\mathbb{Z} + b+\sqrt{\Delta}/2a\ \mathbb{Z}$, so it is always possible to choose a and b such, that the number $b+\sqrt{\Delta}/2a$ is in the standard fundamental domain of $SL_2(\mathbb{Z})$, acting on the upper half plane. This choice gives the following conditions on a, b and c:

$$|\frac{b}{2a}| \leq \frac{1}{2} \quad \text{and} \quad |\frac{b+\sqrt{\Delta}}{2a}| \geq 1$$

i.e.

$$|b| \leq a \leq c.$$



DEFINITION. A binary quadratic form $f = aX^2 + bXY + cY^2$ is called *reduced* if $|b| \leq a \leq c$.

It is obvious that Theorem 2.3 can also be stated in the following form:

THEOREM 2.3'. *Let $\mathcal{O}$ be a complex quadratic order of discriminant Δ. The classes of invertible fractional $\mathcal{O}$-ideals are in 1-1 correspondence with the reduced primitive definite binary quadratic form of discriminant Δ.*

CONVENTION! We will always identify reduced forms (a,b,c) and (a,-b,c), whenever $|b| = a$ or $a = c$. These forms correspond to ideal classes represented by $\mathbb{Z} + \mathbb{Z}\cdot\alpha$, with α on the boundary of the fundamental domain.

It is easily seen, that the conditions $b^2 - 4ac = \Delta$ and $|b| \leq a \leq c$ imply that $a \leq \sqrt{|\Delta|/3}$ and this shows that the class group of $\mathcal{O}$ is finite. By means of the dictionary between ideal classes and quadratic forms, the problem of counting the ideal-classes of a given quadratic order is reduced to a finite problem.

Next, we transport the natural group structure of the group of ideal classes to the finite set of reduced binary quadratic forms.

If $f = aX^2 + bXY + cY^2 = (a,b,c)$ is a primitive positive definite form of discriminant Δ, then the ideal class associated to f consists of ideals $M = (\mathbb{Z} + b+\sqrt{\Delta}/2a\ \mathbb{Z}) \cdot \alpha$, $\alpha \in K^{\times}$; the number α is a so-called primitive point of M, i.e. for all $n \geq 2$ in $\mathbb{Z}$ we have $\alpha/n \notin M$.

Let (a_1,b_1,c_1) and (a_2,b_2,c_2) be two primitive positive definite quadratic forms of discriminant Δ. Let M and N be two fractional ideals in the ideal classes associated to them:

$$M = (\mathbb{Z} + \frac{b_1+\sqrt{\Delta}}{2a_1}\ \mathbb{Z})\alpha \quad \text{and} \quad N = (\mathbb{Z} + \frac{b_2+\sqrt{\Delta}}{2a_2}\ \mathbb{Z})\beta\ .$$

Put

$$MN = (\mathbb{Z} + \frac{b_3+\sqrt{\Delta}}{2a_3}\ \mathbb{Z})\ \gamma$$

where we choose γ such that $\alpha\beta \in \gamma\mathbb{Z}$, say $\alpha\beta = d\gamma$;

$$(\mathbb{Z} + \frac{b_1+\sqrt{\Delta}}{2a_1}\ \mathbb{Z})\ (\mathbb{Z} + \frac{b_2+\sqrt{\Delta}}{2a_2}\ \mathbb{Z})\alpha\beta = (\mathbb{Z} + \frac{b_3+\sqrt{\Delta}}{2a_3}\ \mathbb{Z})\gamma$$

taking norms on both sides gives (cf. [1]):

$$\frac{N(\alpha\beta)}{a_1 \cdot a_2} = \frac{N(\gamma)}{a_3} = \frac{N(\alpha\beta)}{d^2 a_3}\ .$$

So we find

$$(1) \qquad a_3 = \frac{a_1 a_2}{d^2}\ .$$

Multiplying out gives

$$(\mathbb{Z} + \frac{b_1+\sqrt{\Delta}}{2a_1}\ \mathbb{Z} + \frac{b_1+\sqrt{\Delta}}{2a_2}\ \mathbb{Z} + \frac{(b_1 \cdot b_2+\Delta)+(b_1+b_2)\sqrt{\Delta}}{4a_1 a_2})\alpha\beta = (\mathbb{Z} + \frac{b_3+\sqrt{\Delta}}{2a_3}\ \mathbb{Z})\gamma$$

whence, looking at "$\sqrt{\Delta}$ coefficients":

$$(\frac{1}{2a_1}\mathbb{Z} + \frac{1}{2a_2}\mathbb{Z} + \frac{b_1+b_2}{4a_1a_2}\mathbb{Z})\alpha\beta = \frac{1}{2a_3}\mathbb{Z}\frac{\alpha\beta}{d} = \frac{d}{2a_1a_2}\mathbb{Z}\,\alpha\beta$$

$$a_2\mathbb{Z} + a_1\mathbb{Z} + \frac{b_1+b_2}{2}\mathbb{Z} = d\mathbb{Z}.$$

So

$$d = \gcd(a_1, a_2, \frac{b_1+b_2}{2}) \tag{2}$$

and we can easily compute $v_1, v_2, w \in \mathbb{Z}$ such that

$$v_1a_1 + v_2a_2 + w\frac{b_1+b_2}{2} = d.$$

Finally it is easily seen that b_3 can be taken to be

$$b_3 = v_2 \cdot b_1 \cdot \frac{a_2}{d} + v_1 \cdot b_2 \cdot \frac{a_1}{d} + w \cdot \frac{b_1b_2+\Delta}{2d}\,. \tag{3}$$

Formulas (1), (2) and (3) give a form (a_3, b_3, c_3) that corresponds to the ideal class that contains MN.

By Theorem 2.3', these formulas enable us to perform computations in the class group of a complex quadratic order, on condition, that we have a way to compute the *unique* reduced form equivalent to a given form. Fortunately there is a very simple and fast algorithm to do this:

REDUCTION ALGORITHM. Let $f = (a,b,c)$ be a primitive positive definite quadratic form of discriminant Δ.
(i) reduce $b \pmod{2a}$ such that $|b| \le a_3$ and adjust c;
if f is not reduced then
(ii) $f \leftarrow (c,-b,a)$ and start all over.

It is left to the reader to verify that this algorithm terminates and is correct. Perhaps, it is worth noting that $(a,b,c) \leftarrow (c,-b,a)$ corresponds to action of $S = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right) \in SL_2(\mathbb{Z})$ and reducing $b \pmod{2a}$ correspond to action of T^k where $T = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right) \in SL_2(\mathbb{Z})$ and k is some suitable integer. The group $SL_2(\mathbb{Z})$ is generated by S and T.

EXERCISE. In order to reduce a form (a,b,c) no more than $O(\max(1,\log(|a|/\sqrt{|\Delta|}))$ applications of (i) and (ii) are needed.

Now we can calculate in the class group by means of computations with *reduced* quadratic forms. For completeness we give: the inverse of a reduced form (a,b,c) equals (a,-b,c) and the unit element of the class group corresponds to the form $(1,1,\frac{1-\Delta}{4})$ or $(1,0,-\frac{\Delta}{4})$ depending on whether Δ is odd or even.

In the next section we will give Shanks' algorithm to compute class groups of quadratic orders. One of the basic ingredients of the algorithm is the ability to do calculations in the class group itself in an *efficient* way. The formulas given above are sufficiently efficient for these purposes.

Perhaps it is worth quoting the following formulas, which are essentially the formulas (1), (2) and (3), but somewhat more suitable for computation [31]:

Let $f = (a_1,b_1,c_1)$, $g = (a_2,b_2,c_2)$ be two primitive positive definite binary quadratic forms of discriminant Δ. Put $d = \gcd(a_1,a_2,(b_1+b_2)/2)$ and let $v_1,v_2,w \in \mathbb{Z}$ such that $v_1a_1 + v_2a_2 + w(b_1+b_2)/2 = d$. Let

$$a_3 = \frac{a_1a_2}{d^2},$$

$$b_3 = b_2 + 2\frac{a_2}{d}\overbrace{(\frac{b_1-b_2}{2}v_2 - c_2v)}^{(*)};$$

the form (a_3,b_3,c_3) now needs reduction. The term (*) does only matter mod a_1/d.

The algorithm for composition and reduction of binary quadratic forms can easily be programmed on a pocket calculator, like TI58, TI59, HP67, HP41C. In fact it is possible to compute class groups of complex quadratic orders, with the aid of a calculator like that, if the discriminant of the order is not too large, say, $\leq$ 10 decimal digits.

3. SHANKS' ALGORITHM

Let K be a finite abelian extension of $\mathbb{Q}$, then the following formula, the class number formula holds [17]:

$$h = \frac{w\sqrt{|\Delta|}}{2^{r_1}(2\pi)^{r_2}R}\prod_{\chi\neq 1} L(1,\chi) \tag{4}$$

where

$w = w(K) = \#\mu(K)$ = the number of roots of unity in K,

$\Delta = \Delta(K)$ = the discriminant of K,

$r_1 = r_1(K)$ = the number of embeddings $K \hookrightarrow \mathbb{R}$,

$r_2 = r_2(K)$ = half the number of embeddings $K \hookrightarrow \mathbb{C}$ $(\mathrm{im}(K) \not\subset \mathbb{R})$,

R = R(K) = the regulator of K,

χ runs over the non-trivial characters of $\mathrm{Gal}(K/\mathbb{Q})$,

$L(s,\chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} = \prod_{p \text{ prime}} (1 - \chi(p)p^{-s})^{-1}, \quad s \in \mathbb{C}, \ \mathrm{Re}\ s \geq 1.$

Any complex quadratic field K is abelian over $\mathbb{Q}$ and the only non-trivial character of $\mathrm{Gal}(K/\mathbb{Q})$ is the Legendre-symbol $(\frac{\Delta}{\cdot})$, where Δ is the discriminant of K. The class number formula reduces to

$$h = \frac{w(\mathcal{O})}{2\pi} \sqrt{|\Delta|}\, L(1,\chi),$$

and this formula also holds for non-maximal orders [1]. Here $w(\mathcal{O})$ denotes the number of roots of unity contained in $\mathcal{O}$. If $\Delta = -3$ or $\Delta = -4$, the class number of the order of discriminant Δ equals 1, so there is no harm in assuming that $\Delta \neq -3,-4$. Then always $w = 2$ and the class number of formula reduces further to

$$h = \frac{\sqrt{|\Delta|}}{\pi} L(1,\chi) = \frac{\sqrt{|\Delta|}}{\pi} \prod_{p \text{ prime}} (1 - (\tfrac{\Delta}{p})\tfrac{1}{p})^{-1}. \tag{5}$$

The infinite product (5) converges slowly to h. An analysis on assumption of the Generalized Riemann Hypothesis (GRH in the sequel) for this field, shows that only an expansion of this product that uses all primes $\leq c \cdot |\Delta|^{1+\varepsilon}$ for some universal c and ε, gives an approximation of h, accurate enough to determine h.

There are also explicit "finite" formulas for the class number of complex quadratic orders $\mathcal{O}$ with discriminant Δ; for maximal orders it holds that

$$h = \frac{1}{2-\chi(2)} \sum_{\substack{0<x<|\Delta/2| \\ (x,\Delta)=1}} \chi(x), \qquad (\Delta \neq -3,-4)$$

here χ denotes the Legendre symbol $(\frac{\Delta}{\cdot})$, see [1]. However, calculation of the class number of a quadratic order with a discriminant of say 10 decimal digits, using this formula, would be hardly possible.

Using the 1-1 correspondence between $\mathcal{O}$-ideal classes and reduced primitive forms of discriminant $\Delta = \Delta(\mathcal{O})$, one can also determine the class number of $\mathcal{O}$ by counting integral triples (a,b,c) with gcd(a,b,c) = 1, a > 0, $b^2 - 4ac = \Delta$ and $|b| \leq a \leq c$.

EXAMPLE. $\Delta = 691$.
(Recall that if (a,b,c) is reduced, $|b| \leq a < \sqrt{\frac{|\Delta|}{3}}$ and realize that for any form $b \equiv \Delta \pmod 2$).

±b	$\frac{-\Delta+b^2}{4}$	forms
15	229	
13	5·43	
11	7·29	
9	193	
7	5·37	
5	179	
3	$5^2 \cdot 7$	(7,±3,25), (5,±3,35)
1	173	(1,1,173)

So the class number of $\mathbb{Q}(\sqrt{-691})$ is 5. But this method is only efficient for small discriminants.

Counting methods of this sort are very useful to compute tables of class numbers; one then computes forms (a,b,c) with $|b| \leq a \leq c$, a > 0 and counts them, sorting them on discriminant $\Delta = b^2 - 4ac$. This is a very fast method and D.A. BUELL [3] used it, to compile a table of class numbers of complex quadratic number fields with discriminants Δ with $0 < -\Delta < 4000000$.

In 1970, Shanks introduced his algorithm to determine the structure of class groups of complex quadratic orders [31]. His algorithm relies upon an estimate of the class number of the order and computations in the class group itself; it is particularly effective if the discriminant of the order is very large.

Let $\mathcal{O}$ be a complex quadratic order of discriminant Δ, and let h be the class number of $\mathcal{O}$. The starting point in Shanks' algorithm is an approximation of the class number; this is obtained by means of the class number formula

$$(5) \qquad h = \frac{\sqrt{|\Delta|}}{\pi} L(1,\chi) = \frac{\sqrt{|\Delta|}}{\pi} \prod_{p \text{ prime}} (1 - (\frac{\Delta}{p})\frac{1}{p})^{-1}.$$

We approximate h, by simply evaluating

$$\tilde{h} = \frac{\sqrt{|\Delta|}}{\pi} \prod_{\substack{p \text{ prime} \\ p \leq X}} (1 - (\frac{\Delta}{p})\frac{1}{p})^{-1}$$

for some X (which we will take $O(|\Delta|^{\frac{1}{5}})$; we'll say more on choices of particular constants later). Due to convergence of the product (5), we have that

$$(6) \qquad (1-\varepsilon)\tilde{h} < h < (1+\varepsilon)\tilde{h},$$

where ε is a *small* positive number depending on X. This gives us a rough idea of the size of h. Next we choose a form f = (a,b,c) of discriminant Δ (for instance by taking a = p, a prime with $(\frac{\Delta}{p}) = +1$, and $b^2 \equiv \Delta \pmod{4a}$). By group theory, we have that $f^h = 1$ and we use this fact together with the estimate $h \approx \tilde{h}$, to find h by searching in the (relatively short!) interval

$$(7) \qquad ((1-\varepsilon)\tilde{h}, (1+\varepsilon)\tilde{h})$$

for a number h' such that $f^{h'} = 1$. Perhaps h' = h, but this need not be the case. Next we compute the precise order of f, by factoring h', which has size $O(|\Delta|^{\frac{1}{2}+\varepsilon})$ and we put H = the cyclic group generated by f; we keep H by means of a list of (independent) generators of its p-Sylow subgroup. If $(1-\varepsilon)\tilde{h} \leq \#H \leq (1+\varepsilon)h$ we conclude that $H = Cl(0)$; if not, we pick a new form f' and compute its order in the same way, now using that $\#H | \#Cl(0)$ and compute the group generated by H and f', by computing a set of independent generators for its Sylow-subgroup; we call this group H again.
We repeat this procedure until $(1-\varepsilon)\tilde{h} < \#H < (1+\varepsilon)\tilde{h}$ and then we conclude that $H = Cl(0)$.

A few remarks on this algorithm:

- The search for a number h' in the interval (7), such that $f^{h'} = 1$ can be performed effectively, by means of the so-called "baby-giant-step strategy":

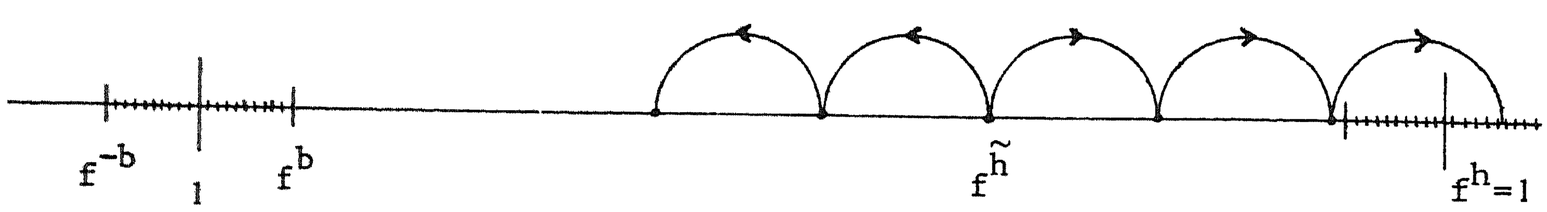
Let $\ell = 2\varepsilon\tilde{h}$ be the length of the interval (7), then compute f^h and search successively for $f^{\tilde{h}}, f^{\tilde{h}}g^{\pm 1}, f^{\tilde{h}}g^{\pm 2}, \ldots$, etc. (the giant steps) in the list of baby-steps. If one finds

$$f^{\tilde{h}}g^{i} = f^{a}$$

with $|a| \leq b$, for some i,

$$f^{\tilde{h}+2ib-a} = 1$$

and $h' = \tilde{h} + 2ib - a$.



The number of calculations needed to perform this strategy is proportional to $\sqrt{\ell}$.

- Determining the precise order of f, knowing that $f^{h'} = 1$, is done by factoring h' ($\sim \sqrt{|\Delta|}$) and by computing suitable powers of f; it is a relatively fast procedure.

- It is possible that many forms are needed to generate the whole class group, but usually, the time consuming baby-giant-step strategy need only be performed once: usually one of the first forms picked generates a large part of the class group; its order n is often divisible by some large primes q such that $qn \nmid h$, since no multiple of qn is in the interval

$$((1-\varepsilon)\tilde{h},\ (1+\varepsilon)\tilde{h}).$$

This implies that the q-Sylow subgroups of the group generated by this form equal the q-Sylow subgroups of $Cl(O)$.

After encountering a form like that, we can raise new forms f to the power m, being the part of n consisting of these large primes q; then we know that f^{m} has a multiple of its order in the interval

$$((1-\varepsilon)\frac{\tilde{h}}{m},\ (1+\varepsilon)\frac{\tilde{h}}{m})$$

which is a very short interval. Usually we need not perform the baby-giant-step strategy and we can avoid computations in the q-Sylow subgroups for the large primes q.

- We will sketch a derivation, under GRH, of the order of the algorithm. For the details we refer to Section 6, where an analysis of the factorization algorithm that is based on Shanks' algorithm is given.

Put

$$\tilde{h} = \tilde{h}(X) = \frac{\sqrt{|\Delta|}}{\pi} \prod_{\substack{p \text{ prime} \\ p \le X}} (1 - (\frac{\Delta}{p})\frac{1}{p})^{-1}$$

then for some effectively computable, universal constant C and for all X large enough:

$$\left|\frac{\tilde{h}(X)}{h} - 1\right| \le C \frac{\log|\Delta X|}{\sqrt{X}}$$

(cf. Section 6).

If we take $X \approx |\Delta|^{\alpha}$ for some α, to be determined, we have that

$$\left|\frac{\tilde{h}(|\Delta|^{\alpha})}{h} - 1\right| = O(|\Delta|^{-\frac{1}{2}\alpha+\varepsilon}), \tag{8}$$

where the O constant depends on ε.

The length of the interval (7) equals

$$\ell \approx |\Delta|^{\frac{1}{2}+\varepsilon} \cdot |\Delta|^{-\frac{1}{2}\alpha+\varepsilon} = |\Delta|^{\frac{1}{2}-\frac{1}{2}\alpha+\varepsilon},$$

where we used that $h(O(\Delta)) = O(|\Delta|^{1/2+\varepsilon})$. [37]. Since evaluating Legendre symbols is logarithmic in the arguments, we have, by the prime number theorem, that the time for evaluating a truncated product

$$\frac{\sqrt{|\Delta|}}{\pi} \prod_{\substack{p \text{ prime} \\ p \le X}} (1 - (\frac{\Delta}{p})\frac{1}{p})^{-1},$$

is $O(X^{1+\varepsilon})$, if we take $X \approx |\Delta|^{\alpha}$. So

$$\text{"time for approximating h"} \sim |\Delta|^{\alpha} \tag{9}$$

The time needed to perform the baby-giant strategy is proportional to $\sqrt{\ell}$, so

$$\text{"baby-giant costs"} \sim |\Delta|^{\frac{1}{4}-\frac{1}{4}\alpha+\varepsilon}.$$

We will have an optimum if

$$\frac{1}{4} - \frac{1}{4}\alpha = \alpha \qquad \text{i.e. } \alpha = \frac{1}{5}.$$

This indicates that Shanks' algorithm has order $|\Delta|^{1/5+\varepsilon}$; however, there are some details:

- Many primes may be needed to generate the whole class group. However, it easily follows from results of LAGARIAS, MONTGOMERY and ODLYZKO [15], obtained under assumption of GRH, that the class group is generated by the classes of the primes with norm $\ll \log^2|\Delta|$, cf. Section 6.
- The computations necessary to compute a presentation of the class group by independent generators, may become time consuming if the structure of the class group is complicated i.e. "highly non-cyclic". At present we cannot estimate the computing time for these calculations better than $|\Delta|^{1/4+\varepsilon}$, but since "almost all" class groups appear to have a large cyclic factor (cf. Section 7), $|\Delta|^{1/5+\varepsilon}$ seems to be a more practical estimate. For bounds on the exponent of class groups see [2,41]. However, computing *the class number* can always be done in time $O(|\Delta|^{1/5+\varepsilon})$. Also determining the isomorphy type of $Cl(\mathcal{O})$ as an abelian group can be done in time $O(|\Delta|^{1/5+\varepsilon})$, without, however, giving a set of independent generators.

4. CLASS GROUPS AND UNITS OF REAL QUADRATIC NUMBER FIELDS

If K is a real quadratic number field, let $\mathcal{O}(K)$ denote its ring of integers and $\Delta(K)$ its discriminant. Real quadratic fields are characterized by their discriminants, which are positive integers congruent to 0 or 1 (mod 4), but, like in the complex case, not every positive integer $\equiv 0$ or 1 (mod 4), is the discriminant of a real quadratic field.

However, every *non-square* positive integer $\Delta \equiv 0$ or 1 (mod 4) is the discriminant of a unique *real quadratic order* $\mathcal{O}$, a subring of a ring of integers of a real quadratic field: Δ can uniquely be written as $\Delta = f^2 D$ where $f \in \mathbb{Z}_{\geq 1}$ and D is the discriminant of a real quadratic field K; then Δ is the discriminant of the unique subring $\mathcal{O}$ of index f in $\mathcal{O}(K)$.

The class group of a real quadratic order $\mathcal{O}$ is defined as the group of invertible fractional $\mathcal{O}$-ideals modulo the principal fractional $\mathcal{O}$-ideals.

If $\Delta = f^2$ is a square, we can consider Δ to be the discriminant of the subring $\mathbb{Z}(1,1) \times \mathbb{Z}(0,f)$ of index f in $\mathbb{Z} \times \mathbb{Z}$; the class group of this ring is isomorphic to $(\mathbb{Z}/f\mathbb{Z})^\times/\{\pm 1\}$. We do not enter into these rather pathological cases. For the "imtermediate case" $\Delta = 0$ see GAUSS [12].

Let $\mathcal{O}$ be a real quadratic order, then

$$\mathcal{O}^\times \simeq \mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z};$$

more precisely: there exists an $\varepsilon \in \mathcal{O}^\times$ such that every unit $u \in \mathcal{O}^\times$ can be written as $\pm\varepsilon^k$, $k \in \mathbb{Z}$. There are four numbers in $\mathcal{O}^\times$, that each, together with -1, generate $\mathcal{O}^\times$; fixing an embedding $K \hookrightarrow \mathbb{R}$, one of these numbers is greater than 1. We denote this number by ε_0 and call it *the fundamental unit of* $\mathcal{O}$.

DEFINITION If ε_0 is the fundamental unit of $\mathcal{O}$ then

$$R(\mathcal{O}) = \log \varepsilon_0$$

is called *the regulator of* $\mathcal{O}$.

If no confusion is likely, we will omit the indices $\mathcal{O}$. Let K be a real quadratic field with discriminant Δ.

DEFINITION. $N: K^\times \to \mathbb{Q}^\times$ by $N\alpha = \alpha\cdot\sigma(\alpha)$ where $1 \neq \sigma \in \mathrm{Gal}(K/\mathbb{Q})$. We call N *the norm map*; it is a homomorphism and if we write $\alpha \in K^\times$, $\alpha = p+q\sqrt{\Delta}$ then

$$N\alpha = p^2 - \Delta q^2.$$

By means of the norm map we can refine the concept of the class group somewhat:

DEFINITION. Let $\mathcal{O}$ be a real quadratic order and let $P(\mathcal{O})^+ = \{$principal ideals generated by elements of positive norm$\}$. We have the following commutative diagram with exact rows and columns:

$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & & & \\
 & & \downarrow & & \downarrow & & & & \\
0 & \longrightarrow & P(\mathcal{O})^{+} & \longrightarrow & I(\mathcal{O}) & \longrightarrow & Cl^{+}(\mathcal{O}) & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & P(\mathcal{O}) & \longrightarrow & I(\mathcal{O}) & \longrightarrow & Cl(\mathcal{O}) & \longrightarrow & 0 \\
 & & & & \downarrow & & \downarrow & & \\
 & & & & 0 & & 0 & &
\end{array}$$

$Cl^{+}(\mathcal{O})$ is called *the narrow class group of* $\mathcal{O}$; it maps surjectively to $Cl(\mathcal{O})$ and it is easy to see, that the kernel of this map has order 1 or 2. By h^{+} we denote the order of $Cl^{+}(\mathcal{O})$: $h^{+} = h$ or $h^{+} = 2h$.

DEFINITION. $\varepsilon^{+} := \varepsilon$ if $N\varepsilon = +1$ and $\varepsilon^{+} := \varepsilon^{2}$ if $N\varepsilon = -1$; $R^{+} := \log \varepsilon^{+}$.

Now the units with positive norm are precisely the numbers $\pm(\varepsilon^{+})^{k}$, $k \in \mathbb{Z}$.

PROPOSITION 4.1.

(i) *if* $N\varepsilon = -1$ *then* $h^{+} = h$ *and* $R^{+} = 2R$,
if $N\varepsilon = +1$ *then* $h^{+} = 2h$ *and* $R^{+} = R$;

(ii) $2hR = h^{+}R^{+}$.

PROOF. □

Next we'll explain the setting, in which the calculation of the class group and the regulator, as discussed in the next section, are performed. The ideas involved are due to LENSTRA and SHANKS [18,33].

DEFINITION. Let $\mathcal{O}$ be a real quadratic order; put

$$F'(\mathcal{O}) = \{(M,\alpha) \mid M \text{ an invertible } \mathcal{O}\text{-submodule of } K;\ \alpha \in M \text{ primitive}\}$$

$$G'(\mathcal{O}) = \{(\beta\mathcal{O},\alpha) \mid \beta \in K^{\times},\ N\beta > 0;\ \alpha \in \beta\mathcal{O} \text{ primitive}\}$$

$$K^{\times}_{N>0} = \{\alpha \in K \mid N\alpha > 0\}.$$

We turn $F'(\mathcal{O})$ into an abelian group, by defining

$$(M,\alpha)(N,\beta) = (MN,\gamma),$$

where $\alpha\beta = d\gamma$, with $d \in \mathbb{Z}_{\geq 1}$ and $\gamma \in \mathfrak{M}\mathfrak{N}$ primitive. We then have the series of subgroups

$$K^{\times}_{N>0} \subset G'(\mathcal{O}) \subset F'(\mathcal{O}).$$

Here $K^{\times}_{N>0} \hookrightarrow G'(\mathcal{O})$ by $\alpha \to (\alpha\mathcal{O},\alpha)$.

DEFINITION.

$$F(\mathcal{O}) = F'(\mathcal{O})/K^{\times}_{N>0},$$

$$G(\mathcal{O}) = G'(\mathcal{O})/K^{\times}_{N>0}.$$

PROPOSITION 4.2. *There is an exact sequence*

$$0 \to G(\mathcal{O}) \to F(\mathcal{O}) \to Cl^{+}(\mathcal{O}) \to 0.$$

PROOF. Define $F'(\mathcal{O}) \to Cl^{+}(\mathcal{O})$ by $(M,\alpha) \to$ class of M; the kernel of this map is precisely $G'(\mathcal{O})$. □

In terms of binary quadratic forms we have that

$$F(\mathcal{O}) = \left\{\begin{matrix}\text{primitive binary quadratic forms} \\ \text{of discriminant } \Delta = \Delta(\mathcal{O})\end{matrix}\right\} \Big/ \begin{pmatrix} 1 & \mathbb{Z} \\ 0 & 1 \end{pmatrix},$$

$$G(\mathcal{O}) = \left\{\begin{matrix}\text{primitive binary quadratic forms of discriminant} \\ \Delta = \Delta(\mathcal{O}) \text{ that are } SL_2(\mathbb{Z})\text{-equivalent to} \\ X^2 + \Delta XY + (\Delta^2-\Delta)/4\ Y^2\end{matrix}\right\} \Big/ \begin{pmatrix} 1 & \mathbb{Z} \\ 0 & 1 \end{pmatrix}.$$

(For definitions and facts on quadratic forms see [1], or Section 2.) A translation between the different descriptions of $F(\mathcal{O})$ and $G(\mathcal{O})$ can be given as follows:

Let Δ be the discriminant of $\mathcal{O}$ and suppose $(M,\alpha) \in F'(\mathcal{O})$; let

$$M = \left(\mathbb{Z} + \frac{b+\sqrt{\Delta}}{2a}\,\mathbb{Z}\right)\alpha$$

with $\operatorname{sgn} a = \operatorname{sgn} N\alpha$. The image of (M,α) in $F(\mathcal{O})$ corresponds to the $\begin{pmatrix} 1 & \mathbb{Z} \\ 0 & 1 \end{pmatrix}$-orbit of $aX^2 + bXY + cY^2$ with $b^2 - 4ac = \Delta$. So we can look at $F(\mathcal{O})$ as consisting

of binary quadratic forms (a,b,c) where we identify forms (a_1,b_1,c_1) and (a_2,b_2,c_2) whenever $a_1 = a_2$ and $b_1 \equiv b_2 \pmod{2a_1}$.

$G(\mathcal{O})$ consists of $\left(\begin{smallmatrix}1 & \mathbb{Z}\\ 0 & 1\end{smallmatrix}\right)$-orbits of quadratic forms that are $SL_2(\mathbb{Z})$-equivalent to those corresponding to the image of $(\mathcal{O},1)$ in $G(\mathcal{O})$; these are precisely the $\left(\begin{smallmatrix}1 & \mathbb{Z}\\ 0 & 1\end{smallmatrix}\right)$- orbits of form that are $SL_2(\mathbb{Z})$-equivalent to $X^2 + \Delta XY + (\Delta^2-\Delta/4)Y^2$.

DEFINITION. Let $\alpha \in K$ and let $i_1: K \to \mathbb{R}$ be a fixed embedding and $i_2: K \to \mathbb{R}$ the other one; then

$$|\alpha|_{\infty 1} := |i_1(\alpha)| \quad \text{and} \quad |\alpha|_{\infty 2} := |i_2(\alpha)|.$$

We define a map, the *distance map*,

$$D: G(\mathcal{O}) \to \mathbb{R}/R^+\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z},$$

by

$$D((\beta\mathcal{O},\alpha)K^{\times}_{N>0}) = (\tfrac{1}{2}\log(|\tfrac{\alpha}{\beta}|_{\infty 1}/|\tfrac{\alpha}{\beta}|_{\infty 2}),\ \operatorname{sgn} N\alpha).$$

Here we use the isomorphism of groups: $\{+1,-1\} \cong \mathbb{Z}/2\mathbb{Z}$.

PROPOSITION 4.3. *D is a well defined homomorphism and* D *is injective.*

PROOF. It is trivial to check, that the value of D on $(\beta\mathcal{O},\alpha)$ and on $\xi\cdot(\beta\mathcal{O},\alpha) = (\xi\beta\mathcal{O}, \frac{\xi\alpha}{d})$, $(d \in \mathbb{Z}_{\geq 1},\ N\xi > 0)$ is the same. If $(\beta\mathcal{O},\alpha) = (\beta'\mathcal{O},\alpha)$ in $G(\mathcal{O})$, we have that β and β' differ by a norm positive unit, say, that $\beta' = \pm(\varepsilon^+)^k\cdot\beta$, for some $k \in \mathbb{Z}$. Then

$$d(\beta'\mathcal{O},\alpha) = \left(\tfrac{1}{2}\log\left(\left|\frac{\alpha}{(\varepsilon^+)^k\beta}\right|_{\infty 1}\Big/\left|\frac{\alpha}{(\varepsilon^+)^k\beta}\right|_{\infty 2}\right),\ \operatorname{sgn} N\alpha\right)$$

$$= (\tfrac{1}{2}\log(|\tfrac{\alpha}{\beta}|_{\infty 1}/|\tfrac{\alpha}{\beta}|_{\infty 2}) - kR^+,\ \operatorname{sgn} N\alpha)$$

$$= d(\beta\mathcal{O},\alpha),$$

and we see that D is well defined. To prove injectivity, let $(\beta\mathcal{O},\alpha) \in G(\mathcal{O})$, with

$$d(\beta\mathcal{O},\alpha) = (\tfrac{1}{2}\log(|\tfrac{\alpha}{\beta}|_{\infty 1}/|\tfrac{\alpha}{\beta}|_{\infty 2}),\ \mathrm{sgn}\ N\alpha) = (0,0).$$

This implies

$$|\tfrac{\alpha}{\beta}|_{\infty 1} = |\tfrac{\alpha}{\beta}|_{\infty 2} \quad \text{and} \quad N\alpha > 0.$$

So $\frac{\alpha}{\beta} \in \mathbb{Q}$ or $\frac{\alpha}{\beta} \in \mathbb{Q}\cdot\sqrt{\Delta}$, whence, since $N\alpha, N\beta > 0$, it follows that $\frac{\alpha}{\beta} \in \mathbb{Q}$ and so, since $\alpha \in \beta\mathcal{O}$ primitive, we have that $\alpha = \pm\beta$ and we find that

$$(\beta\mathcal{O},\alpha) = (\mathcal{O},1) \bmod K^{\times}_{N>0}. \qquad \square$$

NB. The image of D is dense in $\mathbb{R}/R^{+}\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$; however for cardinality reasons, D is not surjective.

DEFINITION 4.4. Let ϕ_1, ϕ_2 be two elements of $F(\mathcal{O})$, that are in the same $G(\mathcal{O})$-coset. We define the *distance from* ϕ_1 *to* ϕ_2 to be the first coordinate of $D(\phi_2\phi_1^{-1})$.

So distances between elements of F, that are in different $G(\mathcal{O})$-cosets, *are not defined*. However, it is possible to define a notion of *absolute* distance, as follows: It is possible to lift the map

$$(\beta\mathcal{O},\alpha) \to \mathrm{sgn}\ N\alpha,$$

to the whole of $F(\mathcal{O})$ in a *canonical* way:

$$(M,\alpha) \to \mathrm{sgn}\ N\alpha.$$

Since $\mathbb{R}/R^{+}\mathbb{Z}$ is a divisible group, one can lift

$$(\beta\mathcal{O},\alpha) \to \tfrac{1}{2}\log(|\tfrac{\beta}{\alpha}|_{\infty 1}/|\tfrac{\beta}{\alpha}|_{\infty 2}),$$

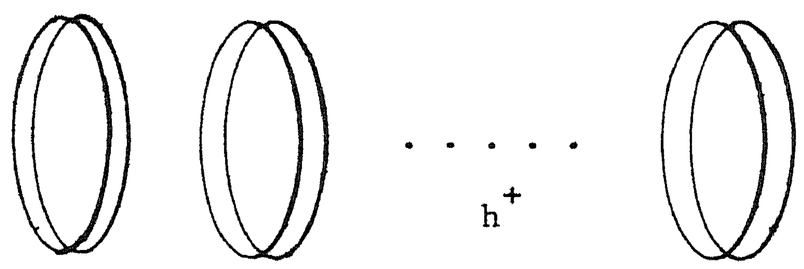
to the whole of F(O) as well (*uncanonically* this time). Combining these maps one finds a lift of D to the whole of F(O):

$$\begin{array}{ccccccccc} & & & & 0 & & & & \\ & & & & \downarrow & & & & \\ 0 & \longrightarrow & G(\mathcal{O}) & \longrightarrow & F(\mathcal{O}) & \longrightarrow & Cl^{+}(\mathcal{O}) & \longrightarrow & 0. \\ & & & & \downarrow D & \swarrow D & & & \\ & & & & \mathbb{R}/R^{+}\mathbb{Z}\oplus\mathbb{Z}/2\mathbb{Z} & & & & \end{array}$$

We will denote this lift by D again, and if $(M,\alpha) \in F(O)$, we will call $D(M,\alpha)$ the *absolute distance* of (M,α). Note, that the absolute distance depends upon the lift of D: $G(O) \to \mathbb{R}/R^+\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ to $F(O)$.

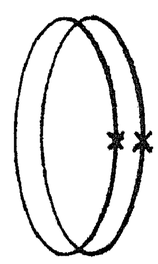
The class group $Cl^+(O)$ can now be viewed as a set of h^+ double circles, each of "circumference" R^+, each point of the image of D on a double circle representing a $\begin{pmatrix}1 & \mathbb{Z}\\ 0 & 1\end{pmatrix}$-orbit, or $\mathbb{Z}$-orbit for short, of a quadratic form. We will call these $\begin{pmatrix}1 & \mathbb{Z}\\ 0 & 1\end{pmatrix}$-orbits forms again.

$Cl^+(O)$:



We will call the ideal classes, pictured as these double circles, also *cycles*.

The double circle, corresponding to the principal ideal class, will be called the *principal cycle*. On the principal cycle, there is always a form $(1,\Delta,\Delta^2-\Delta/4)$ (a $\mathbb{Z}$-orbit!), which we will call the principal form.



Two forms on a double circle that are at the same absolute distance, but on different circles, differ by the sign of a: One circle contains forms (a,b,c) with $a > 0$, the other one contains forms (a,b,c) with $a < 0$.

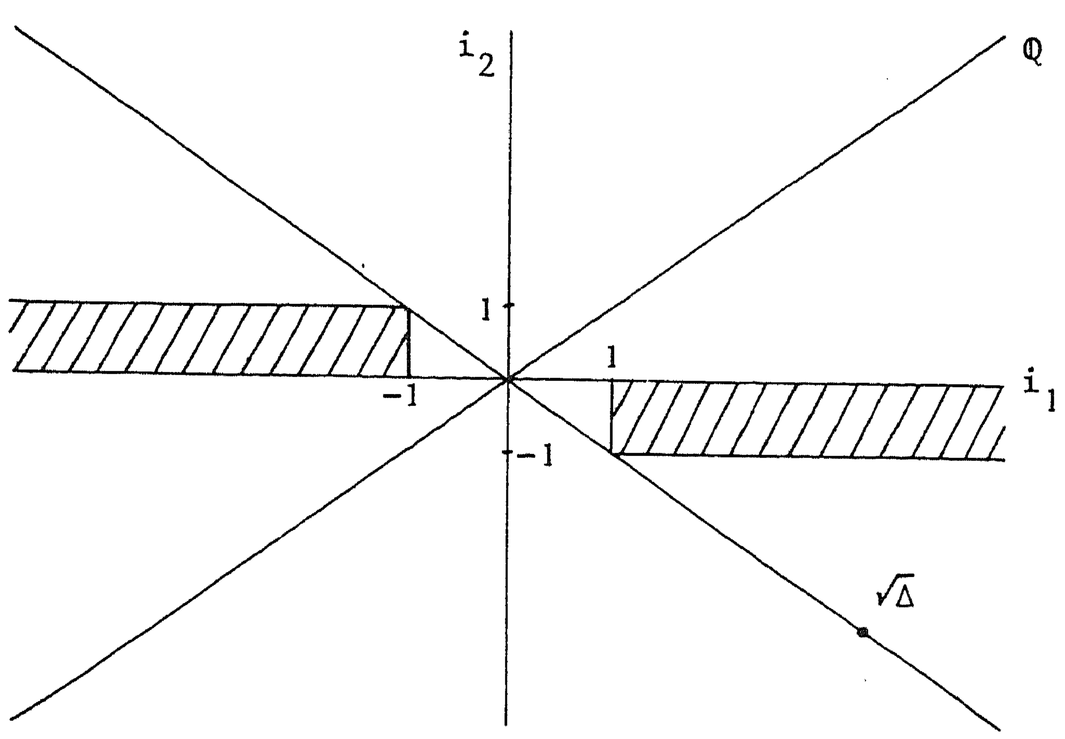
Like in the complex case, it is possible to translate the composition law in terms of quadratic forms (or rather $\mathbb{Z}$-orbits of forms); this yields the same formulas as the formulas (1), (2) and (3) given in Section 2.

The notion of a reduced form is slightly different however:

DEFINITION. Let $f = (a,b,c)$ be a primitive binary quadratic form of discriminant Δ; then f is called *reduced* if

$$|\sqrt{\Delta} - |2a|| < b < \sqrt{\Delta}$$

i.e. if we picture K as embedded in $\mathbb{R} \times \mathbb{R}$ via its embeddings $i_1, i_2 : K \to \mathbb{R}$ via $x \to (i_1(x), i_2(x))$, the point $b+\sqrt{\Delta}/a$ is in the shaded area.



The condition for a form (a,b,c) to be reduced implies that

$$0 < b < \sqrt{\Delta} \quad \text{and} \quad |a| < \sqrt{\Delta};$$

from this it follows easily, that only finitely many reduced forms of discriminant Δ exist; the $\mathbb{Z}$-orbits of these forms form a discrete subset of $F(0)$ and every ideal class (= double circle) contains at least one reduced form.

In view of the applications to the algorithm discussed in Section 5, we like to do our calculations in the *finite* set of reduced forms: We need a *reduction algorithm*, in order to determine a reduced form equivalent to a given form.

Reduction algorithm. Let (a,b,c) be a quadratic form of discriminant Δ:
(i) if $|a| < \sqrt{\Delta}$ reduce b (mod 2a) such that

$$\sqrt{\Delta} - |2a| < b < \sqrt{\Delta}$$

and adjust c;
if $|a| > \sqrt{\Delta}$ reduce b (mod 2a) such that

$$|b| \leq |a|$$

and adjust c;

(ii) if the form is *not* reduced then

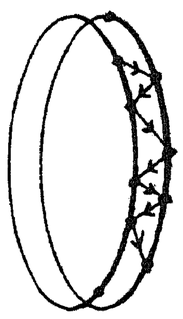
$$(a,b,c) \leftarrow (c,-b,a)$$

and start all over.

It is left to the reader, to verify, that this algorithm terminates and is correct.

EXERCISE: Show that no more than $O(\max(1, \frac{\log|a|}{\sqrt{\Delta}}))$ applications of (i) and (ii) are needed to reduce (a,b,c).

In contrast to the complex case, there can be many more reduced forms in the same $SL_2(\mathbb{Z})$-orbit (= a double circle) and it is possible to jump from one form to another by means of *reduction*: If f is a reduced form in a fixed coset of $G(0)$, say $f = (a,b,c)$, then let $g = (c,b',c')$ with $b' \equiv -b \pmod{2c}$ and $\sqrt{\Delta} - |2c| < b' < \sqrt{\Delta}$ and c' such that $b'^2 - 4cc' = \Delta$. Then g is also reduced and g is on the opposite circle since $ac < 0$ (this follows directly from the fact, that f is reduced). Furthermore, if f is a reduced form on some double-circle, then one finds *all other* reduced forms on this double circle by successive reduction [12].



The distance from the reduced form f, to its successor g equals

$$\frac{1}{2} \log\left(\frac{\sqrt{\Delta}+b}{\sqrt{\Delta}-b}\right). \tag{11}$$

Now we can compute in the set of reduced forms: we can "jump" from one form to the "next" one on the same double-circle, and we can compute the product of two reduced forms: a not necessarily reduced form, which we can reduce

by means of the reduction algorithm. In doing this, we can keep track of the absolute distance of the product. The ultimate reduction gives a form which is one the same double circle as the non-reduced product, but in general at a different position on that double circle.

Fortunately, reduction of the product of two reduced forms "causes only small replacement along the double circle". It can be shown, that the distance between the non-reduced product and the reduced product is at most $\frac{1}{4} \log \Delta + O(1)$, i.e. usually very small compared to the circumference of the cycle, which is often $\sim \sqrt{\Delta}$. So, if f and g are reduced forms, the following "holds".

$$\text{abs.distance}(f) + \text{abs.distance}(g) \approx \text{abs.distance}(\text{reduced}(f \cdot g)).$$

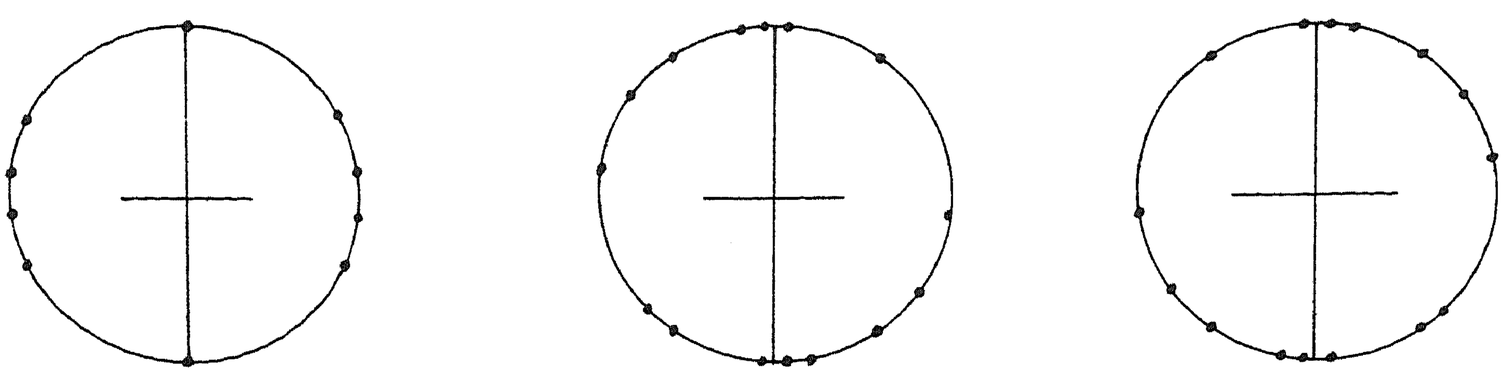
An example: $\Delta = 761$.

The following is a list of all reduced forms of discriminant Δ:

(1,27, -8): 0.0 .	(2,27, -4): 1.704.	(2,25,-17): 13.052.
(-8,21, 10): 2.267.	(-4,21, 20): 3.971.	(-17, 9, 10): 14.558.
(10,19,-10): 3.266.	(20,19, -5): 4.970.	(10,11,-16): 0.141.
(-10,21, 8): 4.112.	(-5,21, 16): 5.815.	(-16,21, 5): 0.563.
(8,27, -1): 5.111.	(16,11,-10): 6.814.	(5,19,-20): 1.562.
(-1,27, 8): 7.378.	(-10, 9, 17): 7.237.	(-20,21, 4): 2.408.
(8,21,-10): 9.645.	(17,25, -2): 7.757.	(4,27, -2): 3.407.
(-10,19, 10): 10.644.	(-2,27, 4): 9.081.	(-2,25, 17): 5.674.
(10,21, -8): 11.489.	(4,21,-20): 11.348.	(17, 9,-10): 7.180.
(-8,27, 1): 12.489.	(-20,19, 5): 12.347.	(-10,11, 16): 7.519.
(1,27, -8): 14.756.	(5,21,-16): 13.193.	(16,21, -5): 7.941.
	(-16,11, 10): 14.192.	(-5,19, 20): 8.940.
	(10, 9,-17): 14.614.	(20,21, -4): 9.786.
	(-17,25, 2): 0.197.	(-4,27, 2): 10.785.
	(2,27, -4): 1.704.	(2,25,-17): 13.052.

The first column lists all reduced forms in the principal cycle with their (approximate) absolute distance, the next two columns list the reduced forms in the cycles, that represent the two other ideal classes: the class number equals 3. The real numbers, given there, are the absolute distances of these forms (depending on the lift of D...).

We can picture the class group as a set of cycles:



Since $(-1,27,8)$ is in the principal cycle, the norm of the fundamental unit is -1. In fact

$$\varepsilon = 800 + 29\sqrt{761}$$

and

$$\varepsilon^{+} = 1280001 + 46400\sqrt{761}.$$

The reader is invited to check the values, given for the absolute distance, by means of composition and reduction.

The example shows, that it is possible to have a different number of forms in the circles; however, the "circumference" is the same for every circle.

Finally, some remarks:

- In the principal cycle there is a reduced form at distance exactly $\frac{1}{2} R^{+}$; if this is the $\mathbb{Z}$-orbit of $(-1,\Delta,\Delta^2-\Delta/4)$, we have that $(0,\varepsilon_0)$ and $(0,1)$ are equal mod $G(0)$ i.e. $N\varepsilon_0 = -1$; if not, we must have, that $N\varepsilon_0 = 1$. So, by computing in the principal cycle, we can find out, whether $N\varepsilon = +1$ or -1.
- In Section 2 we described a method, to compile tables of class numbers of *complex* quadratic orders by means of counting reduced forms. We cannot apply this method straightaway to real quadratic orders; but still an analogous method is possible. One computes positive binary quadratic forms (a,b,c) and counts them, sorting them on $\Delta = b^2 - 4ac$; however, one does not simply count the forms, but one sums their distances to their successors in their double circles i.e.

$$\frac{1}{2}\log\left(\frac{\sqrt{\Delta}+b}{\sqrt{\Delta}-b}\right) .$$

Once this is done, one knows the complete "length" of the class group; after computing R^+ for each Δ, by means of successive reduction of $(1,\Delta,\Delta^2-\Delta/4)$, one divides the length of the class group by R^+; this gives h^+; the norm of the fundamental unit is found as a by-product of the computation of R^+.
- Due to the formulas for composition and reduction we can efficiently calculate in $C\ell(\mathcal{O})$. However, some problems remain hard, it seems. For instance, suppose that one knows, that for a given order $\mathcal{O}$ the class number h^+ equals one, and suppose $(\Delta(\mathcal{O})/2) = +1$; then there must be a form $(2,B,C)$ (some B, C) in the principal cycle. Where to find it? Apart from a rigorous search in the principal cycle, (for instance, by means of a baby-giant-step strategy), there seems to be no way, to find this form in this double circle, which has circumference $\approx \sqrt{\Delta}$ in this case. We'll come back to this problem in the next section.

5. DETERMINATION OF THE CLASS GROUP AND THE REGULATOR OF A REAL QUADRATIC NUMBER FIELD

Let K be a real quadratic number field with discriminant equal to Δ; the class number formula (4) applied to K becomes

$$h = \frac{\sqrt{\Delta}}{2R} L(1,\chi), \tag{12}$$

where $\chi_\Delta = \left(\frac{\Delta}{\cdot}\right)$, the Kronecker symbol; and this formula also holds for non-maximal orders of discriminant Δ. So, in order to derive an estimate of h from (12), one should compute the regulator R. The classical way to do this is, to determine the continued fraction expansion of $\sqrt{\Delta}$. However, experience shows, that the length of the period of this expansion may well be $\sim \sqrt{\Delta}$, so, a straightforward computation of R by this method would take much more time than $c\cdot\Delta^{1/5}$.

By means of the theory, developed in Section 4, we can overcome this difficulty and finally give an algorithm to determine both the regulator and the class number of a real quadratic order, which is similar to Shanks'.

There is not much sense in determining the fundamental unit of a real quadratic order with a large discriminant; for instance suppose $\Delta(\mathcal{O})$ has 20 digits (a *very* reasonable number for our algorithm) and suppose $h^+(\mathcal{O}) = 1$, then $R^+(\mathcal{O}) \approx 10^{10}$ and $\varepsilon^+ = e^{R^+}$ is gigantic. In fact, even writing down this

number, by means of the fastest line printers now available, would take a few weeks!

Let us first mention some simple methods to determine the class number of a real quadratic order $\mathcal{O}$ of discriminant Δ, which are suitable if Δ is not too large, say $\Delta \approx 6$ decimal digits.

One can compute R^+ by successive reduction in the principal cycle: one starts with $(1,b,c)$, the principal form, and reduces it, until for two successive forms (a_1,b_1,c_1) and (a_2,b_2,c_2) one has that $b_1 = b_2$. En route, one sums

$$\log\left(\frac{b+\sqrt{\Delta}}{b-\sqrt{\Delta}}\right) ,$$

for all reduced forms (a,b,c); the sum equals R^+ ($= R$ if $a_2 \neq -1$; $= 2R$ if $a_2 = -1$).

There is the following formula for the class number of a maximal order $\mathcal{O}(\Delta)$:

$$h(\mathcal{O}) = -\frac{1}{R} \sum_{\substack{c<x<\Delta/2 \\ (x,\Delta)=1}} \chi(x) \log \sin \frac{\pi x}{\Delta} ;$$

a similar, but more complicated formula holds for non-maximal orders.

Using the dictionary between $\mathcal{O}$-ideal classes and primitive binary quadratic forms, one can also find the class number by counting all the forms of discriminant $\Delta(\mathcal{O})$ and sorting them by double-circles (by periods). Many investigators determined class numbers and regulators by means of these algorithms; they are completely unfeasible if the discriminants of the orders are very large, say $\Delta = 20$ decimal digits.

Let's explain the algorithm: Let h^+ denote the narrow class number of a real quadratic order $\mathcal{O}$; let $\Delta = \Delta(\mathcal{O})$ and $R^+ = R^+(\mathcal{O})$. By Proposition (4.1) we have that $h^+R^+ = 2hR$; so formula (12) becomes

$$(13) \qquad h^+R^+ = \sqrt{\Delta} \prod_{p \text{ prime}} \left(1 - \left(\frac{\Delta}{p}\right)\frac{1}{p}\right)^{-1} .$$

Like in the complex case, the starting point of the algorithm is an approximation of h^+R^+, obtained from (13): let

$$(14) \qquad \tilde{R} = \sqrt{\Delta} \prod_{\substack{p \text{ prime} \\ p \leq X}} \left(1 - \left(\frac{\Delta}{p}\right)\frac{1}{p}\right)^{-1} ,$$

for some X, which we will take $c \cdot \Delta^{1/5}$. We find, that

$$(15) \qquad (1-\varepsilon)\tilde{R} \leq h^+R^+ \leq (1+\varepsilon)\tilde{R},$$

for some small $\varepsilon \in \mathbb{R}_{>0}$.

Now the principal cycle has length R^+, a number, that we do not know yet. However $\tilde{R}$ is close to h^+R^+, a multiple of R^+, so we can jump to a form f, in the principal cycle at distance $\approx \tilde{R}$ of the principal form (mod R^+ of course; but we do not know R^+ yet) and search for the principal form in the interval

$$(16) \qquad (\tilde{R}(1-\varepsilon),\ \tilde{R}(1+\varepsilon)).$$

If we've found this form, we know a multiple of the narrow regulator R^+.

Then by looking half way the cycle, at $\frac{1}{3}$, at $\frac{1}{5}$ etc., we can determine R^+. This immediately gives us an approximation of h^+:

$$(17) \qquad h^+ \approx \tilde{h} = \frac{\sqrt{\Delta}}{R^+} \prod_{\substack{p \text{ prime} \\ p \leq X}} (1 - (\tfrac{\Delta}{p})\tfrac{1}{p})^{-1},$$

and we can complete the calculations by computing the structure of the class group in a way similar to the complex quadratic case.

Some remarks:

- Finding a multiple of R^+ in the interval (15) can efficiently be done by means of a baby-giant strategy (see Section 3). Here, the baby-steps are, very cheaply, computed by successive reduction of the principal form and for computing the giant-steps, one uses composition of forms. If X in (14) is $O(\Delta^{1/5})$, then the baby-giant computations will also be done in $c \cdot \Delta^{1/5+\varepsilon}$ operations.
- Determining the precise regulator R^+, can also be done in $c \cdot \Delta^{1/5+\varepsilon}$ operations. We won't give the details; suffice it to say, that for small primes p ($<< \Delta^{1/10}$), one jumps at $\frac{1}{p}$ of the principal cycle and looks for the principal form, while for large p one solves the problem, by making more giant steps.
- Knowing R^+, we can copy Shanks' algorithm to compute the class group of $\mathcal{O}$. There are some complications however; first of all: testing for equality of two ideal classes, represented by quadratic forms, is now much harder than in the complex case, since *many* forms may represent the same class. However, if one knows for two forms f_1 and f_2 that for some integer n, the forms f_1^n

and f_2^n are in the principal cycle at absolute distance d_1 resp. d_2, then one can test for equality by computing $f_2 f_1^{-1}$, and checking whether this form is in the principal cycle at distance (d_2-d_1) *modulo* R^+/n. If one does not know any distance, we know nothing better to do than a rigorous search in the principal cycle (e.g. by means of a baby-giant-step strategy). This is a time comsuming operation and turns the algorithm into a $\Delta^{1/4+\varepsilon}$-algorithm. Fortunately, there is some "trade-off": if h^+ is very small, R^+ is very large, and searching for a form in the principal cycle is very expensive. However if R^+ is very large, h^+ is very accurately determined by formula (17); perhaps h^+ is even known with certainty and one can stop the calculations after having determined R^+, if one is satisfied with the class number without knowing the structure of the class group and without knowing explicit generators of the class group. On the other hand. if h^+ is large and R^+ is small, more searching will be necessary, but this is not so expensive since R^+ is small i.e. the principal cycle is short. We can compute the class number in time bounded by $O(\Delta^{1/5+\varepsilon})$; computing the class group can be done in time bounded by $O(\Delta^{1/4+\varepsilon})$.

- Finally, notice that, in contrast to the computed value of the class number, the regulator R, once it is determined, is known with *certainty*, i.e. without any assumption of a generalized Riemann hypothesis. This hypothesis was only used to guarantee termination in $O(\Delta^{1/5+\varepsilon})$ computing time.

6. FACTORIZATION

In this section we will discuss two deterministic factorization algorithms based on computations in class groups of complex quadratic orders and on computations in the principal cycles associated to class groups of real quadratic orders respectively.

If $N \in \mathbb{Z}_{>1}$ denotes the number that will be factored, then, on assumption of certain generalized Riemann hypotheses (GRH), both algorithms run in time bounded by $N^{1/5+\varepsilon}$ for all $\varepsilon > 0$.

First we briefly indicate how the algorithms discussed in the previous sections are related to factorization algorithms.

Let Δ denote the discriminant of a complex quadratic order $\mathcal{O}$. By an *ambiguous* form f in the class group of $\mathcal{O}$ we mean a form f for which $f^2 = 1$ holds. The ambiguous forms make up a subgroup of the class group; they have the following shape:

$$f = (a,\pm a,c) \quad \text{or} \quad (a,b,c) \quad \text{or} \quad (a,0,c).$$

In other words, the ambiguous forms are precisely the forms that correspond to ideal classes

$$\{(\mathbb{Z} + \frac{b+\sqrt{\Delta}}{2a}\,\mathbb{Z})\,\alpha\colon\ \alpha \in K^{\times}\}$$

with $b+\sqrt{\Delta}/2a$ on the imaginary axis or on the edge of the funamental domain. Every ambiguous form gives rise to a factorization of Δ:

$$\begin{aligned} f = (a,\pm a,c) &: \Delta = a(a-4c) \\ f = (a,b,a) &: \Delta = (b+2a)(b-2a) \\ f = (a,0,c) &: \Delta = -4ac. \end{aligned}$$

Conversely it is possible to reveal, in an efficient way, the complete factorization of Δ into prime powers from the subgroup of ambiguous forms in $\mathcal{Cl}(\mathcal{O}(\Delta))$ cf. [18].

Briefly, the algorithm that is based on computations in the class groups of complex quadratic orders consists of computing an ambiguous form in $\mathcal{Cl}(\mathcal{O}(-N))$ of $N \equiv 3 \pmod 4$ resp. in $\mathcal{Cl}(\mathcal{O}(-3N))$ if $N \equiv 1 \pmod 4$. To find this ambiguous form takes about as much effort as it takes to compute $\mathcal{Cl}(\mathcal{O})$ following the strategy discussed in Section 3.

Next, let Δ denote the discriminant of a real quadratic order. Ambiguous forms are defined to be forms of order ≤ 2, in F; ambiguous forms f have the following shape:

$$f = (a,b,c) \quad \text{where } a|b,$$

so like in the complex case, an ambiguous form provides us with a factor of Δ.

In the principal cycle there are two reduced ambiguous forms: the principal form $(1,\Delta,\Delta^2-\Delta/4)$ and one diametrically opposite to it at distance $\frac{1}{2}\,R^+$. The algorithm computes these ambiguous forms at distance $\frac{1}{2}\,R^+$ on the principal cycles $G(\mathcal{O}(\Delta))$ for suitable multiples Δ of N.

Note that if (a,b,c) is any quadratic form on the principal cycle $G(\mathcal{O}(\Delta))$, it holds that there are $X,Y \in \frac{1}{2}\,\mathbb{Z}$ such that $N(X+Y\sqrt{\Delta}) = a$, i.e. $X^2 - \Delta Y^2 = a$, whence for all $q|\Delta$ with $\gcd(q,2a) = 1$ we have that $(\frac{a}{q}) = 1$. For instance, if $(-1,\Delta,\Delta-\Delta^2/4)$ is the reduced ambiguous form at distance

$\frac{1}{2}$ R^+, it holds that $(\frac{-1}{q}) = 1$ for every odd prime dividing Δ i.e., all odd prime dividing Δ are congruent to 1 (mod 4).

Before entering into a more detailed discussion of the algorithms, we quote some theorems from analytic number theory, which at present can only be proved on assumption of certain generalized Riemann hypotheses.

THEOREM 6.1. (GRH). *There exists an absolute, effectively computable constant* $C_1 > 0$ *such that for every finite extension* K *of* $\mathbb{Q}$ *and every Dirichlet character* χ *of* K, *there exists a prime ideal* p *of* K *of degree* 1 *with*

$$\chi(p) \neq 1 \text{ or } 0 \quad \text{and} \quad N_{K/\mathbb{Q}}(p) < C_1 \log^2 (|\Delta_{K/\mathbb{Q}} N_{K/\mathbb{Q}}(\text{cond } \chi)|)$$

PROOF. Cor. 1.3 of Theorem 1.2 of [15].

One needs the Riemann hypothesis for the zeta function of K and for $L(s,\chi)$: if ρ is any zero of $\zeta_K(s)$ or $L(s,\chi)$ with $0 < \text{Re}\,\rho < 1$, we assume that $\text{Re}\,\rho = \frac{1}{2}$. □

COROLLARY 6.2. (GRH). *Let* $\mathcal{O}$ *be a complex quadratic order of discriminant* Δ, *then* $Cl(\mathcal{O})$ *is generated by quadratic forms* (p,b,c) *of discriminant* Δ *and* p prime *with* $(\frac{\Delta}{p}) = 1$ *and* $p < C_1 \log^2 |\Delta|$.

PROOF. Let $\Delta = f^2 D$ with D the discriminant of a complex quadratic number field. Let G_f denote the ray class group mod f of K. We have a surjective map

$$G_f \twoheadrightarrow Cl(\mathcal{O}(\Delta))$$

via

$$[\underline{a}] \longrightarrow [\underline{a} \cap \mathcal{O}].$$

Here we used the correspondence between equivalence classes of primitive quadratic forms of discriminant Δ and classes of invertible $\mathcal{O}(\Delta)$-ideals. We apply Theorem 6.1 to K and all characters of G_f.

Let H denote the subgroup of G_f generated by the image of the classes of prime ideals p of degree 1 in K and for which

$$N(p) < C_1 \log^2 |D \cdot N_{K/\mathbb{Q}} f| = C_1 \log^2 |\Delta|,$$

holds. The group H then equals G_f because, if $H \not\subset G_f$ we can find a nontrivial character χ of G_f with $H \subset \ker \chi$, but this contradicts Theorem 6.1, since all characters of G_f have conductor dividing f. This proves Corollary 6.2. □

The following theorem gives us an estimate of the rate of convergence of the product expansion of L-series at 1. The proof is along the lines of the proofs in [15] but, since the result we need is not explicitly stated there we will give an outline of a proof below.

THEOREM 6.3. (GRH). *There exists absolute, effectively computable positive constants* C_2 *and* C_3 *such that for all* Δ, *discriminants of quadratic orders and for all* $x > C_2 \log^2 |\Delta|$ *it holds that*

$$\left|1 - \prod_{p>x} \left(1 - \left(\frac{\Delta}{p}\right)\frac{1}{p}\right)\right| < \frac{C_3 \log|\Delta x|}{\sqrt{x}} .$$

Theorem 6.3 is a specialization of a more general theorem. In the proof one assumes the Riemann hypothesis for $L(s,\chi)$, where $\chi(p) = \left(\frac{\Delta}{p}\right)$. All O-symbols that occur in the proof below are absolute and effectively computable.

PROOF OF THEOREM 6.3. Let χ denote the Dirichlet character $\left(\frac{\Delta}{\cdot}\right)$.

def. for $n \in \mathbb{Z}_{\geq 1}$, $\Lambda(n,\chi) = \chi(p^k) \log p$ if $1 \neq n = p^k$ a prime power
$= 0$ otherwise.

def. for $x \in \mathbb{R}_{\geq 1}$: $\psi_1(x,\chi) = \sum_{n \leq x} (x-n)\Lambda(n,\chi)$.

Initially we assume that Δ is a fundamental discriminant i.e. Δ is the discriminant of a number field.

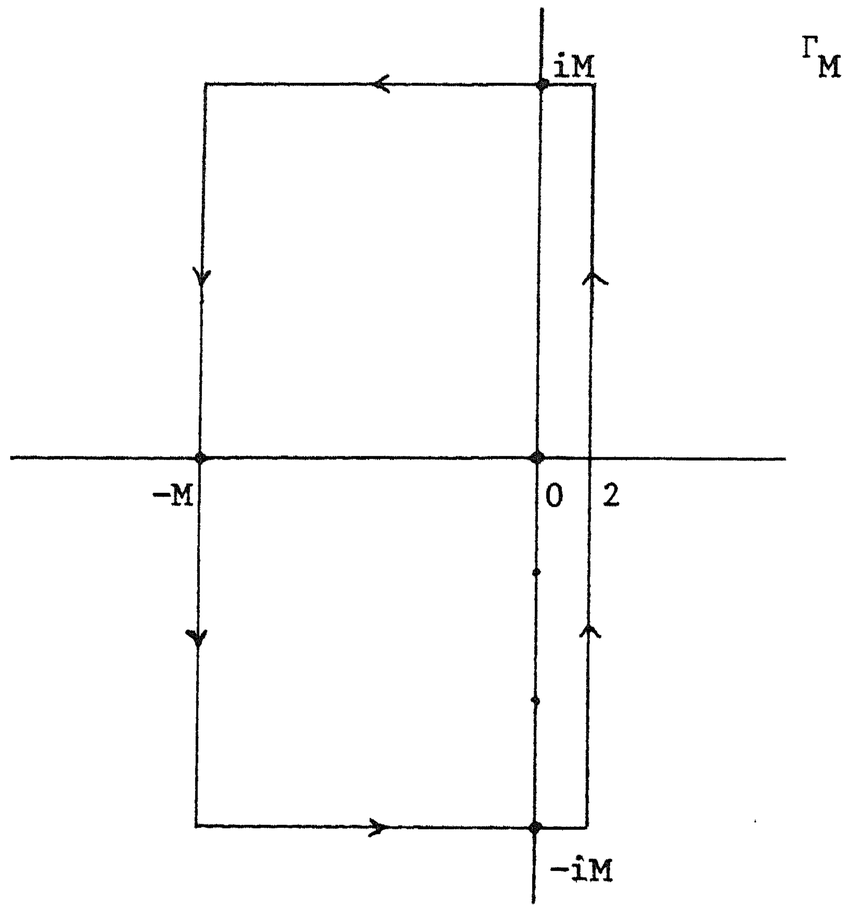
It holds that for $x \in \mathbb{R}_{\geq 1}$

$$\psi_1(x,\chi) = \frac{1}{2\pi i} \int_{2-i\infty}^{2+i\infty} -\frac{L'}{L}(s,\chi) \frac{x^{s+1}}{s(s+1)}\, ds$$

which follows by integrating term by term.

The right hand side of this equation can also be evaluated by computing

$$\frac{1}{2\pi i} \int_{\Gamma_M} -\frac{L'}{L}(s,\chi) \frac{x^{s+1}}{s(s+1)}\, ds$$



for suitable M by applying the residue theorem and by letting $M \to \infty$. One finds

$$\psi_1(x,\chi) = -\sum_{\rho} \frac{x^{\rho+1}}{\rho(\rho+1)} + b_0 - a_0 x + ax$$

$$- \frac{a}{2}\log\left((x-1)^{x-1}(x+1)^{x+1}\right) - \frac{b}{2}\log\left(\frac{(x-1)^{x-1}}{(x+1)^{x+1}}\right) .$$

Here

$$\frac{L'}{L}(s,\chi) = \frac{a}{s} + a_0 + \ldots \qquad \text{near } 0$$

$$\frac{L'}{L}(s,\chi) = \frac{b}{s+1} + b_0 + \ldots \qquad \text{near } -1;$$

it holds that $(a,b) = (1,0)$ if $\Delta > 0$ and $(a,b) = (0,1)$ if $\Delta < 0$. The sum is taken over all ρ, zeros of $L(s,\chi)$ with $0 < \mathrm{Re}\,\rho < 1$.

Subtracting $\psi_1(x,\chi)$ from $\psi_1(x+1,\chi)$ one finds

$$\sum_{n\le x} \Lambda(n,\chi) = (x-[x])\Lambda(n+[x],\chi) - \sum_{\rho \text{ zero}} \frac{(x+1)^{\rho+1}-x^{\rho+1}}{\rho(\rho+1)} - a_0$$

$$- \frac{a}{2}\left(\log \frac{(x+2)^{x+2}x^x}{(x+1)^{x+1}(x-1)^{x-1}} - 2\right) - \frac{b}{2}\left(\log \frac{(x+1)^{x+1}x^x}{(x+2)^{x+2}(x-1)^{x-1}}\right)$$

from which it follows that

$$\left|\sum_{n\le x} \Lambda(n,\chi) + a_0\right| \le \left|\sum_{\rho} \frac{(x+1)^{\rho+1}-x^{\rho+1}}{\rho(\rho+1)}\right| + 2\log(x+1).$$

Let $N(t)$ denote the number of zeros ρ of $L(s,\chi)$ with $0 < \mathrm{Re}\,\rho < 1$ and $t-1 \le \mathrm{Im}\,\rho \le t+1$. Hence, we have

$$N(t) = O(\log(\Delta(|t|+2)))$$

cf. Lemma 5.4 of [15]. Using this estimate and the inequality

$$\left|\sum_{\rho} \frac{(x+1)^{\rho+1}-x^{\rho+1}}{\rho(\rho+1)}\right| \le \sum_{|\mathrm{Im}\rho|\le x} \frac{\sqrt{x+1}}{|\rho|} + \sum_{|\mathrm{Im}\rho|>x} \frac{2(x+1)^2}{|\rho(\rho+1)|}$$

it is not difficult, assuming the Riemann hypothesis for $L(s,\chi)$, to arrive at

$$\left|\sum_{\rho} \frac{(x+1)^{\rho+1}-x^{\rho+1}}{\rho(\rho+1)}\right| = O(\sqrt{x}\,\log|\Delta x|\,\log x).$$

Finally we find that

$$\left|\sum_{n\le x} \Lambda(n,\chi)\right| = O(\sqrt{x}\log|\Delta x|\log x),$$

(disposing of a_0 by using our estimates for $x = \frac{3}{2}$). It is easy to prove that the estimate also holds for non-fundamental Δ (using the estimate for fundamental Δ to be sure; one may arrive at larger (absolute) constants); so from now on we let Δ be an arbitrary discriminant of a quadratic order.

By partial summation we find

$$\sum_{n>x} \frac{\Lambda(n,\chi)}{n\log n} = \sum_{n>x} \left(\sum_{k=2}^{n} \Lambda(k,\chi)\right)\left(\frac{1}{(n+1)\log(n+1)} - \frac{1}{n\log n}\right)$$

and

$$|\sum_{n>x} \frac{\Lambda(n,\chi)}{n \log n}| = O(\frac{\log|\Delta x|}{\sqrt{x}}).$$

We have that

$$|\sum_{n>x} \frac{\Lambda(n,\chi)}{n \log n} - \sum_{p>x} \log(1 - \frac{\chi(p)}{p})|.$$

$$= |\sum_{k=2}^{\infty} \sum_{k\sqrt{x}<p\leq X} \frac{\chi(p)^k}{kp^k}|$$

$$\leq \sum_{k>[^2\log x]} \sum_{p\leq x} \frac{1}{kp^k} + \sum_{k=2}^{[^2\log x]} \sum_{p>k\sqrt{x}} \frac{1}{kp^k} = O(\frac{1}{\sqrt{x}}),$$

and it follows that

$$|\sum_{p>x} \log(1 - \frac{\chi(p)}{p})| = O(\frac{\log|\Delta x|}{\sqrt{x}}).$$

From this estimate one deduces that there exist absolute, effectively computable positive constants C_2 and C_3 such that if $x > C_2 \log^2 |\Delta|$ then

$$|1 - \prod_{p>x} (1 - \frac{\chi(p)}{p})| < \frac{C_3 \log|\Delta x|}{\sqrt{x}}.$$

which proves the theorem. □

Next we present the algorithms.

ALGORITHM 6.4. Factorization algorithm based on computations in class groups of complex quadratic orders.

Let $N \in \mathbb{Z}_{>1}$ denote the number to be factored.

Step 1. Test whether $\gcd(N,6) > 1$ or whether N is a proper power of an integer; if this is the case, we can either factor N or decide it is prime; otherwise, if $N \equiv 3 \pmod 4$ put $\Delta = -N$ and if $N \equiv 1 \pmod 4$ put $\Delta = -3N$. In both cases Δ is the discriminant of a complex quadratic order.

Step 2. Compute

$$\tilde{h} = \frac{\sqrt{|\Delta|}}{\pi} \prod_{p \le X} (1 - (\frac{\Delta}{p})\frac{1}{p})^{-1}$$

with an accuracy of $\log|\Delta|$ significant decimal digits; the product is taken over all primes $p \le X = \max(|\Delta|^{1/5}, C_2 \log^2 |\Delta|)$.

Next for successive primes $p \ge 2$ with $(\frac{\Delta}{p}) = 1$ do the following step until either a factorization of N is found or $p \ge C_1 \log^2 |\Delta|$; if the latter occurs one concludes that N is prime.

Step 3. Compute a quadratic form $f = (p,b,c)$ and compute a multiple of its order using the estimate $\tilde{h} \approx h(\mathcal{O}(\Delta))$ obtained in Step 2 and the baby-giant-step strategy discussed in Section 3. If $N \equiv 3 \pmod 4$, compute the form of order two in the cyclic group generated by f; if a form of order two is actually existing one obtains a nontrivial factorization of N. If $N \equiv 1 \pmod 4$ denote by H a subgroup of the class group of $\mathcal{O}(\Delta)$ which initially, i.e. before entering Step 3, equals $\{(1,1,1-\Delta/4),(3,3,N+3/4)\}$. Compute a form g, a generator of the 2-primary part of the cyclic group generated by f; compute the group generated by H and g and call it H again. If for some p the group H "becomes" non-cyclic, there are three forms of order 2 in H and those different from (3,3,N+3/4) give rise to a nontrivial factorization of N.

This completes the description of the algorithm.

The algorithm is correct by genus theory and Corollary 6.3: If N passes the tests in Step 1 we can be sure that, if N is composite, there exists a form of order 2 in $Cl(\mathcal{O}(\Delta))$ which gives a nontrivial factorization of N; for details see [12]. By Cor. 6.2 the class group is generated by forms (p,b,c) with $(\frac{\Delta}{p}) = 1$ and $p < C_1 \log^2 |\Delta|$, so, if we did never find a form of order 2 in Step 3 of the algorithm we can be sure that no such form exists i.e. that N is a prime.

A brief running time analysis runs as follows: Step 1 is polynomial in $\log N$; Step 2 takes time $O(N^{1/5+\varepsilon})$ as explained in Section 3. Theorem 6.2 and the class number formula imply that $|h(\Delta) - \tilde{h}| = O(N^{2/5+\varepsilon})$, so the baby-giant-step strategy in Step 3 takes $O(N^{1/5+\varepsilon})$; The rest of the computations in Step 3 is polynomial in $\log N$: computing a form (p,b,c) can be done in time $O(p \log N) = O(\log^3 N)$ and computing a generator of the 2-primary part of the cyclic group generated by f and computing a form therein can be done by evaluating certain powers of f which takes time polynomial in $\log N$; all computations concerning the group H can be done in time polynomial in $\log N$.

Since Step 3 is repeated at most $C_1 \log^2|\Delta|$ times, we conclude that the algorithm takes time $O(N^{1/5+\varepsilon})$ for all $\varepsilon > 0$.

The algorithm uses memory proportional to $N^{1/5+\varepsilon}$ to store all the baby-steps.

ALGORITHM 6.5. Factorization algorithm based on computations in the principal cycles $G(\Delta)$ of the groups $F(\Delta)$.

Let $N \in \mathbb{Z}_{>1}$ denote the number to be factored.

Step 1. Test whether N is divisible by the primes $\leq (4C_1 \log^2(8N))^2$ and test whether N is a proper power of an integer. If this is the case we can factor N or decide it is prime, otherwise we know that

$$N > (4C_1 \log^2(8N))^2$$

We distinguish two cases:

Case $N \equiv 3 \pmod 4$: For successive primes $p \equiv 3 \pmod 4$ let $\Delta = pN$: the discriminant of a real quadratic order and do the following steps until either a factorization of N is found or $p > C_3 \log^2 N$; if the latter occurs one concludes that N is prime.

Step 2. Compute $\tilde{R} = \sqrt{\Delta} \prod_{p \leq X} (1 - (\frac{\Delta}{p})\frac{1}{p})^{-1}$ with an accuracy of $\log \Delta$ significant decimal digits; the product is taken over all primes $\geq \max(\Delta^{1/5}, C_2 \log^2 \Delta)$.

Step 3. Find a multiple of $R^+(O(\Delta))$ using the baby-giant-step strategy as discussed in Section 3. Compute the ambiguous form g at distance $\frac{1}{2} R^+$ on the principal cycle; if $g \neq (-1,b,c)$ or $(\pm p,b,c)$ then g gives rise to a nontrivial factorization of N.

Case $N \equiv 1 \pmod 4$:

Step 2. First put $\Delta = N$, the discriminant of a quadratic order; compute $\tilde{R}$ (step 2), find a multiple of $R^+(O(N))$ and compute the ambiguous form g at $\frac{1}{2} R^+$ on G(N); if $g \neq (-1,N,N^2-N/4)$, one obtains a nontrivial factorization of N; otherwise do the following:

For successive pairs of primes $p_1, p_2 \equiv 3 \pmod 4$ and $p_1 < p_2 < 4c_1 \log^2(8N)$ (successive in the sense that the products $p_1 p_2$ form an increasing sequence) put $\Delta = p_1 p_2 N$, the discriminant of a real quadratic order. Do

the following steps until a factorization of N is found; if this does not happen for the finitely many pairs (p_1,p_2) one concludes that N is prime.

Step 3. Compute $\tilde{R} = \sqrt{\Delta}\ \Pi_{p\le X}\ (1 - (\frac{\Delta}{p})\frac{1}{p})^{-1}$ as before.

Step 4. Find a multiple of $R^+(\mathcal{O}(\Delta))$ and the ambiguous form $g = (a,b,c)$ at $\frac{1}{2}R^+$ on the principal cycle; if $a \neq -1, \pm p_1, \pm p_1p_2$ the form g gives rise to a nontrivial factorization of N.

This completes the description of the algorithm.

To prove correctness we distinguish the two cases again:

Case $N \equiv 3 \pmod 4$: Assume that N is composite and that N passed the tests in Step 1. Let q denote a prime congruent to 3 (mod 4) that divides N.

LEMMA (GRH). *There exists a prime* $\equiv 3 \pmod 4$ *satisfying* $(\frac{p}{N/q}) = -1$ *and* $p < C_1 \log^2(4N)$.

PROOF. Apply Theorem 6.1 to the (non-primitive) quadratic character χ of $K = \mathbb{Q}(\sqrt{-N/q})$ belonging to the extension $K(i)/K$ of conductor (2). By Theorem 6.1 there exists a prime $\mathfrak{p}$ of K of degree 1 with $\chi(\mathfrak{p}) = -1$ and

$$N(\mathfrak{p}) < C_1 \log^2(\Delta_K N_{K/\mathbb{Q}}(\text{cond } \chi)) \le C_1 \log^2(4N).$$

Let $p = N(\mathfrak{p})$ then p splits in $\mathbb{Q}(\sqrt{-N/q})$ (since $p \neq 2$) and we have $(\frac{-1}{p}) = (\frac{p}{N/q}) = -1$. □

Let $\Delta = pN$ with p a prime as in the lemma; then the reduced form $g = (a,b,c)$ at $\frac{1}{2}R^+$ in $G(\Delta)$ cannot have $a = \pm N$ since the fact that g is reduced implies that $N < \sqrt{\Delta}$ i.e. $p > N$ which contradicts $p < C_1 \log^2(4N)$ and the fact that N passed the tests in Step 1. Nor can g have $a = -1$ or $\pm p$ since we have that $(\frac{-1}{p}) = -1$ and $(\frac{\pm p}{N/q}) = -1$. We conclude that for this p we will encounter a nontrivial factorization of N in Step 3 of the algorithm.

Case $N \equiv 1 \pmod 4$: Assume that N is composite and that N passed Steps 1 and 2 of the algorithms. This implies inter alia that all divisors of N are congruent to 1 (mod 4): let q_1 and q_2 be *distinct* primes dividing N.

LEMMA (GRH). *There exists two primes* $p_1,p_2 \equiv 3 \pmod 4$ *satisfying* $(p_1/q_1) = -1$; $(p_2/q_1) = 1$ *and* $(p_2/q_2) = -1$ *and* $p_1,p_2 \le 4C_1 \log^2 8N$.

PROOF. For p_1 consider the non-primitive character χ of $K = \mathbb{Q}(\sqrt{-q_1})$ of conductor (2) belonging to K(i)/K and for p_2 consider the non-primitive character χ of $L = \mathbb{Q}(\sqrt{q_1}, \sqrt{-q_2})$ of conductor (2) belonging to L(i)/L. We have $|\Delta_K| = 4q_1$ and $|\Delta_L| = (4q_1q_2)^2$. As in the proof of the lemma in the case $N \equiv 3 \pmod 4$ we can find p_1 and p_2 smaller than

$$C_1 \log^2((4q_1q_2)^2 \cdot 2^4) < 4C_1 \log^2(8N).$$

This proves the lemma.

Let $\Delta = p_1p_2N$ with p_1 and p_2 a pair of primes as in the lemma; the reduced form $g = (a,b,c)$ at $\frac{1}{2}R^+$ in $G(\Delta)$ cannot have $a = \pm N, \pm p_1N, \pm p_2N$ and $\pm p_1p_2N$ since this implies $p_1p_2 > N$ whence $(4C_1 \log^2(8N))^2 > N$ which contradicts the fact that N passed Step 1 of the algorithm. Nor can g have $a = -1, \pm p_1, \pm p_2$ or $\pm p_1p_2$ since $(-1/p_1) = -1$; $(\pm p_1/q_1) = -1$; $(\pm p_2/q_2) = -1$ and $(\pm p_1p_2/q_1) = -1$ respectively. We conclude that for this pair (p_1,p_2) the form g provides us with a nontrivial factorization of N. This finishes the proof of the correctness of the algorithm.

We leave a running rime analysis to the reader; it is analogous to the analysis of the running time of the algorithm based on computations of class groups of complex quadratic orders.

This finishes the description of the algorithms.

7. IRREGULAR CLASS GROUPS

In this section we will consider the structure of the class groups of quadratic orders.

First some terminology: for a finite abelian group A, and a prime p, the minimal number of generators of the p-Sylow subgroup of A is called *the p-rank of* A, *notation* d_pA.

By C(n) we denote the cyclic group of n elements. For instance, $d_p(C(n)) = 1$ for all primes p, that divide n.

DEFINITION. Let $\mathcal{O}$ be a quadratic order and let $Cl(\mathcal{O})$ be its class group. We call $Cl(\mathcal{O})$ *irregular* if $Cl^2(\mathcal{O})$ is non-cyclic, or, equivalently, if $d_pCl^2(\mathcal{O}) \geq 2$. We call $d_pCl^2(\mathcal{O})$ the *exponent* of p-irregularity.

REMARK. If p is an *odd* prime, then $Cl(\mathcal{O})$ is p-irregular iff $d_pCl(\mathcal{O}) \geq 2$.

Although inspection of a list of class groups of orders of small discriminantes might suggest differently, irregular class groups do exist! For instance the class group of the order of discriminant -3299 is isomorphic to $C(3) \times C(3)$. Gauss considered the phenomenon of irregularity to be of great importance [12]:

> *Hoc argumentum, quod ad arithmeticae sublimioris mysteria maxima recondita pertinere, disquisitionibusque difficillimis locum relinquere videtur, paucis tantum observationibus hic illustrare possumus,... .*

In his "Disquisitiones Arithmeticae", Gauss considers irregular class groups of both maximal and non-maximal orders. For non-maximal orders there is for obvious reasons more irregularity, and indeed Gauss found many examples of this kind. Here we will confine ourselves to class groups of maximal orders i.e. class groups of quadratic number fields.

Recently D.A. BUELL [3] made a list of class groups of complex quadratic fields with discriminant > -4000000; it is the largest list available up to now, and it appeared that 95.74% of the listed class groups had a cyclic subgroup of squares, i.e. 95.74% of the class groups were p-regular *for all primes* p. So it seems, that, for complex quadratic fields, irregular class groups are rare, and, as it turns out, even rarer for real quadratic fields.

Let us first consider complex quadratic fields. It is easy to construct 2-irregular class groups with a high exponent of irregularity, e.g. as follows: Let

$$\Delta_1 = -3 \cdot 13,$$

$$\Delta_{k+1} = \Delta_k \cdot p \text{ with } p \text{ the smallest prime } \equiv 1 \pmod 4 \text{ such that } \left(\frac{p}{q}\right) = 1 \ \forall q \mid \Delta_k.$$

It can be proved, that $d_2 Cl^2 \mathbb{Q}(\sqrt{\Delta_k}) = k$. For example

$$\Delta_3 = -3.13.61.601 = -1429779$$

and $\mathbb{Q}(\sqrt{\Delta_3})$ has a class group $\cong C(4) \times C(4) \times C(4) \times C(5)$. So the exponent of 2-irregularity equals 3.

It turns out to be very hard, to construct p-irregular class groups for odd p. The only example. I could find in Gauss' "Disquisitiones", was the

maximal order of discriminant -9748. (determinant = -2437 in his terminology); the class group is isomorphic to $C(3) \times C(3) \times C(2)$.

In the beginning of the 20th century some more examples of p-irregular class groups, with p odd, were known, but, it seems, always 3-irregular of exponent 2.

In 1936, G. PALL [26] seemed to have obtained the first new result on the matter since more than a century: he claimed, that the field $\mathbb{Q}(\sqrt{-12379})$ has a class group isomorphic to $C(5) \times C(5)$.

However, 25 years later, in 1961, LIPPMANN [19] proved that the class group of $\mathbb{Q}(\sqrt{-12379})$ is cyclic of order 25. Lippmann used a computer; he also gave some correct examples of 5-irregular and 7-irregular class groups, viz.

$$Cl(\mathbb{Q}(\sqrt{-12451})) \simeq C(5) \times C(5) \times C(2)$$

$$Cl(\mathbb{Q}(\sqrt{-63499})) \simeq C(7) \times C(7).$$

Lippmann also searched for 11-irregular class groups, but was not successful in this case.

In 1970, YAMAMOTO [42] proved, for all $n \in \mathbb{Z}_{\geq 1}$, the existence of infinitely many complex quadratic fields with $C(n) \times C(n)$ as a subgroup of their class groups. A trivial consequence is, that for all primes p, infinitely many p-irregular class groups exist.

Yamamoto gave his fields explicitly; he parametrized their discriminants by a polynomial of degree 2n. Consequently, the discriminants of his fields, having p-irregular class groups with p large, are huge.

In 1971, SHANKS [32] using the algorithm discussed in Section 3, found that

$$Cl(\mathbb{Q}(\sqrt{-564552759})) \simeq C(3) \times C(3) \times C(3) \times C(604),$$

$(564552759 = 3(3^6 + 4.19^6))$.

It is the first example of a p-irregular class group with p odd and index of p-irregularity > 2. From then on, things go a *bit* faster; some theoretical results are obtained by M. CRAIG [4,5], who proves the existence of infinitely many complex quadratic fields K with $d_3 Cl(K) \geq 4$, and many explicit examples are found by SHANKS and others [6,7,8,9,10,24,28, 34,35,36,38,39]. At present the, perhaps diappointing, state of affairs is:

p	d	−Δ	
3	2	3299	[3]
	3	3321607	[3,7]
	4	653329427	[6]
5	2	11199	[3]
	3	18397407	[11,28]
	4	258559351511807	[28]
7	2	63499	[3]
	3	4805446123032518648268510536	[39]
11	2	65591	[3]
13	2	228679	[3]
17	2	1997799	[3]
19	2	373391	[3]
23	2		
↓	↓		

where

d: an integer ≥ 2, for which a class group $C\ell(O)$ is known to exist with $d_p C\ell(O) = d$.

Δ: smallest known discriminant with $d_p C\ell(O) = d$. For the p-rank = 2 cases and the 3-rank = 3 case, these discriminants have been *proved* to be minimal (in absolute value).

All examples, with p-rank = 2, have been taken from BUELL's list [3]. The 7-rank = 3 example has been found by J. Solderitsch; he used polynomials

$$D_p(s,t) = s^{2p} - 6(st)^p + t^{2p}$$

taking p = 7 and (s,t) = (87,85) gives the example. The 5-rank = 4 case was found by myself by means of ideas of J.F. MESTRE [21].

Perhaps it shouls be indicated that the first examples of class groups with high p-rank, for odd p, were usually very large; for instance, the example of the class group with 7-rank equal to 3, is, at present, the only example known; it is not unlikely, that (say) a 10-digit discriminant exists with the same property, but this one has to be found yet. On the other hand, the 3-rank = 3 example, given here, is much smaller than the first one found by Shanks. At present, many examples of class groups with 3-rank = 3,4 or 5-rank = 3 are known.

As far as the real quadratic fields are concerned, the situation is perhaps even more disappointing.

As in the complex case we'll confine ourselves to p-irregular class groups with p *odd* of *maximal* orders. In [12] GAUSS says, that he did not encounter any example of a real quadratic order which is p-irregular for odd p; he also expresses his firm belief in the existence of these orders, and he was right.

In 1936, PALL [26] gives the first (correct) example: the discriminant 62501 determines a maximal order with class group isomorphic to $C(3) \times C(9)$. In 1972 SHANKS [32] finds the prime $188184253 = 3^6 + 4.19^6$; the field $Q(\sqrt{188184253})$ has class group isomorphic to

$$C(3) \times C(3) \times C(3).$$

In fact, Shanks used this example and an old theorem of SCHOLZ [27], that connects the 3-ranks of the class groups of $\mathbb{Q}(\sqrt{\Delta})$ and $\mathbb{Q}(\sqrt{-3\Delta})$, to construct his example of the complex quadratic field with 3-rank of its class group equal to 3, that was mentioned above. Scholz's theorem implies, that every example of a real quadratic field having a class group with high exponent of 3-irregularity implies an example of a complex quadratic field, with the same property and vice versa. This explains why we know at least some examples of 3-irregular class groups of real quadratic fields. The state of affairs is:

p	d	Δ	
3	2	32009	[30]
3	3	39345017	[6]
3	4	1284062551036124923952823484951333 365764948104727718257285040631602 27	
		161873462515321376471501957999729 57	[10]
5	2	1129841	[16]
7	2	2068117	[16]
11			

In his thesis, Diaz y Diaz announces a proof of the existence of infinity many real quadratic fields, admitting $C(3) \times C(3) \times C(3) \times C(3)$ as a subgroup of their class groups, but his proof has not yet been published cf. [9].

Finally, we'll explain how the example of a complex quadratic field K with $d_5 C\ell(K) = 4$, that is given above, was found. In order to do this, we'll

sketch, how certain polynomials $M_E(t) \in \mathbb{Z}[t]$, may be derived from a Weierstrass equation of an elliptic curve E, which is defined over $\mathbb{Q}$; these polynomials are used to parametrize discriminants of quadratic fields. The computations are based on work of J.F. MESTRE [21].

Let E be an elliptic curve defined over an algebraic number field K; assume that P is a K-rational point on E and that the order of P is n. Let F be the elliptic curve E/<P>; then F is K-rational and there is a K-isogeny $E \overset{\phi}{\to} F$. If $Q \in F$ is K-rational and $R \in \phi^{-1}(Q)$ (not necessarily K-rational) then K(R)/K is an *unramified cyclic extension of degree* n, on the condition that

(i) Q is not singular modulo any prime of K; this condition guarantees that the extension K(R)/K is unramified.

(ii) A rather involved condition, which guarantees that K(R)/K is of degree n; we do not give the precise condition, since, numerically, it is not a very interesting one.

Recall, that by class field theory, the fact that K(R)/K is unramified cyclic of degree n implies that $Cl(K) \twoheadrightarrow C(n)$. We shall apply the above to elliptic curves defined over $\mathbb{Q}$:

$$F: Y^2 + a_1XY + a_3Y = X^3 + a_2X^2 + a_4X^2 + a_4X + a_6.$$

Let $x \in \mathbb{Q}$ and find y such that $(x,y) \in F$; the number y will be in a quadratic number field K, and we'll apply the above to K. Conditions (i) and (ii) boil down to simple congruence conditions on x.

Mestre's idea is, to find two different points, Q_1 and Q_2 on F that satisfy the above conditions. By submitting Q_1 and Q_2 to certain conditions, he can prove that for n prime, the group $C(n) \times C(n)$ is a subgroup of $Cl(K)$. We do not bother about all of these conditions, since computations suggest, that perhaps they are too stringent.

The computation of $M_F(t)$: F is given by

$$\eta^2 = \xi^3 - \frac{c_4}{48}\xi - \frac{c_6}{864}$$

cf. [22]. Assume $\xi_1 \neq \xi_2$ and

$$\eta^2 = \xi_1^3 - \frac{c_4}{48}\xi_1 - \frac{c_6}{864} = \xi_2^3 - \frac{c_4}{48}\xi_2 - \frac{c_6}{864}\,. \tag{19}$$

We take $Q_1 = (\xi_1,\eta)$ and $Q_2 = (\xi_2,\eta)$ the two (different) points on F and we want to compute $\mathbb{Q}(\eta)$. Equation (19) becomes

$$\xi_1^2 + \xi_1\xi_2 + \xi_2^2 = \frac{c_4}{48} \,. \tag{20}$$

Now, if c_4 is the norm of a number in $\mathbb{Q}(\zeta_3)$, the curve (20) is a non-empty rational conic and it can be parameterized e.g. if $\alpha^2 + \alpha\beta + \beta^2 = c_4$, by

$$\xi_1(t) = \frac{1}{12}\,\frac{-\beta t^2+2\alpha t+(\alpha+\beta)}{t^2+t+1}\,, \qquad t \in \mathbb{P}_1(\mathbb{Q}),$$

$$\xi_2(t) = \frac{1}{12}\,\frac{(\alpha+\beta)t^2+2\beta t-\alpha}{t^2+t+1}\,, \qquad t \in \mathbb{P}_1(\mathbb{Q}). \tag{21}$$

Substituting (21) in (19), one easily finds that

$$\mathbb{Q}(\eta) = \mathbb{Q}(\sqrt{M_F(t)})$$

with $M_F(t) \in \mathbb{Z}[t]$ of degree 8.

EXAMPLE. If we take $E \overset{\phi}{\to} F$ to be $X_1(11) \to X_0(11)$, (or $11A \overset{\phi}{\to} 11B$ in the notation of [22]) then this is a 5-isogeny $E \overset{\phi}{\to} F$, with a point of order 5 in $\ker \phi$. We find, that, up to a square,

$$M_{11B}(t) = -(t^2+t+1)(47t^6+21t^5+598t^4+1561t^3+1198t^2+261t+47).$$

Condition (i) now becomes:

$$t \not\equiv 2,-4,4 \pmod{11}.$$

Substituting special values for t we find:

t	Δ	$Cl(\mathbb{Q}(\sqrt{\Delta}))$
1	-11199	$C(5) \times C(5) \times C(4)$
1/4	-18397407	$C(5) \times C(5) \times C(5) \times C(2) \times C(8)$
14/25	-258559351511807	$C(5) \times C(5) \times C(5) \times C(5) \times C(2) \times C(4) \times C(2957)$

These are precisely the examples given above. By taking other elliptic curves, one can search for other types of class groups. In particular, it is possible, to get information on class groups of real quadratic fields as well.

On the other hand, the method is limited in the sense that, using elliptic curves that are defined over $\mathbb{Q}$, one cannot construct p-irregular class groups for $p > 7$; this is a consequence of B. Mazur's classification theorem on the torsion of Mordell-Weil groups of elliptic curves defined over $\mathbb{Q}$, see [20].

8. IMPLEMENTATION

Shanks' algorithm (see Section 3) has been programmed by some people e.g. Solderitsch, Shanks and his collaborators and myself.

The algorithm discussed in Section 5, has been programmed by me on the SARA CDC-Cyber 170-750 computer. At present, four programs are available: SHANKS, LONSH, PODISH and LOPOD.

SHANKS is a program, completely written in PASCAL that determines the class group of a complex quadratic order, given its discriminant Δ, with $|\Delta| < 2.5_{10}14$. It is hard to predict the time needed, to compute the class group of a given quadratic order; apart from the size of Δ, also factors like the accuracy of the approximation of $L(1,\chi)$ and the complexity of the structure of the class group have their influence on the computing time. Roughly speaking, a 10-digit discriminant takes not more than 0.1 seconds and a 15 digit discriminant takes 0.2 seconds. It is possible to give extra data, like an a priori known divisor of the class number, or forms whose order in the class group is known beforehand. In computing an approximation of $L(1,\chi)$, SHANKS uses a file of primes: PRIME, which contains, at present, all primes ≤ 240000.

Apart from the difficulties that arise, when the class group is very complicated, the most time consuming parts of Shanks' algorithm are the computation of the approximation of the class number and the baby-giant-step strategy (both $\sim |\Delta|^{1/5+\varepsilon}$). In order to have an optimum in costs, some care was taken to "balance" the program: the amount of primes used in the evaluation of the approximation of the class number depends upon the size of the current approximation of the class number; the constants involved are chosen in such a way, that the baby-giant-step strategy and the evaluation of the approximation of the class number take about the same amount of time. It

should be remarked, that for the discriminants of the size, that can be handled by SHANKS, only 10 to 15% of the computing time is spent in these "time-comsuming" parts of the algorithm. The reason for this is that discriminants of this size are, in fact, a bit too small for the algorithm (!); most of the time is spent doing "administration" i.e. computations in the class groups, determination of precise orders of forms etc. Considerably larger discriminants can be handled by LONSH and only then, a large part of the computation time is spent in computing an approximation of the class number and in doing the baby-giant-step strategy.

It was suggested by L. Monier to do the search procedures in the baby-giant-step computations by means of hash-coding [14]. SHANKS gives as output:

- the structure of the class group of $O(\Delta)$;
- the complete factorization of Δ;
- the "precise" value of $L(1,\chi)$;
- a lot of information on how the group was computed, how good the approximations were, computing times etc.

LONSH is a double length version of SHANKS; LONSH computes the class groups of orders with discriminant Δ, where $|\Delta| < 10^{29}$. The bulk of LONSH is written in PASCAL, but the composition and reduction algorithms are written in FORTRAN and COMPASS, The CDC assembler language. In fact, since the coefficients of the quadratic forms are $\sim \sqrt{|\Delta|}$, only these parts of LONSH differ essentially from the algorithms used in SHANKS.

LONSH uses the DOUBLE PRECISION facilities of FORTRAN. A 20 digit Δ will take ~ 2 seconds and a 25-digit Δ roughly 20. As indicated before, LONSH displays more clearly the order of the algorithm. Concerning transput facilities: LONSH has exactly the same possibilities as SHANKS.

PODISH computes the regulator and the class group of a real quadratic order, given its discriminant Δ, where $\Delta < 2.5_{10}14$. PODISH uses PRIME and it performs its searching routines, in determining the regulator, by means of hash-coding.

PODISH first computes the regulator and then, depending on an option: "R", PODISH computes the class group. Under the option "R", PODISH only computes the regulator, the norm of the fundamental unit ε, and if $N\varepsilon = 1$, the factorization induced by the non-trivial reduced ambiguous form in the principal cycle, at distance $\frac{1}{2} R^+$. Care has been taken to balance the program, although this is harder to do than in the complex case. Due to the computations that are performed in the principal cycle to determine the precise regulator, *after* a "match" has been found, the algorithm is not as sensitive

to the accuracy of the approximation of the class number, as in the complex case. PODISH gives as output.

- the value of R^+, $L(1,\chi)$ and the norm of the fundamental unit;
- the structure of the class group;
- information on how the regulator and the class group were obtained, like computing times, etc.

LOPOD is a double length version of PODISH.

For a detailed description of the programs mentioned here, see [29].

REFERENCES

[1] BOREVIČ, Z.I. & I.R. ŠAFAREVIČ, *Number Theory*, Academic Press, 1966.

[2] BOYD, D.W. & H. KISILEVSKY, *On the exponent of the ideal class groups of complex quadratic fields*, Proc. Amer. Math. Soc. (1972) 433-436.

[3] BUELL, D.A., *Class groups of quadratic fields*, Math. Comp. 30 (1976) 610-623.

[4] CRAIG, M., *A type of class groups for imaginary quadratic fields*, Acta Arithm. 22 (1973) 449-459.

[5] CRAIG, M., *A construction for irregular discriminants*, Osaka J. of Math. 14 (1977) 365-402.

[6] DIAZ Y DIAZ, F., *Sur le 3-rang des corps quadratiques*, Thèse, Orsay, (1978).

[7] DIAZ Y DIAZ, F., *On some families of imaginary quadratic fields*, Math. Comp. 32 (1978) 637-650.

[8] DIAZ Y DIAZ, F., *Quelques discriminants irreguliers*, Actas de las VII Jornadas Mathematicas Hispano-Lusas, Santander, Iunio, 1979.

[9] DIAZ Y DIAZ, F., *Sur le 3-rang des corps quadratiques*, preprint.

[10] DIAZ Y DIAZ, F., D. SHANKS & H.C. WILLIAMS, *Quadratic fields with 3-rank equal to 4*, Math. Comp. 33 (1979) 836-840.

[11] DIAZ Y DIAZ, F., *Private communication*.

[12] GAUSS, C.F., *Disquisitiones Arithmeticae*.

[13] GUY, R.K., *How to factor a number*, Proc. 5th Manitoba Conf. on Num. Math., October 1975.

[14] KNUTH, D., *The Art of Computer Programming, II*, Addison-Wesley, 1973.

[15] LAGARIAS, J.C., H.L. MONTGOMERY & A.M. ODLYZKO, *A bound for the least prime ideal in the Chebotarev Density Theorem*, Inv. Math. 54 (1979) 271-296.

[16] LAKEIN, R.B., *Computation of the ideal class group of certain complex quadratic fields, II*, Math. Comp. 29 (1975) 137-144.

[17] LANG, S., *Algebraic Number Theory*, Addison-Wesley, 1968.

[18] LENSTRA, H.W., *On the calculation of regulators and class numbers of quadratic fields*, To appear in Proceedings of the Journées Arithmétiques, Exeter 1980.

[19] LIPPMANN, R.A., *Note on irregular discriminants*, J. London Math. Soc. 38 (1963) 385-386.

[20] MAZUR, B., *Rational points on modular curves*, In: *Modular Functions of One Variable*, V. Bonn, J.P. Serre and D.B. Zagier (eds), Lecture Notes in Math. 601, Springer, 1976.

[21] MESTRE, J.F., *Courbes elliptiques et groupes de classes d'idéaux de certaines corps quadratiques*, Sém. de Théorie des nombres, Bordeaux, 1979/1980, Exp. 15.

[22] TATE, J., *Algorithm for determining the type of a singular fiber in an elliptic pencil*, In: *Modular Functions of One Varible, IV*, Anvers, Ed. Birch and Luyk, Lecture Notes in Math. 476, Springer, 1972.

[23] MONIER, L., *Algorithmes de factorisation d'entiers*, Thèse, 3^{me} Cycle, Orsay, 1980.

[24] NEILD, C. & D. SHANKS, *On the 3-rank of quadratic fields and the Euler product*, Math. Comp. 28 (1974) 279-291.

[25] ODLYZKO, A.M., Private Communication.

[26] PALL, G., *Note on irregular determinants*, J. London Math. Soc. 11 (1936) 34-35.

[27] SCHOLZ, A., *Über die Beziehung der Klassenzahlen quadratischer Zahlkörper zu einander*, J. Reine u. Angew. Math. 166 (1932) 201-203.

[28] SCHOOF, R.J., *Class groups of complex quadratic fields*, To appear in Math. Comp..

[29] SCHOOF, R.J., *Two algorithms for determining class groups of quadratic fields*, Report Dept. of Math. Univ. of Amsterdam, to appear.

[30] SHANKS, D., *On Gauss' class number problems*, Math. Comp. 23 (1969) 151-163.

[31] SHANKS, D., *Class number, a theory of factorization and genera*, Proc. Symp. Pure Math. 20 AMS (1971) 415-440.

[32] SHANKS, D. & P. WEINBERGER, *A quadratic field of prime discriminant, requiring three generators for its class group, and related theory*, Acta Arithm. 21 (1972) 71-87.

[33] SHANKS, D., *The infra-structure of a real quadratic field and its applications*, Proc. of the 1972 Number Theory Conf. Boulder, Colorado, 1973, 217-224.

[34] SHANKS, D. & R. SERAFIN, *Quadratic fields with four invariants divisible by 3*, Math. Comp. 27 (1973) 181-187.

[35] SHANKS, D., *New types of quadratic fields having three invariants divisible by 3*, J. Number Theory 4 (1972) 537-556.

[36] SHANKS, D., *Class groups of the quadratic fields, found by F. Diaz y Diaz* Math. Comp. 30 (1976) 173-178.

[37] SCHUR, *Einige Bemerkungen zu den Vorstehenden Arbeit des Herrn G. Polya: über die Verteilung der quadratischen Reste und Nichtreste*, Nachr. Kön. Ges. Wiss. Göttingen, Math.-Phys. Kl. (1918), 30-36.

[38] SOLDERITSCH, J.J., *Imaginary quadratic number fields with special class groups*, Thesis, Lehigh Univ., 1977.

[39] SOLDERITSCH, J.J., *Quadratic fields with special class groups*, To appear in Math. Comp.

[40] WILLIAMS, H.C. & B.K. SCHMID, *A rapid method of evaluating the regulator and class number of a pure cubic field*, to appear in Math. Comp.

[41] WEINBERGER, P., *Exponents of the class group of complex quadratic fields*, Acta. Arithm. 22 (1973) 117-124.

[42] YAMAMOTO, Y., *On unramified Galois extensions of quadratic number fields*, Osaka J. of Math. 7 (1970) 57-76.

[43] ZANTEMA, H., *Class numbers and units*, these proceedings.

MULTI-DIMENSIONAL CONTINUED FRACTION ALGORITHMS

by

A.J. BRENTJES *)

Part one:
Multi-dimensional continued fraction algorithms
and their application to approximation problems

1. INTRODUCTION

Multi-dimensional continued fraction algorithms are generalizations, in a certain sense to be made precise in §2, of the well-known continued fraction algorithm (Euclid's algorithm). They can be used to solve a variety of Diophantine approximation problems and other problems that can be interpreted as such, ranging from the computation of a g.c.d. to the determination of units in algebraic orders. Though their iterative nature makes these algorithms highly suited for computer implementation, the present state of the subject consists mostly of scattered contributions and many questions are still open. Without claiming completeness we sketch the theory (§2), give some historical remarks (§3 and *passim*) and discuss a few techniques (§4-6). In part two we apply these techniques to the unit problem in cubic fields. The material in this paper is part of the author's doctoral thesis [61]. Proofs of several statements in this article can be found there.

2. DEFINITIONS AND EASY FACTS

In the (n+1)-dimensional real vector space $\mathbb{R}^{n+1}$ we consider a lattice Ω (i.e., a discrete additive subgroup of maximal rank) and an arbitrary non-zero vector ℓ_0 defining a line ℓ through the origin O. In several applications Ω will consist simply of all points with integral coordinates and ℓ_0 will be given as $(1,\xi_1,\ldots,\xi_n)$ for some real numbers $\xi_1,\ldots,\xi_n$. Denoting points of Ω (and of $\mathbb{R}^{n+1}$) by capitals we define the *cofactors* $a_0,a_1,\ldots,a_n$ of a lattice base $\{A_0,A_1,\ldots,A_n\}$ by

---

*) Supported by a grant from the Netherlands Organization for the Advancement of Pure Research (Z.W.O.)

(1) $$a_0A_0 + a_1A_1 + \ldots + a_nA_n = \ell_0;$$

they are the coordinates of ℓ_0 with respect to that particular base.

DEFINITION 1. The line ℓ is *dependent on* Ω with *dependence rank* r if there is a lattice base having r cofactors equal to zero whereas no lattice base has r+1 cofactors equal to zero. In case r = 0 we call ℓ an *independent line*.

Notice that r = n if and only if there exists a lattice point different from 0 on ℓ.

We are interested mainly in lattice bases all of whose cofactors are non-negative; when the points of such a base are projected (parallel to ℓ) on the n-dimensional subspace ℓ^* orthogonal to ℓ, the origin is contained in the convex hull of these projections. The projection on ℓ^* of a point A will always be denoted by underlining: $\underline{A}$.

DEFINITION 2. A finite or infinite sequence of lattice bases $\{A_0(i),\ldots$ $\ldots,A_n(i)\}$ (i = 0,1,2,...) all with non-negative cofactors $a_0(i),\ldots,a_n(i)$ is called an (n-dimensional continued fraction) *expansion* of its first element $\{A_0(0),\ldots,A_n(0)\}$ along ℓ, if to each $i \geq 0$ there exist indices s and t ($s \neq t$) and an integer $b \geq 1$ such that

(2) $$A_t(i+1) = A_t(i) + bA_s(i)$$

and

(3) $$A_j(i+1) = A_j(i) \quad \text{for } j \neq t.$$

The letters s, t, b will always be used with this meaning; it is not appropriate to attach the step index i to them.

DEFINITION 3. Any algorithm to expand along a given line a given base with non-negative cofactors relative to that line, is called an n-*dimensional continued fraction algorithm* (n-fraction for short[1]).

From the equality

---

1) The term is from SZEKERES [51]

(4) $$a_s A_s + a_t A_t = (a_s - b a_t) A_s + a_t (b A_s + A_t)$$

it is clear that

(5) $$a_s \geq b a_t$$

is a necessary and sufficient condition for the base $\{A_0(i+1),\ldots,A_n(i+1)\}$ to have non-negative cofactors if $\{A_0(i),\ldots,A_n(i)\}$ has. In view of $a_t \geq 0$ and $b \geq 1$, (5) is equivalent to the combination of

(6) $$a_s \geq a_t$$

and

(7) $$1 \leq b \leq \left[\frac{a_s}{a_t}\right]$$

(interpreted as $b < \infty$ when $a_t = 0$). In general, therefore, the indices s and t can at each step be chosen in $\binom{n+1}{2}$ different ways (and more if some cofactors happen to be equal), after which a number b in the range (7) must be determined. An n-fraction essentially is a procedure to make these choices. An algorithm which always chooses $b = 1$ is called *subtractive*; a *division* algorithm always chooses $b = [a_s/a_t]$. Applying these remarks to the case $n = 1$ we obtain immediately

THEOREM I. *The only one-dimensional continued fraction algorithm is the ordinary continued fraction algorithm (Euclid's algorithm).*

(The twin choice in the case of equal cofactors corresponds to a rational number having two continued fractions associated with it. With $b = 1$ we have the subtractive version as it was geometrically interpreted by KLEIN [37], with $b = [a_s/a_t]$ we have the usual division algorithm as in PERRON [45].)

We shall now discuss the relation between multi-dimensional continued fraction expansions and approximation problems. To that end we introduce a height function and a distance function. The *height function* is a linear function $h: \mathbb{R}^{n+1} \to \mathbb{R}$ such that $h(\ell_0) > 0$; the subspace $h = 0$ does not necessarily coincide with the orthogonal subspace ℓ^*. The distance function d measures the distance of a point of R^{n+1} to ℓ; it must of course have the following properties:

(a) $d(P) \geq 0$ for all P, and $d(P) = 0 \iff P \in \ell$

(b) $d(\lambda P) = |\lambda| d(P)$ for all P and all real λ

(c) $d(P+Q) \leq d(P) + d(Q)$ for all P, Q

(d) $d(P+\lambda \ell_0) = d(P)$ for all P and all real λ

but no further specification is made yet. Notice that (d) is a simple consequence of (a), (b), (c) and is included only to stress the concept. We recall the fact that, given any two distance functions d, d' as above, there exist constants $c_1 > 0$, $c_2 > 0$ depending only on d, d' such that

$$c_1 d'(P) \leq d(P) \leq c_2 d'(P) \tag{8}$$

for all $P \in R^{n+1}$; which is expressed by saying that d and d' are absolutely continuous with respect to each other. Now we give the following definition of convergent expansions.

DEFINITION 4. The expansion $\{A_0(i),\ldots,A_n(i)\}$ $(i = 0,1,\ldots)$ is *strongly convergent* if to any $\varepsilon > 0$ there exists an integer j such that

$$\max_{0\leq k\leq n} d(A_k(j)) < \varepsilon \tag{9}$$

It is *weakly convergent* if to any $\varepsilon > 0$ there is an integer j such that

$$\max_{0\leq k\leq n} d\left(\frac{A_k(j)}{h(A_k(j))}\right) < \varepsilon \tag{10}$$

Notice that, by (8), this definition does not depend on the particular choice of the distance function; also, the choice of the height function is immaterial as long as $\min_{0\leq k\leq n} h(A_k(0)) > 0$.

Not all lines can have strongly convergent expansions; in fact there is the following result translating the dependence problem into an approximation problem.

THEOREM II. *If a line ℓ has dependence rank* r, *then at most* n-r+1 *points of a lattice base can be arbitrarily close to ℓ; in particular, only independent lines can have strongly convergent expansions.*

SKETCH OF PROOF. Let $A_0,\ldots,A_n$ be a base of Ω such that $a_{n-r+1} = \ldots = a_n = 0$; U is the (n-r+1)-dimensional subspace spanned by $A_0,\ldots,A_{n-r}$. By (1), ℓ is contained in U. Let $\delta > 0$ be a lower bound for the Euclidean distances to U of all lattice points not in U; positivity of δ follows from the discreteness of Ω. The distance of any lattice point not in U to ℓ is at least δ. From absolute continuity of any distance function d with Euclidean metric it follows that for sufficiently small $\varepsilon > 0$ at most n-r+1 points of any lattice base $\{B_0,\ldots,B_n\}$ can satisfy $d(B_i) < \varepsilon$.

Whether or not dependent lines have weakly convergent expansions depends on the algorithm these are obtained with. Example: With the algorithm of BRUN [13], in the case n = 2 only independent lines have weakly convergent expansions, whereas from n = 3 onward dependent lines can also have weakly convergent expansions. In view of Theorem II we say that an algorithm *ex* definition 3 is strongly (weakly) convergent if it generates a strongly (weakly) convergent expansion of any independent line compatible with the algorithm's initialization requirements. If a given algorithm is weakly convergent, it is seldom difficult to prove this; on the contrary, strong convergence if it exists is very much harder to prove. In fact, the first algorithm with proven strong convergence was not published until 1979 (FERGUSON & FORCADE [25]).

An important role in the theory of Diophantine approximations is played by the concept of best approximations, whose general definition runs as follows.

DEFINITION 5. A lattice point $B \neq 0$ is a *best approximation of* Ω *to* ℓ (with respect to the height and distance functions h and d) if there exists no lattice point $P \neq 0$ for which either

$$d(P) \leq d(B) \quad \text{and} \quad |h(P)| < |h(B)|$$

or

$$d(P) < d(B) \quad \text{and} \quad |h(P)| = |h(B)|.$$

The sequence of best approximations is infinite in the direction of increasing height if and only if no lattice point difference from 0 exists on ℓ; similarly, the sequence of best approximations in the direction of decreasing absolute height is infinite if and only if no lattice point satisfies

$h = 0$ except for 0. These remarks follow from Minkowski's well-known theorem bounding the volume of convex bodies that are symmetrical about the origin and contain no other lattice points.

In the "classical" theory of Diophantine approximation one works with $\Omega = \mathbb{Z}^{n+1}$ and with $\ell_0 = (1,\xi_1,\ldots,\xi_n)$ where $\xi_i > 0$, $i = 1,\ldots,n$. Best approximations are defined with respect to the height function

$$h_c(x_0,\ldots,x_n) = x_0 \tag{11}$$

and the "sup norm" distance function

$$d_c(x_0,\ldots,x_n) = \max_{1\le k\le n} |x_0\xi_k - x_k| \tag{12}$$

The unit d_c-ball in the space $h_c = x_0 = 0$ is a hyper-cube but in the orthogonal hyperplane ℓ^* it is a skew hyper-parallelepiped whose shape depends on $\xi_1,\ldots,\xi_n$. In spite of its arithmetically natural appearance the function d_c is geometrically often unsatisfactory; this is illustrated by a comparison between recent results of CUSICK [18] and the corollary to Lemma 2 of part two below.

The Euclidean distance function was repeatedly considered in connection with cubic fields of negative discriminant (e.g. by VORONOI [54], DELONE & FADDEEV [21] ch. IV B, BERWICK [9], DUBOIS [22]; for an explanation of this connection see §2 of part two below), but not generally as an alternative to (12) until recent (two-dimensional) work by JURKAT, KRATZ & PEYERIMHOFF [36]. LAGARIAS [38] considered arbitrary distance functions from a non-algorithmic point of view.

3. A FEW HISTORICAL REMARKS

The earliest and most extensively studied multi-dimensional continued fraction algorithm is that of Jacobi-Perron. It was first published as a two-fraction in a post mortem paper of JACOBI [34] in 1868, and received a rigorous foundation in arbitrary dimension in the 1907 thesis of PERRON [44]. Its definition, converted to the notation of §2, runs as follows.

(a) $t := 0$, $s := 1$,

(b) $b := [a_s/a_t]$; $a_s := a_s - ba_t$: $A_t := A_t + b\cdot A_s$;

(c) $s := \sigma(s)$: if $t = s$ then $t := \sigma(t)$, $s := \sigma(t)$: goto (a),

where

$$\sigma(x) = \begin{cases} 0 & \text{if } x = n \\ x+1 & \text{otherwise} \end{cases}$$

Notice that in (b), b = 0 must be allowed; but for fixed t there is at least one s such that $b \geq 1$ (except possibly for the initial t = 0). The cyclical rotation of t corresponds both in JACOBI [33] and in PERRON [44] to a wish for greater formal analytic regularity as compared to the idea of using the index of the smallest non-zero cofactor for t, an idea already implicit in EULER [24].

The Jacobi-Perron algorithm belongs to a general class known as Jacobi-type algorithms; they are characterized by choosing s, t and b from no other information than the relative size of the cofactors. Clearly no such algorithm can effectively be used to obtain best approximations with. For let T be any non-singular linear transformation on R^{n+1} having ℓ_0 as an eigen vector (with positive eigenvalue). If $\{A_0(i),\ldots,A_n(i)\}$ is the expansion obtained from $\{A_0(0),\ldots,A_n(0)\}$ by a Jacobi-type algorithm, then the expansion of $\{TA_0(0),\ldots,TA_n(0)\}$ is $\{TA_0(i),\ldots,TA_n(i)\}$, i = 0,1,2,... . But in general TP need not be a best approximation of TΩ when P is one of Ω, and conversely.

Indeed, the principal subject of study with Jacobi-type algorithms has been the question of periodicity (defined by the existence of k_0, m > 0 such that

$$a_j(k+m) = a_j(k)$$

for $0 \leq j \leq n$ and all $k \geq k_0$), i.e. whether some generalization of Lagrange's theorem for Euclid's algorithm is true. The main question here is, will the expansion of the standard base of $\Omega = \mathbb{Z}^{n+1}$ ultimately become periodical when $\ell_0 = (1,w,\ldots,w^n)$, where w > 0 is algebraic of degree n + 1? In spite of intensive investigations by BERNSTEIN [8] and many others this question has not yet been resolved. Variations of the Jacobi-Perron algorithm have been proposed by BERNSTEIN [5], DAUS [19], GÜTING [31].

Another Jacobi-type algorithm which has received much attention was proposed by BRUN [12] in 1919. Its attractively simple definition is:

(a) choose s and t such that a_s, a_t are the largest and second largest cofactors respectively;

(b) $b := 1$; $A_t := A_t + b \cdot A_s$; $a_s := a_s - ba_t$; goto (a).

Actually this algorithm is so obvious that it has, in various contexts, been reproposed at least five times, always in the division version with b = $[a_s/a_t]$: by PALEY and URSELL [43], ROSSER [47], BARBOUR [3], VAUGHAN [53] and BERGMAN [57]. Note that the division is a mere acceleration of the original algorithm. GREITER [30] was the first to compare Brun's algorithm systematically with that of Jacobi-Perron. In the present state of knowledge the Jacobi-Perron algorithm is weakly convergent in all dimensions; it is not strongly convergent when $n \geq 3$ whereas this is an open question for $n = 2$. Brun's algorithm is strongly convergent for no $n \geq 2$; its weak convergence was proved by GREITER [30].

It was only about 1970 that the interest in multi-dimensional continued fraction algorithms as a means to solve practical approximation problems grew. In that year SZEKERES [51] listed some desiderata which algorithms with good approximation qualities should possess. He also proposed an algorithm intended to obtain sup norm (12) best approximations with. In slightly generalized form with arbitrary distance function his algorithm is as follows: (b = 1 throughout)

(a) $s := 0$

(b) determine t such that $d((A_t/h(A_t)) - (A_s/h(A_s)))$ is maximal

(c) if $a_s < a_t$ interchange s and t

(d) $A_t := A_t + A_s$, $a_s := a_s - a_t$; goto (a).

Weak convergence is quite easily proved for any n; and though the conjecture of Szekeres that the algorithm might never miss a best approximation seems a little too strong, it is probable that strong convergence also holds (at least for small n). On the other hand, the definition is not easy to handle and therefore virtually no theoretical results are known of Szekeres' algorithm; there is a paper by CUSICK [17] who verified a conjecture of Szekeres concerning one particular example. In practice the algorithm exhibits extremely good approximation properties; on the other hand the thoroughly subtractive nature makes the algorithm rather slow. For a comparison between the approximation properties of the algorithms of Jacobi-Perron and Szekeres see JURKAT, KRATZ and PEYERIMHOFF [36] and VAN DE LUNE and TE RIELE [59].

The first algorithm with proven strong convergence was given in 1979 by FERGUSON & FORCADE [25]; it will be discussed in the next section.

4. A STRONGLY CONVERGENT ALGORITHM

In this section we define a strongly convergent algorithm for any dimension n, based essentially upon an idea of FERGUSON & FORCADE [25]. We shall therefore call it the Ferguson and Forcade algorithm (FFA). The method of Ferguson and Forcade, published 1979, is the first n-dimensional algorithm with proven strong convergence (though it is not yet known whether Szekeres' algorithm has this property; see Section 3). A year before, JURKAT, KRATZ and PEYERIMHOFF [36] defined a strongly convergent algorithm for n = 2, which we shall not discuss since its performance is worse than that of the algorithm to be presented in §5.

We shall define the n-dimensional Ferguson & Forcade algorithm FFA_n using induction on n and starting with the ordinary continued fraction algorithm as FFA_1 (see Theorem I). First of all we agree that FFA_n will terminate as soon as a base is obtained one of whose cofactors equals zero. The following theorem will then be seen to hold for every n.

THEOREM III. *The n-dimensional Ferguson and Forcade algorithm* FFA_n, *when applied to a base* $\{A_0(0),\dots,A_n(0)\}$ *with non-negative cofactors, will either terminate or yield, given any* $\varepsilon > 0$, *a base* $\{A_0(i_0),\dots,A_n(i_0)\}$ *with*

$$\max_{0\le k\le n} d(A_k(i_0)) \le \varepsilon.$$

This is obviously true if n = 1, FFA_1 being the ordinary continued fraction algorithm. Now take $n \ge 2$ and suppose FFA_{n-1} has been defined such that theorem III holds for it, and let $\{A_0,A_1,\dots,A_n\}$ be a lattice base with non-negative cofactors $a_0,\dots,a_n$. We denote by $X\cdot Y$ the inner product of two vectors X and Y, and use the notation $|X| = \sqrt{X\cdot X}$ for the Euclidean length of X. Recall also the notation $\underline{A}$ for the projection on ℓ^*, parallel to ℓ, of a point A.

Now if one of $a_0,\dots,a_n$ is zero, then FFA_n terminates by definition.

(i) Otherwise, we choose the index k according to

$$|\underline{A}_k| = \max_{0\le j\le n} |\underline{A}_j|. \tag{13}$$

In the n-dimensional space ℓ^* the base $\{\underline{A}_j \mid j \ne k\}$ defines a lattice Ω^* of rank n; the independence of the $\underline{A}_j$, $j \ne k$, follows from $a_k \ne 0$. In ℓ^*, let m be the line through 0 and $\underline{A}_k$. With $m_0 = -a_k\underline{A}_k \ne 0$ it follows from the

projected cofactor relation,

$$\sum_{j=0}^{n} a_j \underline{A}_j = \underline{\ell}_0 = 0,$$

that the cofactors of $\{\underline{A}_j \mid j \neq k\}$ with respect to m_0 are precisely the a_j $(j \neq k)$. This enables the following construction.

(ii) To $\{\underline{A}_j \mid j \neq k\}$ and m we apply, in ℓ^*, the algorithm FFA_{n-1}. For each step $\underline{A}_t(i+1) = \underline{A}_t(i) + b\underline{A}_s(i)$ (where $s,t \neq k$) the corresponding step of FFA_n will be $A_t(i+1) = A_t(i) + bA_s(i)$, and $\underline{A}_t(i+1)$ is again the projection of $A_t(i+1)$ since projection commutes with vector addition. We continue this until either a termination occurs in FFA_{n-1}, and hence in FFA_n, or a base $\{A'_0, A'_1, \ldots, A'_n\}$ (with $A'_k = A_k$) is reached for which

$$\max_{j \neq k} d_m(\underline{A}'_j) < \delta |\underline{A}_k| \tag{14}$$

and

$$\underline{A}'_j \cdot \underline{A}_k < 0 \quad \text{for all } j \neq k. \tag{15}$$

Here δ is a number in the range $0 < \delta < \frac{1}{2}\sqrt{3}$ (fixed in advance) and d_m denotes Euclidean distance to m. The fact that (14) will be reached of no termination occurs follows from the induction hypothesis. The same is true for (15), because the $\underline{A}_j$, $j \neq k$, were independent.

(iii) Next we replace the A'_j, $j \neq k$, by $A''_j = A'_j + b_j A_k$ with

$$b_j = \left[\frac{-\underline{A}'_j \cdot \underline{A}_k}{|\underline{A}_k|^2}\right]$$

By (15) we have $b_j \geq 0$, and in fact b_j is chosen such that for $j \neq k$,

$$\underline{A}_k \cdot \underline{A}''_j \leq 0 < \underline{A}_k \cdot (\underline{A}''_j + \underline{A}_k). \tag{16}$$

It is easily seen that the cofactor a''_k of A''_k $(=A_k)$ is still non-negative, since (16) gives

$$a''_k |\underline{A}_k|^2 = -\sum_{j \neq k} a''_j \underline{A}''_j \cdot \underline{A}_k \geq 0.$$

Moreover, when B_j is the orthogonal projection on m of $\underline{A}''_j$ we have

$$|B_j| < |\underline{A}_k|$$

by (16), and hence

$$(17) \qquad |\underline{A}''_j| = \sqrt{|B_j|^2+(d_m(\underline{A}''_j))^2} < |\underline{A}_k|\sqrt{1+\delta^2} \quad \text{for } j \neq k$$

in view of (14).

(iv) Now if for some $j \neq k$ $|B_j| \leq \frac{1}{2}|\underline{A}_k|$ we even have

$$|\underline{A}''_j| \leq |\underline{A}_k|\sqrt{\frac{1}{4}+\delta^2}$$

for that value of j, instead of (17). But if $\min_{j\neq k} |B_j| > \frac{1}{2}|\underline{A}_k|$ we choose any $j \neq k$ and replace A''_j or A''_k by $A''_j + A'_k$, depending on whether $a''_k \geq a''_j$ or $a''_k < a''_j$ respectively. For this j we have

$$|\underline{A}_k'' + \underline{A}''_j| < |\underline{A}_k|\sqrt{\frac{1}{4}+\delta^2}.$$

During the cycle (i) (ii) (iii) (iv) - which we shall call one iteration of FFA_n - the index k keeps the same value. After (iv) we return to (i) where k will, in general, receive a new value. Starting the iteration with $\{A_0(i_0),\ldots,A_n(i_0)\}$ and ending with $\{A_0(i_1),\ldots,A_n(i_1)\}$ we have

$$\min_{0\leq j\leq n} |\underline{A}_j(i_1)| \leq \sqrt{\frac{1}{4}+\delta^2} \min_{0\leq j\leq n} |\underline{A}_j(i_0)|$$

Continuing this way, we obtain a sequence of bases $\{A_0(i_0),\ldots,A_n(i_0)\}$, $\{A_0(i_1),\ldots,A_n(i_1)\}$, $\{A_0(i_2),\ldots,A_n(i_2)\}$,... each resulting from its predecessor by an iteration of FFA_n, such that

$$\min_j |\underline{A}_j(i_m)| < (\tfrac{1}{4}+\delta^2)^{m/2} \min_j |\underline{A}_j(i_0)|.$$

If no termination occurs, this implies

$$\lim_{m\to\infty} \min_j |\underline{A}_j(i_m)| = 0,$$

since we chose $\delta < \frac{1}{2}\sqrt{3}$.

The proof of Theorem III for FFA_n is now completed by observing that after step (iii) we have (17), i.e.

$$\max_j |\underline{A}_j(i_{m+1})| < \sqrt{1+\delta^2} \min_j |\underline{A}_j(i_m)|$$
$$< \sqrt{1+\delta^2}\ (\tfrac{1}{4}+\delta^2)^{m/2} \min_j |\underline{A}_j(i_0)|.$$

COROLLARY. *For any* n, FFA_n *is strongly convergent.*

HISTORICAL REMARK. The original algorithm of Ferguson and Forcade differs from the FFA presented here in two respects:
- It allows negative cofactors.
- It measures distance by the maximum of the absolute values of a vector's coordinates.

In the case n = 2 the factor $\sqrt{\frac{1}{4}+\delta^2}$ in (18) can be improved to $\sqrt{1/16+\delta^2}$ for the price of at most two extra steps to be inserted between (ii) and (iii); for general n, this is slightly more difficult, for one has to take care that (14) remains valid. Notice that a very small δ will increase the number of steps in (ii) needed to obtain (14) without substantially improving the factor $\sqrt{1/16+\delta^2}$. Therefore $\delta = 1/5$ seems a good choice, making $\sqrt{1/16+\delta^2} < 1/3$.

We shall now see how we may obtain a good simultaneous approximation to given (positive) $\alpha_1,\alpha_2 \in \mathbb{R}$, by means of FFA_2 (leaving the general case to the reader). For various problems one needs a method to obtain one or two good (though not necessarily best) approximations - see TIJDEMAN [58] and VAN DE LUNE & TE RIELE [59] -, yet FFA_2 in itself does not guarantee that its approximations are good ones. Take $\Omega = \mathbb{Z}^3$ and $\ell_0 = (1,\alpha_1,\alpha_2)$ in $\mathbb{R}^3$ and let the inner product be such that ℓ^* is the yz-plane. The projection of $A = (q,p_1,p_2)$ then is

$$\underline{A} = (0, p_1-\alpha_1 q,\ p_2-\alpha_2 q)$$

and we define the number $|q|((q\alpha_1-p_1)^2+(q\alpha_2-p_2)^2)$ to be the *quality* of A. It is obvious that for any α_1,α_2 there are infinitely many approximations with quality ≤ 2; but when $\varepsilon < 2/\sqrt{23}$ the quality $\le \varepsilon$ is not always obtainable infinitely often, as the example $\alpha_1^3-\alpha_1-1=0$ and $\alpha_2=\alpha_1^2$ shows. Now let $\{A_0,A_1,A_2\}$ be a base of Ω with non-negative cofactors such that, say, $a_0 > \max(a_1,a_2)$. Then we have (Lemma 4.8 of [61])

$$\min_{0\le j\le 2} |\underline{A}_j|^2 < \frac{1/\sin\phi}{q_0+x_1q_1+x_2q_2}$$

where $x_1 = a_1/a_0$, $x_2 = a_2/a_0$, ϕ is the angle between $\underline{A}_1$ and $\underline{A}_2$, and q_j is the x-coordinate of A_j. Therefore we may expect one of A_0, A_1, A_2 to be a good approximation if we control $\sin\phi$, say by $\sin\phi \ge \frac{1}{2}\sqrt{3}$. Once FFA_2 has given us one point A_2 close enough to ℓ for our purpose, we complete it to a base of Ω with A_1 such that $\sin\phi \ge \frac{1}{2}\sqrt{3}$ and with A_0 such that $a_0 \ge a_1, a_2$; this is easily done in the obvious way.

EXAMPLE. Applied to $\ell_0 = (1,e,\pi)$ FFA_2 (with $\delta = 1/12$) stopped after 19 steps with the approximation

$$(26804611,\ 72862487,\ 84209169) \quad \text{(quality 1.34...)}.$$

The above procedure then yielded the excellent approximation

$$(286786708,\ 779567097,\ 900967015) \quad \text{(quality 0.031...)}.$$

Another application of strongly convergent algorithms such as FFA_n concerns the question of linear dependence. Firstly, we can find a dependence relation if we know that one exists. (Example: The determination of the minimum polynomial of an algebraic number whose degree is known, by applying the algorithm to its powers.) Secondly, we can decide whether numbers $1,\xi_1,\ldots,\xi_n$ are independent if a number M is known such that either $1,\xi_1,\ldots,\xi_n$ are independent or $c_0 + \Sigma_{i=1}^{n} \xi_i c_i = 0$ for some $c_i \in \mathbb{Z}$, with $0 < \max|c_i| < M$. (Example: If a polynomial $g(x)$ of degree ≥ 1 divides a polynomial $f(x)$ of degree n with at least one real root α, then height $(g) < 2^{n-1}(1 + \text{height}(f))^{n-2}$. Using this number as M we can decide whether $f(x)$ is irreducible, and if not, find a factor.) For the second application one needs a quantitative version of Theorem II, such as:

THEOREM IV. *If* $\{A_0,A_1,\ldots,A_n\}$ *is a base of* $\mathbb{Z}^{n+1}$ *and*

$$\max_{0\le k\le n} |\underline{A}_k| < \varepsilon$$

then any vector $C \in \mathbb{Z}^{n+1}$ *such that* $C\cdot\ell_0 = 0$ *satisfies* $C = 0$ *or* $|C| > 1/\varepsilon$.

PROOF. By elementary linear algebra along the line of proof of Theorem II.

5. BEST APPROXIMATIONS

The question whether best approximations with respect to given height and distance functions can be calculated by means of multi-dimensional continued fraction algorithms is yet unresolved. In fact the complexity of this problem in the case $n = 2$ is already such that nobody has made a serious study for $n \geq 3$. (As remarked in §3, there is the unverified claim about the Szekeres algorithm yielding all best approximations.) Therefore most of this section concerns the case $n = 2$.

Given a best approximation B one can, using analytical methods, determine all points $P \in \Omega$ satisfying

$$d(P) < d(B), \qquad |h(P)| < c \tag{18}$$

where c is chosen such that at least one $P \neq 0$ satisfies (18) (such a c follows from the Minkowski theorem). Ω being discrete, the number of points P in the body (18) is finite and the one of least absolute height clearly is the best approximation next to B (in the direction of increasing height). Various versions of this method, all for $n = 2$ and some in the reversed direction of decreasing absolute height, have been proposed and elaborated, e.g. by MINKOWSKI [40], and FURTWÄNGLER [26] for the sup norm distance function (12) and by VORONOI [54], DUBOIS [23] for the Euclidean distance function; they are all based additionally on the fact that two consecutive best approximations can be completed with a third point to form a lattice base. But this method is not a continued fraction algorithm in the sense of definitions 2, 3, since it lacks the strictly additive nature.

Recently the present author introduced a new method in which one does not examine the points that are found by an algorithm, but, to the contrary, focuses attention on the points that are *not* found. When we say for convenience that a point P is a positive combination of a lattice base $\{A_0, A_1, A_2\}$ if $P = pA_0 + qA_1 + rA_2$ with $p,q,r \geq 0$, then this method amounts, in view of the additive nature of continued fraction algorithms, to studying points that are positive combinations of the i-th but not of the (i+1)th base in an expansion. We illustrate the power of the method on the best appproximation problem with Euclidean distance function. As before, $\underline{A}$ is the projection on ℓ^* of a point $A \in \mathbb{R}^3$, parallel to ℓ. The inner product of two vectors X and Y is $X \cdot Y$, and we write

(19) $\qquad |X| = \sqrt{X \cdot X};$

furthermore the angle $\measuredangle(X,Y)$ is defined by $0 \le \measuredangle(X,Y) \le \pi$ and

(20) $\qquad |X||Y| \cos \measuredangle(X,Y) = X \cdot Y.$

The Euclidean distance function is then explained as

$$d(P) = |\underline{P}| \quad \text{for all } P \in \mathbb{R}^3.$$

The main tool is the following lemma, which at once illustrates the method.

LEMMA 1. *Let the base* $\{A,B,C\}$ *of* Ω *have non-negative cofactors* a, b, c, *and suppose* $a \ge b$ *and* $\min(h(A),h(B),h(C)) > 0$. *Also suppose that*

(21) $$\measuredangle(\underline{B},\underline{C}) \le \frac{2}{3}\pi, \quad \measuredangle(\underline{B},\underline{A+B}) \le \frac{2}{3}\pi$$

or

(22) $$\forall k \in \mathbb{N}, \quad |\underline{B}+k\cdot\underline{C}| \ge |\underline{C}|, \quad |\underline{B}+k\cdot(\underline{A}+\underline{B})| \ge |\underline{A}+\underline{B}|.$$

If a lattice point $P \neq B$ *is a positive combination of* $\{A,B,C\}$, *but not of* $\{A,A+B,C\}$, *then* P *is not a best approximation.*

SKETCH OF PROOF. Let P be a positive combination of $\{A,B,C\}$ but not of $\{A,A+B,C\}$ and write $P = pA + qB + rC = (p-q)A + q(A+B) + rC$ $(p,q,r \in \mathbb{Z})$ to see that this means precisely $q > p \ge 0$, $r \ge 0$. From $P = (q-p)B + p(A+B) + rC \neq B$ it follows at once that

(23) $\qquad h(P) > h(B).$

Using the projected cofactor relation $a\underline{A} + b\underline{B} + c\underline{C} = 0$ we write

$$\underline{P} = (q - p + \frac{r(a-b)}{c})\underline{B} + (p - r\,\frac{a}{c})(\underline{A}+\underline{B})$$

$$= (q - p + \frac{p(a-b)}{a})\underline{B} + (r - p\,\frac{c}{a})\underline{C}.$$

Therefore we have, either with $D = C$ or with $D = A + B$,

$$\underline{P} = \tau\underline{B} + \sigma\underline{D}$$

with $\tau \geq 1$, $\sigma \geq 0$.

If (21) holds we remark that P is not a best approximation when $|\underline{P}| \geq |\underline{B}|$ (by (23)); if $|\underline{P}| < |\underline{B}|$ one proves, using $\sphericalangle(\underline{B},\underline{D}) \leq \frac{2}{3}\pi$, that in the triangle with vertices 0, $\underline{P}$ and $\underline{B}$ the angle at 0 is smallest, whence $|\underline{P}-\underline{B}| < |\underline{P}|$, and again P is not a best approximation.

If (22) holds, one trivially has $|D| \leq |P|$, $h(D) < h(P)$ if P is of the form $B + k\cdot D$; otherwise one puts $Q = B + [\frac{\sigma}{\tau}]D$ (so that $h(Q) < h(P)$ because $\tau \geq 1$) and proves $|\underline{P}-\underline{Q}| < |\underline{P}|$. In either case P is not a best approximation.

Now we define a two-fraction as follows (always b = 1).

I. If

$$\min_{i\neq j} \sphericalangle(\underline{A}_i,\underline{A}_j) \geq \frac{1}{3}\pi, \tag{24}$$

define the permutation f, g, h of 0, 1, 2 by

$$|\underline{A}_f| \leq \min(|\underline{A}_g|, |\underline{A}_h|), \quad a_h \geq a_g. \tag{25}$$

Then

Ia. if $a_g = 0$: s := f, t := h; else

Ib. if $|\underline{A}_g + k\cdot\underline{A}_f| \geq |\underline{A}_f|$ for all $k \in \mathbb{N}$: s := h, t := g; else

Ic. if $\sphericalangle(\underline{A}_f,\underline{A}_g) \geq \sphericalangle(\underline{A}_g,\underline{A}_h)$: s := f, t := g; else

Id. s := g, t := f.

II. If (24) does not hold, define the permutation f, g, h by

$$\sphericalangle(\underline{A}_g,\underline{A}_h) < \tfrac{1}{3}\pi, \quad |\underline{A}_g| \geq |\underline{A}_h| \tag{26}$$

and take s := f, t := g.

REMARK. If one of the points A_0, A_1, A_2 is on ℓ, i.e. is projected into O, the necessary angles do not exist; therefore we agree that the algorithm terminates when this happens.

Using Lemma 1 the following theorem about the above algorithm can be proved by separate treatment of the five cases. Alongside one proves that the algorithm indeed always chooses s and t satisfying (6).

THEOREM V. *Let the initial base* $\{A_0(0), A_1(0), A_2(0)\}$ *have non-negative cofactors and let* $h(A_i(0)) > 0$ $(i = 0,1,2)$. *Also suppose that the initial base satisfies* (24). *If the above algorithm does not terminate and the best approximation* P *is a positive combination of the initial base, then* $P \in \{A_0(j), A_1(j), A_2(j)\}$ *for some* $j \geq 0$.

In practice (24) will always be satisfied. For instance with $\Omega = \mathbb{Z}^3$, $\ell_0 = (1,\xi_1,\xi_2)$ $(\xi_1,\xi_2 > 0)$ one has

$$\min_{i\neq j} \sphericalangle(\underline{A}_i,\underline{A}_j) > \frac{1}{2}\pi$$

if $A_0 = (1,0,0)$, $A_1 = (0,1,0)$, $A_2 = (0,0,1)$ form the standard base. It must be noted that the best approximations are not necessarily found in increasing order of height.

The speed of our algorithm (which is subtractive in the form presented above) can be adequately improved by the following remarks. If a base does not satisfy (24), but the specifications (25) and (26) do give the same f, g, h, then we can apply a Ib. step instead of a II-step when the condition for Ib is satisfied. Furthermore, in the two most frequently used steps (Ib and II) we can replace the subtractive $b = 1$ by

$$b = \max\left(1, \left[-\frac{|\underline{A}_g|}{|\underline{A}_f|}\left(\xi + \frac{1}{3}\sqrt{3-3\xi^2}\right)\right]\right)$$

in case II (with $\xi := \cos\sphericalangle(\underline{A}_f,\underline{A}_g) < 0$), and by

$$b = \max\left(1, \left[\frac{a_h}{a_g}\right] - \left[\frac{|\underline{A}_f|}{|\underline{A}_h|\sin\sphericalangle(\underline{A}_f,\underline{A}_h)}\right]\right)$$

in case Ib, the latter provided that $h(A_g+A_h) \geq h(A_f)$.

Application of these remarks turns our algorithm into a kind of "careful FFA_2"; the validity of Theorem V remains unaffected.

An investigation, with the aid of Lemma 1, of some simpler devices such as the greatest angle algorithm defined by

$$(27) \qquad \sphericalangle(\underline{A}_s,\underline{A}_t) = \max_{i\neq j} \sphericalangle(\underline{A}_i,\underline{A}_j)$$

or the inner product algorithm

$$(28) \qquad \underline{A}_s \cdot \underline{A}_t = \min_{i\neq j} \underline{A}_i \cdot \underline{A}_j,$$

stresses the point that the steps must be chosen exceedingly carefully if one does not want to miss a best approximation; for though easily proved to be strongly convergent, the greatest angle and inner product two-fractions occasionally do miss a best approximation. This contrasts with the rougher requirement of strong convergence, which was sufficient in §4.

As to the case $n \geq 3$ no results whatsoever are known; for instance one does not know at all if there exists an n-fraction providing all best approximations. Of course one could, on any given example, try the method outlined above and look at all points that during each step cease to be positive combinations. But in the absence of a helpful criterium like Lemma 1 that takes case of most situations, this can hardly be an effective algorithm.

6. RELATIVE MINIMA

Another type of approximation problem concerns the notion of a relative minimum of lattice. In $\mathbb{R}^{n+1}$ let a system of coordinates $(x_0,\ldots,x_n)$ be given (which we may assume to be orthogonal). For simplicity we assume that no point of Ω except for the origin has a coordinate equal to zero.

DEFINITION 6. A lattice point $R = (r_0,\ldots,r_n) \neq 0$ is a *relative minimum of* Ω if there exists no lattice point $P = (p_0,\ldots,p_n) \neq 0$ such that

$$|p_i| < |r_i|, \quad i = 0,\ldots,n.$$

When R is such a minimum, its x_j-successor $(R)_j$ is the relative minimum in the region

$$|x_i| < |r_i|, \quad i \neq j \tag{29}$$

with least value of $|x_j|$ (one may fix $x_j > 0$ to define $(R)_j$ uniquely, but the other choice is merely symmetrical about 0). Continuing the process of constructing x_j-successors, the x_j-*chain* $\{R\}_j = R,R_1,R_2,\ldots$ of a relative minimum $R = R_0$ is defined by $R_{i+1} = (R_i)_j$. Such chains play a role in certain problems of algebraic number theory (cf. §3 of part two).

The problem how to calculate the x_j-successor of a given minimum can, in principle, be solved by a Gaussian elimination process applied to the inequalities (29) completed with $0 < x_j < c$, where the x-es are expressed

through a lattice base and n+1 integral parameters; the constant c must be so that at least one integral solution exists (take the Minkowski bound for instance). From the resulting finite (though perhaps not uniformly bounded) number of solutions the x_j-successor is the one with least value of x_j. Faddeev gave, in the case n = 2, a series of stereometrical considerations reducing the number of points from which the successor must be chosen according to lowest x_j, to at most five (DELONE & FADDEEV [21], Ch. IVA). No other methods are known as yet.

The present author has, in the case of n = 2, investigated the question whether a continued fraction expansion can be constructed that contains all points of a chain of minima, using the same method as explained in §5. Though an additional difficulty arises because the points of the chain have to be found strictly in increasing order of $|x_j|$, posing the problem to determine at once if a newly found point is a point of the chain, the question was answered affirmatively. We sketch the algorithm, writing x,y,z-coordinates instead of x_0, x_1, x_2, for the construction of an x-chain.

When $A \in \Omega$ we denote by $\Sigma(A)$ the rectangle in the yz-plane with $\underline{A}$ as a vertex:

$$|y| < |a_y|, \quad |z| < |a_z|$$

if $A = (a_x, a_y, a_z)$. A base {A,B,C} of Ω is called A-*positive* when $P \in \Omega$, $P \in \Sigma(A)$ imply that P = (or -P) is a positive combination of {A,B,C}. Note that A-positivity implies that the cofactors of the base are non-negative. We call {A,B,C} A-*regular* if $\underline{B}$ and $\underline{C}$ are not in the same quadrant of the yz-plane as $\underline{A}$.

Now assume we have a lattice base {A,B,C} with A a relative minimum whose x-chain we seek; suppose {A,B,C} is A-positive and A-regular. By A-positivity, all points of the x-chain $\{A\}_x$ are still positive combinations of {A,B,C}. It is obviously sufficient if we can construct new A-positive, A-regular bases until we have a base of which we are sure that it contains the successor $(A)_x$ and is $(A)_x$-positive and $(A)_x$-regular; we then proceed to find the x-chain of $(A)_x$. For clearness we define new $\eta\zeta$-coordinates by

$$\eta = -y/a_y$$
$$\zeta = -z/a_z$$

so that $\underline{A}$ conveniently becomes (-1,-1); and in $\eta\zeta$-coordinates we put

$\underline{B} = (a,b)$
$\underline{C} = (c,d)$

where, because of A-positivity, we may assume $a > b$, $c < d$. The cofactors of A, B, C are ad-bc, d-c, a-b respectively. By I through IV we mean the quadrants of the $\eta\zeta$-plane; thus $\underline{A} \in$ III. We then have the easy criterium

LEMMA 2.

(a) *If* $\underline{B} \in \Sigma(A)$ *and* $\underline{C} \in \Sigma(A)$, *then the lowest (least* $|x|$*) of* B *and* C *is* $(A)_x$ *and* {A,B,C} *is* $(A)_x$*-positive and* $(A)_x$*-regular.*

(b) *If* $\underline{C} \in \Sigma(A)$, $\underline{B} \in IV\backslash\Sigma(A)$, *then* $C = (A)_x$ *and* {A,B,C} *is* $(A)_x$*-positive and* $(A)_x$*-regular.*

(c) *If* $\underline{C} \in \Sigma(A)$, $\underline{B} \in I\backslash\Sigma(A)$ *and* ad-bc < d-c *then* $(A)_x$ *is either* C *or* A+B, *and* {A+B,B,C} *is* $(A)_x$*-positive and* $(A)_x$*-regular.*

Furthermore, nearly all steps can be done by one of the following lemmata, easily proved by looking at the points that cease to be positive combinations and showing that their projections are not in $\Sigma(A)$.

LEMMA 3. *If* $d-c \geq \max(a-b,2)$ *then* {A,B,B+C} *is* A*-positive and* A*-regular.*

LEMMA 4. *If* $a \geq 1$ *and* $ad-bc \geq d-c$ *then* {A,A+B,C} *is* A*-positive and* A*-regular.*

The remaining situations, those were none of Lemmata 2, 3, 4 apply, all have $\underline{B} \notin \Sigma(A)$, $\underline{C} \notin \Sigma(A)$. By an argument of symmetry we may suppose $a \geq 1$, whence ad-bc < d-c because Lemma 4 does not apply. This in turn easily yields $0 < b < 1$, $c < 0$. Below we list under what conditions we take what new base, and whether the base is still A-positive and A-regular or already $(A)_x$-positive and $(A)_x$-regular. The extensive number of special cases (which could be avoided in §5) is due to the need already mentioned to verify instantly if a certain point whose projection is in $\Sigma(A)$ is indeed the lowest such point.

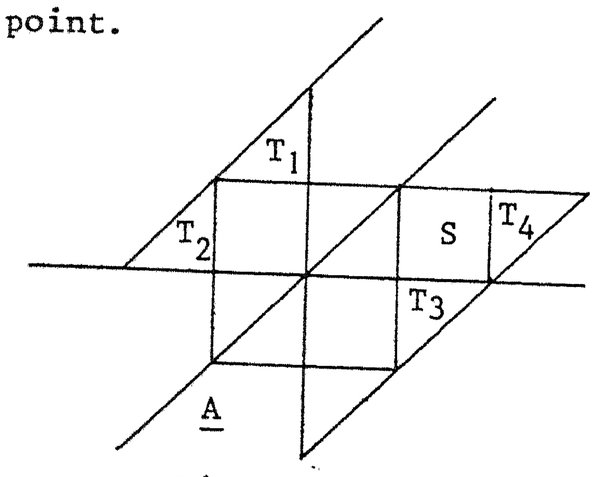


Figure 1

In Figure 1 four triangles and a square have received names referred to in the table.

if	and	then $(A)_x$ =	new base, is pos. and reg. for	
$B \in S$	$b+d > 1$	A+B	A+B, B, C	$(A)_x$
	$b+d < 1$, $a+c < 0$	A+B or B+C	A, B, B+C	A
	$b+d < 1$, $a+c > 0$	A+B or B+C	A+B, B+C, C	$(A)_x$
$\underline{B} \in T_3$, $\underline{C} \in T_2$	$\underline{B}+\underline{C} \in II$	B+C	A, B, B+C -	$(A)_x$
	$\underline{B}+\underline{C} \in I$	B+C	A+B, B+C, C	$(A)_x$
	$\underline{B}+\underline{C} \in IV$	B+C	A, B+C, C	$(A)_x$
$\underline{B} \in T_3$, $\underline{C} \in T_1$	$d-c \geq a-b$, $\underline{B}+\underline{C} \in \Sigma(A)$	B+C	A+B, B+C, C	$(A)_x$
	$d-c \geq a-b$, $\underline{B}+\underline{C} \notin \Sigma(A)$	A+B+C	A+B, B+C, A+B+C	$(A)_x$
$\underline{B} \in T_4$, $\underline{C} \in T_2$	$\underline{B}+\underline{C} \in \Sigma(A)$	B+C	A+B, B+C, C	$(A)_x$
	$\underline{B}+\underline{C} \notin \Sigma(A)$, $b+d < 1$	A+B+C or B+2C	one of A+B+C, B+C, B+2C A+B+C, B+2C, C A+B+C, B+2C, A+B+2C	$(A)_x$
	$\underline{B}+\underline{C} \notin \Sigma(A)$, $b+d \geq 1$	A+B+C	A+B, B+C, A+B+C	$(A)_x$

Part two:
Units in cubic number fields

1. CUBIC NUMBER FIELDS

A root θ of a rationally irreducible polynomial $f(x) \in \mathbb{Z}[x]$ of the third degree defines a *cubic number field* $\mathbb{Q}(\theta)$, consisting of all numbers $\alpha = x + y\theta + z\theta^2$ where x, y, z range over the rationals. The other roots of $f(x)$, to be denoted θ', θ'', are the *conjugates* of θ; the conjugates of α are $\alpha' = x + y\theta' + z\theta'^2$, $\alpha'' = x + y\theta'' + z\theta''^2$.

The *discriminant* $D(\theta)$ of θ is

$$D(\theta) = (\theta-\theta')^2(\theta'-\theta'')^2(\theta''-\theta)^2.$$

As to the sign of $D(\theta)$ there are two cases. When $D(\theta) > 0$ all roots of $f(x)$ are real; when $D(\theta) < 0$ one root is real and two are complex numbers conjugate to each other (we shall always assume then, that θ is real, and θ', θ'' are complex).

The numbers α in $\mathbb{Q}(\theta)$ satisfying polynomials in $\mathbb{Z}[x]$ with leading coefficient 1 form the ring $O(\theta)$ of *algebraic integers* in $\mathbb{Q}(\theta)$. In particular their *norm* $N(\alpha) = \alpha\alpha'\alpha''$ is an integer, being the quotient of the constant term and the leading coefficient. The only element of norm zero is 0 itself.

Every cubic field contains a number θ satisfying

$$(1) \qquad f(\theta) \equiv \theta^3 - q\theta - n = 0$$

for some $q,n \in \mathbb{Z}$. For if quite generally ρ satisfies $r\rho^3 - s\rho^2 - t\rho - u = 0$, then the minimal equation for $3r\rho - s$ is of the form (1). Moreover, in (1) we may assume that no integer $k > 1$ exists whose square divides q and whose cube divides n; for the equation satisfied by θ/k would also be of the form (1), namely $(\theta/k)^3 - q/k^2(\theta/k) - n/k^3 = 0$.

Lastly, when $D(\theta) < 0$ we may suppose $n > 0$, otherwise we could switch to $-\theta$. From now on we shall suppose that θ is given by (1) with the restrictions just explained. The discriminant is now calculated to be

$$D(\theta) = 4q^3 - 27n^2;$$

and a base of the $\mathbb{Z}$-module $O(\theta)$ is, according to a theorem of Voronoi (DELONE & FADDEEV [21], §17),

$$(2) \qquad 1, \ \frac{\theta - t}{\delta}, \ \frac{\theta^2 + t\theta + t^2 - q}{\delta^2 a}$$

Here $\delta = 3$ if $q \equiv 3 \pmod 9$ and $n \pm (q-1) \equiv 0 \pmod{27}$ for one of the sign choices, otherwise $\delta = 1$; and a is the greatest integer whose square divides $D(\theta)\delta^{-6}$ and for which the simultaneous congruences

$$(3) \qquad \begin{aligned} 3t^2 - q &\equiv 0 \pmod{\delta^2 a} \\ t^3 - qt - n &\equiv 0 \pmod{\delta^3 a^2} \end{aligned}$$

have a solution t with $-a\delta < t \leq 0$. The field discriminant then is $D(\theta)\delta^{-6}a^{-2}$. In the purely cubic case, $q = 0$, the base (2) can be simplified to

$$(4) \qquad 1, \ \theta, \ \frac{1 + t_1\theta + t_2\theta^2/k}{\delta}$$

with $\delta = 3$ if $n \equiv \pm 1 \pmod 9$, $\delta = 1$ otherwise; k is the greatest integer whose square divides n and t_1, t_2 are congruent (mod δ) to nk^{-2} and to k respectively.

Of special interest are the numbers in $O(\theta)$ whose norms are ± 1. They are the *units* of $O(\theta)$ and form a group under multiplication. The famous unit theorem of Dirichlet states that

- if $D(\theta) < 0$ there is one unit ε_0 such that $\pm\varepsilon_0^m$, $m \in \mathbb{Z}$ are all the units;
- if $D(\theta) > 0$ there are two units ε_1, ε_2 such that $\pm\varepsilon_1^m\varepsilon_2^\ell$, $m, \ell \in \mathbb{Z}$, are all the units, with $\varepsilon_1^m\varepsilon_2^\ell = 1$ only if $m = \ell = 0$.

The unit ε_0 and the pair of units ε_1, ε_2 are called *fundamental* (in fact, with ε_0 also $-\varepsilon_0$ and $\pm\varepsilon_0^{-1}$ are fundamental; the pair $\varepsilon_1^a\varepsilon_2^b$, $\varepsilon_1^c\varepsilon_2^d$ is fundamental if and only if $ad - bc = \pm 1$).

The determination of the fundamental unit in the case $D(\theta) < 0$ is the subject of §2; the case $D(\theta) > 0$ is treated

2. THE CASE OF NEGATIVE DISCRIMINANT

In this section $D(\theta) < 0$, so that there is one fundamental unit ε_0, fixed by $\varepsilon_0 > 1$. We embed $O(\theta)$ into $\mathbb{R}^3$ by putting

$$\Omega(\theta) = \{(\alpha, \text{Re } \alpha', \text{ Im } \alpha') \mid \alpha \in O(\theta)\}.$$

By the additive structure of the ring $O(\theta)$, $\Omega(\theta)$ is a lattice; the point $(\alpha, \text{Re } \alpha', \text{ Im } \alpha')$ will also be denoted by α. Choosing the height function

$$h(x,y,z) = x$$

and the Euclidean distance function

$$d(x,y,z) = \sqrt{y^2+z^2}$$

we have

(5) $$h(\alpha) = \alpha$$

and

(6) $$d(\alpha) = |\alpha'|.$$

LEMMA 1. *If ε is a unit, then ε is a best approximation of $\Omega(\theta)$ to the x-axis (with respect to the height and distance functions just introduced).*

PROOF. Suppose $\alpha \in \Omega(\theta)$ satisfies $d(\alpha) \leq d(\varepsilon)$, $|h(\alpha)| < |h(\varepsilon)|$. [1] By (5) and (6) this implies $|N(\alpha)| = |\alpha| \cdot |\alpha'|^2 < |\varepsilon| \cdot^2 |\varepsilon'|^2 = |N(\varepsilon)| = 1$, whence $N(\alpha) = 0$, i.e. $\alpha = 0$.

LEMMA 2. *If β is a best approximation, ε a unit, then $\beta\varepsilon$ is a best approximation.*

---

1) Note that no two points of $\Omega(\theta)$ have equal height.

PROOF. Suppose $\alpha \in \Omega(\theta)$ satisfies $d(\alpha) \le d(\beta\varepsilon)$, $|h(\alpha)| < |h(\beta\varepsilon)|$. By (5) and (6), this means $|\alpha'| \le |\beta'\varepsilon'|$, $|\alpha| < |\beta\varepsilon|$, whence $|\alpha'\varepsilon'^{-1}| \le |\beta'|$, $|\alpha\varepsilon^{-1}| < |\beta|$, i.e. $d(\alpha\varepsilon^{-1}) \le d(\beta)$ and $|h(\alpha\varepsilon^{-1})| < |h(\beta)|$. But, β being a best approximation, this implies $\alpha = 0$.

COROLLARY. *If* $1 < \beta_1 < \ldots < \beta_k = \varepsilon_0$ *are the best approximations with heights between* 1 *and* ε_0, *then every best approximation is of the form* $\pm\beta_\ell\varepsilon_0^m$ *for some* $m \in \mathbb{Z}$, $\ell \in \{1,\ldots,k\}$.

PROOF. If β is a best approximation, then so is $|\beta|\varepsilon_0^{-m}$ with m defined by $\log|\beta| > m\log\varepsilon_0 \ge \log|\beta| - \log\varepsilon_0$. But this choice gives $1 < |\beta|\varepsilon_0^{-m} \le \varepsilon_0$.

The fundamental unit, and in fact all best approximations, can thus be calculated by a best approximation two-fraction such as presented in §5 of part one.

SOME HISTORICAL REMARKS. In the past some small and not always correct tables of fundamental units were computed by hand, chiefly by trial and error methods. The oldest one seems to be given by MARKOV [39] in 1891, for cubic fields $Q(\sqrt[3]{n})$, $n \le 70$ cubefree (reproduced on p. 304 of DELONE & FADDEEV [21]). Other such tables exist from DEDEKIND [20] and CASSELS [15] (for $n \le 23$, $n \le 50$ respectively). REID [46] gave an incomplete table of bases, discriminants, units and class numbers for many fields (1) with $|q| \le 9$, $1 \le n \le 9$ (reproduced in DELONE & FADDEEV [21], p. 141). WOLFE [56] calculated the fundamental unit of the module $M(\theta)$ with base $\{1,\theta,\theta^2\}$ for $\theta = \sqrt[3]{n}$, $2 \le n \le 100$ cubefree. - This module is not always $O(\theta)$; in fact the fundamental unit of $M(\theta)$ might be any power of the fundamental unit of $O(\theta)$ (see NAGELL [41], note I), unlike the real quadratic case where the fundamental unit of the module with base $\{1,\sqrt{d}\}$ is either the fundamental unit of $O(\sqrt{d})$ or its third power (NARKIEWICZ [42], p. 112). We note, however, that it is relatively easy to decide whether some unit in $O(\theta)$ is ε_0 or not: Clearly it suffices to check if $\varepsilon^{1/p}$ is an algebraic integer in $O(\theta)$ for primes p, and an upper bound for p can be given e.g. from the fact that $\varepsilon_0 > 2$ for all fields with discriminant not equal to -23, -31 (so that $p < {}^2\log\varepsilon$); a somewhat more sophisticated version of this argument was developed recently by JEANS & HENDY [35] for the purely cubic case -.

It was recognized early that the theory of two-fractions provided a more systematic approach: the problem has been used as an example of the application of nearly all algorithms that have been proposed.

The oldest method is to apply a two-fraction to $\ell_0 = (1,w_1,w_2)$ (this being a base of $O(\theta)$) and $\Omega = \mathbb{Z}^3$. The cofactors of the initial triple $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ are 1, w_1 and w_2 so that all later cofactors belong to $O(\theta)$; and one of them may happen to be a unit. This method is specially favourite with Jacobi-type algorithms. For example, GÜTING [31] used it for 255 values of $\theta = \sqrt[3]{n}$, $2 \le n \le 999$; he missed several fundamental units that should be obtainable with his 16-digit computer precision (e.g. $n = 34,51$). A better result was obtained by SVED [50] who used the algorithm of SZEKERES [51]. Often using a hundred or more digits she compiled a table of units for $\theta = \sqrt[3]{n}$, $2 \le n \le 199$. She did not stop after the first unit was found, but continued to find more; if all were powers of the first one, it was reasonable to expect the first one to be ε_0. By means of the algorithm presented in §5 of part one, Mr. te Riele of the Mathematical Centre, Amsterdam, checked several of Sved's largest units (which go up to about 6.10^{95} in the case of $n = 167$) and found them all fundamental indeed.

A second method is, again to use $\ell_0 = (1,w_1,w_2)$ and $\Omega = \mathbb{Z}^3$, and hope to obtain a periodical expansion with some algorithm. The unimodular transformation connected with one period has a unit of $\mathbb{Q}(\theta)$ for an eigenvalue. Results of this method are found, e.g. in BERNSTEIN [7] (who used the Jacobi-Perron algorithm). But very often the units obtained this way are not fundamental; for example, the eigenvalue of the expansion for $\theta = \sqrt[3]{2}$ with Brun's algorithm is ε_0^5. DAUS [19] gave a completer version of Reid's table forcing periodicity on the Jacobi-Perron algorithm by a clever choice of the integer b (achieved seemingly through trial and error).

The characterization of units as best approximations in an appropriate setting was first found in the work of VORONOI [54] and BERWICK [9]; later DUBOIS [22] also used it, but none of these authors employed a two-fraction for actual calculations.

ANGELL [1] compiled a table of all fields with negative discriminants between -20000 and 0 (there are 3169 such fields), using Voronoi's method for the units. In [62] the present author uses the new algorithm of this paper, part one, §5, to present a table of fundamental units of all cubic fields $x^3 - qx - n = 0$ with negative discriminant and $|q|,|n| \le 10$.

USPENSKY [52] gave a method, based on successive minima of certain quadratic forms associated with $\mathbb{Q}(\theta)$, yielding usually the fundamental unit, occasionally its square.

3. THE CASE OF POSITIVE DISCRIMINANT

In this section, $D(\theta) > 0$. The following technique which guarantees that a pair of fundamental units will be found, is due to VORONOI [54] and involves chains of relative minima. We embed $O(\theta)$ into $\mathbb{R}^3$ by putting

$$\Omega(\theta) = \{(\alpha,\alpha',\alpha'') \mid \alpha \in O(\theta)\}.$$

The point $(\alpha,\alpha',\alpha'')$ will also be denoted by α. The additive ring structure of $O(\theta)$ makes $\Omega(\theta)$ a lattice. If α is a relative minimum of $\Omega(\theta)$ it follows from the Minkowski theorem that $|N(\alpha)| \leq \sqrt{D}$ where D is the discriminant of the field $Q(\theta)$; for the parallelepiped

$$|x| < |\alpha|, \quad |y| < |\alpha'|, \quad |z| < |\alpha''|$$

whose volume is $8|\alpha\alpha'\alpha''| = 8|N(\alpha)|$, must not contain other lattice points than 0. Furthermore, we have

LEMMA 3. *If α is a relative minimum, ε a unit, then $\alpha\varepsilon$ is a relative minimum; moreover if β is the x-successor of α, then $\beta\varepsilon$ is the x-successor of $\alpha\varepsilon$.*

PROOF. Analogous to Lemma 2.

Now take any relative minimum α and construct its x-chain $\{\alpha\}_x$ of relative minima; let it be

$$\alpha = \alpha_0 < \alpha_1 < \alpha_2 \ldots . \tag{7}$$

Using the formula $|N(\alpha_i)| \leq \sqrt{D}$ and the fact that for given norm there exist only finitely many non-associated elements in $O(\theta)$ of that norm, one infers that the chain contains two points $\alpha_k < \alpha_\ell$ whose quotient $\varepsilon = \alpha_\ell/\alpha_k$ is a non-trivial unit in $O(\theta)$. By Lemma 3 the chain (7) must be periodical at least from α_k onward, i.e. it reads

$$\alpha_k < \ldots \alpha_{\ell-1} < \alpha_\ell = \varepsilon\alpha_k < \varepsilon\alpha_{k+1} \ldots < \varepsilon\alpha_\ell = \varepsilon^2\alpha_k < \ldots . \tag{8}$$

But (8) can also be extended to the left, using Lemma 3. The two-sided chain of minima

(9) $$\ldots\ \alpha_k\varepsilon^{-2} < \ldots < \alpha_k\varepsilon^{-1}\ \ldots < \alpha_k < \ldots < \alpha_k\varepsilon < \ldots$$

has the property that each element is its left side neighbour's x-successor. (Note that (9) does not necessarily contain α_0.)

Through stereometric considerations (DELONE & FADDEEV [21], Ch. IVA) the following facts can be shown:

LEMMA 4. *Two two-sided chains of minima of different directions (e.g. an x-chain and a z-chain) always have a common element.*

LEMMA 5. *Two two-sided x-chains have either no or all elements in common.*

Using these facts, the following algorithm due to Voronoi guarantees that a pair of fundamental units will be found:

(a) Choose an arbitrary relative minimum α_0 and construct its x-chain $\alpha_0 < \alpha_1 < \ldots$.

(b) Determine ε_1 as the chain's first automorphism, that is

(10) $$\varepsilon_1 = \alpha_\ell/\alpha_k$$

where $\ell > k$ are such that α_j/α_i is not a unit when $\ell > j > i$.

(c) Construct the y- (or z-) chain of α_k, say $\alpha_k = \beta_0 < \beta_1 < \ldots$.

(d) Determine the least $m \geq 1$ such that β_m is associated to one of the numbers $\alpha_k, \alpha_{k+1}, \ldots, \alpha_{\ell-1}$, and supposing this to be α_j, put

(11) $$\varepsilon_2 = \beta_m/\alpha_j.$$

Now ε_1 and ε_2 from (10), (11) form a fundamental pair.

In §6 of part one it was seen how the calculations in (a) and (c) can be done by means of a two-fraction. As to (b) and (d), it is clear that one only needs to compare elements of equal norm.

HISTORICAL REMARKS. In 1976, ANGELL [2] made use of Voronoi's method when compiling a unit table for all cubic fields with positive discriminant less than 100.000 (there are 4794 such fields). So did WILLIAMS and ZARNKE [55] for a table of all fields $\theta^3 - q\theta - n = 0$ with $|q|, |n| \leq 50$.

BERWICK [10] proved in 1932 that any two of the units ε_1, ε_2, ε_3 defined by

$\lvert\varepsilon_1'\rvert < 1$,	$\lvert\varepsilon_1''\rvert < 1$	$\lvert\varepsilon_1\rvert$ minimal
$\lvert\varepsilon_2\rvert < 1$,	$\lvert\varepsilon_2''\rvert < 1$	$\lvert\varepsilon_2'\rvert$ minimal
$\lvert\varepsilon_3\rvert < 1$,	$\lvert\varepsilon_3'\rvert < 1$	$\lvert\varepsilon_3''\rvert$ minimal

form a fundamental pair. However, these cannot be found by a two-fraction algorithm; the algorithm proposed by BILLEVICH [11] does the calculation in a very unpractical way (see RUDMAN & STEINER [48]). Another process, rather related to Voronoi's, was proposed by BERGMANN [4]. DAUS [19] has some success by his variation of the Jacobi-Perron algorithm. Fundamentality of some units of BERNSTEIN [6] was disproved by STENDER [49]. To conclude we mention some work on units in cubic fields of positive discriminant that has no connection with continued fraction algorithms: GODWIN [27], GODWIN & SEMET [28], BRUNOTTE & HALTER-KOCH [14].

The special (and easier) case of a cyclic field (where θ', θ'' belong to $\mathbb{Q}(\theta)$, or, equivalently, $D(\theta)$ is a square) was considered by HASSE [32], GRAS-MONTOUCHET [29], COHN & GORN [16].

REFERENCES

[1] ANGELL, I.O., *A table of complex cubic fields*, Bull. London Math. Soc. 5 (1973), 37-38.

[2] ANGELL, I.O., *A table of totally real cubic fields*, Math. Comp. 30 (1976), 184-187.

[3] BARBOUR, J.M., *Music and ternary continued fractions*, Am. Math. Monthly 55 (1948), 545-555.

[4] BERGMANN, G., *Ein Beispiel numerischer Einheitenberechnung*, Math. Ann. 167 (1966), 143-168.

[5] BERNSTEIN, L., *Rational approximations of algebraic irrationals by means of a modified Jacobi-Perron algorithm*, Duke Math. J. 32 (1965), 161-176.

[6] BERNSTEIN, L. & H. HASSE, *An explicit formula for the units of an algebraic number field of degree* $n \geq 2$, Pac. J. Math. 30 (1969), 293-365.

[7] BERNSTEIN, L., *Einheitenberechnung in kubischen Körpern mittels des Jacobi-Perronschen Algorithmus aus der Rechenanlage*, J. reine angew. Math. 244 (1970) 201-220.

[8] BERNSTEIN, L., *The Jacobi-Perron algorithm, its theory and application*, Lecture Notes in Mathematics 207, Springer Verlag, 1971.

[9] BERWICK, W.E.H., *The classification of ideal numbers that depend on a cubic irrationality*, Proc. London Math. Soc. Ser. 2, XII (1913), 393-429.

[10] BERWICK, W.E.H., *Algebraic number fields with two independent units*, Proc. London Math. Soc. Ser. 2, XXXIV (1932), 360-378.

[11] BILLEVICH, K.K., *On units of algebraic number fields of third and fourth degrees*, (Russian) Mat. Sb. 40 (82) (1956), 123-136.

[12] BRUN, V., *En generalisation av kjedebrøken*, Skr. Vid. Selsk. Kristiania, Mat. Nat. Kl., 1919, nr. 6 og 1920, nr. 6.

[13] BRUN, V., *Algorithmes euclidiens pour trois et quatre nombres*, XIII^{e} Congr. Math. Scand., Helsinki 1957, 45-64.

[14] BRUNOTTE, H. & F. HALTER-KOCH, *Zur Einheitenberechnung in totalreellen kubischen Zahlkörpern nach Godwin*, J. Number Th. 11 (1979), 552-559.

[15] CASSELS, J.W.S., *The rational solutions of the Diophantine equation* $y^2 = x^3 - D$, Acta Math. 82 (1950), 244-273.

[16] COHN, H. & S. GORN, *A computation of cyclic cubic units*, J. Research, Nat. Bur. Stand., 59 (1967), 155-168.

[17] CUSICK, T.W., *The Szekeres multi-dimensional continued fraction*, Math. Comp. 31 (1977), 280-317.

[18] CUSICK, T.W., *Best Diophantine approximations for ternary linear forms*, J. reine angew. Math. 315 (1980), 40-52.

[19] DAUS, P.H., *Normal ternary continued fraction expansions for cubic irrationalities*, Amer. J. Math. 51 (1929), 67-98.

[20] DEDEKIND, R., *Ueber die Anzahl der Idealklassen in reinen kubischen Zahlkörpern*, J. reine angew. Math. 121 (1900), 40-123.

[21] DELONE, B.N. & D.K. FADDEEV, *The theory of irrationalities of the third degree*, Am. Math. Soc. Transl. of Monographs, 10, 1964.

[22] DUBOIS, E., *Meilleures approximations diophantiennes. Applications à la recherche de l'unité fondamentale des corps cubiques non totalement reéls*, C.R. Acad. Sci. Paris, t. 289 (1979) 39-41.

[37] KLEIN, F., *Ausgewählte Kapittel der Zahlentheorie*, pp. 17 sqq, Teubner, Leipzig 1895.

[38] LAGARIAS, J.C., *Some new results in simultaneous Diophantine approximation*, Proc. Queen's Univ. Conf. on Numb. Th., 1979 (preprint).

[39] MARKOV, A., *Sur les nombres entiers dépendants d'une racine cubique d'un nombre entier ordinaire*, Mem. Acad. Imp. Sci., Petersburg, (7), 38 (1892) no. 9, 1-37.

[40] MINKOWSKI, H., *Zur Theorie der Kettenbrüche*, Gesammelte Abh. Band I 278-292, Teubner, Leipzig 1911.

[41] NAGELL, T., *Solution complète de quelques équations cubiques à deux indeterminées*, J. Math. Pure Appl. Ser. 9 Vol. 4 (1925), 209-270.

[42] NARKIEWICZ, W., *Elementary and analytic theory of algebraic numbers*, Warszawa 1974.

[43] PALEY, R.E.A.C. & H.D. URSELL, *Continued fractions in several dimensions*, Proc. Camb. Phil. Soc. XXVI (1930) 127-144.

[44] PERRON, O., *Grundlagen für eine Theorie des Jacobischen Kettenbruchalgorithmus*, Math. Ann. 64 (1907), 1-76.

[45] PERRON, O., *Die Lehre von den Kettenbrüchen*, Teubner, Stuttgart 1954.

[46] REID, L.W., *Tafel der Klassenanzahlen für kubische Zahlkörper*, Dissertation, Göttingen 1899.

[47] ROSSER, B., *A note on the linear Diophantine equation*, Amer. Math. Monthly 48 (1941), 662-666.

[48] RUDMAN, R.J. & R.P. STEINER, *On an algorithm of Billevich for finding units in algebraic number fields*, Math. Comp. 30 (1976), 598-609.

[49] STENDER, H.J., *Einheiten für eine allgemeine Klasse total reeller algebraischer Zahlkörper*, J. reine angew. Math. 257 (1972), 151-178.

[50] SVED, M., *Units in pure cubic number fields*, Ann. Univ. Sci. Budapest Eötvös, Sect. Math., 13 (1970), 141-149.

[51] SZEKERES, G., *Multidimensional continued fractions*, Ann. Univ. Sci. Budapest Eötvös, Sect. Math., 13 (1970), 113-140.

[23] DUBOIS, E., *Calculation of F-best approximations of zero by a ternary form. Computation of units*, subm. to Math. Comp. (preprint).

[24] EULER, L., *De relatione inter ternas pluresve quantitates instituenda*, Comm. Arithm. Coll., vol. 2, p. 99 sqq, Petersburg 1849.

[25] FERGUSON, H.R.P. & R.W. FORCADE, *Generalization of the Euclidean algorithm for real numbers to all dimensions higher than two*, preprint (Res. Ann. Bull. A.M.S., Nov. 1979).

[26] FURTWÄNGLER, Ph., *Ueber die simultane Approximation von Irrationalzahlen*, Math. Ann. 99 (1927), 71-83.

[27] GODWIN, H.J., *The determination of units in totally real cubic fields*, Proc. Camb. Phil. Soc., 56 (1960), 318-321.

[28] GODWIN, H.J. & P.A. SEMET, *A table of real cubic fields*, J. London Math. Soc. 34 (1959), 108-110.

[29] GRAS-MONTOUCHET, M.N., *Méthodes et algorithmes pour le calcul numérique du nombre de classes et des unités des extensions cubiques cycliques de* $\mathbb{Q}$, J. reine angew. Math. 277 (1975), 89-116.

[30] GREITER, G., *Mehrdimensionale Kettenbrüche*, Dissertation, Techn. Univ. München, Aug. 24, 1977.

[31] GÜTING, R., *Zur Verallgemeinering des Kettenbruch-Algorithmus I, II*, J. reine angew. Math. 278/279 (1975), 165-173, 281 (1976), 184-198.

[32] HASSE, H., *Arithmetische Bestimmung von Grundeinheiten und Klassenzahl in zyklischen kubischen und biquadratischen Zahlkorpern*, Abh. Deutsche Akad. Wiss., Berlin, 2 (1948), 1-95.

[33] JACOBI, C.G.J., *Über die Auflösung der Gleichung* $a_1x_1+\ldots+a_nx_n$ = f.u., J. reine angew. Math., 69 (1868), 1-28.

[34] JACOBI, C.G.J., *Allgemeine Theorie der Kettenbruchähnlichen Algorithmen*, J. reine angew. Math. 69 (1868), 29-64.

[35] JEANS, N.S. & M.D. HENDY, *Some inequalities related to the determination of the fundamental unit of a pure cubic field*, Math. Comp. 32 (1978), 925-935.

[36] JURKAT, W., W. KRATZ & A. PEYERIMHOFF, *On best two-dimensional Dirichlet approximations and their algorithmic calculation*, Math. Ann. 244 (1979), 1-32.

[52] USPENSKY, J.V., *A method for finding units in cubic orders of a negative discriminant*, Transactions Amer. Math. Soc. 33 (1931).

[53] VAUGHAN, T.P., *A generalization of the simple continued fraction algorithm*, Math. Comp. 32 (1978), 537-558.

[54] VORONOI, G.F., *On a generalization of the algorithm of continued fractions*, (Russian), Dissertation, Warszawa 1896; review in Jahrb. Fortschr. Math. 27 (1896), 170-174; also in [21].

[55] WILLIAMS, H.C. & C.R. ZARNKE, *A table of fundamental units for cubic fields*, Sci. Report 63, Univ. of Manitoba, 1973.

[56] WOLFE, C., *On the indeterminate cubic equation* $x^3+Dy^3+D^2z^3-3Dxyz = 1$, Univ. Calif. Publ. Math. 1, 16 (1923) 359-369.

[57] BERGMAN, G.M., *Notes on Ferguson and Forcade's generalized Euclidean algorithm*, preprint.

[58] STROEKER, R.J. & R. TIJDEMAN, *Diophantine equations*, these proceedings.

[59] VAN DE LUNE, J. & H.J.J. TE RIELE, *Explicit computation of special zeros of partial sums of Riemann's zeta function*, these proceedings.

[60] WILLIAMS, H.C., G. CORMACK & E. SEAH, *Calculator of the regulator of a pure cubic field*, Math. Comp. 34 (1980), 567-611.

[61] BRENTJES, A.J., *Multi-dimensional continued fraction algorithms*, MC Tract 145, Mathematical Centre, Amsterdam, 1981.

[62] BRENTJES, A.J., *A two-dimensional continued fraction algorithm for best approximations with an application in cubic number fields*, J. reine angew. Math. 326 (1981), 18-44.

DIOPHANTINE EQUATIONS

by

R.J. STROEKER & R. TIJDEMAN

1. INTRODUCTION

A *diophantine equation* is usually defined as an equation in integers or in rationals, viz.

$$f(x_1,x_2,\ldots,x_n) = 0 \tag{1}$$

in the variables $x_1,x_2,\ldots,x_n$. Sometimes a more general definition is adopted by asking for solutions taken from other algebraic structures like an algebraic number field, the ring of integers of such a field or a finite field. But, anyway one should always restrict the solutions to those one could rightfully call rational or integral, in some sense. In the present paper we shall restrict ourselves to diophantine equations in rational integers.

It is almost impossible to classify diophantine equations in some sensible way. The ad hoc character of the subject, especially of the period before 1930, is shown very clearly in Dickson's famous history on the theory of numbers [23]. Some sort of classification, which is useful in practice, is based upon methods and techniques used for solving diophantine equations. Roughly speaking, diophantine analysis borrows mainly from the following fields:

(a) Elementary Number Theory, (b) Algebraic Number Theory, (c) Algebraic Geometry, (d) p-adic Analysis, (e) Diophantine Approximation Theory, and (f) Miscellaneous Theories (like Logic, Combinatorics, etc.).

The most general results are obtained by combining (a), (b), (d) and (e). Certain classes of diophantine equations are introduced in Section 2 and some existence results about their solutions are given in Section 3. The remainder of the paper is devoted to methods for effectively determining the complete solution of diophantine equations. There are two parts which can be read independently of each other.

A. *Algebraic methods*. Some constructive techniques taken from the fields (b), (a), (c) and (d) are illustrated by specific equations. We indicate for which parts computers have been successfully used.

B. *Approximation methods*. We give a survey of the equations which have been completely solved by the Gelfond-Baker method. Here methods from (e), (a), (b) and (d) have been combined.

Since rounding errors make the use of computers particularly dangerous when approximation methods are being applied, we add a contribution of P.L. Cijsouw, A. Korlaar and R. Tijdeman. This appendix describes the numerical treatment of the inequality

$$|p^x - q^y| \leq p^{x/2}$$

which is applied in part B. It might also serve as a general indication as to how one can make sure that the complete solution of some equation will indeed be found.

Information on diophantine equations in general, also of a historical nature, can be found in the books written by BAŠMAKOVA [10], MORDELL [46] and SKOLEM [60].

We thank Prof. P.L. Cijsouw, Prof. H.W. Lenstra and Prof. A.J. van der Poorten for their valuable remarks on preliminary versions of this paper.

2. CLASSES OF DIOPHANTINE EQUATIONS

Most attention has been paid to *polynomial equations* where the function f in (1) is a polynomial with rational integer coefficients. If $n = 2$, we call f a *binary* polynomial. If f is homogeneous, it is said to be a *form*. If f is a binary form and m is a rational integer, then the equation

$$f(x,y) = m, \quad x,y \in \mathbb{Z}, \tag{2}$$

is called a *Thue-equation*. If f is irreducible, then this equation is an example of a so-called *norm form equation*. Indeed, let ξ be a root of $f(t,1) = 0$ and let K be the number field generated by ξ over $\mathbb{Q}$. Then

$$f(x,y) = \mathrm{Norm}_{K/\mathbb{Q}}(x - \xi y).$$

Norm form equations play an important rôle in solving the so-called *Weierstrass-equation*

$$y^2 = x^3 + ax + b, \quad x,y \in \mathbb{Z}$$

where $a,b \in \mathbb{Z}$ with $4a^3 + 27b^2 \neq 0$. This equation represents an elliptic curve defined over the rationals. On the other hand, any elliptic curve defined over $\mathbb{Q}$ can be represented by such an equation. A wealth of information on this equation may be found in CASSELS' survey article [19]; see also ZIMMER [77]. An important special case, obtained by setting $a = 0$, is the so-called *Mordell-equation*,

$$y^2 = x^3 + k, \quad x,y \in \mathbb{Z}, \tag{3}$$

where $k \neq 0$ is a fixed integer. This is an example of a *hyperelliptic equation*,

$$y^m = f(x), \quad x,y \in \mathbb{Z},$$

where m is an integer with $m \geq 2$ and f is a polynomial with integer coefficients.

If exponents are considered variable as well, we speak of an *exponential equation*. Famous examples are the *Fermat equation*,

$$x^n + y^n = z^n, \quad n,x,y,z \in \mathbb{N}, \ n \geq 3,$$

the *Pillai-equation*

$$ax^m - by^n = c, \quad m,n,x,y \in \mathbb{N}, \ m \geq 3, \ n \geq 2,$$

where a, b, c are fixed integers with $abc \neq 0$, and the *Thue-Mahler equation*,

$$f(x,y) = p_1^{z_1} \ldots p_s^{z_s}, \quad x,y \in \mathbb{Z}; \ z_1,\ldots,z_s \in \mathbb{N} \cup \{0\},$$

where f is a binary form and $p_1,p_2,\ldots,p_s$ are fixed primes.

If f in (1) has the special form

$$X_1 + X_2 + \ldots + X_n = 0,$$

where n is a fixed integer and $X_1, X_2, \ldots, X_n$ are integers composed of primes taken from some fixed set S, then we speak of a *purely exponential equation*. Examples of such equations are

$$2^x - 2^y = 3^z - 3^w, \quad x,y,z,w \in \mathbb{N}, \ x > y, \ z > w,$$

and

$$2^x + 3^y = 5^z 7^w, \quad x,y,z,w \in \mathbb{N} \cup \{0\}.$$

There are many equations occurring in the literature which we have not yet classified, like the equation $x^{-1} + y^{-1} + z^{-1} = 4w^{-1}$ in $x,y,z,w \in \mathbb{N}$ and the Goldbach-equation $p_1 + p_2 = 2n$ in primes p_1, p_2 and integer $n \in \mathbb{N}$. We shall not deal with unclassified equations in this paper.

3. SOME RESULTS OBTAINED BY GENERAL METHODS

A method is called *ineffective* if it provides a demonstration of the finiteness of the number of solutions (possibly by giving an explicit upper bound for this number), *without* providing an algorithm to determine the complete set of solutions. The most important ineffective method is due to Thue and Siegel, whereafter important extensions were developed by Mahler, Roth, Schmidt and others.

The main theorem of Thue is as follows:

THEOREM 3.1. [72]. *Let* f *be a binary form with coefficients in* $\mathbb{Z}$ *and of degree at least* 3. *If* f *is irreducible over* $\mathbb{Q}$, *then for any* $m \in \mathbb{Z}$ *the equation* $f(x,y) = m$ *has at most a finite number of solutions in integers* x *and* y.

If $m \neq 0$, then the condition "f is irreducible over $\mathbb{Q}$" can be replaced by "f has at least three distinct factors in its factorization in linear forms", cf. [55]. MAHLER [42] introduced a p-adic analogue of Thue's method and proved that under the conditions of Theorem 3.1 the greatest prime factor of $f(x,y)$ tends to infinity if $\max(|x|,|y|) \to \infty$ subject to $(x,y) = 1$. A straightforward generalization yields the following result on the Thue-Mahler equation.

THEOREM 3.2. [55]. *Let* $f \in \mathbb{Z}[x,y]$ *be a binary form such that among the linear factors in the factorization of* f *at least three are distinct. Let* $p_1, p_2, \ldots, p_r$ *be primes. Then the equation*

$$f(x,y) = p_1^{z_1} p_2^{z_2} \dots p_r^{z_r}$$

has only finitely many solutions in integers $x,y,z_1,z_2,\dots,z_r$ *with* $z_1 \geq 0$, $z_2 \geq 0,\dots,z_r \geq 0$.

COROLLARY.[42]. *The purely exponential equation*

$$X_1 + X_2 = X_3$$

in integers composed of primes taken from some fixed set S, *has only finitely many solutions with* $(X_1,X_2) = 1$.

SIEGEL [56,57] extended Thue's method to hyperelliptic equations. He generalized results of Mordell and himself by proving:

THEOREM 3.3. [57]. *The equation* $f(x,y) = 0$ *for* $f \in \mathbb{Z}[x,y]$ *has only a finite number of integral solutions* x, y *if the curve it represents has genus* ≥ 1, *or has genus* 0 *and at least three infinite valuations.*

COROLLARY.[36]. *Let* m *be a positive integer,* $m \geq 2$. *Let* $f \in \mathbb{Z}[x]$. *Put*

$$f(x) = a_0x^n + a_1x^{n-1} + \dots + a_n = a_0 \prod_i (x - \alpha_i)^{r_i},$$

with $a_0 \neq 0$, $\alpha_i \neq \alpha_j$ *for* $i \neq j$. *Then the equation*

$$y^m = f(x)$$

has only finitely many solutions in integers x, y, *unless*

(i) *all but one number* r_i *are multiples of* m; *or*

(ii) *all but two numbers* r_i *are multiples of* m *and all of them are multiples of* m/2.

This result covers the Weierstrass-equation and hence the Mordell-equation.

We end with a result on purely exponential equations which is proved by a p-adic analogue of the extension of Roth and Schmidt of the Thue-Siegel method. The result is due to SCHLICKEWEI [52] and to DUBOIS and RHIN [24].

THEOREM 3.4. *Let* n *be a positive integer. The equation* $X_1 + X_2 + \dots + X_n = 0$ *in integers composed of primes taken from some fixed set* S *has only finitely*

many solutions with $(X_i,X_j) = 1$ *for all* i *and* j *with* $i \neq j$.

Of course, an algorithm is called *effective* if it provides an algorithm to determine the complete set of solutions of a diophantine equation. The most important effective approximation method is due to Gelfond and Baker. It can be used to derive upper bounds for the absolute values of the solutions of equations of different types. For a survey of such results we refer to T.N. SHOREY et al. [55]. In particular, there are now effective proofs of Theorems 3.1 (see BAKER [7]) and 3.2 (see COATES [20]), but not of Theorems 3.3 and 3.4. BAKER [8] gave an effective proof of the corollary of Theorem 3.3 under the extra condition that f has at least two (three if $m = 2$) simple roots. Obviously there is an effective proof of Theorem 3.4 if $n = 3$, this being a corollary of Theorem 3.2.

Part A. ALGEBRAIC METHODS

4. SOME RESULTS FROM ALGEBRAIC NUMBER THEORY

The purpose of the following example is to suggest that often the relation between the variables occurring in a diophantine equation can be made transparent by simple factorization. By this we mean application of the *Fundamental Theorem of Arithmetic*: *any positive integer may be written in one way only as a product of primes, except for the order in which the primes occur in the product.*

EXAMPLE 4.1. For given $k \in \mathbb{Z}$, $k \neq 0$ consider the equation $x^4 = y^2 + k$. If $x,y \in \mathbb{Z}$ gives a solution of this equation, then $(x^2-y)(x^2+y) = k$ and a divisor d of k exists such that

$$x^2 - y = d \quad \text{and} \quad x^2 + y = \frac{k}{d} .$$

Here we may assume that $\frac{k}{d} \geq d > 0$, because there is no loss of generality in taking $y \geq 0$. Thus

$$x^2 = \frac{1}{2}\,(d + \frac{k}{d}) \quad \text{and} \quad y = \frac{1}{2}\,(\frac{k}{d} - d) .$$

The number of divisors of k is finite and so it should be immediately clear from the above whether solutions do exist and if so how they can be computed.

If in addition one requires k to be prime (k = p), then d can have no value other than 1 and consequently

$$x^2 = \frac{1}{2}(p+1) \quad \text{and} \quad y = \frac{1}{2}(p-1).$$

This shows that at most one solution in positive integers x and y can exist. The prime numbers $p < 100$ for which the equation is soluble are p = 7, 17, 31, 71 and 97. □

Most constructive methods used in diophantine problems apply at some stage factorization in certain algebraic number fields. Therefore, we intend to formulate a few theorems from the realm of Algebraic Number Theory, which in our view are of fundamental importance in the process of solving diophantine equations. We shall give no proofs, but confine ourselves to indicating the relevant places in the literature. (See also the introductory sections of the expositions by H. Zantema and R.J. Schoof in these proceedings, [76] and [53].)

Let K be a number field (a number field is a finite - and thus algebraic - extension of the field $\mathbb{Q}$) with ring of algebraic integers $\mathcal{O} = \mathcal{O}(K)$. An ideal of $\mathcal{O}$ has a finite basis. A *fractional* ideal of $\mathcal{O}$ is a finitely generated $\mathcal{O}$-module $a \neq 0$, contained in K. Hence, each ideal $a \neq 0$ of $\mathcal{O}$ is also a fractional ideal of $\mathcal{O}$; in this context ideals of $\mathcal{O}$ are sometimes called *integral* ideals. In the set of fractional ideals of $\mathcal{O}$ we define multiplication as follows: the *product* $a \cdot b$ of the fractional ideals a and b is the fractional ideal generated by all products $\alpha\beta$ with $\alpha \in a$ and $\beta \in b$. In this way, the set of fractional ideals of $\mathcal{O}$ becomes a group, the so-called *ideal group* of K. This group is denoted by $\mathcal{I} = \mathcal{I}(K)$.

A direct generalization of the fundamental theorem of arithmetic is given in Dedekind's theorem.

THEOREM 4.2. (see JANUSZ [31], Theorem I.4.2). *Each fractional ideal a of $\mathcal{O}$ can be written as a product* $a = \mathfrak{p}_1^{a_1} \ldots \mathfrak{p}_n^{a_n}$, *where the* $\mathfrak{p}_i$*'s are distinct prime ideals of $\mathcal{O}$ and the* a_i*'s are non-zero integers. This product representation is unique, except for the order in which the prime ideals occur.*

From this theorem it easily follows that the ideal group $\mathcal{I}$ is a free abelian group generated by the prime ideals of $\mathcal{O}$. An important subgroup of $\mathcal{I}$ is the group of all fractional ideals generated by one element only. This is the subgroup $\mathcal{H} = \mathcal{H}(K)$ of fractional *principal* ideals. The factor group

I/H =: Cl = $Cl(K)$, the so-called *class group* of K, has the following property, discovered by Dirichlet:

THEOREM 4.3. (See JANUSZ [31], Theorem I.11.10). *The class group* Cl *is finite.* □

The order of Cl is known as the *class number* of K, notation: h = h(K).

COROLLARY 4.4. *Let a be a fractional ideal. Then* a^h *is principal. Moreover, if* k *and* h *are relatively prime then a is principal whenever* a^k *is principal.* □

A widely used theorem is Dirichlet's *unit* theorem:

THEOREM 4.5. (See JANUSZ [31], Theorem I.11.19). *The unit group in the ring* O *is the direct product of a finite cyclic group of roots of unity and a free abelian group of rank* r + s - 1. *Here* r *is the number of real conjugate fields of* K *and* s *is the number of pairs of complex conjugate fields of* K.

The meaning of the theorem is the following: in O a set of units $\{\varepsilon_1,\ldots,\varepsilon_{r+s-1}\}$ may be found, so-called *fundamental units*, with the property that for any unit $\eta \in O$ rational integers a_i exist such that the quotient of η and the product $\varepsilon_1^{a_1} \ldots \varepsilon_{r+s-1}^{a_{r+s-1}}$ is one of a finite number of roots of unity contained in O:

$$\eta = \zeta \cdot \prod_{i=1}^{r+s-1} \varepsilon_i^{a_i}, \qquad \zeta^m = 1.$$

The next theorem may be considered the most fundamental tool in the process of solving diophantine equations, at least from an algebraic point of view.

THEOREM 4.6. (See also LONDON & FINKELSTEIN [41], Theorem 25, p.70). *Let* $O = O(K)$ *be the ring of integers of the number field* K. *Further, let* a_0 *be a fixed ideal of* O *and let* m *be a fixed positive rational integer. Then there exist finite sets* $E \subset O$ *and* $F \subset O$ *with the following property: If* $x,y,z \in O$ *satisfy the requirements*

(i) $x \cdot y = z^m$, $xyz \neq 0$ *and*

(ii) *the ideal generated by* x *and* y *divides* a_0,

then there are units $\varepsilon_1,\varepsilon_2,\varepsilon_3 \in E$, *elements* $\alpha,\beta,\gamma \in F$ *and elements* $a,b \in O$ *such that*

$$x = \varepsilon_1 \alpha a^m, \quad y = \varepsilon_2 \beta b^m, \quad z = \varepsilon_3 \gamma ab \quad \textit{and} \quad \varepsilon_1 \varepsilon_2 \alpha\beta = \varepsilon_3^m \gamma^m.$$

PROOF. From the assumption, together with Theorem 4.2 we deduce that the principal ideals generated by x and y may be written as

$$(x) = \mathfrak{a}_1 A^m \quad \text{and} \quad (y) = \mathfrak{a}_2 B^m,$$

where $\mathfrak{a}_1$, $\mathfrak{a}_2$, A and B are ideals of O. Moreover, $\mathfrak{a}_1$ and $\mathfrak{a}_2$ are elements of a fixed finite set of integral ideals with the property that any common ideal divisor of $\mathfrak{a}_1$ and $\mathfrak{a}_2$ also divides $\mathfrak{a}_0$. Now, the number of ideal classes is finite. Let A belong to class C, and suppose the ideal A' is a fixed ideal of the inverse class C^{-1}. Consequently, A' belongs to a finite set. Further $A \cdot A' = (a)$ for some $a \in O$. Thus

$$(A')^m (x) = \mathfrak{a}_1 (a)^m,$$

which shows that $(A')^m$ and $\mathfrak{a}_1$ belong to the same ideal class. It follows that $\mathfrak{a}_1/(A')^m$ is a (fractional) principal ideal, generated by $\alpha \in K$, say. Then $(x) = (\alpha)(a)^m$ and hence

$$x = \varepsilon_1 \alpha a^m,$$

where ε_1 is a unit. Now ε_1 may be written as $\varepsilon_1 = \eta_1 \cdot \eta_2^m$, where the unit η_1 can assume only finitely many different values. Moreover, η_2^m may be absorbed by a^m. From its definition it is clear that α may be chosen from a finite subset of K. The remainder of the proof now follows easily. □

REMARK 4.7. If in addition $\mathfrak{a}_0 = (1)$ and m and h are relatively prime, then

$$x = \varepsilon_1 a^m, \quad y = \varepsilon_2 b^m, \quad z = \varepsilon_3 ab$$

with $a,b \in O$ and finitely many possible values for the units ε_1, ε_2 and ε_3. Indeed, from $(x) = A^m$ and $(m,h) = 1$ it follows that A is principal by 4.4. □

EXAMPLE 4.8. We return to Example 4.1, but now we take $k = p^2$, where p is a given prime number. The positive divisors of k are 1, p and p^2. For $d = 1$ we find $x^2 = \frac{1}{2}(p^2 + 1)$ and $y = \frac{1}{2}(p^2 - 1)$. This is the only possible value for d, since $d = p$ yields $x^2 = p$. Hence

$$p^2 - 2x^2 = -1,$$

which also may be written as

$$\mathrm{Norm}_{\mathbb{Q}(\sqrt{2})/\mathbb{Q}}(p + x\sqrt{2}) = -1.$$

This expression means that $p + x\sqrt{2}$ is a unit of $\mathcal{O}(K)$ with norm -1; here $K = \mathbb{Q}(\sqrt{2})$. The group of units of $\mathcal{O}(K)$ is obvious: 1 and -1 are the only roots of unity and $\varepsilon = 1 + \sqrt{2}$ is a fundamental unit (i.e. ε generates the free abelian unit group; $r + s - 1 = 1$). Since we may assume x to be positive, we find that

$$p + x\sqrt{2} = (1 + \sqrt{2})^{2k+1}$$

for some non-negative rational integer k. This means also that the prime p can be written as

$$p = \sum_{j=0}^{k} \binom{2k+1}{2j} 2^j \qquad (k \in \mathbb{Z},\ k \geq 0).$$

A different formulation of the problem may be given as follows. Let the sequence $\{a_k\}$ and $\{b_k\}$ be defined by

$$a_k + b_k\sqrt{2} = (1 + \sqrt{2})^{2k+1}, \qquad k \in \mathbb{Z}.$$

The sequence $\{a_k\}$ complies with the recurrence relation

$$a_{k+1} = 6a_k - a_{k-1}, \qquad k \in \mathbb{Z}$$

with initial conditions $a_0 = 1$, $a_1 = 7$. (The sequence $\{b_k\}$ satisfies the same recurrence relation, however with a different set of initial values.) Because the sequence $\{a_k\}_{k\geq 0}$ is increasing, we conclude that the original diophantine equation is soluble (with a single solution only) if and only if the prime p appears in the sequence $\{a_k\}_{k\geq 0}$. The prime numbers $p < 1000$ for which a solution exists are $p = 7$, 41 and 239.

We could also treat the equation

$$x^4 = y^2 + p^2, \qquad p \text{ prime}$$

in a different way. If x, y gives a solution, we write

$$x^4 = (y + pi)(y - pi).$$

Thus we factorize the right hand side in $O(L)$ with $L = \mathbb{Q}(i)$. Here $h(L) = 1$, the cyclic group of roots of unity is $\{1,-1,i,-i\}$ and the free abelian unit group is trivial, because $r + s - 1 = 0$. It is not difficult to prove that p cannot possibly divide both x and y. This implies that the only possible common prime ideal divisor of $(y + pi)$ and $(y - pi)$ is $(1 + i)$. Now by 4.6 with $m = 4$ and $\alpha_0 = (1 + i)$ it is easy to show that

$$y + pi = \varepsilon(1 + i)^a A^4$$

with $\varepsilon \in \{1,-1,i,-i\}$; $a \in \{0,1,2,3\}$ and $A \in O(L)$. From $\mathrm{Norm}_{L/\mathbb{Q}}(y+pi) = x^4$ it then follows that $a = 0$. Thus

$$y + pi = \varepsilon(u + iv)^4 = \varepsilon\{u^4 - 6u^2v^2 + v^4 + i(4u^3v - 4uv^3)\}$$

for certain $u,v \in \mathbb{Z}$. Because of the primality of p, we must have $\varepsilon = \pm i$. Then equating coefficients of 1 and i gives

$$\pm p = u^4 - 6u^2v^2 + v^4, \quad \mp y = 4uv(u^2 - v^2), \quad x = u^2 + v^2$$

where the ± signs correspond as indicated. This gives rise to the generally very difficult representation problem of type $(F = \mathbb{Q}(\theta))$

$$\mathrm{Norm}_{F/\mathbb{Q}}(u - v\theta) = m, \quad 0 \neq m \in \mathbb{Z}$$

with $[F:\mathbb{Q}] = 4$. We shall discuss such problems in Section 6.

The positive values of u and v corresponding with the solutions (x,y) of the original equation with p = 7, 41 and 239 are (u,v) = (1,2), (2,5) and (5,12) respectively.

5. THE MORDELL EQUATION

A fundamental problem when studying diophantine equations is the question of solvability. And further, assuming a given equation is solvable, how many solutions are there? A very important problem, closely related to the

previous one, is the question of the actual (and practical!) computation of the existing solutions, or in case infinitely many solutions exist, can they be characterized in a simple way (such as parametrization)?

Very little is known about solvability criteria: on the one hand it is quite often easy to show the insolubility of a given equation by means of impossible congruences (see: NAGELL [48], Chapter VII and MORDELL [46], Chapters 2, 26; see also BOREVICH [14] Problem 4 on p.3), and, on the other hand is the proof of the existence of solutions nearly always constructive.

For the sake of simplicity, we shall only consider binary polynomial equations in this part. Of all such equations the homogeneous ones (Thue equations, see Section 2) can be more systematically dealt with than the inhomogeneous equations. Very often an inhomogeneous equation can be reduced to one or more (but finitely many) norm form equations. These latter equations are discussed in the next section. For the moment we intend to give an outline of this reduction process by considering the Mordell equation (Section 2, equation (3)).

THEOREM 5.1. (see MORDELL [46], Chapters 24, 25 and 26). *Solving the equation* $x^3 = y^2 + k$ ($k \in \mathbb{Z}$, $k \neq 0$) *in rational integers* x *and* y *is equivalent to each of the following:*

(i) *solving finitely many equations of type* $f_3(u,v) = m$ *in rational integers* u *and* v, *where the* f_3 *are binary cubic forms of negative or positive discriminant as* k *is negative or positive respectively.*

(ii) *solving finitely many equations of type* $f_4(u,v) = m$ *in rational integers* u *and* v, *where the* f_4 *are binary quartic forms of negative discriminant.*

From Thue's theorem (Theorem 3.1) it follows immediately that the Mordell equation admits of at most finitely many solutions. Although we shall not attempt to prove Theorem 5.1, it may help to know that equations of type $f_3(u,v) = m$ are obtained by the factorization of $y^2 + k$ into prime ideals of a quadratic number field, whereas the factorization of $x^3 - k$ into prime ideals of a cubic extension of $\mathbb{Q}$ yields equations of type $f_4(u,v) = m$.

In the following example we shall go into more detail.

EXAMPLE 5.2. (See STROEKER [64]). In this example we intend to give a rather sketchy proof of the assertion (note that we follow Theorem 5.1 to the letter): the solutions in integers x and y of the equation

$$x^3 - 7y^2 = 1$$

are determined by the solutions in integers u and v of

$$\text{(i)}\quad \begin{aligned} u^3 - 21uv^2 &= 1, \quad \text{and} \\ u^3 - 42uv^2 + 98v^3 &= 1, \end{aligned}$$

but also by those of the equations

$$\text{(ii)}\quad \begin{aligned} u^4 - 84u^2v^2 - 392uv^3 - 588v^4 &= 1, \\ u^4 - 168u^2v^2 - 1{,}176uv^3 - 2{,}352v^4 &= 1, \quad \text{and} \\ u^4 - 924u^2v^2 - 15{,}288uv^3 - 71{,}148v^4 &= 1. \end{aligned}$$

Note that a solution (x,y) of $x^3 - 7y^2 = 1$ gives rise to a solution $(X,Y) = (7x, 7^2y)$ of the Mordell equation $X^3 - Y^2 = 7^3$.

Firstly, we factorize $7y^2 + 1$ in prime ideals of $\mathcal{O}(K)$ where $K = \mathbb{Q}(\sqrt{-7})$. Thus

$$(1 + y\sqrt{-7})(1 - y\sqrt{-7}) = x^3.$$

The number field K has the following fundamental properties: the class number $h(K) = 1$ and $\{1,\omega\}$ with $\omega = \frac{1}{2} + \frac{1}{2}\sqrt{-7}$ is a basis for $\mathcal{O}(K)$ and $2 = \omega\cdot\bar{\omega}$. Common prime ideal divisors of $(1 + y\sqrt{-7})$ and $(1 - y\sqrt{-7})$ are possibly (ω) or $(\bar{\omega})$ and no others. Hence

$$1 - y + 2y\omega = 1 + y\sqrt{-7} = (\omega)^{\alpha}(\bar{\omega})^{\beta}(a + b\omega)^3$$

with $\alpha,\beta \in \{0,1,2\}$ and $a,b \in \mathbb{Z}$. Taking also the conjugate equation into consideration, we see immediately that $\alpha + \beta = 0$ or 3. If $\alpha = \beta = 0$, then comparison of coefficients of 1 and ω left and right, and subsequently elimination of y from the resulting equations, yields

$$u^3 - 21uv^2 = 1.$$

Here u and v are defined by $2u = 2a + b$, $2v = b$. If $\alpha = 1$ and $\beta = 2$, then similarly we obtain the equation

$$u^3 - 42uv^2 + 98v^3 = 1,$$

where $u = a + 4b$, $v = b$. Analogously, the assumption $\alpha = 2$, $\beta = 1$ leads to the same equation in $u = a - 3b$ and $v = -b$. This proves the first part of our assertion.

Secondly, factorization in $\mathbb{Z}$ of $x^3 - 1$ yields

$$(x-1)(x^2+x+1) = 7y^2.$$

This furnishes the three possibilities:

$$\left.\begin{aligned} x - 1 &= \lambda a^2 \\ x^2 + x + 1 &= \mu b^2 \end{aligned}\right\} \text{ with } (\lambda,\mu) = (1,7),\ (3,21) \text{ or } (21,3).$$

Note that $\mathrm{hcf}(x-1, x^2+x+1) = 1$ or 3. Now the particulars of the number field $L = \mathbb{Q}(\rho)$, where ρ is the third root of unity $\rho = \frac{1}{2} + \frac{1}{2}\sqrt{-3}$, are: $h(L) = 1$ and $\{1,\rho\}$ is a basis for $\mathcal{O}(L)$, the cyclic group of roots of unity is generated by ρ and the free abelian group of units is trivial, because $r + s - 1 = 0$. For $\lambda = 1$, $\mu = 7$ we write

$$x - 1 = a^2, \qquad (x+\rho)(x-\rho^2) = 7b^2.$$

From Theorem 4.6 we deduce

$$x + \rho = \pm\alpha(1-2\rho)^s(c+d\rho)^2,$$

where $\alpha \in \{2+\rho, 3-\rho\}$, $s \in \{0,1\}$ and $(c,d) \in \mathbb{Z}^2$. Note that units may be absorbed in the square $(c+d\rho)^2$. From $\mathrm{Norm}_{L/\mathbb{Q}}(x+\rho) = 7.3^s(c^2+cd+d^2)^2$ and also $\mathrm{Norm}_{L/\mathbb{Q}}(x+\rho) = 7b^2$, one deduces immediately $s = 0$. The choice $\alpha = 3 - \rho$ leads to a contradiction when considering congruences mod 4 and mod 3 successively. On the other hand, if $\alpha = 2 + \rho$ then equating coefficients yields

$$x = 2c^2 - 2cd - 3d^2 \quad \text{and} \quad 1 = c^2 + 6cd + 2d^2.$$

Because of $a^2 = x - 1 = c^2 - 8cd - 5d^2 = (c - 4d)^2 - 21d^2$, we may write

$$21d^2 = (c - 4d - a)(c - 4d + a).$$

Further, from $\text{hcf}(c-4d-a, c-4d+a) = 2$, we deduce the existence of co-prime integers u and v such that

$$c-4d+a = 2u^2, \quad c-4d-a = 42v^2, \quad d = 2uv.$$

The second possibility, namely $c-4d+a = 6u^2$, $c-4d-a = 14v^2$ and $d = 2uv$, gives rise to an impossible congruence mod 5. Now, on substitution of $c = u^2+8uv+21v^2$ and $d = 2uv$ into $c^2+6cd+2d^2 = 1$, we find

$$u^4+28u^3v+210u^2v^2+588uv^3+441v^4 = 1.$$

The unimodular transformation given by the matrix $\begin{pmatrix} 1 & 7 \\ 0 & -1 \end{pmatrix}$ carries this equation to the equation

$$u^4-84u^2v^2-392uv^3-588v^4 = 1.$$

Similar arguments are used to obtain the other two norm form equations. □

Assertions like those stated in Theorem 5.1 are true for a larger class of equations than merely the Mordell equations. This becomes evident in the following example.

EXAMPLE 5.3. (See STROEKER [68]). In this example we consider the equation

$$(2y^2-3)^2 = x^2(3x^2-2).$$

We shall show, at least in outline, that solutions of this equation in integers x and y are determined by those of the equation

$$U^4-24UV^3+24V^4 = 1$$

in integers U and V. To be precise, this connection is given by

$$|x| = U^2-2UV+4V^2 \quad \text{and} \quad |y| = U^2+2UV-6V^2.$$

Suppose $x,y \in \mathbb{N}$ solve the original equation. Then a positive integer z exists such that

$$3x^2-2 = z^2 \quad \text{and} \quad 2y^2-3 = xz.$$

It is easy to see that both x and z must be odd and $1 \le x \le z$. On setting $u = \frac{1}{2}(z+x)$ and $v = \frac{1}{2}(z-x)$, one finds the relations

$$u^2 - 4uv + v^2 = 1 \quad \text{and} \quad u^2 - v^2 + 3 = 2y^2.$$

Hence, also $2u^2 - 6uv + v^2 = y^2$ and this equation may be written as

$$\left(\frac{v-3u-y}{2}\right)\left(\frac{v-3u+y}{2}\right) = 7\left(\frac{u}{2}\right)^2,$$

where u is even, v is odd and the factors of the left hand side are relatively prime. Moreover, these factors both have negative sign. Thus

$$v - 3u = \frac{1}{2}(v - 3u - y) + \frac{1}{2}(v - 3u + y) = -a^2 - 7b^2 \quad \text{and} \quad u = 2ab$$

for certain co-prime integers a and b. On substituting of $u = 2ab$, $v = -a^2 + 6ab - 7b^2$ into $u^2 - 4uv + v^2 = 1$, the equation

$$(a-b)^4 - 24(a-b)b^3 + 24b^4 = 1$$

is obtained, from which the required result follows.

Finally, we note that the original equation in x and y, represents an elliptic curve defined over $\mathbb{Q}$. The group of rational points on this curve is generated by the point (x,y) = (3,3). There is only one other solution in positive integers, namely (x,y) = (1,1). □

6. NORM FORM EQUATIONS

Let f_n be an irreducible (over $\mathbb{Q}$) binary form of degree n. A root θ of $f_n(t,1) = 0$ gives the extension $K = \mathbb{Q}(\theta)$ of $\mathbb{Q}$ of degree n. The other roots of $f_n(t,1) = 0$ are the field conjugates of θ and $f_n(x,y) = \mathrm{Norm}_{K/\mathbb{Q}}(x-y\theta)$. Solving an equation of type

$$\mathrm{Norm}_{K/\mathbb{Q}}(x-y\theta) = m \quad (m \in \mathbb{Z},\ m \neq 0)$$

in rational integers x and y generally boils down to solving a finite number of equations of the form

$$x - y\theta = \varepsilon \cdot \alpha \tag{4}$$

where α takes only finitely many different values in $\mathcal{O} = \mathcal{O}(K)$ (this number depends on the factorization of (m) into prime ideals of $\mathcal{O}$), and ε runs through the unit group of $\mathcal{O}$. What makes equation (4) so special is the fact that the left hand side does not contain the basis elements $\theta^2,\ldots,\theta^{n-1}$. Hence, for each value of α, the expression (4) asks for units ε of a very special type. Consequently, each equation (4) is equivalent to finitely many sets (depending on the number of roots of unity contained in $\mathcal{O}$) of $n-2$ equations in the exponents of fundamental units. In case the number of fundamental units (which is the rank of the free abelian unit group viz. r+s-1), agrees with the number of exponential equations mentioned above (this number is $n-2 = r+2s-2$), then SKOLEM's p-adic method [61] is applicable; see also LEWIS [38]. In Example 6.3 we shall give a brief discussion of this method. We shall illustrate these contemplations by some examples.

EXAMPLE 6.1. (See NAGELL [48], Chapter VI). Suppose $(x,y) \in \mathbb{Z}^2$ gives a solution of the quadratic equation

$$15x^2 - 20xy + 6y^2 = 1.$$

On setting $u = 10x - 6y$, $v = x$ this equation becomes

$$u^2 - 10v^2 = 6.$$

If $K = \mathbb{Q}(\sqrt{10})$ then $h(K) = 2$ and $\{1,\omega\}$ with $\omega = \sqrt{10}$ is a basis for $\mathcal{O} = \mathcal{O}(K)$ and $\varepsilon = 3+\omega$ is a fundamental unit of norm -1 (see the tables of quadratic number fields in BOREVICH [14], pp. 422-427). Further, 2 and 3 factor into ideals of $\mathcal{O}$ as follows: $(2) = p^2$ and $(3) = q\cdot q'$ where $p = (2,\omega)$, $q = (3,1+\omega)$ and $q' = (3,1-\omega)$. Hence

$$u^2 - 10v^2 = \mathrm{Norm}_{K/\mathbb{Q}}(u+v\omega) = 6$$

and this gives in terms of ideals of $\mathcal{O}$

$$(u+v\omega) = p\cdot q \text{ or } p\cdot q'.$$

It is not difficult to prove that $p\cdot q = (4+\omega)$ and $p\cdot q' = (4-\omega)$ and consequently

$$u + v\omega = \pm(4\pm\omega)(3+\omega)^{2k}$$

with $k \in \mathbb{Z}$ and independent ± signs. If we assume both u and v to be positive (this is no loss of generality) then we may drop the first ± sign. As in Example 4.8 the solutions can be determined by means of recurrences of order two. It turns out that there are infinitely many. The first few values of u and v are: (u,v) = (4,1), (16,5), (136,43), (604,191) etc., and the corresponding values of x (> 0) and y are: (x,y) = (1,1), (5,11), (43,49), (191,419) etc.

Continued fractions are also used quite frequently when dealing with quadratic equations (see LeVEQUE [37], Chapters 8 and 9). □

EXAMPLE 6.2. We return to example 5.2. (i). The equation $u^3 - 21uv^2 = 1$ is trivially solvable: the only solution is u = 1, v = 0. The cubic equation

$$f_3(u,v) = u^3 - 42uv^2 + 98v^3 = 1$$

is anything but trivial. The discriminant D of f_3 is positive, to be precise $D = 2^3 3^3 7^3$ and this means that the equation $f_3(t,1) = 0$ has three real roots θ_1, θ_2 and θ_3 say. For each i = 1, 2, 3 the number field $K_i = \mathbb{Q}(\theta_i)$ has a free abelian unit group of rank 2. Hence

$$u^3 - 42uv^2 + 98v^3 = 1 \quad \text{and} \quad \mathrm{Norm}_{K_i/\mathbb{Q}}(u-v\theta_i) = 1$$

is equivalent with

$$u - v\theta_i = \pm\varepsilon_1^{m_1}\cdot\varepsilon_2^{m_2},$$

where $\{\varepsilon_1,\varepsilon_2\}$ is a set of fundamental units of $O(K_i)$. This gives rise to only *one* equation in the *two* unknown exponents m_1 and m_2; Skolem's method, referred to above, is not applicable in this case. Considering also the conjugate equations, one may try factorization in an extension of K_i. That this could get very complicated is apparent from LJUNGGREN [40], where the similar equation $x^3 - 3xy^2 - y^3 = 1$ is treated.

The fact that $f_3(u,v) = 1$ can be solved after all, is a consequence of the relation which exists between the solutions of this equation and those of $x^3 - 7y^2 = 1$; the solutions of the latter equation are in turn related to those of the three norm form equations of 5.2 (ii), which can be

solved by Skolem's p-adic method. The equations $f_4(u,v) = 1$ are found to have the solution $(u,v) = (1,0)$ and only the third equation has the additional solution $(u,v) = (13,-1)$. Further, the only solutions of $x^3 - 7y^2 = 1$ are $(x,y) = (2,1)$, $(4,3)$ and $(22,39)$. For all this and the corresponding relations we refer to [65].

The implication of these results is that the equation

$$f_3(u,v) = u^3 - 42uv^2 + 98v^3 = 1$$

has no other than the following three solutions: $(u,v) = (1,0)$, $(-3,-1)$ and $(9,2)$. □

EXAMPLE 6.3. (See STROEKER [67]). Now we shall give an example of the use of p-adic arguments. We consider the quartic norm equation

$$f_4(u,v) = u^4 + 2u^2v^2 - 2v^4 = 1.$$

The discriminant of f_4 equals $-2^9 3^3$ and thus $f_4(t,1) = 0$ has two real roots and one pair of complex conjugate roots; $r + 2s = 4$, $r = 2$ and $s = 1$. Let θ be a real root of $f_4(t,1) = 0$. Then the ring $\mathcal{O}(K)$ of $K = \mathbb{Q}(\theta)$ has a free abelian unit group of rank 2. Since K is a quadratic extension of $\mathbb{Q}(\sqrt{3})$, it easily follows that $\{1,\theta,\theta^2,\theta^3\}$ is a basis for $\mathcal{O}(K)$. It is also reasonably easy to establish that $\{1+\theta,1-\theta\}$ is a fundamental set. From

$$u^4 + 2u^2v^2 - 2v^4 = 1 \quad \text{or} \quad \mathrm{Norm}_{K/\mathbb{Q}}(u-v\theta) = 1$$

we deduce

$$u - v\theta = \pm(1+\theta)^e(1-\theta)^f \tag{5}$$

with $e,f \in \mathbb{Z}$. If we do not specify the sign of u and v, then the $\pm$ sign in (5) may be dropped. Further, it is no restriction to assume $e \geq f$. Now

$$u - v\theta = (1+\theta)^{e-f}(1-\theta^2)^f. \tag{6}$$

We have mentioned Skolem's method on previous occasions. In the setting of the present example, this p-adic method may be described as follows. Expand the right-hand side of equation (6) in a power series in θ. Since the

coefficients of θ^2 and θ^3 of the left-hand side of (6) are zero, this leads to two equations in the unknowns e and f:

$$(7) \qquad \sum_{i,j} c_{ij}^{(k)} e^i f^j = 0 \quad (k = 1,2)$$

with rational integer coefficients $c_{ij}^{(k)}$. Suppose that for some rational prime p the power series mentioned above converges for all integer p-adic values and further suppose that for this prime p the congruence equations

$$(8) \qquad \sum_{i,j} c_{ij}^{(k)} e^i f^j \equiv 0 \pmod{p} \quad (k = 1,2)$$

are independent. Then there can only be finitely many pairs of p-adic values (e,f) satisfying (8). And this means that there are only finitely many rational integer values for e and f satisfying (7). In practice, all but a finite number of values of the exponents e and f in (6) may be eliminated by considering congruences modulo a power of p. See SKOLEM [60], [61].

We return to our example. For reasons of simplicity we assume v to be odd. Because of

$$u^2 - v^2\theta = (1-\theta^2)^{e+f} = 1 - (e+f)\theta^2 + 2(\ldots),$$

e - f is also odd. Put $2n + 1 = e - f$. We intend to show that $n = 0$. Define a_i, b_i, c_i and d_i for each $i \in \mathbb{Z}$ by

$$(1+\theta)^{2i+1} = a_i + b_i\theta + c_i\theta^2 + d_i\theta^3.$$

Then from

$$u - v\theta = (a_n+b_n\theta+c_n\theta^2+d_n\theta^3)(1-\theta^2)^f$$

we deduce

$$(9) \qquad a_n d_n = b_n c_n.$$

Knowing that there can be at most finitely many values for n, we consider only this equation (9) instead of the two equations one would obtain using Skolem's method.

Let α_i and β_i be given by

$$\theta^{2i} = \alpha_i + \beta_i\theta^2 \quad (i \in \mathbb{Z}).$$

Then after some calculations, we obtain the expressions

$$a_n = 2\sum_{j=0}^{n} \binom{2n+1}{2j}\beta_{j-1}, \qquad b_n = 2\sum_{j=0}^{n} \binom{2n+1}{2j+1}\beta_{j-1},$$

$$c_n = \sum_{j=0}^{n} \binom{2n+1}{2j}\beta_j \quad \text{and} \quad d_n = \sum_{j=0}^{n} \binom{2n+1}{2j+1}\beta_j.$$

Substituting these expressions for a_n, b_n, c_n and d_n into the relation $a_nd_n = b_nc_n$, yields, after dividing through by $4(n+1)(2n+1)^2$,

$$\sum_{i,j=0}^{n} r_{ij}(n)\binom{2n}{2i}\binom{2n}{2j}\beta_{i-1}\beta_j = 0,$$

where the rational numbers $r_{ij}(n)$, defined by

$$r_{ij}(n) := (j-i)/(2i+1)(2j+1)(2n-2i+1)(2n-2j+1)$$

are 2-adic integers, i.e. they have odd denominators.

Now suppose $n \geq 1$ with 2-adic value m (this means that n contains precisely m factors 2 in its prime decomposition). Then it is easy to show that for any pair (i,j) with $i \geq 0$ and $j \geq 0$ ($i = j = 0$ is excluded) the (i,j)th term in the double sum above has 2-adic value at least $m+1$, with the single exception of the (0,1)th term, which has 2-adic value m. This is a clear contradiction, because the total sum equals zero. Hence $n = 0$. Then

$$u - v\theta = (1+\theta)(1-\theta^2)^f,$$

and this is only possible when $f = 0$. Consequently, $(u,v) = (1,1)$ is the only solution of the original equation $f_4(u,v) = 1$ with positive u, v and odd v. For a nice application of Skolem's method, see MORDELL [46], p.207. See also STROEKER [65], [66], [67]. □

7. COMPUTATIONAL CONSIDERATIONS

From the previous sections it is clear that in the process of solving a diophantine equation one is often confronted with the necessity of computing:

(i) *The class number of a number field.*
There are computer programs for calculating the class number of quadratic number fields (tables can be found in BOREVICH [14]), pp. 422-427 and cubic number fields (cf. the tables by SELMER [54] and ANGELL [4], [5]). In case one is dealing with a norm form equation of type $f(x,y) = 1$, one only needs to have information on units; knowledge of class numbers of number fields involved is of little importance here. But when studying equations of type $f(x,y) = m \neq 1$, the prime ideal decomposition of (m) plays an important part; in particular, one needs information on the class group in such cases. Most practical methods for calculating the class number of a number field $K = \mathbb{Q}(\theta)$ use the fact that each ideal class contains an integral ideal of bounded norm (this bound $M(K)$ only depends on K). By inspection of principal ideals of small norm, generated by elements of type $u + v\theta$ $(u,v \in \mathbb{Z})$, it is often possible to select a maximal set of inequivalent ideals representing all classes, and such that the norm of each ideal is bounded by $M(K)$. In this way one may find the class number of K. For further information the reader should consult the relevant parts of BOREVICH [14] and JANUSZ [31]. See also ZANTEMA [76].

(ii) *A basis for the ring $O(K)$ of a number field* K.
Usually, this is not very hard. A well written description of the computation of a canonical basis is given in HOLZER [29], pp. 119-130. See also ZANTEMA [76].

(iii) *A set of generators of the free abelian group of units (a fundamental set) in the ring $O(K)$ of the number field* K.
This is a very important, and often difficult part of the methods described in this exposition.

We shall briefly mention a method due to BERWICK [12]. For more information we refer the reader to BRENTJES [17] and ZANTEMA [76]. Let K be a number field with a fundamental set of cardinality 2. For instance, let $n = 4$, $r = 2$ and $s = 1$. According to BERWICK [12], p. 367, the free abelian unit group of $O(K)$ is generated by each couple of units defined by:

(1) $\varepsilon_1 > 1$ and minimal, $|\varepsilon_1'| < 1$, $\varepsilon_1''\overline{\varepsilon_1''} < 1$

(2) $|\varepsilon_2| < 1$, $\varepsilon_2' > 1$ and minimal, $\varepsilon_2''\overline{\varepsilon_2''} < 1$

(3) $|\varepsilon_3| < 1$, $|\varepsilon_3'| < 1$, $|\varepsilon_3''| = |\overline{\varepsilon_3''}| > 1$ and minimal.

(ε_i', ε_i'' and $\overline{\varepsilon_i''}$ are the field conjugates of ε_i.)

In addition we have $\varepsilon_1\varepsilon_2\varepsilon_3 = 1$. An algorithm for computing the units ε_i can be devised as follows: let each of the restrictions (1), (2) and (3) successively be imposed on

$$\varepsilon = a\omega_1 + b\omega_2 + c\omega_3 + d\omega_4,$$

where $\{\omega_1,\omega_2,\omega_3,\omega_4\}$ is a basis for $O(K)$. Since the ω_i have known values, we get conditions on the rational integers a, b, c and d. So this provides ε with something like an "ideal ratio" a: b: c: d for the $\mathrm{Norm}_{K/\mathbb{Q}}(\varepsilon)$ to be small (this process also can be used when calculating class numbers; see under (i) at the beginning of this section). A very clear exposition, with many examples, is given in LONDON & FINKELSTEIN [41], p.81 etc.; here the algorithm in question is called the *scaling algorithm*.

Sometimes it is sufficient to use a full set of independent units instead of a fundamental set; these independent units should behave like fundamental ones modulo powers of a suitable rational prime (see STROEKER [65]). COGHLAN and STEPHENS [21] essentially used Berwick's method to deal with the remaining 20 difficult cases of the Mordell equation (3) with $0 < |k| \leq 100$.

In [75] WILLIAMS and ZARNKE use Voronoi's algorithm (see DELONE and FADDEEV [22]) to compute the fundamental unit of $\mathbb{Q}(d^{1/3})$ for $2 \leq d \leq 15{,}000$. Hence the equation $x^3 + dy^3 = 1$ is solved in the range indicated. In WILLIAMS and HOLTE [74] these results are extended to $d \leq 50{,}000$. Finally, BEACH and WILLIAMS [11] solved the equation $x^4 - dy^4 = 1$ for $2 \leq d \leq 10^6$ by computing the fundamental unit of $\mathbb{Q}(d^{\frac{1}{2}})$. This was done by employing the continued fraction algorithm.

Part B. APPROXIMATION METHODS

There exist algorithms which can be employed to obtain all solutions of certain diophantine inequalities. For example, the continued fraction algorithm enables one to parametrize all solutions $x,y \in \mathbb{Z}$ of the inequality $|x^2-dy^2| < \sqrt{d}$, where d is some fixed positive integer. Furthermore, W.J. Ellison proved by Baker's method that

$$|2^x-3^y| > \exp(x(\log 2 - 1/10)) \quad \text{for all } x,y \in \mathbb{N} \text{ with } x > 27$$

and F. Beukers showed by a variant of the hypergeometric method of Thue, Siegel and Baker that $|2^{2x+1} - y^2| > 2^{-50+x/5}$ for all $x,y \in \mathbb{N}$. These results

are based on approximation properties of certain numbers; in the examples: $\sqrt{d}$, log 3/log 2 and $\sqrt{2}$, respectively. These results make it possible to determine all solutions of such diophantine equations as $2^x - 3^y = 5$ and $x^2 + 7 = 2^y$. The advantage of solving diophantine equations by approximation results is that the methods are general and not especially dependent on the particular value of the constants. However, it is often true that the more general the method is, the larger are the upper bounds for the solutions. Sometimes these upper bounds can be reached by aid of computers, but in many cases they are too large. We shall give some examples of equations which have been completely solved by approximation methods.

8. POLYNOMIAL EQUATIONS

We consider equations $P(x_1,x_2,\ldots,x_n) = 0$ where $P \in \mathbb{Z}[x_1,x_2,\ldots,x_n]$ is fixed and $x_1,x_2,\ldots,x_n \in \mathbb{Z}$ are variables. In all but one case $n = 2$. For algebraic methods to solve such equations we refer to Part A of this paper.

The first application of approximation methods to polynomial equations was to the equation $x^2 - dy^2 = 1$, and more generally to $x^2 - dy^2 = k$ in variables $x,y \in \mathbb{Z}$. Here d and k are fixed integers with $d \in \mathbb{N}$, not a square and $k \neq 0$. This equation has been named after Pell for some two hundred years, but the idea of applying the continued fraction algorithm to find solutions is more than a thousand years old. The equation has either none or infinitely many solutions. However, it is possible to compute a number $X = X(d,k)$ such that every solution can be expressed in a simple manner in terms of some solutions (x,y) with $\max(|x|,|y|) \leq X$. Hence all solutions may be determined by computing the finitely many solutions which are bounded by X. If $|k| < \sqrt{d}$, then all solutions are directly provided by the continued fraction algorithm. In his exposition R.J. SCHOOF [53] discussed recent fast techniques to find the fundamental solution. (The fundamental solution is the smallest solution of $x^2 - dy^2 = 1$ in positive integers and knowing it is crucial for solving $x^2 - dy^2 = k$.) For the numerical treatment of the Pell equation we refer to this paper. It is easy to reduce the general equation $ax^2 + bxy + cy^2 = k$ with $a,b,c,k \in \mathbb{Z}$ fixed and $x,y \in \mathbb{Z}$ variable to an equation of Pell.

The continued fraction algorithm gives, for any real number α, all positive integers p and q with $|q\alpha - p| \leq q^{-1}$. Hence, it also enables us to compute any solution of the equation $ax^n - by^n = k$, where a, b, k, n are fixed integers with $k \neq 0$, $n \geq 3$ and $x,y \in \mathbb{Z}$ are variables. It follows from

Theorem 3.1 that this equation has only finitely many solutions.

As remarked before, Thue's proof of Theorem 3.1 is ineffective. Thue's hypergeometric method was extended by C.L. SIEGEL [58] and A. BAKER [6] and in their work one can find some particular equations which can be solved in this way. Siegel showed for example that if $|ab| > 17000n^2(a-b)^8$, then the only solution of $ax^n - by^n = a-b$ is given by $x = y = 1$ and it follows from Baker's irrationality measure for $\sqrt[3]{2}$ that each solution (x,y) of $x^3 - 2y^3 = k$ satisfies $\max(|x|,|y|) < (3\cdot 10^5|k|)^{23}$.

BAKER [7] showed in 1968 how his estimates for linear forms of logarithms can be used to give effective upper bounds for the solutions of the Thue-equation. In general the upper bounds are so large that they cannot be used to determine all solutions in practice. We shall describe here the few instances where Baker's method has been used to solve polynomial equations.

The four numbers 1, 3, 8, 120 have the property that the product of each pair plus 1 is a square. J.H. VAN LINT [39] wondered whether this property could also hold when 120 is replaced by a larger integer. A. BAKER and H. DAVENPORT [9] answered this question by solving the system of Pell equations

$$\begin{cases} 3x^2 - 2 = y^2 \\ 8x^2 - 7 = z^2 \end{cases}$$

in integers x, y, z. (Note that the system of polynomial equations $P_1=0$, $P_2=0,\ldots,P_h=0$ is equivalent to the single polynomial $P_1^2+P_2^2+\ldots+P_h^2 = 0$.) It follows from the theory of Pell - equations that each solution (x,y,z) corresponds to a pair (m,n) such that

$$(10) \qquad 0 < m\log(2+\sqrt{3}) - n\log(3+\sqrt{8}) + \log\frac{(1+\sqrt{3})\sqrt{8}}{(\pm 1+\sqrt{8})\sqrt{3}} < \frac{0.11}{(2+\sqrt{3})^{2m}}.$$

This is a linear form of logarithms of algebraic numbers. Baker's estimate implied that $m < 10^{487}$. To cover these cases, Davenport introduced a simple, but ingenious lemma. Here $\|q\beta\|$ denotes the distance of $q\beta$ to the nearest integer.

LEMMA. *Let* C, K *and* θ *be real numbers with* $K > 6$. *For any positive integer* M, *let* p *and* q *be integers satisfying*

$$1 \le q \le KM, \quad |\theta q - p| < 2(KM)^{-1}.$$

Then if

$$\|q\beta\| \geq 3K^{-1} \tag{11}$$

there is no solution of

$$|m\theta - n + \beta| < C^{-m}$$

in the range

$$\frac{\log K^2 M}{\log C} < m < M.$$

The simple proof rests on the fact that

$$\|q\beta\| = \|m\theta q - nq + \beta q - m\theta q + mp\| \leq q\|m\theta - n + \beta\| + m\|\theta q - p\|$$

has to be small, in contradiction with (11). After dividing through by $\log(3+\sqrt{8})$ the inequalities (10) look like $0 < m\theta - n + \beta < C^{-m}$, with

$$\theta = \log(2+\sqrt{3})/\log(3+\sqrt{8}),$$

$$\beta = \log\left(\frac{(1+\sqrt{3})/\sqrt{8}}{(\pm 1+\sqrt{8})/\sqrt{3}}\right)\Big/\log(3+\sqrt{8}) \quad \text{and} \quad C = (2+\sqrt{3})^2 \approx 14.$$

The lemma was applied with $K = 10^{33}$, $M = 10^{487}$. In order to check (11) both β's were computed to 600 decimal places. Then suitable values of p and q were found and it followed that $m < 500$. The remaining values of m could then be computed by hand and it turned out that 120 is indeed the unique positive integer with the required property. Later P. KANGASABAPATY and T. PONNUDURAI [32] and G. SANSONE [51] have given elementary proofs of this result.

W.J. ELLISON et al. [27] used the same method to solve the equations

$$x^3 - 12xy^2 - 12y^3 = \pm 1. \tag{12}$$

In following Baker's treatment of the Thue equation they obtained algebraic numbers α_1, α_2, α_3 such that every solution corresponds to a pair of integers b_1, b_2 satisfying

$$|b_1 \log \alpha_1 + b_2 \log \alpha_2 - \log \alpha_3| < \exp(-.404H)$$

where $H = \max(|b_1|,|b_2|)$. Baker's estimate gave $H \leq 10^{563}$. On applying Davenport's lemma twice they obtained $H < 50$. It was now easy to conclude that $(\pm 1,0)$ and $(\pm 1,\pm 1)$ are the only solutions of (12).

Ellison et al. used this result to solve the elliptic equation $y^2 = x^3 - 28$. At the time they started their research this was the smallest value of $|k|$ for which $y^2 = x^3 + k$ had not yet been solved and they wanted to show that any such equation can be solved, not only effectively, but also in a practical sense, by Baker's method. BAKER [7] and later H.M. STARK [62] had given general, very large, upper bounds for the solutions, following ideas of Mordell who had pointed out the existence of a correspondence between the solutions of the equation $y^2 = x^3 + k$ and those of certain Thue equations of degree 3 (see Theorem 5.1(i)). The equations corresponding to $y^2 = x^3 - 28$ are given by (12). In this way Ellison et al. proved that $y^2 = x^3 - 28$ has no solutions other than

$$(x,y) = (4,\pm 6),(8,\pm 22),(37,\pm 225).$$

The amount of computational work seems to have been slightly more than in the case of Baker and Davenport.

9. EXPONENTIAL EQUATIONS

The most famous exponential equation is the Fermat-equation $x^n + y^n = z^n$ in integers $n \geq 3$, $x,y,z \in \mathbb{Z}_{>0}$. S.S. WAGSTAFF Jr. [73] has proved that there are no solutions with $n \leq 125{,}000$. We shall deal here with exponential equations

$$P(x_1,x_2,\ldots,x_n) = a_1^{z_1} a_2^{z_2} \ldots a_m^{z_m} \tag{13}$$

in integer variables $x_1,x_2,\ldots,x_n$, $z_1,z_2,\ldots,z_m$, where $P \in \mathbb{Z}[x_1,x_2,\ldots,x_n]$ is a fixed polynomial and $a_1,a_2,\ldots,a_m$ are fixed integers. If P is a form, then (13) represents the Thue-Mahler equation. Another classical example of equation (13) is named after S. Ramanujan and T. Nagell, namely $x^2 + 7 = 2^z$.

Already in 1897 C. STØRMER [64] had studied equation (13) with $n = 1$ and $P(x) = 1 + x^2$. By using the theory of the Pell-equation he was able to give a constructive method to determine all solutions and to give the upper bound $3^m - 2^m$ for the number of solutions. Note that $x^2 + 1$ can neither be divisible by 4 nor by a prime $\equiv 3 \pmod 4$. As an example Størmer showed that the only positive integers x such that all prime factors of $x^2 + 1$ are less

than 14 are given by

$$1,2,3,5,7,8,18,57,239.$$

In 1915 S. RAMANUJAN [50] posed the problem of finding all values of z other than 3, 4, 5, 7 and 15 for which $2^z - 7$ is a perfect square. NAGELL [47] proved in 1948 that these five values of z exhaust all possibilities and several authors gave different algebraic proofs for this and related results. F. BEUKERS [13] employed a variant of the hypergeometric method of Thue, Siegel and Baker to deduce small upper bounds for the solutions x, z of the equation $x^2 + d = 2^z$, where d is any non-zero integer. Similarly upper bounds for the equation $x^2 + d = a^z$ can be computed, but only in cases in which an odd power of a is exceptionally close to a square. This is the case for $a = 2$ and $a = 3$, since $|2^{15} - 181^2|$ and $|3^{15} - 3788^2|$ are sufficiently small. On the other hand, there does not seem to be such an exceptional power of 5. Beukers proved that if (x,z) is a solution of $x^2 + d = 2^z$ with $d \neq 0$, then

$$z < 435 + 10 \log |d|/\log 2$$

and even that

$$z < 18 + 2 \log |d|/\log 2$$

when $|d| < 2^{96}$.

Since it suffices to solve at most 3^m equations $P(x_1,x_2,\ldots,x_n) = ay^3$ in integers $x_1,x_2,\ldots,x_n,y$ with fixed P and a in order to solve (13), it is obvious from what we said about the elliptic equation $y^2 = x^3 + k$, that any equation

$$bx^2 + k = a_1^{z_1} a_2^{z_2} \ldots a_m^{z_m}$$

with small values of $m,a_1,a_2,\ldots,a_m,b$ and k can be solved by Baker's method. It is more efficient to use estimates for the p-adic values of linear forms. This was done by D.C. HUNT and A.J. VAN DER POORTEN [30] for the equation $x^2 - 11 = 5^z$. Estimates for linear forms yielded $z < 10^{20}$. By using a PDP 11/70 at the University of New South Wales they checked these values in an intelligent way. It turned out that all solutions are given by (z,x) = (1,4), (2,6) and (5,56).

It is even possible to solve more complicated diophantine equations by Baker's method. In order to find all rational elliptic curves of conductor 11 M.K. AGRAWAL et al. [1] wanted to solve the equations

$$x^3 - x^2y + xy^2 + y^3 = \pm 11^z \tag{14}$$

in integers x, y, z with $x \equiv 0 \pmod 2$, $x - y \equiv 1 \pmod 4$ and $x \not\equiv 3y \pmod{11}$. On using estimates for both the complex and the 11-adic case they found the upper bound 10^{15} for a certain parameter H (cf. the treatment of equation (12)). Next they applied a generalization of the lemma of Davenport due to W.J. ELLISON [25]. For this a simultaneous approximation algorithm was needed, and they employed the algorithm of G. SZEKERES [70]. They further used an idea of K. Mahler to obtain 11-adic approximants. By computing to an accuracy of 98 decimal places on a PDP 11/70, they reduced the upper bound for H to 20. All solutions of small size were already known and so they succeeded in solving their original problem. As a by-product they found that the only solutions (x,y) of (14) with $|x| > 1$ are $(\pm 2, \mp 3)$, $(\pm 4, \mp 3)$, $(\pm 56, \mp 103)$.

10. PURELY EXPONENTIAL EQUATIONS

We consider the equations

$$X_1 + X_2 + \ldots + X_n = 0, \tag{15}$$

where $X_1, X_2, \ldots, X_n$ are integers composed of primes taken from a finite, fixed set S. For primes $p_1, p_2, \ldots, p_r$ we shall denote the set of integers composed of these primes by $S(p_1, p_2, \ldots, p_r)$.

Equations of type (15) occur in the theory of finite simple groups. R. BRAUER [15] used the solution of $X_1 = X_2 + 1$ with $X_1, X_2 \in S(2,3)$ in classifying all simple groups of order $5 \cdot 2^a \cdot 3^b$. Some generalizations were given by L.J. ALEX [2,3]. His main result on exponential equations is the determination of all 62 solutions of the equation $X_1 = X_2 + X_3$ with $X_1, X_2, X_3 \in S(2,3,5,7)$ The proof is elementary, by combining conclusions on residue classes containing solutions. Others who applied such techniques to exponential equations successfully are S.S. PILLAI [49] to the equation $2^x \pm 2^y = 3^z \pm 3^w$, Wm.J. LeVEQUE [35] and J.W.S. CASSELS [18] to the equation $a^x - b^y = 1$, where a and b are fixed integers, W. SIERPIŃSKI [59] to $3^x + 4^y = 5^z$, A. MAKOWSKI

[43,44] to $13^x - 3^y = 10$ and $2^x + 11^y = 5^z$ and R.J. STROEKER [70] to $3^x + 3^y = 5^z + 5^w$. In the last two instances some known results on the equation $y^2 = x^3 + k$ (k fixed) were used. Equations of type (15) were solved by J.L. BRENNER & L.L. FOSTER [16]. A related equation occurs in the Syracuse problem, see R.P. STEINER [63].

One of the first persons who solved exponential equations by approximation methods was C. STØRMER [64]. In 1897 he used the theory of the Pell-equation to prove that (15) has only finitely many solutions when $n = 3$ and $X_3 = 1$. In particular he gave all 23 solutions of the equation $X_1 = X_2 + 1$ with $X_1, X_2 \in S(2,3,5,7)$. This yields all triangular numbers composed of 2, 3, 5 and 7. His work was extended by D.H. LEHMER [33,34] who was interested in nearly dependent logarithms of primes. He found for example,

$$13\log 2 - 3\log 3 - 3\log 5 - 7\log 7 + 4\log 11 + \log 13 - \log 23 + \log 41$$
$$= \log \frac{63927525376}{63927525375} < 2.10^{-11}.$$

Gelfond realized that his irrationality measure for the quotient of two logarithms of algebraic numbers had consequences for exponential equations, but as far as we know, he never solved an equation (15). After Baker had generalized Gelfond's method, W.J. ELLISON [26] used Baker's estimates to prove that for any $\delta > 0$ and $a,b,m,n,x,y \in \mathbb{N}$ either $am^x - bn^y = 0$ or $|am^x - bn^y| \geq m^{(1-\delta)x}$ for $x \geq x_0$ where he gave $x_0 = x_0(a,b,m,n,\delta)$ explicitly. (His value of x_0 can be improved by employing more recent estimates.) In particular Ellison proved

$$(16) \qquad |2^x - 3^y| > \exp(x(\log 2 - \tfrac{1}{10}))$$

for all $x,y \in \mathbb{N}$ with $x > 11$, $x \neq 13,14,16,19,27$. We give an example to demonstrate the applicability of inequality (16).

EXAMPLE 1. In 1945 S.S. PILLAI conjectured that all solutions of

$$(17) \qquad 2^x - 2^y = 3^z - 3^w, \quad x > y > 0,\ z > w > 0,$$

in integers x, y, z, w are given by (3,1,2,1), (5,3,3,1) and (8,4,5,1). In the paper [49] he solved all equations $2^x \pm 2^y = 3^z \pm 3^w$ except for (17). He

achieved this by solving systems of congruence equations. This approach does not work in the case of (17), since this equation has solutions $x = y$, $z = w$ common to every modulus.

We shall solve (17) by the method of Baker. Suppose x, y, z, w is a solution not given by Pillai. We have

$$2^y(2^{x-y} - 1) = 3^w(3^{z-w} - 1).$$

Hence, $2^{y-2} \mid (z-w)$ if $y > 1$ and $2 \cdot 3^{w-1} \mid (x-y)$ if $w \geq 1$. It follows that

$$2^{y-2} \leq z - w, \quad 2 \cdot 3^{w-1} \leq x - y. \tag{18}$$

Since $2^{x-1} \leq 2^x - 2^y < 2^x$ and $2 \cdot 3^{z-1} \leq 3^z - 3^w < 3^z$, we further have

$$\frac{2}{3} < 2^x/3^z < 2. \tag{19}$$

Thus for every value of x the value of z is uniquely determined. On using (19) and (18) it is easy to check that there are no solutions (x,y,z,w) with $x \leq 11$ or $x \in \{13,14,16,19,27\}$ other than those mentioned by Pillai. It follows from (16) that $|2^x - 3^z| > \exp(x(\log 2 - \frac{1}{10}))$. Hence, by (17),

$$|2^y - 3^w| > 2^{5x/6}. \tag{20}$$

If $2^y > 2^{5x/6-1}$, then $y > 5x/6-1$. By (19) and (18)

$$(x-1)\log 2/\log 3 \geq z - 1 \geq z - w \geq 2^{y-2} > 2^{5x/6-3},$$

a contradiction for $x \geq 12$. If, on the other hand, $3^w > 2^{5x/6-1}$, then by (18) and (20)

$$x \geq x - y \geq 2 \cdot 3^{w-1} > \frac{1}{3}\, 2^{5x/6},$$

again a contradiction for $x \geq 12$. Hence there are no solutions other than those given by Pillai. □

Example 1 is the solution of a purely exponential equation with more than three terms, but with only two primes involved. The basic idea is that if there is a large solution, then two prime powers with fixed bases, p^x

and q^y say, have to have a relatively small difference. Hence $|x \log p - y \log q|$ is very small, and this contradicts estimates on linear forms obtained by Baker's method. Since we have a linear form of only two logarithms, we need not apply Baker's general method, but we can also use the older methods of Gelfond and Schneider. Recently M. MIGNOTTE and M. WALDSCHMIDT [45] gave an estimation by Schneider's method for such a form where the constants in the final lower bound are rather small. P.L. Cijsouw and Ā. Korlaar have written a computer program for solving the inequality $|p^x - q^y| < p^{x/2}$ in positive integers x, y for given integers p and q based on this estimate. A description of their method and a discussion on the completeness of the set of solutions obtained is given in an appendix of this paper. A computer run for all primes less than 20 had the following outcome:

The only solutions of the inequality

$$|p^x - q^y| < p^{x/2} \tag{21}$$

in positive integers x, y *and primes* p, q *with* $p < q < 20$ *are given by*

$$\begin{aligned}(p,q,x,y) = {} & (2,3,1,1),(2,3,2,1),(2,3,3,2),(2,3,5,3),(2,3,8,5)\\ & (2,5,2,1),(2,5,7,3),(2,7,3,1),(2,11,7,2),\ (2,13,4,1)\\ & (2,17,4,1),(2,19,4,1),(3,5,3,2),(3,7,2,1),(3,11,2,1)\\ & (3,13,7,3),(5,7,1,1),(5,11,3,2),(7,19,3,2),\\ & (11,13,1,1),(17,19,1,1).\end{aligned}$$

We illustrate the applicability of this result by another example:

EXAMPLE 2. In a letter to the second named author L.J. Alex asked for the set of solutions of the equation

$$1 + 3^x = 5^y + 3^z 5^w. \tag{22}$$

If $y = 0$, then $x = z$, $w = 0$. We may therefore assume $y > 0$ and hence $x > z$. We have $5^y - 1 = 3^z(3^{x-z} - 5^w)$. Since $y > 0$ and $5^y - 1$ has z factors 3, it follows that $2 \cdot 3^{z-1} | y$ and hence

$$3^z \leq 3y/2. \tag{23}$$

We treat the cases (i) $y \geq w$ and (ii) $y < w$ separately.

(i) We have $1 + 3^x = 5^w(5^{y-w} + 3^z)$. Since $3^x + 1$ has w factors 5, the number x is divisible by $2 \times 5^{w-1}$ and hence $x \geq 2 \times 5^{w-1}$. Thus, by (23),

$$0 < 3^x - 5^y < 3^z 5^w \leq \frac{15xy}{4} .$$

Hence $y \leq x \log 3/\log 5$ and

$$0 < 3^x - 5^y \leq 2.6x^2. \tag{24}$$

From the list of Cijsouw and Korlaar we see that

$$|3^x - 5^y| > 3^{x/2} \quad \text{if } (x,y) \neq (3,2).$$

Hence $x \leq 10$, $y \leq 6$, $z \leq 2$, $w \leq 2$. It follows from (24) and the inequality $y > 0$ that no pairs (x,y) are possible other than $(x,y) = (2,1)$, $(3,1), (3,2)$. This yields the solutions $(x,y,z,w) = (2,1,0,1), (3,2,1,0)$.

(ii) We have $1 + 3^x = 5^y(1 + 3^z 5^{w-y})$. Since $3^x + 1$ has y factors 5, the number x is divisible by $2 \times 5^{y-1}$ and hence $5^y \leq 5x/2$. Thus, by (23),

$$0 < 3^x - 3^z 5^w < 5^y \leq 5x/2$$

and

$$0 < 3^{x-z} - 5^w \leq \frac{5x}{2 \times 3^z} .$$

From the list we see that

$$|3^{x-z} - 5^w| > 3^{(x-z)/2}, \quad \text{if } (x-z,w) \neq (3,2).$$

If $x - z = 3$ and $w = 2$, then we infer $y = 1$, $z = 0$ by (23) and $x = 3$. This not being the case, we have $3^{x/2} \leq 5x/2$, which implies $x \leq 4$. Thus $y = 1$, and $w \geq 2$. Since $5 | (3^x + 1)$, we obtain $x = 2$ and this contradicts (22).

The only solutions of (22) therefore are given by $(x,y,z,w) = (x,0,x,0)$ for $x \in \mathbb{N}_0$, $(2,1,0,1)$ and $(3,2,1,0)$. □

APPENDIX

by

P.L. Cijsouw, A. Korlaar & R. Tijdeman

1. INTRODUCTION

In this appendix, we give a description of the proof of the following property (cf. Section 10 of the previous paper):

The only solutions of the inequality

$$|p^x - q^y| < p^{x/2} \tag{1}$$

in positive integers x, y *and primes* p, q *with* $p < q < 20$ *are given by*

$$\begin{aligned}(p,q,x,y) = {} & (2,3,1,1),(2,3,2,1),(2,3,3,2),(2,3,5,3),(2,3,8,5),\\ & (2,5,2,1),(2,5,7,3),(2,7,3,1)(2,11,7,2),(2,13,4,1),\\ & (2,17,4,1),(2,19,4,1),(3,5,3,2),(3,7,2,1),(3,11,2,1),\\ & (3,13,7,3),(5,7,1,1),(5,11,3,2),(7,19,3,2),\\ & (11,13,1,1),(17,19,1,1).\end{aligned}$$

In outline, the proof is as follows. We treat all combinations (p,q) with p, q prime and $p < q < 20$ consecutively. First, we prove that the linear form

$$\Lambda = x \log p - y \log q \tag{2}$$

has a value close to zero when (x,y) is a solution of (1). Then we distinguish three cases: x is "very large", x is "medium large" and x is "small"; these cases correspond approximately to $x \geq 2^{43}$, $10 < x < 2^{43}$ and $x \leq 10$ respectively. A result from transcendental number theory, stating that Λ cannot be close to zero, implies that there are no solutions with "very large" x. When x is "medium large", we translate the smallness of $|\Lambda|$ into

$$\left|\frac{\log p}{\log q} - \frac{y}{x}\right| < \frac{1}{2x^2}, \tag{3}$$

so that y/x is a convergent of the continued fraction of $(\log p)/(\log q)$. In order to solve (1) for medium large x, it clearly suffices to make a

check of all numbers x in the relevant range which are denominators of the convergents. To avoid excessively large numbers, this check is executed in a logarithmic form. Finally, for "small" values of x, the solutions of (1) can be found by direct substitution of (x,y) into (1). In fact, all solutions of (1) have "small" x.

Since the computer program is part of the proof, attention must be paid to the correctness of the program. We shall not describe the entire program, but in order to indicate our style of programming, we present the algorithm by which logarithms with base 2 of positive numbers have been computed.

2. THE LINEAR FORM

Let (x,y), $x \geq 4$, be a solution of (1) for some pair of primes (p,q) with $p < q < 20$. Then $x \geq y$ and

$$\frac{3}{4} p^x \leq q^y \leq \frac{5}{4} p^x. \tag{4}$$

Since

$$0 < |x \log p - y \log q| \leq \max(p^x q^{-y} - 1, q^y p^{-x} - 1),$$

(1) and (4) imply

$$0 < |x \log p - y \log q| < \frac{4}{3} p^{-x/2}. \tag{5}$$

The result from transcendental number theory that we shall use is the main theorem of Mignotte, Waldschmidt [45]. For our purpose, the advantage of this theorem is that the occurring constant 5×10^8 is pretty small. We only quote a simplified version of the theorem, corresponding to positive rational integers α_1, α_2, principal values of $\log \alpha_1$ and $\log \alpha_2$, and a rational number β.

THEOREM. (Mignotte-Waldschmidt). *Let α_1, α_2 be positive rational integers and put $S_1 = 1 + \log \alpha_1$, $S_2 = 1 + \log \alpha_2$. Let β be a rational number and take $B \geq e$ as an upper bound for the absolute values of the numerator and the denominator of β. Put $S_0 = 1 + \log B$ and $T = 4 + S_0 + \log S_1 S_2$. Let $E \geq e$ be a real number such that*

$$E \le \min(e^{2T/5}, e^{S_1}, e^{S_2}, e\frac{S_1}{\log \alpha_1}, e\frac{S_2}{\log \alpha_2}).$$

Then $\Lambda = \beta \log \alpha_1 - \log \alpha_2$ *satisfies* $\Lambda = 0$ *or*

$$(6) \qquad |\Lambda| > \exp\{-5 \times 10^8 S_1 S_2 T^2 (\log E)^{-3}\}.$$

On applying this theorem to $\Lambda = \frac{y}{x}\log q - \log p$ with $S_0 = I + \log x$, $S_1 = 1 + \log q$, $S_2 = 1 + \log p$,

$$T = 5 + \log x + \log(1+\log 19)(1+\log 17) < 7{\cdot}8 + \log x$$

and $E = e\frac{1+\log 19}{\log 19}$, thus $\log E > 5/4$, we find

$$|\frac{y}{x}\log q - \log p| > \exp\{-2.56 \times 10^8 (1+\log p)(1+\log q)(7.8+\log x)^2\}.$$

Hence, by (5),

$$2.56 \times 10^8 \frac{1+\log p}{\log p}(1+\log 19)(7.8+\log x)^2 > \frac{x}{2} - \frac{1}{2}$$

and, since $(1+\log p)/\log p \le (1+\log 2)/\log 2 < 2.5$,

$$5.2 \times 10^9 (7.8+\log x)^2 > x - 1.$$

This inequality implies $x < 2^{43}$. Hence, there is no solution (x,y) with $x \ge 2^{43}$, that is, there is no "very large" solution.

3. THE CONTINUED FRACTION

Let $x_0(p)$ be the smallest positive integer x satisfying $p^x \ge 2^{17}$. For the values of p and q under consideration, $5 \le x_0(p) \le 17$. The values of x with $x_0(p) \le x < 2^{43}$ will be called the "medium large" values. Note that for these values $p^x > 453x^2$, since

$$p^{x_0(p)} \ge 2^{17} > 453.17^2 \ge 453(x_0(p))^2.$$

It follows that

$$(7) \qquad p^{x/2}\log q > (\sqrt{453}\cdot\log 3)x > 23x.$$

Let (x,y) be a solution of (1) with x "medium large". Then (5) and (7) imply

$$(8) \qquad \left|\frac{\log p}{\log q} - \frac{y}{x}\right| < \frac{1}{17x^2},$$

so that y/x must be a convergent of the continued fraction $1 \int b_1 + 1 \int b_2 + \ldots$ of (log p)/(log q). Let us consider a convergent r_m/s_m of this continued fraction with $x_0(p) \le s_m < 2^{43}$. Recall that

$$\frac{1}{s_m\{(b_{m+1}+1)s_m + s_{m-1}\}} < \left|\frac{\log p}{\log q} - \frac{r_m}{s_m}\right| < \frac{1}{s_m\{b_{m+1}s_m + s_{m-1}\}},$$

whence

$$(9) \qquad \frac{1}{(b_{m+1}+2)s_m^2} < \left|\frac{\log p}{\log q} - \frac{r_m}{s_m}\right| < \frac{1}{b_{m+1}s_m^2}.$$

When $b_{m+1} \le 15$, the left hand side of (9) contradicts (8). In order to find all "medium large" solutions of (1), we therefore may restrict ourselves to the selection of the solutions from the (s_m, r_m) subject to $b_{m+1} \ge 16$ and $x_0(p) \le s_m < 2^{43}$.

Let u with $0 < u < 1$ be an approximation to (log p)/(log q) with the continued fraction $1 \int a_1 + 1 \int a_2 + \ldots$ and the convergents p_n/q_n $(n = 1,2,\ldots)$. We assume that the continued fraction of u does not break off before or on the n-th level, so that a_{n+1} exists. Suppose that

$$(10) \qquad \left|u - \frac{\log p}{\log q}\right| < \varepsilon, \quad \text{with } \varepsilon = 2^{-117}.$$

Then

$$(11) \qquad \left|u - \frac{\log p}{\log q}\right| < \frac{1}{2^{31}s_m^2},$$

for all "medium large" solutions (s_m, r_m) of (1). Together with (8) this implies

$$\left|u - \frac{r_m}{s_m}\right| < \frac{1}{16s_m^2},$$

for these solutions, so that r_m/s_m is one of the convergents of the continued fraction of u; say, $r_m/s_m = p_n/q_n$ (usually, n = m). From $b_{m+1} \ge 16$ and the inequality

$$(12) \qquad (a_{n+1}+2)^{-1}q_n^{-2} < |u - p_n/q_n| < a_{n+1}^{-1}\, q_n^{-2},$$

which is the analogue of (9), we see that

$$\frac{1}{(a_{n+1}+2)q_n^2} < \left|u-\frac{p_n}{q_n}\right| = \left|u-\frac{r_m}{s_m}\right| < \frac{1}{16s_m^2} = \frac{1}{16q_n^2}\ .$$

Hence, $a_{n+1} \geq 15$. We can therefore restrict ourselves to the (q_n,p_n) subject to: a_{n+1} exists, $a_{n+1} \geq 15$ and $x_0(p) \leq q_n < 2^{43}$.

In order to obtain a number u for which (10) holds, we computed numerical approximations P and Q to $\log_2 p$ and $\log_2 q$:

$$\log_2 p = P + \delta_p, \quad \log_2 q = Q + \delta_q$$

with errors δ_p and δ_q for which $0 \leq \delta_p < 2^{-117}$ and $0 \leq \delta_q < 2^{-117}$. Then $u = P/Q$ satisfies (10).

The convergents p_n/q_n are alternating around u, in the sense that $p_n/q_n < u$ when n is even and $p_n/q_n > u$ when n is odd. In order to know the sign of some round-off errors we want to secure that the p_n/q_n are also alternating around $(\log p)/(\log q)$. By (12), this is certainly the case if

$$(13) \qquad \left|u-\frac{\log p}{\log q}\right| < \frac{1}{(a_{n+1}+2)q_n^2}\ .$$

This inequality holds when

$$(14) \qquad \varepsilon < \frac{1}{4a_{n+1}^2 q_n^2}\ ,$$

a condition that can be checked easily in the form

$$(15) \qquad 2\log_2 a_{n+1} + 2\log_2 q_n < 115.$$

When (14) is true, it follows from (12) and (13) that

$$\left(\frac{1}{a_{n+1}+2}-\frac{1}{4a_{n+1}^2}\right)\frac{1}{q_n^2} < \left|\frac{\log p}{\log q}-\frac{p_n}{q_n}\right| < \left(\frac{1}{a_{n+1}}+\frac{1}{4a_{n+1}^2}\right)\frac{1}{q_n^2}\ .$$

Because of $a_{n+1} \geq 15$ this implies

$$(16) \qquad \frac{1}{(a_{n+1}+3)q_n^2} < \left|\frac{\log p}{\log q}-\frac{p_n}{q_n}\right| < \frac{1}{(a_{n+1}-1)q_n^2}\ .$$

Now let n be *even*, so that we can omit the absolute value signs in the middle term of (16). Multiplying (16) by $q_n \log q$, taking exponentials, subtracting 1 from all terms and finally multiplying by q^{p_n}, we obtain

$$q^{p_n}\{q^{((a_{n+1}+3)q_n)^{-1}} - 1\} < p^{q_n} - q^{p_n} < q^{p_n}\{q^{((a_{n+1}-1)q_n)^{-1}} - 1\}.$$

Hence,

$$((a_{n+1}+3)q_n)^{-1} q^{p_n+\xi} \log q < p^{q_n} - q^{p_n} < ((a_{n+1}-1)q_n)^{-1} q^{p_n+\eta} \log q$$

for suitable numbers ξ with $0 < \xi < ((a_{n+1}+3)q_n)^{-1}$ and η with $0 < \eta < ((a_{n+1}-1)q_n)^{-1}$. Note that $(a_{n+1}-1)q_n \geq 70$ when $a_{n+1} \geq 15$ and $q_n \geq x_0(p)$.

Consider a pair (q_n,p_n) which could be a solution, i.e. $x_0(p) \leq q_n < 2^{43}$ and $a_{n+1} \geq 15$. Then clearly (p_n,q_n) is a solution when

$$((a_{n+1}-1)q_n)^{-1} q^{p_n} q^{1/70} \log q < p^{q_n/2}$$

thus, when

$$(17) \qquad p_n \log_2 q - \frac{1}{2} q_n \log_2 p < \log_2((a_{n+1}-1)q_n) - \frac{1}{70}\log_2 q - \log_2 \log q,$$

and (q_n,p_n) is certainly no such solution when

$$((a_{n+1}+3)q_n)^{-1} q^{p_n} \log q > p^{q_n/2},$$

i.e.

$$(18) \qquad p_n \log_2 q - \frac{1}{2} q_n \log_2 p > \log_2((a_{n+1}+3)q_n) - \log_2 \log q.$$

In the same way, we obtain in the case of *odd* n that (q_n,p_n) is a solution when

$$(19) \qquad \frac{1}{2} q_n \log_2 p < \log_2((a_{n+1}-1)q_n) - \frac{1}{70}\log_2 q - \log_2 \log q$$

and no solution when

$$(20) \qquad \frac{1}{2} q_n \log_2 p > \log_2((a_{n+1}+3)q_n) - \log_2 \log q.$$

The checks (17), (18), (19) and (20) can be evaluated very easily. They have been executed in single machine precision, rounding off in the right directions.

For each convergent p_n/q_n, the check concerning the existence of a_{n+1} and the check on the correctness of (15) have to be positive; otherwise, the described technique provides no answer to the question whether p_n/q_n is a solution of (1) or not. The same happens when both (17) and (18), resp. both (19) and (20) appear to be false. In these cases, one should try once more with ε replaced by a smaller value (for this, $\log p$ and $\log q$ must be computed more accurately) or with $x_0(p)$ replaced by a larger value.

4. DIRECT CHECKING

The remaining pairs (x,y) with x "small" can be substituted directly into the inequality

$$(21) \qquad (p^x - q^y)^2 < p^x.$$

For $1 \le x \le 3$, all pairs (x,y) with $1 \le y \le x$ have been checked. For each value of x with $4 \le x \le x_0(p)$ there is, by (4), only one possibility for y, namely the highest value for which $4q^y \le 5p^x$. These pairs (x,y) have been checked in (21) too.

5. SOME REMARKS ON THE PROGRAM

The execution of our program starts with the computation of the numerical approximations of the numbers $\log_2 p$ with p prime, $2 \le p \le 19$. After that, all combinations of (p,q) are treated consecutively. For each (p,q), the solutions with x "small" are detected by direct checking. All "medium large" solutions are found by the technique described in Section 3 (in fact, there were none).

As a consequence of the choice of ε and $x_0(p)$, there appeared to be no failing checks.

Finally, we present the algorithm that has been used for the computation of the logarithms with base 2. Let a be a positive real number of which we know the integral part and the first m binary digits of the fractional part. We want to compute as many binary digits of $\log_2 a$ as possible from this information. Multiply a by 2^m to obtain an integer X (supposed to be positive) for which

$$X \le Z = 2^m a < X + 1.$$

Consider all numbers $\log_2 z$ for $X \le z < X+1$. The binary expansions of these logarithms coincide up to a certain number of digits. The algorithm generates this common part. By that, $\log_2 Z$ and so $\log_2 a$ become known. Note that the algorithm leads to only very few digits for certain values of X. For the values of X we used, the number of generated digits of the logarithm was nearly the number of digits of X.

In what follows, we pay attention only to the fractional part of $\log_2 Z$. We shall use the following simple lemmas:

LEMMA 1. *Let* x *and* s *be positive integers with* $x \ge 2^s$. *Then* $\lfloor 2^{-s-1} x^2 \rfloor \ge 2^{s-1}$.

PROOF. By $x \ge 2^s$ we have $2^{-s-1} x^2 \ge 2^{s-1}$ where the right hand side is an integer.

LEMMA 2. *Let* x *and* s *be positive integers with* $x < 2^{s+1}$. *Then* $\lceil 2^{-s-1} x^2 \rceil < 2^{s+1}$.

PROOF. Since $x \le 2^{s+1} - 1$, we have

$$x^2 \le 2^{2s+2} - 2^{s+2} + 1 < 2^{2s+2} - 2^{s+1}$$

so that

$$\lceil 2^{-s-1} x^2 \rceil \le 2^{s+1} - 1 < 2^{s+1}.$$

Let $\log_2 Z = S + (0.b_0 b_1 \ldots)_2$, where S is the integral part of $\log_2 Z$, and let T be the number of binary digits $b_0, b_1, \ldots, b_{T-1}$ we want to generate. Further, z_0, z_1 and s are integers, q is a Boolean and z is a real "ghost variable" that serves only in the proof of the correctness. As invariant we use

$$I: (\log_2 z = s + (0.b_j b_{j+1} \ldots)_2) \wedge (j \ge 0) \wedge$$

$$\wedge (b_0, \ldots, b_{j-1} \text{ have been computed}) \wedge (q = Q)$$

where

$$Q: (2^s \le z_0 \le z < z_1 < 2^{s+1}) \wedge (s \ge 1).$$

The postcondition will be

$$R: I \wedge (j = T \vee \neg q).$$

The next algorithm describes the transformation of the precondition with given Z, X, S and T into the desired postcondition R. Some remarks by which

the reader will be able to understand the correctness of the algorithm are added between braces.

```
integer j,s,z0,z1; boolean q; {real z;}
j:=0; s:=S; z0:=X; z1:=X+1; {z:=Z;}
q:=(z0≥2^s) ∧ (z1<2^(s+1)) ∧ (s≥1); {I}
do j≠T ∧ q →
   z0 := ⌊2^(-s-1)*z0^2⌋;
         {z0 ≥ 2^(s-1) by Lemma 1; z0 ≤ 2^(-s-1)*z^2}
   z1 := ⌈2^(-s-1)*z1^2⌉;
         {z1 < 2^(s+1) by Lemma 2; z1 > 2^(-s-1)*z^2}
   if z1 < 2^s →
            {2^(s-1) ≤ z0 ≤ 2^(-s-1)*z^2 < z1 < 2^s, hence
             s ≤ log2 z < s+1/2; thus bj = 0}
      bj = 0; {log2 z = s + (0.0bj+1bj+2...)2}
      j := j+1; {log2 z = s + (0.0bjbj+1...)2,
            hence log2 z^2 = 2s + (0.bjbj+1...)2}
      s := s-1; {log2 z^2 = 2s + 2 + (0.bjbj+1...)2,
            hence log2(2^(-s-2)*z^2) = s + (0.bjbj+1...)2;
            2^s ≤ z0 ≤ 2^(-s-2)*z^2 < z1 < 2^(s+1);
            z := 2^(-s-2)*z^2;}
      q := s ≥ 1  {I}
   □ z0 ≥ 2^s →
            {2^s ≤ z0 ≤ 2^(-s-1)*z^2 < z1 < 2^(s+1), hence
             s+1/2 ≤ log2 z < s+1; thus bj = 1}
      bj := 1; {log2 z = s + (0.1bj+1bj+2...)2}
      j := j+1 {log2 z = s + (0.1bjbj+1...)2}, hence
            log2 z^2 = 2s + 1 + (0.bjbj+1...)2, thus
            log2(2^(-s-1)*z^2) = s + (0.bjbj+1...)2;
            z := 2^(-s-1)*z^2} {I}
   □ (z1 ≥ 2^s) ∧ (z0 < 2^s) → q := false {I}
   fi {I}
od {I ∧ (j = T ∨ ¬q)}
```

When $j = T$, the desired number of digits have been obtained; when $\neg q$ holds, there is not enough information left for the determination of the next digit.

All arithmetical operations in the algorithm (adding, subtracting, squaring, multiplying by a power of two, comparing with a power of two, taking "floor" and taking "ceiling") can be executed in an easy way. It leads to exact integer arithmetic on any binary computer.

The complete program has been executed on the Burroughs B7700 computer of the Eindhoven University of Technology. This computer has a word length of 39 binary bits. We operated up to 5-fold precision for-p, leading to the desired 117 bits for the fractional part of $\log_2 p$.

The compilation took about 6 seconds (mainly because of the numerous multilength arithmetic procedures), the computation of the logarithms involved needed 14 seconds and the remaining part of the program 10 seconds.

REFERENCES

[1] AGRAWAL, M.K. J.H. COATES, D.C. HUNT & A.J. VAN DER POORTEN, *Elliptic curves of conductor 11*, Math. Comp. 35 (1980), 991-1002.

[2] ALEX, L.J., *On simple groups of order* $2^a3^b7^cp$, J. Algebra 25 (1973), 113-124.

[3] ALEX, L.J., *Diophantine equations related to finite groups*, Comm. in Algebra 4 (1976), 77-100.

[4] ANGELL, I.O., *A table of complex cubic fields*, Bull. London Math. Soc. 5 (1973), 37-38.

[5] ANGELL, I.O., *A table of totally real cubic fields*, Math. Comp. 30 (1976), 184-187.

[6] BAKER, A., *Rational approximations to* $\sqrt[3]{2}$ *and other algebraic numbers*, Quart. J. Math. Oxford (2), 15 (1964), 375-383.

[7] BAKER, A., *Contributions to the theory of diophantine equations, I : On the representation of integers by binary forms, II: The diophantine equation* $y^2 = x^3+k$, Phil. Trans. Roy. Soc. London A 263 (1967/68), 173-208.

[8] BAKER, A., *Bounds for the solutions of the hyperelliptic equations*, Proc. Cambr. Phil. Soc. 65 (1969), 439-444.

[9] BAKER, A. & H. DAVENPORT, *The equations* $3x^2-2 = y^2$ *and* $8x^2-7 = z^2$, Quart. J. Math. Oxford (2), 20 (1969), 129-137.

[10] BAŠMAKOVA, I.G., *Diophant und Diophantische Gleichungen*, Uni-Taschenbücher 360, Birkhäuser Verlag, Basel und Stuttgart, 1974.

[11] BEACH, B.D. & H.C. WILLIAMS, *A computer algorithm for determining the solution of the diophantine equation* $x^4 - dy^4 = 1$, Proc. Manitoba Conf. Num. Math., Winnipeg (1971), 663-670.

[12] BERWICK, W.E.H., *Algebraic number fields with two independent units*, Proc. London Math. Soc. 34 (1932), 360-378.

[13] BEUKERS, F., *The generalised Ramanujan-Nagell equation*, Acta Arith. I: 38 (1980/81), 389-410, II: 39 (1981), 113-123.

[14] BOREVICH, Z.I. & I.R. SHAFAREVICH, *Number theory*, Pure and Appl. Math. Ser., vol. 20, Academic Press, London and New York, 1966.

[15] BRAUER, R., *On simple groups of order* $5.3^a.2^b$, Bull. Amer. Math. Soc. 74 (1968), 900-903.

[16] BRENNER, J.L. & L.L. FOSTER, *Exponential diophantine equations*, Pac. J. Math., to appear.

[17] BRENTJES, A.J., *Multi-dimensional continued fraction algorithms*, these Proceedings.

[18] CASSELS, J.W.S., *On the equation* $a^x-b^y = 1$, Amer. J. Math. 75 (1953), 159-162.

[19] CASSELS, J.W.S., *Diophantine equations with special reference to elliptic curves*, J. London Math. Soc. 41 (1966), 193-291.

[20] COATES, J., *An effective p-adic analogue of a theorem of Thue, The greatest prime factor of a binary form*, Acta Arith. I: 15 (1969), 279-305, II: 16 (1970), 399-412.

[21] COGHLAN, F.B. & N.M. STEPHENS, *The diophantine equation* $x^3-y^2 = k$, Computers in Number Theory (Proc. Sci. Res. Council Atlas Symp. No. 2 No. 2, Oxford 1969), Academic Press, 1971, pp. 199-206.

[22] DELONE, B.N. & D.K. FADDEEV, *The theory of irrationalities of the third degree*. Transl. Math. Monogr. No. 10, Amer. Math. Soc., 1964.

[23] DICKSON, L.E., *History of the theory of numbers, Vol. II: Diophantine analysis*, Chelsea Publ., New York, 1971. (repr. from orig. 1920 ed.).

[24] DUBOIS, E. & G. RHIN, *Sur la majoration de formes linéaires à coefficients algébriques réels et p-adiques*, Démonstration d'une conjecture de K. Mahler, C.R. Acad. Sc. Paris A 282 (1976), 1211-1214.

[25] ELLISON, W.J., *Recipes for solving diophantine problems by Baker's method*, Sém. Th. Nombr., 1970-1971, Exp. No. 11, Lab. Théorie Nombres, C.N.R.S., Talence, 1971.

[26] ELLISON, W.J., *On a theorem of S. Sivasankaranarayana Pillai*, Sém. Th. Nombr. 1970-1971, Exp. No. 12, Lab. Théorie Nombres, C.N.R.S., Talence, 1971.

[27] ELLISON, W.J., F. ELLISON, J. PESEK, C.E. STAHL & D.S. STALL, *The diophantine equation* $y^2+k = x^3$, J. Number Th. 4 (1972), 107-117.

[28] FINKELSTEIN, R. & H. LONDON, *On Mordell's equation* $y^2-k=x^3$: *an interesting case of Sierpinski*, J. Number Th. 2 (1970), 310-321.

[29] HOLTZER, L., *Zahlentheorie, Teil I.* Math. Naturw. Bibl. 13, B.G. Teubner Verlag, Leipzig, 1958.

[30] HUNT, D.C. & A.J. VAN DER POORTEN, *Solving diophantine equations* $x^2+d=a^u$, unpublished.

[31] JANUSZ, G.J., *Algebraic number fields*, Pure Appl. Math. Ser. Vol. 55, Academic Press, 1973.

[32] KANGASABAPATY, P. & T. PONNUDURAI, *The simultaneous diophantine equations* $y^2-3x^2=-2$ and $z^2-8x^2=-7$, Quart. J. Math. Oxford Ser. (2), 26 (1975), 275-278.

[33] LEHMER, D.H., *On a problem of Størmer*, Illinois J. Math. 8 (1964), 57-79.

[34] LEHMER, D.H., *The prime factors of consecutive integers*, Amer. Math. Monthly 72 (1965), No. 2, Part II, 19-20.

[35] LeVEQUE, W.J., *On the equation* $a^x-b^y=1$, Amer. J. Math. 74 (1952), 325-331.

[36] LeVEQUE, W.J., *On the equation* $y^m=f(x)$, Acta Arith. 9 (1964), 209-219.

[37] LeVEQUE, W.J., *Fundamentals of Number Theory*, Addison-Wesley, 1977. (This is a revised version of Volume I of Topics in Number Theory, Addison-Wesley, 1956.)

[38] LEWIS, D.J., *Diophantine equations, p-adic methods*, In: *Studies in number theory*, W.J. LeVeque (ed.), Math. Ass. Amer., 1969, pp. 25-75.

[39] LINT, J.H. VAN, *On a set of diophantine equations*, Rep. 68-WSK-03, TH Eindhoven, 1968.

[40] LJUNGGREN, W., *Einige Bemerkungen über die Darstellung ganzer Zahlen durch binäre kubische Formen mit positiven Diskriminante*, Acta Math. 75 (1942), 1-21.

[41] LONDON, H. & R. FINKELSTEIN, *On Mordell's equation* $y^2-k=x^3$, Bowling Green State Univer. Press, Bowling Green, Ohio, 1973.

[42] MAHLER, K., *Zur Approximation algebraischer Zahlen*, Math. Ann. 107 (1933), 691-730 and 108 (1933), 37-55.

[43] MAKOWSKI, A., *On the diophantine equation* $2^x+11^y = 5^z$, Nordisk Mat. Tidskr. 7 (1959), 81.

[44] MAKOWSKI, A., *On the equation* $13^x-3^y = 10$, The Mathematics Student, 28 (1960), 87.

[45] MIGNOTTE, M. & M. WALDSCHMIDT, *Linear forms in two logarithms and Schneider's method*, Math. Ann. 231 (1978), 241-267.

[46] MORDELL, L.J., *Diophantine equations*, Pure Appl. Math. Ser. Vol. 30, Academic Press, 1969.

[47] NAGELL, T., *The diophantine equation* $x^2+7 = 2^n$, Norsk. Mat. Tidskr. 30 (1948), 62-64; Ark. Mat. 4 (1960), 185-187.

[48] NAGELL, T., *Introduction to number theory*, Chelsea Publ., New York, 1964 (repr. of 2nd ed. 1951).

[49] PILLAI, S.S., *On the equation* $2^x-3^y = 2^X+3^Y$, Bull. Calcutta Math. Soc. 37 (1945), 15-20.

[50] RAMANUJAN, S., *Question 464*, J. Indian Math. Soc. 5 (1915) 120; Coll. Papers, Chelsea Publ., New York, 1962, p. 327.

[51] SANSONE, G., *Il sistema diofanteo* $N+1 = x^2$, $3N+1 = y^2$, $8N+1 = z^2$, Ann. Mat. Pura Appl. (4) 111 (1976), 125-151.

[52] SCHLICKEWEI, H.-P., *Über die diophantische Gleichung* $x_1+x_2+\ldots+x_n = 0$, Acta Arith. 33 (1977), 183-185.

[53] SCHOOF, R.J., *Quadratic number fields and factorizations*, these Proceedings.

[54] SELMER, E.S., *Tables for the purely cubic field* $K(m^{1/3})$, Avh. Norske Vid. Akad. Oslo I, Mat. Naturv. Klasse, 1955, No. 5.

[55] SHOREY, T.N., A.J. VAN DER POORTEN, R. TIJDEMAN & A. SCHINZEL, *Applications of the Gelfond-Baker method to diophantine equations*, In: *Transcendence Theory: Advances and Applications*, A. Baker and D.W. Masser (eds), Academic Press, 1977, pp. 59-77.

[56] SIEGEL, C.L., *The integer solutions of the equation* $y^2 = ax^n+bx^{n-1}+\ldots+k$, J. London Math. Soc. 1 (1926), 66-68 (under the pseudonym X).

[57] SIEGEL, C.L., *Über einige Anwendungen diophantischer Approximationen*, Abh. Preuss. Akad. Wiss. No. 1 (1929), 1-70.

[58] SIEGEL, C.L., *Die Gleichung* $ax^n-by^n = c$, Math. Ann. 114 (1937), 57-68.

[59] SIERPIŃSKI, W., *On the equation* $3^x+4^y=5^z$ (Polish), Wiadom. Mat. (2) 1 (1955/56), 194-195.

[60] SKOLEM, T., *Diophantische Gleichungen*, Erg. Math. Grenzgeb. Bd. 5, Heft 4, Springer, Berlin, 1938 (repr. by Chelsea, 1950).

[61] SKOLEM, T., *The use of p-adic methods in the theory of diophantine equations*, Bull. Soc. Math. Belg. 7 (1955), 83-95..

[62] STARK, H.M., *Effective estimates on solutions of some diophantine equations*, Acta Arith. 24 (1973), 251-259.

[63] STEINER, R.P., *A theorem on the Syracuse problem*, Proc. Seventh Manitoba Conf. Numer. Math. Comp. 1977, 553-559.

[64] STØRMER, C., *Quelques théorèmes sur l'équation de Pell* $x^2-dy^2=\pm 1$ *et leurs applications*, Skr. Norske Vid. Akad. Oslo, I No. 2 (1897).

[65] STROEKER, R.J., *On the diophantine equation* $x^3-Dy^2=1$, Nieuw Arch. Wisk. (3), 24 (1976), 231-255.

[66] STROEKER, R.J., *On a diophantine equation of E. Bombieri*, Indag. Math. 39 (1977), 131-139.

[67] STROEKER, R.J., *Triangular-square-pentagonal numbers*, Rep. 7701/M, Econometric Inst., Erasmus Univ., Rotterdam, 1977.

[68] STROEKER, R.J., *A class of diophantine equations connected with certain elliptic curves over* $\mathbb{Q}(\sqrt{-13})$, Compositio Math. 38 (1979), 329-346.

[69] STROEKER, R.J., *On the diophantine equation* $(2y^2-3)^2=x^2(3x^2-2)$ *in connection with the existence of non-trivial tight 4-designs*, Indag. Math. 43 (1981), 353-388.

[70] STROEKER, R.J., *Solution of problem 500*, Nieuw Arch. Wiskunde (3), 26 (1978), 476-478.

[71] SZEKERES, G., *Multidimensional continued fractions*, Ann. Univ. Sci. Budapest Eötvös, Sect. Math., 13 (1970), 113-140.

[72] THUE, A., *Über Annäherungswerte algebraischer Zahlen*, J. reine angew. Math. 135 (1909), 284-305.

[73] WAGSTAFF, Jr., S.S., *The irregular primes to 125000*, Math. Comp. 32 (1978), 583-591.

[74] WILLIAMS, H.C. & R. HOLTE, *Computation of the solution of* $x^3+Dy^3=1$, Math. Comp. 31 (1977), 778-785.

[75] WILLIAMS, H.C. & C.R. ZARNKE, *Computation of the solutions of the diophantine equation* $x^3+dy^3 = 1$, Proc. Manitoba Conf. Num. Math. (1971), 671-676.

[76] ZANTEMA, H., *Class numbers and units*, these Proceedings.

[77] ZIMMER, H.G., *Computational problems, methods, and results in algebraic number theory*, Lecture Notes Math. 262, Springer, 1972.

NUMERICAL COMPUTATION OF SPECIAL ZEROS OF PARTIAL SUMS OF RIEMANN's ZETA FUNCTION

by

J. VAN DE LUNE & H.J.J. TE RIELE

0. INTRODUCTION

In 1948 TURÁN [8] showed that the Riemann hypothesis for $\zeta(s)$ is true if there exist positive numbers N_0 and c such that for all $N > N_0$ the functions

$$\zeta_N(s) := \sum_{n=1}^{N} n^{-s}, \quad (s \in \mathbb{C},\ s = \sigma + it)$$

have no zeros in the halfplane $\sigma \geq 1 + cN^{-\frac{1}{2}}$.

In 1958 HASELGROVE [2] implicitly showed (cf. SPIRA [6; Section 3]) that there exist $N \in \mathbb{N}$ such that $\zeta_N(s) = 0$ for some s with $\sigma > 1$.

In 1968 SPIRA [6] proved, computationally, that, for $N = 19,\ 22\ (1)\ 27,\ 29\ (1)\ 50$, $\zeta_N(s)$ has zeros with $\sigma > 1$.

In this paper a zero $s = \sigma + it$ of $\zeta_N(s)$ with the property $\sigma > 1$ will be called a *special zero*. As far as we know, up till now no special zero of any $\zeta_N(s)$ has been located numerically.

We shall present two different methods for the explicit numerical computation of special zeros of $\zeta_N(s)$.

The first method is believed to produce *all* special zeros of $\zeta_N(s)$ with imaginary part in a given interval (Sections 2, 3 and 4). In the second method (Section 5) we first compute several (vertical) "almost-periods" of $\zeta_N(s)$ and then try to find special zeros of $\zeta_N(s)$ by adding these almost-periods to zeros of $\zeta_N(s)$ with real part very close to $\sigma = 1$ (not necessarily in the halfplane $\sigma > 1$). Of course, the second method is by no means exhaustive, but it is much less time consuming than the first one.

Finally, we present a selection of the special zeros of $\zeta_N(s)$ for $N = 19,\ 22(1)27,\ 29(1)35,\ 37(1)41,\ 47$, which were actually computed by the two methods.

The zero-search methods described in this paper may be applied to other functions in analytic number theory as well. For an application to Flett's function we refer to VAN DE LUNE [4]. The idea of using multidimensional continued fraction algorithms (see Section 5) was applied to Mertens's conjecture by TE RIELE [5].

1. SOME GENERALITIES ON THE ZERO CURVES OF THE REAL AND IMAGINARY PARTS OF $\zeta_N(s)$

Before explaining the heuristic principle for finding special zeros of $\zeta_N(s)$ we give a global description of the zero curves of the real and imaginary parts of $\zeta_N(s)$ in the complex plane.

Defining

$$R_N(\sigma,t) := \mathrm{Re}\ \zeta_N(s) = \sum_{n=1}^{N} \frac{\cos(t \log n)}{n^{\sigma}}$$

and

$$I_N(\sigma,\ t) := \mathrm{Im}\ \zeta_N(s) = -\sum_{n=2}^{N} \frac{\sin(t \log n)}{n^{\sigma}}\ ,$$

we obviously have $\zeta_N(s) = 0$ if and only if both $R_N(\sigma,t) = 0$ and $I_N(\sigma,t) = 0$. It is easy to see that

$$R_N(\sigma,t) > 0 \quad \text{for } \sigma \geq 2$$

so that the entire zero set of $\zeta_N(s)$ is contained in the halfplane $\sigma < 2$. Now let N (≥ 3) be fixed and consider the zero-set of $R_N(\sigma,t)$ in the halfplane $\sigma < 0$. If $R_N(\sigma_0,t_0) = 0$ then

$$-N^{-\sigma_0} \cos(t_0 \log N) = \sum_{n=1}^{N-1} n^{-\sigma_0} \cos(t_0 \log n)$$

so that

$$|\cos(t_0 \log N)| \leq \sum_{n=1}^{N-1} \left(\frac{n}{N}\right)^{-\sigma_0} < N \int_0^1 x^{-\sigma_0}\, dx = \frac{N}{1-\sigma_0}\ .$$

Choose a small $\varepsilon > 0$ ($\varepsilon = \frac{1}{N}$ is sufficient) and take $\sigma_0 < 1 - \frac{N}{\varepsilon}$. Then we have $|\cos(t_0 \log N)| < \varepsilon$ so that we must have $t_0 \log N \sim \frac{\pi}{2} + k\pi$, for some $k \in \mathbb{Z}$ or, equivalently, $t_0 \sim (2k+1)\pi/(2 \log N)$, for some $k \in \mathbb{Z}$.

From this it follows that the zero set of $R_N(\sigma,t)$ in the halfplane $\sigma < 1 - \frac{N}{\varepsilon}$ consists of simple zero curves having $-\infty + (2k+1)\pi i/(2\log N)$, $(k \in \mathbb{Z})$ as asymptotical points.

For $\sigma = 1$ (and any other fixed $\sigma \in \mathbb{R}$) we have that $R_N(\sigma,t)$ is an almost periodic function of t and since

$$\max_{t\in\mathbb{R}} R_N(1,t) = R_N(1,0) = \sum_{n=1}^{N} \frac{1}{n}$$

there exist arbitrarily large values t^* of t for which

$$R_N(1,t^*) > -\varepsilon + \sum_{n=1}^{N} \frac{1}{n}$$

or, equivalently,

$$\sum_{n=1}^{N} \frac{1}{n} \cos(t^* \log n) > -\varepsilon + \sum_{n=1}^{N} \frac{1}{n} . \tag{1}$$

Choosing $\varepsilon > 0$ small enough it follows that all cosines in (1) are close to 1 and hence positive so that for these particular values t^* of t we have

$$R_N(\sigma,t^*) = \sum_{n=1}^{N} n^{-\sigma} \cos(t^* \log n) > 0 \quad \text{for all } \sigma \in \mathbb{R}.$$

Hence, these horizontal lines $t = t^*$ act as barriers for the zero-lines of $R_N(\sigma,t)$. Since the zero lines of any harmonic function on the entire plane cannot have endpoints, it follows that a zero line of $R_N(\sigma,t)$ starting at a point

$$-\infty + \frac{(2k+1)\pi i}{2\log N}$$

must return to some other asymptotical point of the same form (possibly not a neighbouring one).

Next, we consider the zero lines of $I_N(\sigma,t)$. If $I_N(\sigma_0,t_0) = 0$ then

$$N^{-\sigma_0} \sin(t_0 \log N) = -\sum_{n=2}^{N-1} n^{-\sigma_0} \sin(t_0 \log n)$$

so that for $\sigma_0 < 0$

$$|\sin(t_0 \log N)| \le \sum_{n=2}^{N-1} \left(\frac{n}{N}\right)^{-\sigma_0} < \frac{N}{1-\sigma_0} .$$

Similarly as before, we choose a small $\varepsilon > 0$ and take $\sigma_0 < 1 - \frac{N}{\varepsilon}$ so that $|\sin(t_0 \log N)| < \varepsilon$. Consequently, $t_0 \log N \sim k\pi$, for some $k \in \mathbb{Z}$, or, equivalently, $t_0 \sim k\pi/\log N$, for some $k \in \mathbb{Z}$. Hence, the zero set of $I_N(\sigma,t)$ in the halfplane $\sigma < 1 - \frac{N}{\varepsilon}$ consists of a system of simple zero curves having the points $-\infty + k\pi i/\log N$, $(k \in \mathbb{Z})$ as asymptotical points.

For large positive σ we have in case of a zero of $I_N(\sigma,t)$

$$2^{-\sigma_0} \sin(t_0 \log 2) = -\sum_{n=3}^{N} n^{-\sigma_0} \sin(t_0 \log n)$$

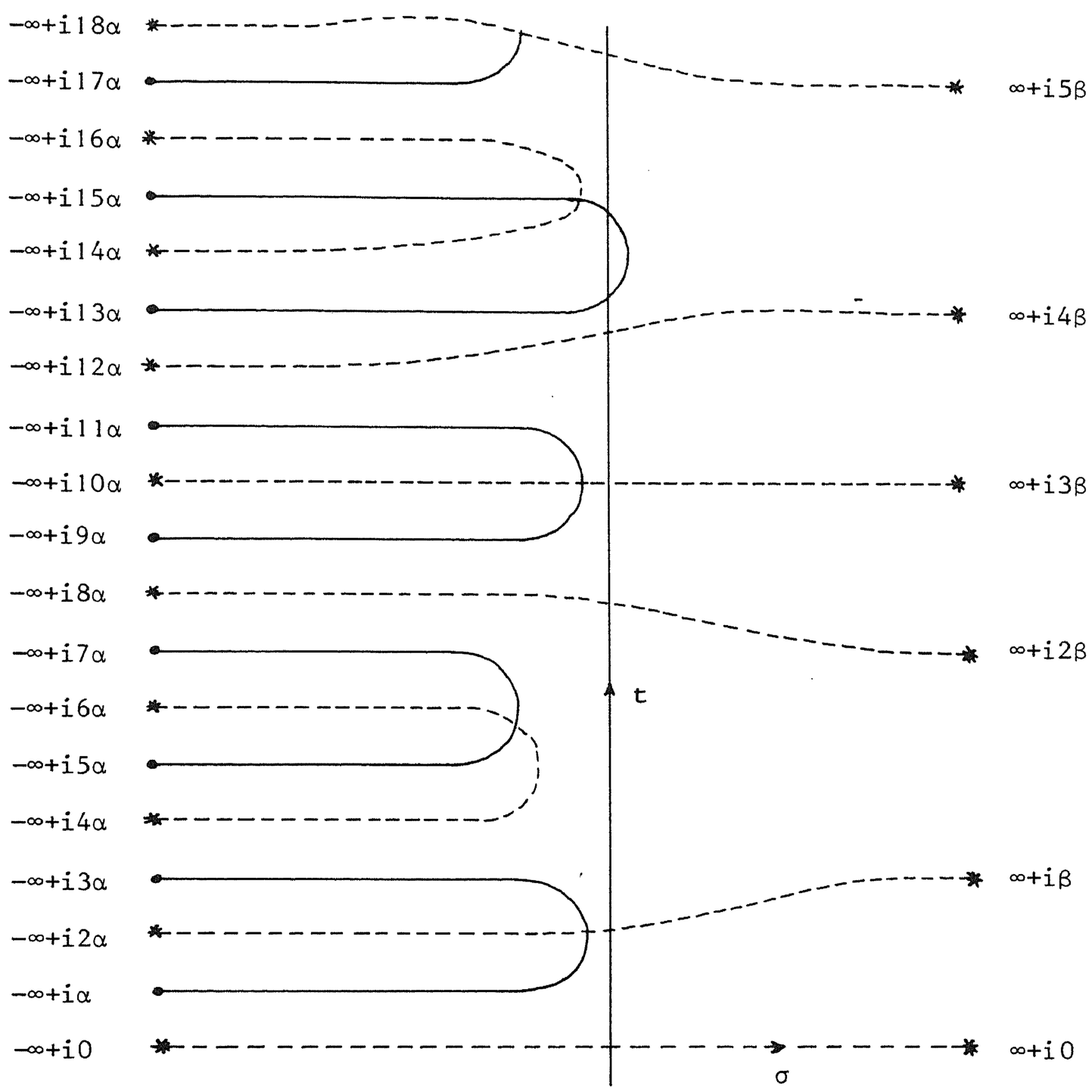
and hence $|\sin(t_0 \log 2)| \leq \Sigma_{n=3}^{N} (\frac{2}{n})^{\sigma 0} < N(\frac{2}{3})^{\sigma 0}$. Choosing a small $\varepsilon > 0$ and taking $\sigma_0 > \log(N/\varepsilon)/\log(3/2)$ we thus have $|\sin(t_0 \log 2)| < \varepsilon$ so that $t_0 \log 2 \sim k\pi$, for some $k \in \mathbb{Z}$, or, equivalently, $t_0 \sim k\pi/\log 2$, for some $k \in \mathbb{Z}$. It follows that the zero set of $I_N(\sigma,t)$ in the halfplane $\sigma > \log(N/\varepsilon)/\log(3/2)$ consists of simple zero curves having the points $+\infty + k\pi i/\log 2$ $(k \in \mathbb{Z})$ as asymptotical points.

It can be shown that every zero curve of $I_N(\sigma,t)$ starting at some asymptotical point $+\infty + k\pi i/\log 2$ is somehow connected with some asymptotical point $-\infty + \ell\pi i/\log N$. In other words: such a zero curve traverses the s-plane more or less horizontally.

Moreover, every zero curve of $I_N(\sigma,t)$ starting at $-\infty + k\pi i/\log N$ is either connected with an asymptotical point $+\infty + \ell\pi i/\log 2$ or with an asymptotical point of the form $-\infty + m\pi i/\log N$.

Drawing the zero curves of $I_N(\sigma,t)$ as dotted lines, the zero curves of $I_N(\sigma,t)$ and $R_N(\sigma,t)$ have a typical pattern as sketched in Figure 1. This sketch is based on actual computations of the signs of R_N and I_N for various values of N.

It may be noted here that in some earlier reports ([4A] and [4B]) the authors showed that for $N \leq 10$, $N \neq 7$, the zero curves of $R_N(\sigma,t)$ do not intersect the vertical $\sigma = 1$. Hence, in order to find a special zero of $\zeta_N(s)$ we should take N fairly large (see Section 4).



$\alpha = \frac{\pi}{2 \log N}$ $\qquad$ $\beta = \frac{\pi}{\log 2}$

——— zero curve of $R_N(\sigma,t)$

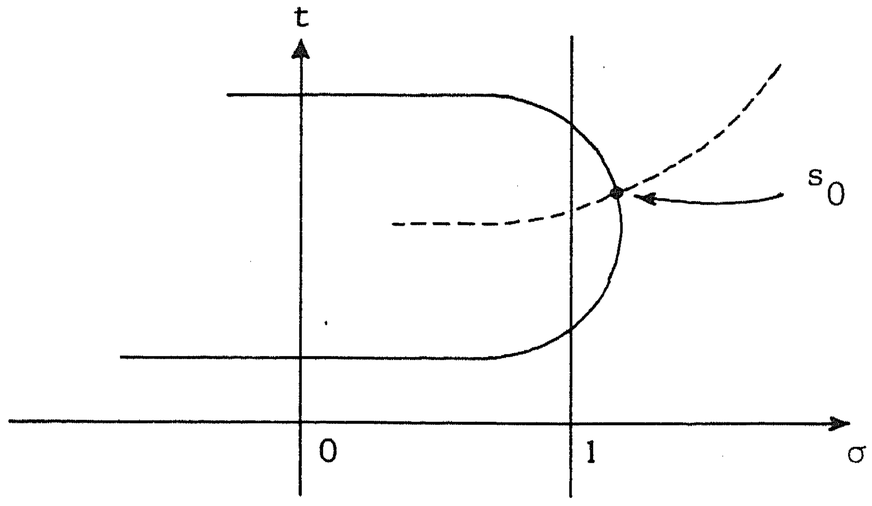
- - - - - zero curve of $I_N(\sigma,t)$

Figure 1

2. THE HEURISTIC PRINCIPLE

In case of a special zero s_0 of $\zeta_N(s)$ we expect to have a pattern as sketched in Figure 2. Here (in accordance with our numerical observations) we have tacitly assumed that all zeros of $\zeta_N(s)$ are simple, so that in s_0

the zero curves of R_N and I_N are perpendicular.



——— zero curve of R_N (σ,t)

- - - - - zero curve of I_N (σ,t)

Figure 2.

In order to detect such a pattern of the zero curves of R_N and I_N we have computed zeros of $R_N(1,t)$ for $t > 0$, yielding the increasing sequence $t_1 < t_2 < \ldots$ of zeros of $R_N(1,t)$. Once the zeros $t_{2\ell-1}$ and $t_{2\ell}$ were located it was checked whether $I_N(1,t)$ had a zero between $t_{2\ell-1}$ and $t_{2\ell}$. If so, it was a simple matter to locate the corresponding (special) zero of $\zeta_N(s)$.

A slight modification of this procedure may be used to obtain zeros of $\zeta_N(s)$ with real part just less than 1.

3. THE MAXIMAL SLOPE PRINCIPLE

For the systematic search of zeros of $R_N(1,t)$ we apply what we call the *maximal slope* principle, which we shall describe first. Let $f(t)$ be a differentiable function for $t \geq t_0$, and suppose that

$$|f'(t)| \leq M_0 \quad \text{for } t \geq t_0.$$

The maximal slope principle is the simple observation that if $f(t_0) > 0$ and f is not linear than $f(t) > 0$ for $t \in [t_0, t_1]$, where

$$t_1 := t_0 + \frac{f(t_0)}{M_0}.$$

If t^* is the smallest zero of f which is larger than t_0, this principle provides us with a new lower bound t_1 for t^*. By repeating this principle, t^* may be approximated as close as desired. Note that every application of this principle requires an evaluation of f. If the definition of f contains functions like sin, cos or log (see the definitions of R_N and I_N) it is considerably more efficient to apply the following modification of the maximal slope principle.

Suppose that for some $k > 0$

$$|f^{(k+1)}(t)| \le M_k \quad \text{for } t \ge t_0.$$

Then, from the Taylor expansion of f(t) around t_0, we have

$$P_k(t_0,t) := \sum_{r=0}^{k} \frac{f^{(r)}(t_0)}{r!}(t-t_0)^r - \frac{M_k}{(k+1)!}(t-t_0)^{k+1} \le f(t),$$

for all $t \ge t_0$.

If an evaluation of $P_k(t_0,t)$ is considerably cheaper than an evaluation of f(t), then it is preferable to apply the maximal slope principle to P_k rather than to f. This yields an increasing sequence of points $t_{0,j}$ $(j = 0,1,2,\ldots)$ defined by

$$t_{0,0} := t_0 \quad \text{and} \quad t_{0,j+1} := t_{0,j} + \frac{P_k(t_0,t_{0,j})}{M_0}, \qquad j = 0,1,2,\ldots.$$

We interrupt the procedure at $t = t_{0,n}$ if $P_k(t_0,t_{0,n}) \le \varepsilon$ $(= 10^{-6}$, say). Note that $t_{0,j} < t^*$ for $j = 0,1,\ldots,n$. Now we compute $f(t_{0,n})$. If $f(t_{0,n}) > \varepsilon$ then we put $t_1 := t_{0,n}$ and set up a *new* polynomial $P_k(t_1,t)$ and continue as above. This yields a finite sequence $t_{1,0} := t_1, t_{1,1}, t_{1,2}, \ldots$, and at the next repetition we get $t_{2,0} := t_2, t_{2,1}, \ldots$. We continue until we find an m and a corresponding $n = n(m)$ such that $P_k(t_m,t_{m,n}) \le \varepsilon$ *and* $f(t_{m,n}) \le \varepsilon$. If so, we compute $f(t_{m,n} + \delta)$ (with $\delta = 10^{-2}$, say). The values of ε and δ given above were determined experimentally such that always $f(t_{m,n}) \cdot f(t_{m,n} + \delta) < 0$ (see Figure 3). The next sign change of f(t) is determined similarly, starting at $t_0 := t_{m,n} + \delta$.

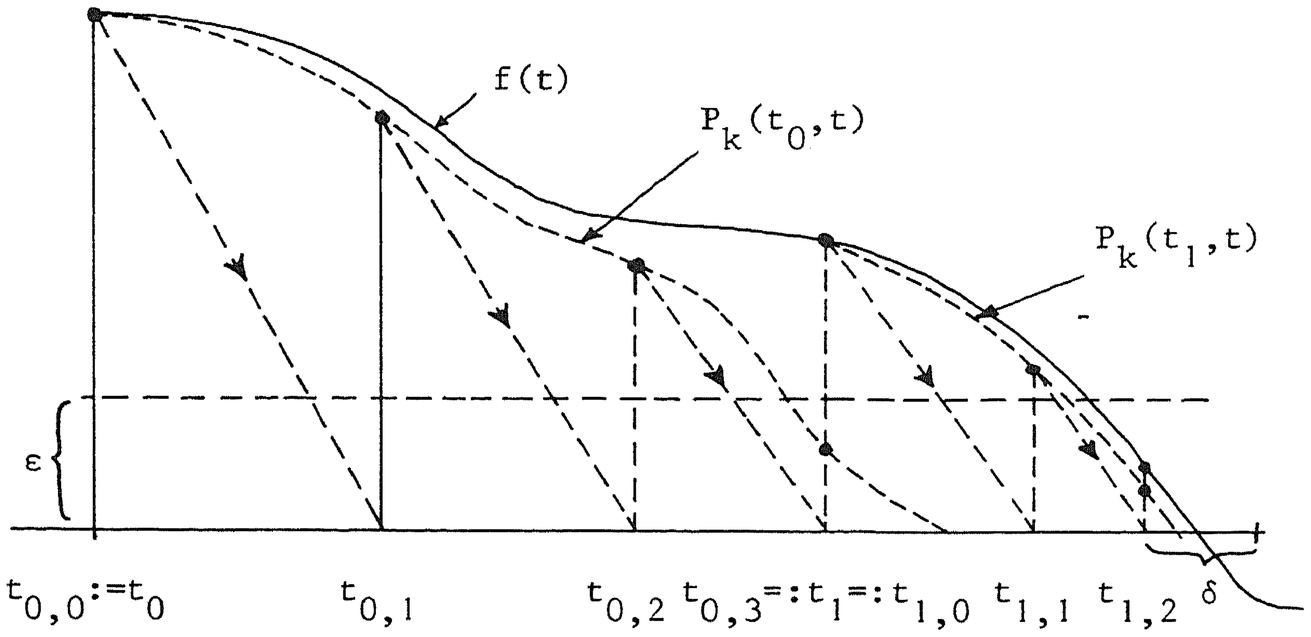


Figure 3.

4. THE SYSTEMATIC SEARCH FOR ZEROS OF $R_N(1,t)$

In order to apply the maximal slope principle to $R_N(1,t)$ we need suitable estimates for $\sup_t |R_N'(1,t)| =: M_{0,N}$ and $\sup_t |R_N^{(k+1)}(1,t)| =: M_{k,N}$. Since

$$R_N'(1,t) = -\sum_{n=2}^{N} \frac{\log n}{n} \sin(t \log n),$$

we have

$$M_{0,N} = \sup_{t\in\mathbb{R}} \left| \sum_{n=2}^{N} \frac{\log n}{n} \sin(t \log n) \right| \le \sum_{n=2}^{N} \frac{\log n}{n},$$

which yields, e.g., $M_{0,22} < 4.78$. By using the prime decomposition of all $n \in [2,N]$ and the linear independence of the logarithms of the primes over the rationals, it was possible to derive the improved bound $M_{0,22} < 4.275$. However, this improvement did not speed up the systematic search considerably.

For the higher derivatives we used the straightforward estimates

$$M_{k,N} := \sup_t |R_N^{(k+1)}(1,t)| \le \sum_{n=2}^{N} \frac{(\log n)^k}{n}.$$

Similar estimates were used to find zeros of $I_N(1,t)$.

For N = 23 our procedure led very quickly to the special zero

$$\sigma = 1.008\ 496\ 93, \quad t = 8645.524\ 423\ 32.$$

It took considerably more time to find a special zero for N = 19. We found

$$\sigma = 1.001\ 095\ 51, \quad t = 600\ 884.203\ 427\ 78.$$

SPIRA's investigations [6] show that N = 19,22 and 23 are the first candidates for having special zeros. We did not succeed in finding a special zero of $\zeta_{22}(s)$ in the range $0 \le t \le 75{,}000{,}000$. (Note that 22 is composite!) Various experiments showed that k = 14 was the optimal choice in this case. Anticipating the results of the next section we already note here that by the method described there we have found the following special zero for N = 22:

$$\sigma = 1.002\ 890\ 95, \quad t = 558\ 159\ 406.148\ 225\ 57.$$

5. SECOND METHOD: SEARCH BY USE OF ALMOST-PERIODS

In this section we describe a second method for the computation of special zeros of ζ_N. In fact, by this method we are able to construct (finite) sequences of zeros of ζ_N, all with real part close to one, some of them with real part *greater* than one.

The starting point is the supposition that already a zero s_0 of ζ_N is available, for which $|\mathrm{Re}\ s_0 - 1|$ is small. Such a zero may be found, for instance, by applying our first method to a line $\sigma = 1 - \varepsilon$. Let $T_1 \in \mathbb{R}$ be such that $|\zeta_N(s) - \zeta_N(s + iT_1)|$ is small for all s on the line $\sigma = 1$. Such a T_1 exists since $\zeta_N(1 + it)$ is an almost-periodic function of t. Then one may expect that also $|\zeta_N(s_0) - \zeta_N(s_0 \pm iT_1)|$ is small, and there may be a zero, s_1 say, of ζ_N in the neighbourhood of $s_0 \pm iT_1$. If $\mathrm{Re}\ s_1 > \mathrm{Re}\ s_0$, we look for another zero, s_2 say, of ζ_N in the neighbourhood of $s_1 \pm iT_1$, and so on. In order to cross the line $\sigma = 1$, we always demand that $\mathrm{Re}\ s_j > \mathrm{Re}\ s_{j-1}$. If $\mathrm{Re}\ s_j \le \mathrm{Re}\ s_{j-1}$ we continue with another almost-period T_2. After crossing the line $\sigma = 1$ we may still continue this procedure in order to find more and more special zeros of ζ_N.

The crucial point in the above procedure is, of course, the availability of sufficiently many almost-periods of ζ_N on the line $\sigma = 1$. We have

LEMMA 5.1. *Almost-periods of* $\zeta_N(s)$ *may be computed from "sufficiently go(*
(to be specified later) approximations of the $\pi(N)$ (>1) *numbers* $\log p_j/\ln c$
$(j = 1,2,\ldots,\pi(N);\ j_0 \in \{1,2,\ldots,\pi(N)\})$ *by rational numbers with the sam(*
denominator.

PROOF. Let k be such a common denominator, i.e., $k \log p_j/\log p_{j_0} \equiv \varepsilon_j$ (m(
where $\varepsilon_{j_0} = 0$ and the other ε_j's are small or close to 1 (but not zero, si(
the logarithms of the primes are independent over $\mathbb{Q}$). Let the canonical fac(
ization of n $(\le N)$ be given by $n = \prod_{j=1}^{\pi(N)} p_j^{\alpha_j(n)}$. Then for $T := k\cdot 2\pi/\log$ (
and for any fixed $s \in \mathbb{C}$ we have

$$\zeta_N(s+iT) = \sum_{n=1}^{N} n^{-s} \exp(-iT \log n) = \sum_{n=1}^{N} n^{-s} \exp(-i\theta_n),$$

where

$$\theta_n = T \log n = (k\cdot 2\pi/\log p_{j_0}) \log \prod_{j=1}^{\pi(N)} p_j^{\alpha_j(n)}$$
$$= 2\pi \sum_{j=1}^{\pi(N)} \alpha_j(n) k \log p_j/\log p_{j_0}$$
$$\equiv \Big(\sum_{j=1}^{\pi(N)} \varepsilon_j \alpha_j(n)\Big) \pmod{2\pi}.$$

If the ε_j's are small enough, we may expect the value of $\zeta_N(s+iT)$ to be close to the value of $\zeta_N(s)$. Hence, T is an almost-period of ζ_N. The same argument holds, if one replaces T by -T. □

We have used the well-known modified Jacobi-Perron algorithm [1] and the less-known Szekeres algorithm [7] for the computation of the rational approximations of $\log p_j/\log p_{j_0}$ $(j = 1,2,\ldots,\pi(N);\ j \neq j_0)$. We first give description of both algorithms in the style of KNUTH [3]. Both algorithms are simplified and put in a form suitable for our purpose.

ALGORITHM JP (Jacobi-Perron). Given $n \ge 1$ positive irrational numbers $\alpha_1,\alpha_2,\ldots,\alpha_n$. In step JP2 a positive integer k is computed such that $\{k\alpha_i\}$ is small, for $i = 1,2,\ldots,n$ (where $\{x\}$ means the distance of x to the nea(
est integer). Auxiliary vectors $\vec{a} = (a_1,a_2,\ldots,a_n)$, $\vec{b} = (b_1,\ldots,b_n)$ and $\vec{c}$
$(c_0,c_1,\ldots,c_n)$ are used. The algorithm terminates when $k > k_{max}$.

JP1. [Initialize]. Set $c_0 \leftarrow 0$ and set $a_i \leftarrow \alpha_i$ and $c_i \leftarrow 0$, for $i = 1,2,..$

JP2. [Take integer part of a and compute new k]. Set $b_i \leftarrow [a_i]$ for $i =$ $1,2,\ldots,n$ and set $k \leftarrow c_0 + \Sigma_{i=1}^{n} c_i b_i$. If $k > k_{max}$ then stop.

JP3. [Compute new $\vec{c}$ and $\vec{a}$]. Set $c_0 \leftarrow c_1$, $c_i \leftarrow c_{i+1}$ and $a_i \leftarrow (a_{i+1}-b_{i+1})/(a_1-b_1)$, for $i = 1,2,\ldots,n-1$ and set $c_n \leftarrow k$ and $a_n \leftarrow 1/(a_1-b_1)$. Go to JP2.

Note that for $n = 1$, this algorithm produces the denominators of the convergents of the regular continued fraction expansion of α_1.

The Szekeres algorithm is more complicated than JP, but it will appear to produce much better approximations than JP.

ALGORITHM SZ (Szekeres). Given $n \geq 1$ positive irrational numbers $\alpha_1,\alpha_2,\ldots,\alpha_n$, with $1 > \alpha_1 > \alpha_2 > \ldots > \alpha_n$. In Step SZ6 a positive integer k is computed such that $\{k\alpha_i\}$ is small, for $i = 1,2,\ldots,n$. An auxiliary vector $\vec{\gamma} = (\gamma_0,\gamma_1,\ldots,\gamma_n)$, auxiliary arrays $A = (a_{ij})$, $i,j = 0,1,\ldots,n$ and $V = (v_{ij})$, $i,j = 1,2,\ldots,n$, and an auxiliary scalar h are used. The algorithm terminates, when $k > k\max$. In order to explain the notation in SZ3, we define a partial ordering of n-component vectors as follows: let $\vec{x} = (x_1,\ldots,x_n)$ and $\vec{y} = (y_1,\ldots,y_n)$ and let $i_1,i_2,\ldots,i_n$ be a permutation of $1,2,\ldots,n$ such that $|x_{i_1}| \geq |x_{i_2}| \geq \ldots \geq |x_{i_n}|$; similarly, let $|y_{j_1}| \geq |y_{j_2}| \geq \ldots \geq |y_{j_n}|$. We write $\vec{x} \simeq \vec{y}$ if $|x_{i_\mu}| = |y_{j_\mu}|$, for $\mu = 1,2,\ldots,n$ and $\vec{x} \prec \vec{y}$ if $\exists\nu$, $1 \leq \nu \leq n$ such that $|x_{j_\nu}| < |y_{j_\nu}|$, and $|x_{j_\mu}| = |y_{j_\mu}|$, for $1 \leq \mu < \nu$.

SZ1. [Initialize]. Set $\gamma_0 \leftarrow 1-\alpha_1$, $\gamma_i \leftarrow \alpha_i - \alpha_{i+1}$, $i = 1,2,\ldots,n-1$, $\gamma_n \leftarrow \alpha_n$. Set $a_{ij} \leftarrow 1$, $i = 0,1,\ldots,n$ and $j = 0,1,\ldots,i$ and $a_{ij} \leftarrow 0$, $i = 0,1,\ldots,n-1$ and $j = i+1,i+2,\ldots,n$.

SZ2. [Compute the differences v_{ij}]. Set $v_{ij} \leftarrow |a_{ij}/a_{i0} - a_{0j}/a_{00}|$, $i,j = 1,2,\ldots,n$.

SZ3. [Select index μ]. Let $\vec{v}_i$ be the i-th row of V, so $\vec{v}_i = (v_{i1},v_{i2},\ldots,v_{in})$. Find the largest index μ such that for every $1 \leq i \leq n$

$$\text{either } \vec{v}_i \prec \vec{v}_\mu, \quad \text{or} \quad \vec{v}_i \simeq \vec{v}_\mu.$$

If $\gamma_0 < \gamma_\mu$, then go to SZ5.

SZ4. [$\gamma_0 \geq \gamma_\mu$]. Set $\gamma_0 \leftarrow \gamma_0 - \gamma_\mu$ and $a_{\mu j} \leftarrow a_{\mu j} + a_{0j}$, $j = 0,1,\ldots,n$. Go to SZ6.

SZ5. [$\gamma_0 < \gamma_\mu$]. Set $h \leftarrow \gamma_0$ and $\gamma_0 \leftarrow \gamma_\mu - \gamma_0$, $\gamma_\mu \leftarrow h$. Set $h \leftarrow a_{0j}$ and $a_{0j} \leftarrow a_{\mu j}$, $a_{\mu j} \leftarrow a_{\mu j} + h$, for $j = 0,1,\ldots,n$.

SZ6. [New k]. Set $k \leftarrow a_{\mu 0}$. If $k \leq k\max$, then go to SZ2, else stop.

For $n = 1$, this algorithm not only produces the denominators of the convergents of the regular continued fraction expansion of α_1, but also the denominators of the *intermediary* convergents.

Both algorithms were coded in FORTRAN, and run on a CDC 6600 computer, in double precision (28 significant digits) with $k\max = 10^{20}$, $n = 6$ and for α_i the six irrationals $\log 3/\log 2$, $\log 5/\log 2$, $\log 7/\log 2$, $\log 11/\log 2$, $\log 13/\log 2$, and $\log 17/\log 2$. Let $k_1, k_2, \ldots$ be the sequence of k's produced by one of the algorithms. Define $m_i := \max_{1\le j\le 6} \{k_i\alpha_j\}$. In Table 1, for both algorithms we give the values of k_j and m_j, such that $m_j < m_i$, for $1 \le i \le j-1$. Clearly the results of SZ are much better than those of JP, so that we decided to choose the Szekeres algorithm for our further computations.

As indicated in Section 4, we first applied the method of this section to N = 22. In order to find almost periods for N = 22, we ran the SZ algorithm with N = 19, i.e. $\pi(N) = 8$ and $j_0 = 1$, 2, 3 and 4. This yielded sufficiently many almost periods, and with the strategy described in the beginning of this section, we found many special zeros of $\zeta_{22}(s)$.

Although we already had found a few special zeros of ζ_{19} by the systematic method, we also applied the almost period method to ζ_{19}. As an illustration of the power of this method, we select the following result:

$\zeta_{19}(s) = 0$ for $s = \sigma_0 + it_0$, where

$\sigma_0 = 1.002\ 793\ 85$, $t_0 = 987\ 047\ 804\ 990\ 437\ 138.210\ 000\ 67$

and for $k = 1,2,\ldots,58$ the numbers $t_k = t_0 + kP$, where

$P = 119\ 473\ 414\ 699\ 017\ 719\ 233.343\ 2$,

are approximations, with absolute error of, at most, 0.1, of the imaginary parts of special zeros of ζ_{19}. These zeros are listed in Table 2 (σ rounded to 8, t to 5 decimals). We have also listed the first zero in this "almost-arithmetic progression" with real part < 1 (namely the zero with imaginary part $\approx t_0 + 59P$).

Table 1

Results of runs with the Jacobi-Perron Algorithm
and the Szekeres Algorithm

ALG.	j	k_j	m_j
JP	1	1	.460
	3	2	.401
	8	168	.365
	9	877	.331
	10	882	.219
	17	278575	.164
	25	1170241231	.158
	26	18158873714	.0675
	31	9176933208351	.0654
	35	259812674489863	.0349
SZ	1	2	.401
	8	4	.350
	19	9	.304
	30	31	.289
	49	311	.201
	57	764	.181
	71	2414	.139
	80	5855	.111
	83	14348	.0910
	113	88209	.0871
	116	119365	.0798
	125	272356	.0483
	149	2316275	.0276
	169	23993538	.0221
	218	890512495	.0184
	225	2039172447	.0178
	234	2929684942	.0167
	239	5312742147	.0115
	246	9640622028	.0106
	263	69123516771	.00715
	296	1903569470016	.00704
	297	2244797172219	.00615
	399	1740704456733	.00548
	300	2907809851158	.00522
	325	13059799506657	.00353
	339	61833456490027	.00344
	343	65818958118979	.00180
	392	71641948032572 68	.00167
	407	381014737150800 26	.00115
	419	1020255017592578 46	.00107
	447	1778599299350212805	.00053
	448	1485640231520813937	.00046

Table 2

59 special zeros of ζ_{19}, the imaginary parts of which form an "almost" arithmetic progression, and the first "non-special" zero in this progression.

σ	t
1.00279385	9870478049904371 38.21000
1.00287891	1204604625040081563 71.55227
1.00295917	2399338772030258756 04.89453
1.00303464	3594072919020435948 38.23680
1.00310532	4788807066010613140 71.57906
1.00317121	5983541213000790333 04.92133
1.00323237	7178275359990967525 38.26360
1.00328876	8373009506981144717 71.60587
1.00334038	9567743653971321910 04.94813
1.00338727	1076247780096149910 238.29040
1.00342941	1195721194795167629 471.63267
1.00346685	1315194609494185348 704.97495
1.00349959	1434668024193203067 938.31722
1.00352756	1554141438892220787 171.65949
1.00355087	1673614853591238506 405.00176
1.00356948	1793088268290256225 638.34404
1.00358339	1912561682989273944 871.68631
1.00359263	2032035097688291664 105.02859
1.00359720	2151508512387309383 338.37086
1.00359712	2270981927086327102 571.71314
1.00359237	2390455341785344821 805.05542
1.00358294	2509928756484362541 038.39770
1.00356893	2629402171183380260 271.73997
1.00355030	2748875585882397979 505.08225
1.00352700	2868349000581415698 738.42453
1.00349914	2987822415280433417 971.76681
1.00346660	3107295829979451137 205.10910
1.00342954	3226769244678468856 438.45138
1.00338783	3346242659377486575 671.79366
1.00334159	3465716074076504294 905.13595
1.00329071	3585189488775522014 138.47823
1.00323534	3704662903474539733 371.82052
1.00317535	3824136318173557452 605.16280
1.00311082	3943609732872575171 838.50509
1.00304179	4063083147571592891 071.84738
1.00296821	4182556562270610610 305.18966
1.00289013	4302029976969628329 538.53195
1.00280750	4421503391668646048 771.87424
1.00272038	4540976806367663768 005.21653
1.00262865	4660450221066681487 238.55883
1.00253266	4779923635765699206 471.90112
1.00243208	4899397050464716925 705.24341
1.00232686	5018870465163734644 938.58570
1.00221735	5138343879862752364 171.92800
1.00210347	5257817294561770083 405.27029
1.00198488	5377290709260787802 638.61259

Table 2 (cont'd)

1.00186194	549676412395980552187 1.95489
1.00173467	561623753865882324110 5.29718
1.00160285	573571095335784096033 8.63948
1.00146665	585518436805685867957 1.98178
1.00132607	597465778275587639880 5.32408
1.00118127	609413119745489411803 8.66638
1.00103183	621360461215391183727 2.00868
1.00087808	633307802685292955650 5.35098
1.00071993	645255144155194727573 8.69329
1.00055737	657202485625096499497 2.03559
1.00039068	669149827094998271420 5.37789
1.00021931	681097168564900043343 8.72020
1.00004367	693044510034801815267 2.06250
.99986388	704991851504703587190 5.40481

In order to find almost periods for ζ_N, $23 \leq N \leq 28$, we ran the SZ algorithm with N = 23, i.e. $\pi(N) = 9$, and $j_0 = 1$, 2, 3 and 4.

Unfortunately the SZ algorithm did not produce satisfactory results for $\pi(N) \geq 10$, unless we extended the precision of the calculations. Instead of doing this we decided to try to find zeros of ζ_N, $N \geq 29$ with the use of the almost periods found with the SZ algorithms, for the cases $\pi(N) = 8$ and $\pi(N) = 9$. This led to satisfactory results.

In Table 3 we give a selection of special zeros found by means of the two methods described above. σ and t are rounded to 8 decimals. All zeros with imaginary part greater than 5.10^8 were found by the method of almost periods described in this section.

Table 3

A selection of special zeros of ζ_N, N = 19,22(1)27,29(1)35,37(1)41,47, computed with the systematic or with the almost period method

N	σ	t
19	1.00109551	600884.20342778
19	1.00235653	11771253.22839263
22	1.00289095	558159406.14822557
22	1.00159434	46892766540.42816696
23	1.00849693	8645.52442332
23	1.00519091	938296.18122556
23	1.01338428	32520751.77163493
24	1.00404187	32520751.78599510
24	1.00266176	558159406.14677888
25	1.00044920	32520751.80223907
25	1.00281451	1948209609528.90258422
26	1.00147172	3202110.43537085
26	1.00515827	32520751.81725186
27	1.00041028	61242054160408938.59968064
29	1.00370506	2589158977352418.11781520
29	1.00263365	31626643541569868.61843369
30	1.00035753	2589158977352418.10546556
31	1.00710369	52331955.65876128
31	1.01237852	2589158977352418.10678941
31	1.01213846	31626643541569868.60340243
32	1.00165867	2589158977352418.10218851
32	1.00064974	31626643541569868.59995286
33	1.00311308	2589158977352418.09084140
33	1.00006912	31626643541569868.58813015
34	1.00224271	2589158977352418.07991295
34	1.00231563	31626643541569868.57704514
35	1.00271904	2589158977352418.06938499
35	1.00632459	31626643541569868.56710359
37	1.00386526	2589158977352418.06806263
38	1.00612140	2589158977352418.05885220
39	1.00801942	2589158977352418.04998790
40	1.00138033	2589158977352418.04412159
41	1.00099738	2589158977352418.05290762
47	1.00039216	20749499.96408269

REFERENCES

[1] BERNSTEIN, L., *The modified algorithm of Jacobi-Perron*, Memoirs of the Amer. Math. Soc., Number 67, 1966.

[2] HASELGROVE, C.B., *A disproof of a conjecture of Pólya*, Matematika, 5 (1958), 141-145.

[3] KNUTH, D.E., *The art of computer programming*, Vol. 1, Fundamental Algorithms, Addison-Wesley, 1968.

[4] LUNE, J. VAN DE, *A note on the zeros of Flett's function*, Report ZW 167/81, Mathematical Centre, Amsterdam.

[4A] LUNE, J. VAN DE, *A note on the partial sums of* $\zeta(s)$, Report ZW 53/75, Mathematical Centre, Amsterdam.

[4B] LUNE, J. VAN DE & H.J.J. TE RIELE, *A note on the partial sums of* $\zeta(s)$, Reports ZW 58/75 (part II) and ZW 84/76 (part III), Mathematical Centre, Amsterdam.

[5] RIELE. H.J.J. TE, *Computations concerning the conjecture of Mertens*, J. reine angew. Math., 311/312 (1979), 356-360.

[6] SPIRA, R., *Zeros of sections of the zeta function, II*, Math. Comp., 22 (1968), 163-173.

[7] SZEKERES, G., *Multidimensional continued fractions*, Ann. Univ. Sci. Budapest. de Rolando Eötvös nom., 13(1970), 113-140.

[8] TURÁN, P., *On some approximative Dirichlet polynomials in the theory of the zeta function of Riemann*, Danske Vid. Selsk. Mat. Fys. Medd., 24 (1948), 3-36.

THE FIRST 200,000,001 ZEROS OF RIEMANN's ZETA FUNCTION

by

R.P. BRENT, J. VAN DE LUNE, H.J.J. TE RIELE & D.T. WINTER

1. INTRODUCTION

This paper contains a description of extensive computations carried out by Brent at the Department of Computer Science of the Australian National University (Canberra) and by van de Lune, te Riele and Winter at the Mathematical Centre (Amsterdam, The Netherlands). The main results will appear in 1982 in Mathematics of Computation. The details of the computations by van de Lune, te Riele and Winter have been described in the Mathematical Centre Report NW 113/81 [12].

Riemann's zeta function is the meromorphic function $\zeta: \mathbb{C}\setminus\{1\} \to \mathbb{C}$, which, for $\mathrm{Re}(s) > 1$, may be represented explicitly by

$$\zeta(s) = \Sigma_{n=1}^{\infty} n^{-s}, \quad (s = \sigma + it).$$

It is well-known (see TITCHMARSH [17, Chapters II and X]) that

$$\xi(s) := \tfrac{1}{2}s(s-1)\pi^{-\frac{s}{2}}\Gamma(\tfrac{s}{2})\zeta(s)$$

is an entire function of order 1, satisfying the functional equation

$$\xi(s) = \xi(1-s),$$

so that

$$\Xi(z) := \xi(\tfrac{1}{2} + iz), \quad (z \in \mathbb{C}),$$

being an even entire function of order 1, has an infinity of zeros. The Riemann Hypothesis is the statement that all zeros of $\Xi(z)$ are real, or, equivalently, that all non-real zeros of $\zeta(s)$ lie on the "critical" line $\sigma = \frac{1}{2}$. Since $\zeta(\bar{s}) = \overline{\zeta(s)}$ we may restrict ourselves to the half plane $t > 0$. To this day, Riemann's Hypothesis has neither been proved nor disproved.

Numerical investigations related to this unsolved problem were initiated by Riemann himself and later on continued more systematically by the

writers listed below (including their progress).

Investigator	Year	The first n complex zeros of $\zeta(s)$ are simple and lie on $\sigma = \frac{1}{2}$
GRAM [6]	1903	n = 15
BACKLUND [1]	1914	n = 79
HUTCHINSON [7]	1925	n = 138
TITCHMARSH [16]	1935/6	n = 1,041

Those listed above utilized the Euler-Maclaurin summation formula and performed their computations by hand or desk calculator whereas those listed below applied the Riemann-Siegel formula in conjunction with electronic computing devices.

LEHMER [10,11]	1956	n = 25,000
MELLER [13]	1958	n = 35,337
LEHMAN [9]	1966	n = 250,000
ROSSER, YOHE & SCHOENFELD [15]	1968	n = 3,500,000
BRENT [2]	1979	n = 81,000,001

An excellent explanatory account of most of these computations may be found in EDWARDS [4].

In this paper (which presupposes the knowledge of BRENT [2]) we report on extensive computations by which the first named author has extended his former result to n = 156,800,001 and by which the remaining three authors (LR & W, for short) have extended this bound to n = 200,000,001. Independently of Brent, LR & W have also checked the range $[g_{81,000,000}, g_{120,000,000})$.

In practice, the numerical verification of the Riemann hypothesis in a given range consists of separating the zeros of the well-known real function Z(t) (see formula (2.6) of BRENT [2] or formula (3.1) in Section 3 of this paper), or, equivalently, of finding sufficiently many sign changes of Z(t). Our programs (aiming at a fast separation of these zeros) are based, essentially, on the modification of LEHMER's [11] method introduced by ROSSER et al. [15]. However, LR & W have developed a more efficient strategy of searching for sign changes of Z(t) in Gram blocks of length $L \geq 2$. Brent's average number of Z-evaluations, needed to separate a zero from its

predecessor, amounts to about 1.41 (compare BRENT [2]) whereas LR & W have brought this figure down to about 1.21. It may be noted here that in the most recent version of the program of LR & W this figure has been reduced further to about 1.185. From the statistics in Section 4 it follows that in the range $[g_{156,800,000}, g_{200,000,000})$ this average number of Z-evaluations could not have been reduced below 1.135 by any program which evaluated Z(t) at all Gram points. We also note that about 98 percent of the running time of the LR & W - program was spent on evaluating Z(t). This program was executed on a CDC CYBER 175 computer and ran about ten times as fast as the UNIVAC 1100/42 program of Brent. This is roughly what could be expected, given the relative speeds of the different machines.

2. THE STRATEGY FOR FINDING SUFFICIENTLY MANY SIGN CHANGES OF Z(t)

We recall some definitions. Let $\theta(t)$ be the real continuous function defined by

$$(2.1) \qquad \theta(t) = \arg[\pi^{-\frac{1}{2}it}\Gamma(\tfrac{1}{4}+\tfrac{1}{2}\,it)], \quad \theta(0) = 0.$$

The j-th Gram point g_j is defined as the unique number satisfying $\theta(g_j) = j\pi$ $(j = -1,0,1,2,\ldots)$. A Gram point g_j is called good if $(-1)^j Z(g_j) > 0$ and bad otherwise. A Gram block of length L (≥ 1) is an interval $B_j = [g_j, g_{j+L})$ such that g_j and g_{j+L} are good Gram points and $g_{j+1},\ldots,g_{j+L-1}$ are bad Gram points. An interval $[g_j, g_{j+1})$ is called a Gram interval. A Gram block B_j of length L is said to satisfy "Rosser's rule" if Z(t) has at least L zeros in B_j.

The strategy of Brent for finding the required number of sign changes of Z(t) is based on this rule. LR & W refined this strategy in order to reduce the number of Z-evaluations as much as they could. This will be described here in some detail.

In order to reduce the number of Z-evaluations as much as possible, we first observe that after having determined a Gram block B_j of length $L \geq 2$, we already have implicitly detected L - 2 sign changes of Z(t). Hence, the problem reduces to finding the "missing two" sign changes. Next we observe that these missing two (if existing) must both lie in one and the same Gram interval of the block B_j. Some preliminary experiments with the LR & W-program revealed that in the majority of cases the missing two are situated in one of the *outer* Gram intervals of B_j. Therefore, we first search in (g_j, g_{j+1})

or (g_{j+L-1}, g_{j+L}) according to which of $\mathrm{abs}(Z(g_j) + Z(g_{j+1}))$ and $\mathrm{abs}(Z(g_{j+L-1}) + Z(g_{j+L}))$ is the smallest. In the selected interval an efficient parabolic interpolation search routine is invoked. (Here is the main improvement over Brent's method, which used random search rather than parabolic interpolation.) If this routine terminates without having found the missing two sign changes, the other outer Gram interval of the block is treated in the same manner. In case the missing two are still not found, another search routine is called, depending on the length L of the block $B_j = [g_j, g_{j+L})$.

If L = 2, the interval (g_j, g_{j+2}) is scanned again, and if L > 2 we continue to search in the interval (g_{j+1}, g_{j+L-1}). In both cases, the search is performed by means of a refinement of a search routine described by LEHMAN [9]. For more details we refer the reader to the source text of the LR & W - program in [12].

If at some instant one of the search routines has detected the missing two, a new Gram block is set up and we continue as described above. In the opposite case (which occurs very rarely) the program prints a message and a "plot" of Z(t) corresponding to the whole Gram block under investigation and proceeds by pretending (!) that the missing two were found indeed. These plots of Z(t) were inspected afterwards (if necessary) "by hand". So far, the missing two were always easily found either in the Gram block under consideration or in an adjacent Gram block (compare BRENT [2, Section 4]).

After having covered the range $[g_{156,800,000}, g_{200,000,000})$ we ran the computation a little further, and found 4 Gram blocks in $[g_{200,000,000}, g_{200,000,004})$, all of them satisfying Rosser's rule. By applying Theorem 3.2 of BRENT [2] we completed the proof of our claim that the first $n = 200,000,001$ zeros of $\zeta(s)$ are simple and lie on $\sigma = \frac{1}{2}$.

3. COMPUTATION OF Z(t) AND ERROR ANALYSIS

3.0. Introduction

In principle, Brent and LR & W's methods of computing Z(t) and error analysis are exactly as described in Section 5 of BRENT [2]. We will only mention here some details of LR & W's computations and error analysis. The full details are given in [12].

The unambiguous determination of the sign of Z(t) requires a rigorous bound for the error, committed when one actually computes Z(t) on a computer.

In our program we actually used two methods (A and B) for evaluating $Z(t)$.

Method A is a very fast and efficient method which usually gives the correct sign of $Z(t)$.

Method B is a comparitively slow, but very accurate method which is invoked when $|Z(t)|$ is too small for method A.

We used the well-known Riemann-Siegel formula (with two correction terms in either case):

$$(3.1)\qquad Z(t) = 2\sum_{k=1}^{m} k^{-\frac{1}{2}}\cos[t.\ell n(k) - \theta(t)] + (-1)^{m-1}\tau^{-\frac{1}{4}}\sum_{j=0}^{l}\Phi_j(z)\tau^{-j/2} + R_l(t),$$

where $m = \lfloor\tau^{\frac{1}{2}}\rfloor$, $\tau = t/(2\pi)$, $z = 1 - 2(\tau^{\frac{1}{2}} - m)$,

$$(3.2)\qquad \theta(t) = \arg[\pi^{-\frac{1}{2}it}\Gamma(\tfrac{1}{4}+\tfrac{1}{2}\,it)], \quad \theta(t) \text{ continuous and } \theta(0) = 0,$$

$$(3.3)\qquad \Phi_0(z) = \cos[\pi(4z^2+3)/8]/\cos(\pi z) =: \sum_{k=0}^{\infty} c_{2k}^{(0)} z^{2k}, \quad -1 < z \le 1,$$

$$(3.4)\qquad \Phi_1(z) = \frac{d^3}{dz^3}\,\Phi_0(z)/(12\pi^2) =: \sum_{k=0}^{\infty} c_{2k+1}^{(1)} z^{2k+1}.$$

The last term $R_1(t)$ will be dropped in our actual computation. GABCKE [5] and BRENT & SCHOENFELD [4] have given bounds on $R_n(t)$ (here, n+1 denotes the number of terms in the second sum in (3.1), hence n = 1 in our situation). We used the bound (GABCKE [5])

$$(3.5)\qquad |R_1(t)| < 0.053t^{-5/4} < 0.0054\tau^{-5/4}, \quad \text{for } t \ge 200.$$

The floating point machine approximations of Z by means of methods A and B will be denoted by $\tilde{Z}_A$ and $\tilde{Z}_B$, respectively. Throughout this section, the result of the floating point machine approximation of some quantity q will be denoted by $\tilde{q}$.

We present an error analysis which accounts for *all* possible errors in $\tilde{Z}$, for any t (resp. τ) in the range,

$$(3.6)\qquad 3.5\times 10^7 < t < 3.72\times 10^8 \quad (\text{resp. } 5.5\times 10^6 < \tau < 5.92\times 10^7).$$

This covers the range of zero #81,000,001 till zero #1,000,000,000 of $\zeta(s)$ in the critical strip, which we had originally planned to investigate ($\gamma_{81,000,001} \approx 35,018,261.166$, $\gamma_{1,000,000,000} \approx 371,870,203.837$).

The computations were carried out on a CDC CYBER 175 computer having a 60-bit word, and single-precision (SP) and double-precision (DP) floating-point arithmetic using 48- and 96- bit binary fractions, respectively. In the sequel we will frequently work with the unit roundoffs $\varepsilon_s := 2^{-47}$ and $\varepsilon_d := 2^{-95}$.

3.1. Computation of Z(t)

At the start of the program four tables are precomputed:

(i) $\ell n(k)$ for $1 \le k \le m_0$ in DP, where m_0 is large enough to cover the range of the current job;

(ii) $k^{-\frac{1}{2}}$ for $1 \le k \le m_0$ in DP, truncated to SP;

(iii) $\cos(2\pi k.2^{-13})$ for $0 \le k \le 2^{13}+1$ in DP, truncated to SP;

(iv) $\cos(2\pi(k+1)2^{-13}) - \cos(2\pi k2^{-13})$ for $0 \le k \le 2^{13}$ in DP, truncated to SP.

Methods A and B run essentially as follows.

Method A. Given a τ as a DP floating point number, $t = 2\pi\tau$ and $\theta(t)$ are computed in DP; $f^{(1)} := \mathrm{frac}\{\theta(t)(2\pi)^{-1}\}$ is computed in DP, and truncated to SP. Next, the main loop (corresponding to the first sum in (3.1)) is executed. This loop has been written in COMPASS (machine language of the CYBER) and optimized using the specific properties of the CYBER's central processing units. One cycle of the loop executes in about 2.1 μ sec. $f^{(2)} := \mathrm{frac}\{\tau\,\ell n(k)\}$ (where $\ell n(k)$ is looked up in the precomputed table) is computed as follows: the DP product of τ and $\ell n(k)$ is decreased with the integer part of the SP product of τ and $\ell n(k)$ and the result is truncated to SP. This programming "trick" saves a considerable amount of time in the main loop. $x = \mathrm{abs}(f^{(1)}-f^{(2)})$ is computed in SP, and $\cos(2\pi x)$ is approximated by linear interpolation in the precomputed cosine-table, using the precomputed cosine-difference table. The result is multiplied by the precomputed $k^{-\frac{1}{2}}$ and the product is accumulated in an SP sum. End of the main loop. Next, the two terms in the asymptotic expansion of the Riemann-Siegel formula (3.1) are approximated using the truncated Taylor series expansions

$$\Phi_0(z) \cong \sum_{k=0}^{N_0} c_{2k}^{(0)} z^{2k} \quad \text{and} \quad \Phi_1(z) \cong \sum_{k=0}^{N_1} c_{2k+1}^{(1)} z^{2k+1}. \tag{3.7}$$

For the values of N_0 and N_1 actually used, see Section 3.4. The total correction is computed and added to 2 times the sum obtained in the main loop. The computations after the main loop are carried out in SP.

Method B. The same as Method A, with *all* computations in DP. The value of $\cos(2\pi x)$ is computed using the available standard FORTRAN DP library function DCOS.

3.2. Error analysis

In our error analysis we assume that τ is exactly representable as a floating point number. The positive integer m ($= \lfloor \tau^{\frac{1}{2}} \rfloor$) is *exactly* computed from τ by testing the inequalities $m^2 \le \tau < (m+1)^2$ and by correcting the machine-computed value, if necessary. Now let

$$s(t) := 2 \sum_{k=1}^{m} k^{-\frac{1}{2}} \cos[t.\ell n(k) - \theta(t)], \quad (t = 2\pi\tau). \tag{3.8}$$

By $\tilde{s}(\tilde{t})$ we denote the computed value of $s(t)$, where errors may be made in the computation of t, $\ell n(k)$, $\theta(t)$, $t.\ell n(k) - \theta(t)$, $\cos(.)$, $k^{-\frac{1}{2}}$ and the final inner product. The following lemma accounts for *all* these errors.

LEMMA 3.1. *Suppose that* $|t-\tilde{t}| \le \delta_0 t$, $|\ell n(k) - \tilde{\ell n}(k)| \le \delta_1 \ell n(k)$ *for* $k = 1,2,\ldots,m$, *and* $|\theta(u) - \tilde{\theta}(u)| \le \delta_2 \theta(u)$; *let* $f_k := \mathrm{frac}\{\tau \tilde{\ell n}(k) - \tilde{\theta}(\tilde{t})(2\pi)^{-1}\}$ *and suppose that* $|f_k - \tilde{f}_k| \le \delta_3$ *for* $k = 1,2,\ldots,m$. *Moreover, suppose that* $|\cos(x) - \tilde{\tilde{\cos}}(x)| \le \delta_4$ *for* $0 \le x \le 2\pi+h$, *where* h *is fixed**), $|k^{-\frac{1}{2}} - \tilde{k}^{-\frac{1}{2}}| \le \delta_5 k^{-\frac{1}{2}}$ *for* $k = 1,2,\ldots,m$, *and that the inner product of the two vectors with components* $\tilde{k}^{-\frac{1}{2}}$ *and* $\tilde{\tilde{\cos}}(2\pi\tilde{f}_k)$, *respectively,* $(1 \le k \le m)$ *is computed with a relative error in the basic arithmetic operations* (+, −, * *and* /) *bounded by* ε. *Then we have*

$$\begin{aligned} |s(t) - \tilde{s}(\tilde{t})| \le\ & 4\pi\tau^{5/4}\ell n(\tau)[2\delta_0 + \delta_1(1+\delta_0) + \delta_2] + \\ & + 4\tau^{\frac{1}{4}}[2\pi\delta_3 + \delta_4 + (1+\delta_4)\{\delta_5 + (1+\delta_5)((1+\varepsilon)^m - 1)\}]. \end{aligned} \tag{3.9}$$

This lemma is similar to Lemma 5.3 of BRENT [2], the difference being that we explicitly account for *all* possible errors in the computation of s(t). The proof is routine and uses the technique of backward error analysis (cf. WILKINSON [18]) for the inner product computation (cf. PARLETT [14, pp. 30-32]) and for the other basic arithmetic operations.

Let

$$\chi(\tau) := (-1)^{m-1}\tau^{-\frac{1}{4}}[\Phi_0(z) + \tau^{-\frac{1}{2}}\Phi_1(z)]. \tag{3.10}$$

---

*) The reason for the occurrence of this (small) number h in this lemma will be clarified in Section 3.3.

By $\tilde{\chi}(\tau)$ we denote the computed value of $\chi(\tau)$ where errors may be made in the computation of $\tau^{-\frac{1}{2}}$, $\tau^{-\frac{1}{4}}$, z, $\Phi_0(z)$, $\Phi_1(z)$, and in the other arithmetic operations. The following lemma accounts for *all* these errors.

LEMMA 3.2. *Let ε be as in Lemma* 3.1 *and let the relative error in the square root computation be bounded by* $a\varepsilon$. *Moreover, suppose that* $|z-\tilde{z}| \le \delta_6$ *and that* $\Phi_0(z)$ *and* $\Phi_1(z)$ *are approximated by* $\tilde{\Phi}_0(z) := \Sigma_{k=0}^{N_0} \widetilde{c_{2k}^{(0)}}\, z^{2k}$ *and* $\tilde{\Phi}_1(z) := \Sigma_{k=0}^{N_1} \widetilde{c_{2k+1}^{(1)}}\, z^{2k+1}$, *respectively, where* $|c_{2k}^{(0)} - \widetilde{c_{2k}^{(0)}}| \le \delta_7$ *and* $|c_{2k+1}^{(1)} - \widetilde{c_{2k+1}^{(1)}}| \le \delta_7$. *Then*

$$|\chi(\tau) - \tilde{\chi}(\tau)| \le \tau^{-\frac{1}{4}}\left[2\delta_6 + 2\delta_7(N_0+1) + \frac{1}{(N_0+1)!}\left(\frac{\pi}{2}\right)^{N_0+1} + (5N_0+2a+4)\varepsilon\right] +$$

$$+ \tau^{-3/4}\left[\frac{1}{4}\delta_6 + 2\delta_7(N_1+1) + \frac{1}{6}\,\frac{N_1+5/2}{(N_1+1)!}\left(\frac{\pi}{2}\right)^{N_1+1} + (5N_1+3a+7)\varepsilon\right]. \tag{3.11}$$

In the proof of this lemma, which we omit, use is made of the inequalities $|\Phi_0(z)| \le 1$, $|\Phi_1(z)| \le 1$, $|\Phi_0'(z)| \le 1$ and $|\Phi_1'(z)| \le \frac{1}{4}$ for $|z| \le 1$ (see GABCKE [5, Theorem 1, p. 60]) and of the bounds given in GABCKE [5, Theorem 2, p. 62] on the error induced by truncating the infinite series in (3.3) and (3.4).

3.3. Estimates for $\delta_0,\ldots,\delta_7$ for methods A and B

Because of the programming "trick" mentioned in 3.1 we must take into account the possibility that the *computed* value of $f^{(2)}$ may be (slightly) larger than 1 by an amount which is bounded by $2.5\varepsilon_s\tau\tilde{\ell}n(k)$. In the t-range (3.6) this excess is bounded by 10^{-5}. Instead of correcting $f^{(2)}$ by subtracting 1, which is needed only very rarely, we use *one* extra element in the cosine interpolation table beyond $\cos(2\pi)$, viz. $\cos(2\pi+h)$, where $h = 2\pi.2^{-13} \approx$ $\approx 7.7\times10^{-4}$ $(> 10^{-5})$.

In [12] we have given an account of our computation of the values of $\delta_0,\ldots,\delta_7$. The results are summarized in Table 3.1.

Table 3.1.

Values of $\delta_0,\ldots,\delta_7$ for methods A and B

method	δ_0	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7
A	1.01×10^{-28}	5.1×10^{-29}	3.6×10^{-27}	5×10^{-14}	7.36×10^{-8}	7.2×10^{-15}	7.2×10^{-15}	5×10^{-14}
B	1.01×10^{-28}	5.1×10^{-29}	3.6×10^{-27}	1.2×10^{-17}	1.5×10^{-27}	1.01×10^{-28}	2.0×10^{-24}	5×10^{-28}

3.4. The error bounds on $\tilde{Z}$ for methods A and B

To complete the error analysis we apply Lemmas 3.1 and 3.2 with $\delta_0, \ldots, \delta_7$ as given in Table 3.1, $\varepsilon = \varepsilon_s = 2^{-47}$, $a = 10$, $N_0 = 16$ and $N_1 = 17$ for method A, and $\varepsilon = \varepsilon_d = 2^{-95}$, $a = 10$, $N_0 = N_1 = 29$ for method B; including the inherent error (3.5) this yields

$$|Z(t) - \tilde{Z}_A(\tilde{t})| \leq 3 \times 10^{-7} \tau^{1/4} \tag{3.13}$$

and

$$|Z(t) - \tilde{Z}_B(\tilde{t})| \leq 5.4 \times 10^{-3} \tau^{-5/4} + 3.1 \times 10^{-16} \tau^{1/4} + 4.1 \times 10^{-24} \tau^{-1/4} + 5 \times 10^{-26} \tau^{5/4} \ell n(\tau), \tag{3.14}$$

for any t $(= 2\pi\tau)$ in the interval $(3.5 \times 10^7, 3.72 \times 10^8)$. In this interval, safe upper bounds for the errors are 2.7×10^{-5} and 2.0×10^{-11}, respectively. In the LR & W-program (see [12]) the extremely conservative *fixed* bounds $\varepsilon_1 = 10^{-4}$ and $\varepsilon_2 = 2.5 \times 10^{-6}$ were used, respectively. In case $|\tilde{Z}_A(\tilde{t})|$ was less than ε_1, a few rather small shifts with $\tilde{t}$ were tried. If still no "clear" value was found with method A, method B was invoked. Until now not a single t occurred for which method B could not determine the sign of $Z(t)$ rigorously.

4. STATISTICS

The LR & W-program was organized in such a way that in case the value of $Z(t)$, obtained with method A, was too small for a rigorous sign determination, a few small shifts of the argument were tried before method B was invoked. Therefore, the LR & W-program uses, in relatively few cases, an approximation of the Gram point g_j instead of g_j itself. (In a run of 2,500,000 zeros, with error bound 10^{-4} for method A, the *total* number of shifts was always less than 370. Most of them were made when separating the zeros *inside* the Gram blocks. Only a few of them were made in Gram points. Also see the text introducing Table 4.3.) Consequently, the statistics found by LR & W cannot, strictly speaking, be accumulated to those, found by Brent. Nevertheless, just for convenience, we have put together all results. This should be kept in mind when reading the tables.

In Table 4.1 we present a list of 104 exceptions to Rosser's rule up to $g_{200,000,000}$ found by Brent and LR & W, including the 15 exceptions up to $g_{75,000,000}$ from [2], for completeness. Moreover, the types (see Table 4.2) are given in parentheses, followed by the local extreme values of S(t) (see BRENT [2]) near B_n. It is possible that for n ≥ 156,800,000 the LR & W-program has not detected *all* exceptions to Rosser's rule, due to

Table 4.1

(extension of Table 3 of BRENT [2])

104 exceptions to Rosser's rule up to $g_{200,000,000}$

Notation: n (type) extreme S(t)

where n is the index of the Gram block B_n containing no zeros.

13,999,525(1)	-2.0041	100,788,444(1)	-2.0230	146,130,246(2)	2.0005	173,737,614(2)	2.0221
30,783,329(1)	-2.0026	106,236,172(1)	-2.0184	147,059,770(1)	-2.0498	174,102,513(1)	-2.0180
30,930,927(2)	2.0506	106,941,328(2)	2.1559	147,896,100(2)	2.0391	174,284,990(1)	-2.0181
37,592,215(1)	-2.0764	107,287,955(1)	-2.0786	151,097,113(1)	-2.0043	174,500,513(1)	-2.0125
40,870,156(1)	-2.0038	107,532,017(2)	2.0728	152,539,438(1)	-2.0026	175,710,609(1)	-2.0193
43,628,107(1)	-2.0242	110,571,044(1)	-2.0458	152,863,169(2)	2.0459	176,870,844(2)	2.0125
46,082,042(1)	-2.0311	111,885,254(2)	2.0247	153,522,727(2)	2.0027	177,332,733(2)	2.0146
46,875,667(1)	-2.0046	113,239,783(1)	-2.0306	155,171,525(2)	2.0437	177,902,862(2)	2.0223
49,624,541(2)	2.0018	120,159,903(1)	-2.0589	155,366,607(1)	-2.0277	179,979,095(1)	-2.0182
50,799,238(1)	-2.0288	121,424,392(2)	2.0515	157,260,687(2)	2.0363	181,233,727(2)	2.1018
55,221,454(2)	2.0242	121,692,932(2)	2.0616	157,269,224(1)	-2.0329	181,625,435(1)	-2.0401
56,948,780(2)	2.0177	121,934,171(2)	2.1719	157,755,123(1)	-2.0205	182,105,257(6)	2.0084
60,515,663(1)	-2.0081	122,612,849(2)	2.0072	158,298,485(2)	2.0273	182,223,560(2)	2.0156
61,331,766(3)	-2.0543	126,116,567(1)	-2.0106	160,369,051(2)	2.0071	191,116,405(2)	2.0195
69,784,844(2)	2.0637	127,936,513(1)	-2.1105	162,962,787(1)	-2.0115	191,165,600(2)	2.0283
75,052,114(1)	-2.0045	128,710,278(2)	2.0444	163,724,709(1)	-2.0163	191,297,535(5)	-2.1490
79,545,241(2)	2.0113	129,398,903(2)	2.0431	164,198,114(2)	2.0235	192,485,616(1)	-2.0416
79,652,248(2)	2.0066	130,461,097(2)	2.0963	164,689,301(1)	-2.1579	193,264,636(6)	2.0055
83,088,043(1)	-2.1328	131,331,948(2)	2.0047	164,880,229(2)	2.0308	194,696,968(1)	-2.0664
83,689,523(2)	2.0775	137,334,072(2)	2.0239	166,201,932(1)	-2.0024	195,876,805(1)	-2.0143
85,348,958(1)	-2.0095	137,832,603(1)	-2.0134	168,573,836(1)	-2.0159	195,916,549(2)	2.0546
86,513,820(1)	-2.0154	138,799,472(2)	2.0135	169,750,763(1)	-2.1036	196,395,161(2)	2.0326
87,947,597(2)	2.0523	139,027,791(1)	-2.0031	170,375,507(1)	-2.0009	196,676,303(1)	-2.0135
88,600,095(1)	-2.1394	141,617,806(1)	-2.1253	170,704,880(2)	2.0249	197,889,883(2)	2.0034
93,681,183(1)	-2.0165	144,454,931(1)	-2.0380	172,000,993(2)	2.0608	198,014,122(1)	-2.0333
100,316,552(2)	2.0233	145,402,380(2)	2.0012	173,289,941(1)	-2.0378	199,235,289(1)	-2.0205

possible shifts in Gram points. For instance, an exception of type 2 (see Table 4.2) may have been detected as a Gram block of length 3 with "2 1 0" zero-pattern. It may be noted, however, that in the range $[g_{81,000,000}, g_{120,000,000})$ LR & W have found exactly the same exceptions to Rosser's rule as Brent.

In addition to the types 1, 2 and 3 introduced by BRENT [2] we have defined the types 4, 5 and 6, the meaning of which should be clear from Table 4.2. This table also gives the frequencies of the occurrences of the various types in $[g_{-1}, g_{200,000,000})$. Note that an exception of type 4 has not yet been found, so that at the time of writing we still know only one Gram interval with four zeros, viz. $G_{61,331,768}$, found by BRENT [2].

Table 4.2

Various types of exceptions to Rosser's rule and their frequencies in $[g_{-1}, g_{200,000,000})$.

Gram block of length 2 without any zeros									type	frequency
g_{n-2}	g_{n-1}	g_n	g_{n+1}	g_{n+2}	g_{n+3}	g_{n+4}				
		0	0	3					1	53
	3	0	0						2	47
		0	0	4	0				3	1
0	4	0	0						4	0
		0	0	2	2				5	1
2	2	0	0						6	2

Very recently, KARKOSCHKA and WERNER [8] have developed a method for detecting exceptions to Rosser's rule with relatively small computational effort, i.e., by searching in certain selected small ranges of a given t-interval. A comparison of their results with Table 4.1 shows the power of their method: in $[g_{3,500,000}, g_{50,000,000})$ they found all 9 exceptions to Rosser's rule, and in $[g_{100,000,000}, g_{120,000,000})$ they found 6 of the 9 exceptions.

Table 4.3 is a continuation of Table 1 of BRENT [2]. Six Gram blocks of length 8 were found. The average block length up to n = 200,000,000 is 1.1951. We have compared the results of LR & W with those of Brent in the range $[g_{110,000,000}, g_{120,000,002})$. Brent's program counted 7,011,482 Gram blocks of length 1, 1,055,511 of length 2 and 230,234 of length 3. The

Table 4.5 continues Table 4 of BRENT [2]. As yet, no Gram block of type (7,1) was found. Due to the shifts, we may have missed earlier occurrences of blocks of types (7,7), (8,3) and (8,7), although we consider this unlikely.

Table 4.5

(continuation of Table 4 of BRENT [2])

First occurrences of Gram blocks of various types

j	k	n
7	7	195,610,937 (LR & W)
8	2	112,154,948 (BRENT)
8	3	175,330,804 (LR & W)
8	6	145,659,810 (BRENT)
8	7	165,152,519 (LR & W)

In Table 4.6 we list the number of Gram blocks of type (j,k), $1 \le j \le 8$, $1 \le k \le j$, in the interval $[g_{156,800,000}, g_{200,000,000})$, as they were actually counted by the LR & W-program. On the line with $j = 2$ we also mention the numbers of Gram blocks of length 2 with zero-pattern "0 0" and those with pattern "2 2" which could, of course, neither be classified as type (2,1) nor as (2,2). The 43 blocks with "0 0"-pattern correspond to the exceptions to Rosser's rule in $[g_{156,800,000}, g_{200,000,000})$ and the 3 blocks with "2 2"-pattern correspond to the exceptions of types 5 and 6 (cf. Table 4.2). The entries in parentheses give the approximate percentages with respect to the total number of blocks of length j, given in the final column.

Our main purpose of presenting this table is to render support to the LR & W-strategy of dealing with Gram blocks of length $j \ge 2$. The table shows that this strategy is successful for $2 \le j \le 5$. However, for $j \ge 6$ the missing two zeros in B_n show an increasing tendency to lie either in (g_{n+1}, g_{n+2}) or in (g_{n+j-2}, g_{n+j-1}). Only one of the 93 blocks of length $j = 7$ has its missing two zeros in one of the outer Gram intervals!

Table 4.6

Number of Gram blocks of type (j,k), $1 \le j \le 8$, $1 \le k \le j$, *in the interval* $[g_{156,800,000}, g_{200,000,000})$

↓j \ k→ 1	2	3	4	5	6 - 7	8	total
1 30,162,315							30,162,315
2 2,279,942 (50)	2,281,053 (50)	43 blocks with 0 0 zero-pattern 3 blocks with 2 2 zero-pattern					4,561,041
3 479,720 (47)	53,497 (5)	480,613 (47)					1,013,830
4 87,367 (46)	8,592 (4)	8,499 (4)	87,150 (45)				191,608
5 7,581 (38)	1,811 (9)	948 (5)	1,882 (9)	7,678 (39)			19,900
6 156 (12)	337 (27)	119 (10)	126 (10)	366 (29)	147 (12)		1,251
7 0	29	17	3	17	26	1	93
8 0	0	1*)	0	0	1*)	1*) 0	3

*) viz. B_n, for n = 175,330,804, 181,390,731 and 165,152,519.

REFERENCES

[1] BACKLUND, R., *Sur les zéros de la fonction* ζ(s) *de Riemann*, C.R. Acad. Sci. Paris, 158 (1914), pp. 1979-1982.

[2] BRENT, R.P., *On the zeros of the Riemann zeta function in the critical strip*, Math. Comp., 33 (1979), pp. 1361-1372.

[3] BRENT, R.P. & L. SCHOENFELD, *Numerical approximation of the Riemann zeta function*, Technical Report, Dept. of Computer Science, The Australian National University, to appear.

[4] EDWARDS, H.M., *Riemann's zeta function*, Academic Press, New York, 1974.

[5] GABCKE, W., *Neue Herleitung und explizite Restabschätzung der Riemann-Siegel-Formel*, Dissertation, Universität Göttingen, 1979.

[6] GRAM, J., *Sur les zéros de la fonction* ζ(s) *de Riemann*, Acta Math., 27 (1903), pp. 289-304.

[7] HUTCHINSON, J.I., *On the roots of the Riemann zeta-function*, Trans. Amer. Math. Soc., 27 (1925), pp. 49-60.

[8] KARKOSCHKA, E. & P. WERNER, *Einige Ausnahmen zur Rosserschen Regel in der Theorie der Riemannschen Zetafunktion*, Computing, 27 (1981), pp. 57-69.

[9] LEHMAN, R.S., *Separation of zeros of the Riemann zeta-function*, Math. Comp., 20 (1966), pp. 523-541.

[10] LEHMER, D.H., *On the roots of the Riemann zeta function*, Acta Math., 95 (1956), pp. 291-298.

[11] LEHMER, D.H., *Extended computation of the Riemann zeta function*, Mathematika, 3 (1956), pp. 102-108.

[12] LUNE, J. VAN DE, H.J.J. TE RIELE & D.T. WINTER, *Rigorous high speed separation of zeros of Riemann's zeta function*, Report NW 113/81, October 1981, Mathematical Centre, Amsterdam.

[13] MELLER, N.A., *Computations connected with the check of Riemann's hypothesis*, Doklad. Akad. Nauk SSSR, 123 (1958), pp. 246-248 (Russian).

[14] PARLETT, B.N., *The symmetric eigenvalue problem*, Prentice-Hall, 1980.

[15] ROSSER, J.B., J.M. YOHE & L. SCHOENFELD, *Rigorous computation and the zeros of the Riemann zeta-function*, Proc. IFIP Congress, Edinburgh, 1968.

[16] TITCHMARSH, E.C., *The zeros of the Riemann zeta function*, Proc. Roy. Soc. London, 151 (1935), pp. 234-255, 157 (1936), pp. 261-263.

[17] TITCHMARSH, E.C., *The theory of the Riemann zeta-function*, Oxford, Clarendon Press, 1951.

[18] WILKINSON, J.H., *Rounding errors in algebraic processes*, Prentice-Hall, 1963.

Added in Proof. In the meanwhile VAN DE LUNE & TE RIELE have extended the computations so far that we can now (December 1982) say that the first 307 000 000 non-trivial zeros of $\zeta(s)$ are all simple and lie on $\sigma = \frac{1}{2}$.