MATHEMATICAL CENTRE TRACTS 145

MULTI-DIMENSIONAL CONTINUED FRACTION ALGORITHMS

A.J. BRENTJES

MATHEMATISCH CENTRUM AMSTERDAM 1981

1980 Mathematics subject classification: 10F20

ISBN 90 6196 231 5

ACKNOWLEDGEMENTS

The research which led to the results of this book, was done at the Mathematical Institute of the State University of Leiden and supported by a three-year grant from the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

I feel greatly indebted to prof. dr. R. Tijdeman for his inspiration and encouragement during this period. His continuing interest in my work cannot be overestimated and has left its marks throughout the book. I am also very grateful to prof. dr. P.L. Cijsouw and dr. F. Beukers, whose valuable remarks led to important improvements in the text and the inclusion of the appendix. I owe much to drs. W. Pelt for his advice concerning the English language.

I thank the Mathematical Centre for the opportunity to publish this monograph in their series Mathematical Centre Tracts and all those at the Mathematical Centre who have contributed to its technical realization.

CONTENTS

LIST OF SYMBOLS AND NOTATIONS 3

INTRODUCTION 5

CHAPTER 1. The continued fraction algorithm

1A. The algorithm 9

1B. Convergents 10

1C. A geometrical interpretation 12

1D. Quadratic irrationals 14

CHAPTER 2. Multi-dimensional continued fraction algorithms

2A. Notations and preliminaries 16

2B. Multi-dimensional continued fraction algorithms 18

2C. Convergence 21

2D. The linear Diophantine equation 22

2E. Outline 24

CHAPTER 3. Vectorial algorithms

3A. Introduction 25

3B. Division or subtraction? 26

3C. Periodicity of vectorial algorithms 27

3D. The Jacobi-Perron algorithm 30

3E. Convergence of the J.P.A. 33

3F. Periodicity of the J.P.A. 35

3G. Variations of the J.P.A. 37

3H. Brun's algorithm 40

3I. A non-convergent 2-dimensional Brun expansion 44

3J. Some other algorithms 49

CHAPTER 4. Introduction to some approximation problems

4A. Linear dependence as an approximation problem 52

4B. Best approximations 53

4C. Periodicity of best approximations 57

4D. Good approximations 62

4E. Simultaneous Diophantine approximation 63

CHAPTER 5. The 2-dimensional case

5A. The computation of best approximations 66

5B. Preliminary results 67

5C. The 2-dimensional best approximation algorithm. 71

5D. Further discussion of the algorithm and an example 85

5E. Other convergent algorithms 94

5F. Application: Units in cubic number fields 105

CHAPTER 6. The multi-dimensional case.

6A. The Szekeres and inner product algorithms 117

6B. Convergent algorithms by induction 123

CHAPTER 7. An algorithm for relative minima

7A. Relative minima 130

7B. Notations and easy cases 131

7C. A continued fraction algorithm for relative minima 138

7D. Application: Units in cubic number fields 148

APPENDIX. On computer implementation 155

REFERENCES 165

REGISTER 181

LIST OF SYMBOLS AND NOTATIONS, OTHER THAN THOSE OF LOCAL USE

$n+1$	the dimension
$\mathbb{R}$, $\mathbb{Q}$	the real and rational numbers
$\mathbb{Z}$, $\mathbb{N}$	the rational integers and natural numbers
X, Y	points of $\mathbb{R}^{n+1}$
$X.Y$	inner product of X and Y
$\measuredangle(X,Y)$	angle between vectors X and Y
$\lvert X\rvert = \sqrt{X.X}$	Euclidean length of X
O	the origin in $\mathbb{R}^{n+1}$
ℓ	a line through O
ℓ^*	the orthoplement of ℓ
Ω	a lattice of rank $n+1$ in $\mathbb{R}^{n+1}$
$\mathcal{A}$	a lattice base, consisting of $A_0, A_1, \ldots\ldots, A_n$
$a = (a_0, \ldots\ldots, a_n)$	cofactors of $\mathcal{A} = \{A_0, \ldots\ldots, A_n\}$ (p. 16)
B, C,.., P, Q,	points of $\mathbb{R}^{n+1}$, usually lattice points
$\underline{A}$, $\underline{P}$, etc.	projection, parallel to ℓ, of A, P, etc. on ℓ^*
$d(X,\mathcal{W})$	Euclidean distance between point X and set $\mathcal{W}$
$\{\mathcal{A}\}^+$	$= \{x_0A_0 + \ldots\ldots + x_nA_n \mid x_k \in \mathbb{R},\ x_k \geq 0\ (k=0,1,\ldots\ldots,n)\}$, the "first quadrant" of $\mathcal{A}$ (p. 18)
$s(i)$, $t(i)$	indices involved in i-th step (p. 19)
$b(i)$	size of i-th step (p. 19)
I^b_{st}	step matrix (p. 19)
Δ	basic volume of lattice (p. 54)
h	height function (p. 54)
$h(\mathcal{A}) > 0$	short for $h(A_k) > 0$ for $k = 0,1,\ldots\ldots,n$
ρ	radius function (p. 54)
$c_{\rho h}$	volume constant for ρ and h
Ia-d, II	step types of 2-dimensional best approximation algorithm (p. 72)
$(R)_{x_j}$	x_j-successor of relative minimum R (p. 130)
$\Sigma(A)$	the region $\lvert y\rvert < \lvert y_A\rvert$, $\lvert z\rvert < \lvert z_A\rvert$ if $A = (x_A, y_A, z_A)$ (p. 132)
$\mathcal{O}(\theta)$	ring of integers in an algebraic number field $\mathbb{Q}(\theta)$
$\Omega(\theta)$	lattice embedding of $\mathcal{O}(\theta)$
$\mathcal{E}(\theta)$	unit group of $\mathcal{O}(\theta)$

INTRODUCTION

In this book we study multi-dimensional continued fraction algorithms and their application to Diophantine approximation problems. The approach is more of a geometric nature than in most of the related literature. As a concise introduction we consider in chapter 1 the ordinary continued fraction algorithm from this point of view.

In chapter 2 multi-dimensional continued fraction algorithms are introduced as follows.

Let Ω be a lattice of maximal rank in an (n+1)-dimensional real vector space R^{n+1}, and let ℓ be a line through the origin O. Along ℓ a non-zero vector ℓ_0 is given. We define the cofactors of a lattice base $A = \{A_0, \ldots\ldots, A_n\}$ to be the coordinates $a_0, \ldots\ldots, a_n$ of ℓ_0 with respect to this base, so that $\ell_0 = a_0 A_0 + \ldots\ldots + a_n A_n$. We are mainly interested in bases with non-negative cofactors, which we express by

$$\ell_0 \in \{A\}^+ := \{ \sum_{k=0}^{n} x_k A_k \mid x_0 \geq 0, \ldots\ldots, x_n \geq 0 \}.$$

When $A(0)$ is a base of Ω such that $\ell_0 \in \{A(0)\}^+$, an n-dimensional continued fraction expansion of $A(0)$ along ℓ is a sequence of bases $A(i)$, $i = 0,1,2,..$ such that for every $i \geq 0$

(1) $\ell_0 \in \{A(i)\}^+$;

(2) there exist indices $s = s_i$, $t = t_i$ $(s \neq t)$ and a natural number b_i such that

$$\begin{cases} A_t(i+1) = A_t(i) + b_i A_s(i), \text{ and} \\ A_k(i+1) = A_k(i) \text{ for } k \neq t. \end{cases}$$

Given that $\ell_0 \in \{A(i)\}^+$, the cofactor inequality $a_s(i) \geq b_i a_t(i)$ is a necessary and sufficient condition for $\ell_0 \in \{A(i+1)\}^+$. An n-dimensional continued fraction algorithm is an algorithm which determines for each step the indices s_i and t_i and then the number b_i in accordance with this condition. Therefore the algorithm can in general choose between $\binom{n+1}{2}$ pairs s,t for each step. It follows that there is only one 1-dimensional algorithm: the ordinary continued fraction algorithm.

A continued fraction expansion $A(i) = \{A_0(i),\ldots\ldots,A_n(i)\}$, $i = 0,1,2,\ldots$, is called convergent if for every given $\varepsilon > 0$ there is an i_0 such that

$$\max_{0\le k\le n} d(A_k(i_0),\ell) < \varepsilon,$$

where $d(X,\ell)$ is the distance between a point X and the line ℓ.

The main part of chapter 3 is a historical survey of vectorial algorithms: Those which base their choice of s_i and t_i only on the cofactors of $A(i)$. All continued fraction algorithms which have been proposed since the beginning (Jacobi, 1868) and up to about 1970 belong to this class. The best known algorithms are those of Jacobi and of Brun (1919). A great disadvantage is that the expansions of vectorial algorithms often converge too slowly or not at all. Therefore they are less well suited for the approximation problems which are the subject of chapter 4. This holds in particular for the problem of finding "good" or "best" approximations in the following sense. Let h be a linear function on $\mathbb{R}^{n+1}$ with $h(\ell_0) > 0$ (the height function) and ρ a function which measures the distance to ℓ in some norm (the radius function). Then a point $B \in \Omega$, $B \neq 0$, is called a best approximation with respect to ρ and h if no lattice point $P \neq 0$ satisfies

$$\begin{cases} |h(P)| < |h(B)| \text{ and } \rho(P) \le \rho(B) \\ \text{or } |h(P)| = |h(B)| \text{ and } \rho(P) < \rho(B). \end{cases}$$

A good approximation is a lattice point $B \neq 0$ for which the product $|h(B)(\rho(B))^n|$ is small (compare Minkowski's theorem on lattice points in convex symmetrical bodies).

Up to now no continued fraction algorithms were known which could find all best approximations to a given line, with the exception of the case $n = 1$. The main part of chapter 5 consists of the construction of such an algorithm for the special case $n = 2$ with the Euclidean radius function. For the choice of s_i and t_i this algorithm distinguishes between five cases, the conditions of which can easily be checked by a computer. We show that some simpler algorithms, though convergent, do not find all best approximations. As an application of the new algorithm we show how it can be used to calculate the fundamental unit of a cubic field of negative discriminant.

Chapter 6 deals with what scarce knowledge we have for higher

dimensions. For $n \geq 3$ it is still far out of reach to find all best approximations. However, in most applications one is chiefly interested in mere good approximations. To this end we have at our disposition a few algorithms which give good results in general, even when this is not yet guaranteed by theory. Among these algorithms we meet the first multi-dimensional continued fraction algorithm especially designed for good approximations, that of Szekeres (1970). We show that (a variant of) an algorithm of Ferguson and Forcade (1979) is convergent, but its inductive definition makes it less significant for application in practice.

Another approximation problem is the determination of relative minima of a lattice. With respect to a coordinate system $x_0, \ldots, x_n$, a point $R = (r_0, \ldots, r_n) \in \Omega$, $R \neq 0$, is called a relative minimum of Ω if the region $|x_j| < |r_j|$, $j = 0, \ldots, n$, contains no lattice point except for 0. In 1896 Voronoi gave a method, based on chains of relative minima in $\mathbb{R}^3$, for the determination of a set of fundamental units in a cubic field of positive discriminant. We show in chapter 7 that these chains of minima can be calculated by means of a 2-dimensional continued fraction algorithm.

For the sake of completeness, the list of references at the end of the book contains nearly all publications on multi-dimensional continued fraction algorithms which the author has been able to find.

CHAPTER 1. THE CONTINUED FRACTION ALGORITHM

In this chapter we recollect, as a brief introduction, some properties of the ordinary continued fraction algorithm. We give no proofs: These can be found in the excellent books of PERRON[2] and KHINTCHINE[2].

1A. The algorithm.

Let ξ_{-1}, ξ_0 be two positive real numbers. Repeated application of the division algorithm determines a unique sequence of integers $b_0, b_1, b_2, \ldots$ and a unique sequence of real numbers $\xi_0 > \xi_1 > \ldots \geq 0$ according to the following scheme:

$$
\begin{aligned}
\xi_{-1} &= b_0\xi_0 + \xi_1 \quad , \; 0 \leq \xi_1 < \xi_0 \\
\xi_0 &= b_1\xi_1 + \xi_2 \quad , \; 0 \leq \xi_2 < \xi_1 \\
&\\
\xi_{i-1} &= b_i\xi_i + \xi_{i+1}, \; 0 \leq \xi_{i+1} < \xi_i .
\end{aligned}
\tag{1.1}
$$

This represents the homogeneous form of the continued fraction expansion of the ratio $\xi_{-1} : \xi_0$. The partial denominators b_i remain unaltered when ξ_{-1}, ξ_0 are replaced by $\lambda\xi_{-1}, \lambda\xi_0$ for some $\lambda > 0$. Note that

$$b_i = \left[\frac{\xi_{i-1}}{\xi_i}\right],$$

whence $b_i \geq 1$ for $i \geq 1$. It is clear that the sequence of partial denominators is necessarily finite if, for some $m \geq 0$, $\xi_{m+1} = 0$; and this happens if and only if $\xi_{-1} : \xi_0$ is rational. The last equation of (1.1) then reads

$$\xi_{m-1} = b_m\xi_m$$

and implies $b_m \geq 2$, because $\xi_m < \xi_{m-1}$.

When $\xi_{-1} : \xi_0$ is irrational, (1.1) defines infinite sequences $b_0, b_1, \ldots$ and $\xi_1, \xi_2, \ldots$ We can also write (1.1) in inhomogeneous form:

$$\alpha_0 = b_0 + \frac{1}{\alpha_1}, \quad \alpha_1 > 1$$

$$(1.2) \qquad \alpha_1 = b_1 + \frac{1}{\alpha_2}, \quad \alpha_2 > 1$$

$$\alpha_i = b_i + \frac{1}{\alpha_{i+1}}, \quad \alpha_{i+1} > 1$$

with $\alpha_i = \frac{\xi_{i-1}}{\xi_i}$. The usual abbreviation of (1.1) and (1.2) is

$$(1.3) \qquad \alpha_0 = \frac{\xi_{-1}}{\xi_0} = [b_0; b_1, b_2, \ldots\ldots],$$

called the continued fraction of α_0. In the case of a finite expansion we write $\alpha_0 = [b_0; b_1, \ldots\ldots, b_m]$. The equality sign in (1.3) is justified by the uniqueness theorem:

<u>THEOREM 1.1.</u> *If an infinite sequence of integers $b_0, b_1, \ldots.$ satisfies $b_0 \geq 0$ and $b_i \geq 1$ when $i \geq 1$, then there is a unique irrational number α_0 whose partial denominators are precisely $b_0, b_1, \ldots.$ If a finite sequence of integers $b_0, b_1, \ldots., b_m$ satisfies $b_0 \geq 0$ and $b_i \geq 1$ when $i \geq 1$, then there is a unique rational number α_0 whose partial denominators are precisely $b_0, b_1, \ldots., b_m$ (if $b_m \geq 2$) or $b_0, b_1, \ldots., b_{m-2}, b_{m-1}+1$ (if $b_m = 1$).*

It is easy to calculate the (rational) value α_0 of a finite continued fraction, by working through (1.2) from below upwards, starting with $\alpha_m = b_m$. Elimination of α_m, α_{m-1}, ..., α_1 produces the familiar expression

$$(1.4) \qquad \alpha_0 = [b_0; b_1, \ldots\ldots, b_m] = b_0 + \cfrac{1}{b_1 + \cfrac{1}{\ddots + \cfrac{1}{b_{m-1} + \cfrac{1}{b_m}}}}$$

Though finite fractions ending with $b_m = 1$ cannot formally be obtained through (1.1), we shall take them into account with their meaning explained by (1.4).

<u>1B. Convergents.</u>

Let $\alpha = \alpha_0 = [b_0; b_1, \ldots\ldots]$ be a given continued fraction; if α is

rational let b_m be its last partial denominator. The k-th convergent to α is the rational number

$$(1.5) \qquad \frac{p_k}{q_k} = [b_0; b_1, \ldots\ldots, b_k] \qquad (k \geq 0)$$

where p_k, q_k are integers, $q_k > 0$ and $(p_k, q_k) = 1$. In case α is rational we must take $k \leq m$ and α itself is its last convergent p_m/q_m.

By induction one shows that the numbers p_k and q_k satisfy the recurrences

$$(1.6) \qquad \begin{cases} p_k = b_k p_{k-1} + p_{k-2} \\ q_k = b_k q_{k-1} + q_{k-2} \end{cases}$$

for $k \geq 0$, where we use the initial values

$$(1.7) \qquad \begin{pmatrix} p_{-2} & p_{-1} \\ q_{-2} & q_{-1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

From (1.6) and (1.7) we obtain the determinant relation

$$(1.8) \qquad \begin{vmatrix} p_{k-1} & p_k \\ q_{k-1} & q_k \end{vmatrix} = (-1)^k$$

for $k \geq -1$, and also, in combination with (1.1),

$$(1.9) \qquad \begin{cases} p_{k-1}\xi_{k+1} + p_k\xi_k = \xi_{-1} \\ q_{k-1}\xi_{k+1} + q_k\xi_k = \xi_0. \end{cases}$$

Note that (1.8) enables us to solve the linear Diophantine equation $ax - by = \pm 1$ in integers x, y when $(a,b) = 1$: The last but one convergent to a/b provides a solution.

Regarding the approximation of α by its convergents one has the following result, which justifies the name convergent.

<u>THEOREM 1.2.</u> *(a) (Lagrange) Each convergent p_k/q_k satisfies $|q_k\alpha - p_k| < \frac{1}{q_k}$.*
(b) (Legendre) If integers p, q with $q > 0$ satisfy $|q\alpha - p| < \frac{1}{2q}$, then p/q is one of the convergents of α, after cancellation of common factors.

THEOREM 1.3. *(a) (Borel) When $c \le \sqrt{5}$ at least one of every three consecutive convergents to α satisfies $|q_k\alpha - p_k| < \frac{1}{cq_k}$.*
(b) (Hurwitz) When $c > \sqrt{5}$ there exist numbers α for which $|q\alpha - p| < \frac{1}{cq}$ has only finitely many solutions in relatively prime integers p,q with $q > 0$ (for example $\alpha = \frac{1}{2}(1+\sqrt{5})$ has this property).

The meaning of the convergents can be explained still further if we introduce the concept of best approximations. We say that p/q $(q > 0)$ is a best approximation to α if there exists no solution in integers x,y of

$$(1.10) \qquad \begin{cases} |x\alpha - y| \le |q\alpha - p| \text{ and } |x| < q \\ \text{or } |x\alpha - y| < |q\alpha - p| \text{ and } |x| = q \end{cases}$$

except for $x = y = 0$. Then we have the following characterization due to Lagrange.

THEOREM 1.4. *Every best approximation to α is a convergent of α. Every convergent is a best approximation with the possible exception of p_0/q_0 (which is a best approximation only if $b_1 > 1$).*

In connection with (1.6) it is natural to consider, if $b_k > 1$, the fractions p/q given by

$$(1.11) \qquad \begin{cases} p = cp_{k-1} + p_{k-2} \\ q = cq_{k-1} + q_{k-2} \end{cases}$$

where c is a natural number in the range $1 \le c \le b_k - 1$. These fractions are called intermediate convergents or one-sided convergents. The latter name refers to the property that

$$0 \le \sigma(x\alpha - y) \le \sigma(q\alpha - p)$$

(where $\sigma = \operatorname{sign}(q\alpha - p)$) has no integer solutions (x,y), different from $(0,0)$ and (q,p), for which $0 \le x \le q$. The only solutions of (1.10) are $(x,y) = \pm(q_{k-1}, p_{k-1})$, but $\operatorname{sign}(q_{k-1}\alpha - p_{k-1}) = -\sigma$.

1C. A geometrical interpretation.

Following KLEIN[1], pp. 17-25, the foregoing can be represented

geometrically in the xy-plane, if we let a fraction p/q correspond to the point (q,p) of the lattice $\mathbb{Z}^2$, and the number α to the line ℓ: $y = \alpha x$ that passes through the origin and the point (ξ_0, ξ_{-1}). The relation (1.8) shows that any two consecutive convergents form a base of $\mathbb{Z}^2$, and (1.9), now reading

$$(1.12) \qquad \xi_{k+1}\begin{pmatrix} q_{k-1} \\ p_{k-1} \end{pmatrix} + \xi_k \begin{pmatrix} q_k \\ p_k \end{pmatrix} = \begin{pmatrix} \xi_0 \\ \xi_{-1} \end{pmatrix},$$

tells us that these points lie on either side of ℓ, because $\xi_k, \xi_{k+1} \geq 0$. Given such a base (q_{k-1}, p_{k-1}), (q_k, p_k), the next convergent must be a linear combination of this base, and this is expressed in the recurrence relation (1.6). The point (q_{k+1}, p_{k+1}) results from vector addition of b_{k+1} times (q_k, p_k) to (q_{k-1}, p_{k-1}) and lies on the same side of ℓ as (q_{k-1}, p_{k-1}). In fact, b_{k+1} is the largest multiple of (q_k, p_k) that can be added to (q_{k-1}, p_{k-1}) without crossing ℓ. It follows from (1.11) that the intermediate convergents are obtained by adding b_{k+1} copies of (q_k, p_k) to (q_{k-1}, p_{k-1}), one at a time.

The quantity $|q\alpha - p|$ measures the distance between ℓ and the point (q,p) along the vertical line $x = q$. Its ratio to the shortest distance depends only on α and in fact equals $\sqrt{1+\alpha^2} : 1$. We see that the point (q,p) (not the origin) is a best approximation if and only if the parallelogram given by

$$|x| \leq |q|, \quad |x\alpha - y| \leq |q\alpha - p|$$

contains no lattice point except at the origin and at its vertices.

Without reference to a coordinate system the continued fraction algorithm now takes the following form, where we anticipate the notations and definitions in chapter 2. In $\mathbb{R}^2$, let Ω be a lattice, whose points we shall denote by capitals. Let ℓ be a line through the origin, along which we choose a non-zero vector ℓ_0 that will play the part of (ξ_0, ξ_{-1}) in the coordinate version. The cofactors a_0, a_1 of a base $\{A_0, A_1\}$ of Ω are defined by

$$(1.13) \qquad \ell_0 = a_0 A_0 + a_1 A_1 .$$

Now suppose $a_0 > 0$, $a_1 > 0$, and think of (1.13) as the coordinate-free

version of (1.12). One step of the continued fraction algorithm applied to $\{A_0, A_1\}$ yields the new base $\{A_0 + bA_1, A_1\}$ with $b = [a_1/a_0]$ (if $a_1 \geq a_0$), or $\{A_0, A_1 + b'A_0\}$ with $b' = [a_0/a_1]$ (if $a_1 \leq a_0$). In either case the new base has non-negative cofactors, since

$$\ell_0 = a_0(A_0 + bA_1) + (a_1 - ba_0)A_1 = (a_0 - b'a_1)A_0 + a_1(A_1 + b'A_0).$$

The process can be continued as long as both cofactors are positive and stops as soon as one cofactor is zero (which by (1.13) implies that the other point lies on ℓ). If we also wish to obtain the intermediate convergents, we replace $\{A_0, A_1\}$ by $\{A_0 + A_1, A_1\}$ or $\{A_0, A_1 + A_0\}$, according to whether $a_1 \geq a_0$ or $a_1 \leq a_0$. The twin choice in the case $a_0 = a_1$ corresponds to the fact that a rational number has two continued fractions (cf. theorem 1.1).

1D. Quadratic irrationals.

The continued fraction $\alpha = [b_0; b_1, b_2, \ldots\ldots]$ is said to be periodic if there exist integers $k_0 \geq 0$ and $m > 0$ such that

$$b_{k+m} = b_k \quad \text{when } k \geq k_0. \tag{1.14}$$

This is equivalent to

$$\frac{\xi_{k+m-1}}{\xi_{k+m}} = \frac{\xi_{k-1}}{\xi_k} \quad \text{when } k \geq k_0. \tag{1.15}$$

The smallest possible m is called the length of the period, the corresponding smallest k_0 that of the pre-period. When $k_0 = 0$ we speak of a purely periodic continued fraction. The notation is

$$\alpha = [b_0; b_1, \ldots\ldots, \overline{b_{k_0}, \ldots\ldots, b_{k_0+m-1}}].$$

Inserting (1.15) into (1.9) one obtains the result, already known to Euler, that a periodic continued fraction has a quadratic irrational value. Lagrange proved the converse of this.

THEOREM 1.5. *The continued fraction of α is periodic if and only if α is irrational of quadratic type.*

Therefore, the continued fraction algorithm gives a nice criterion for quadratic irrationalities by becoming ultimately periodic, just as it characterizes rationals by termination. It also allows us to calculate all solutions of Pell's equation, $x^2 - dy^2 = N$, when $0 < |N| < \sqrt{d}$, because x/y must be a convergent of $\sqrt{d}$ (we assume that d is not a square). This follows if we write $x^2 - dy^2 = (x - y\sqrt{d})(x + y\sqrt{d})$ and use theorem 1.2(b). In the case of $d \equiv 2$ or $3 \pmod 4$, the numbers 1 and $\sqrt{d}$ form an integral base of the real quadratic field $\mathbb{Q}(\sqrt{d})$, and $x^2 - dy^2$ is precisely the norm of the element $x + y\sqrt{d}$. Thus every element whose norm is less than $\sqrt{d}$ in absolute value can be obtained from the convergents of $\sqrt{d}$. In particular this is true for the fundamental unit of the field. When $d \equiv 1 \pmod 4$, an integral base of $\mathbb{Q}(\sqrt{d})$ is given by 1 and $\frac{1}{2}(1 + \sqrt{d})$. In a similar way all integer elements of $\mathbb{Q}(\sqrt{d})$ with norms less than $\frac{1}{2}\sqrt{d}$ in absolute value can be obtained from the convergents of $\frac{1}{2}(1 + \sqrt{d})$.

<u>Summary.</u>

The continued fraction algorithm

- indicates that a number is rational, by terminating;
- can solve a linear Diophantine equation in two unknowns;
- calculates all best approximations to a given number;
- indicates that a number is quadratic irrational, by becoming periodic;
- can determine the fundamental unit and elements of small norm of the integer ring of a real quadratic number field.

CHAPTER 2. MULTI-DIMENSIONAL CONTINUED FRACTION ALGORITHMS

2A. Notations and preliminaries.

In the remainder of this book we shall work in the (n+1)-dimensional real vector space $\mathbb{R}^{n+1}$, where $n \geq 1$. About this space we assume that an inner product has been fixed which we can use without further introduction. The inner product of two vectors X and Y will be written as X.Y. It introduces a Euclidean metric on $\mathbb{R}^{n+1}$ by giving a vector X the length $|X| = \sqrt{X.X}$. It also defines the angle $\measuredangle(X,Y)$ between two vectors X and Y by $0 \leq \measuredangle(X,Y) \leq \pi$ and

$$\cos \measuredangle(X,Y) = \frac{X.Y}{|X|\,|Y|} .$$

The distance between a point X and a subset W of $\mathbb{R}^{n+1}$ is

$$d(X,W) = \inf_{Y \in W} |X-Y| .$$

When necessary we shall use coordinates and denote them by $x_0, x_1, \ldots, x_n$.

We denote by Ω a lattice in $\mathbb{R}^{n+1}$. A lattice is a discrete additive subgroup of rank n+1, such as $\mathbb{Z}^{n+1}$ or the collection of all points $(\alpha_0, \alpha_1, \ldots, \alpha_n)$ where α_0 runs through the algebraic integers in a totally real number field of degree n+1 and $\alpha_1, \ldots, \alpha_n$ are the conjugates of α_0. Usually we shall use capitals to denote points of Ω. For a lattice base we use the notation $A = \{A_0, A_1, \ldots, A_n\}$. A will be both the set of n+1 vectors $A_0, A_1, \ldots, A_n$ and the matrix with these vectors as its columns.

Let ℓ be a line through the origin O and choose a fixed non-zero vector ℓ_0 along ℓ. The n-dimensional subspace orthogonal to ℓ will be called ℓ^*.

DEFINITION 2.1. When $A = \{A_0, A_1, \ldots, A_n\}$ is a lattice base of Ω, the real numbers $a_0, a_1, \ldots, a_n$ defined by

$$\ell_0 = \sum_{j=0}^{n} a_j A_j \tag{2.1}$$

are called its *cofactors* with respect to ℓ.

When an indexed capital letter denotes a point of a lattice base, the corresponding lower case letter with the same index will always be the corresponding cofactor. Note that another choice of ℓ_0 along ℓ affects the size but not the ratio of the cofactors. As compared to chapter 1, formula (2.1) generalizes (1.12) and (1.13). Formula (2.1) can also be written as

$$\ell_0 = Aa \tag{2.2}$$

where a is the column vector with entries $a_0, a_1, \ldots, a_n$. It expresses the vectorial relationship between ℓ and Ω.

Let us now consider the specific case that some cofactors are equal to zero.

DEFINITION 2.2. 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. When $r = n$ we call ℓ a *rational* line.

Note that ℓ is a rational line if and only if there exists on ℓ a lattice point different from O. We also have:

LEMMA 2.1. *If ℓ is an independent line and A is a base of Ω, then no two cofactors of A are equal.*

PROOF. Let A consist of $A_0, A_1, \ldots, A_n$. If, say, $a_0 = a_1$, then $\ell_0 = a_0(A_0 + A_1) + 0.A_1 + a_2A_2 + \ldots + a_nA_n$, so that the base $\{A_0 + A_1, A_1, \ldots, A_n\}$ has a cofactor equal to zero, which contradicts the independence. □

The following theorem concerns the approximation of a dependent line by points of Ω.

THEOREM 2.2. *Suppose that ℓ has dependence rank $r > 0$ and that $\{A_0, A_1, \ldots, A_n\}$ is a base of Ω whose last r cofactors are zero. Denote by U the $(n-r+1)$-dimensional subspace spanned by $A_0, \ldots, A_{n-r}$. Then there is a number $\delta > 0$ such that for any $P \in \Omega$ with $P \notin U$ we have $d(P, \ell) \geq \delta$.*

PROOF. Since $\ell \subset U$ we always have $d(P,\ell) \geq d(P,U)$. Therefore it is sufficient to prove that

$$\delta := \inf_{P\in\Omega, P\notin U} d(P,U) > 0.$$

To arrive at a contradiction, assume that $\delta = 0$, and let $Q_1, Q_2, \ldots.$ be a sequence of lattice points not in U for which $\lim_{j\to\infty} d(Q_j,U) = 0$, and, say, $d(Q_j,U) < 1$.
Let R_j be points of U such that $|Q_j - R_j| < 1 + d(Q_j,U)$ $(j = 1,2,\ldots)$. The R_j are not necessarily lattice points. Write

$$R_j = \sum_{k=0}^{n-r} q_{jk}A_k \quad \text{with } q_{jk} \in \mathbb{R},$$

and define

$$Q_j' = Q_j - \sum_{k=0}^{n-r} [q_{jk}]A_k.$$

Then $Q_j' \in \Omega$ and $d(Q_j',U) = d(Q_j,U)$. We now find that

$$|Q_j'| = |Q_j - R_j + \sum_{k=0}^{n-r} (q_{jk} - [q_{jk}])A_k| \le 1 + d(Q_j,U) + \sum_{k=0}^{n-r} |A_k|$$
$$\le 2 + \sum_{k=0}^{n-r} |A_k|.$$

This means that the sequence of lattice points Q_1', Q_2', ... has an accumulation point, which contradicts the discreteness of Ω. □

2B. Multi-dimensional continued fraction algorithms.

We are interested now in lattice bases $A = \{A_0, A_1, \ldots, A_n\}$ with non-negative cofactors. When

$$\{A\}^+ = \{ \sum_{j=0}^{n} x_jA_j \mid \text{all } x_j \ge 0 \}$$

denotes the "first octant" of A, the fact that A has non-negative cofactors can be written as $\ell_0 \in \{A\}^+$.

DEFINITION 2.3. A finite or infinite sequence of lattice bases $A(i) = \{A_0(i), \ldots, A_n(i)\}$ is called an *n-dimensional continued fraction expansion* (or merely: expansion) of its first element $A(0)$ along ℓ if for each $i \ge 0$
(1) $\ell_0 \in \{A(i)\}^+$;

(2) there exist indices $s = s(i)$, $t = t(i)$ $(s \neq t)$ and an integer $b = b(i) \geq 1$ such that

(2.3) $$A_t(i+1) = A_t(i) + bA_s(i)$$

and

(2.4) $$A_j(i+1) = A_j(i) \quad \text{for } j \neq t.$$

The transition from $A(i)$ to $A(i+1)$ is called the $(i+1)$-th *step* of the expansion.

DEFINITION 2.4. 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.*

As a shorthand notation for (2.3), (2.4) we use

(2.5) $$A(i+1) = A(i)\ I^b_{st}$$

where I^b_{st} is the identity matrix I with the zero in row s and column t replaced by the number b. (Note that $I^b_{st} = (I_{st})^b$ when $I_{st} = I^1_{st}$, and $(I^b_{st})^{-1} = I^{-b}_{st}$).

Since $A(i)$ and its cofactor vector $a(i)$ are connected by (2.2), (2.5) immediately gives

(2.6) $$a(i+1) = I^{-b}_{st}\ a(i).$$

The letters $s(i)$, $t(i)$ and $b(i)$ will be reserved for the meaning they have in definition 2.3. To avoid a confusion of indices we shall write s and t instead of $s(i)$ and $t(i)$ if it is clear which step we are speaking about. Instead of the monstrous $I^{b(i)}_{s(i)t(i)}$ we write $I^b_{st}(i)$.

From the equality

$$a_sA_s + a_tA_t = (a_s - ba_t)A_s + a_t(A_t + bA_s)$$

it is clear that

(2.7) $\quad a_s(i) \geq b a_t(i)$

is a necessary and sufficient condition for $\ell_0 \in \{A(i+1)\}^+$, given that $\ell_0 \in \{A(i)\}^+$. In view of $a_t(i) \geq 0$ and $b \geq 1$ [1]), (2.7) is equivalent to the combination of

(2.8) $\quad a_s(i) \geq a_t(i)$

and

(2.9) $\quad 1 \leq b \leq \left[\frac{a_s(i)}{a_t(i)} \right]$,

where (2.9) is to be interpreted as $1 \leq b < \infty$ when $a_t(i) = 0$. In general, therefore, we can at each step determine the pair of indices s,t in $\binom{n+1}{2}$ different ways (and more if some cofactors happen to be equal), after which an integer b in the range (2.9) must be determined. Essentially, an n-dimensional continued fraction algorithm is a procedure to make these choices. When we are designing an algorithm to solve a particular problem we are completely free to let this procedure take into account any information we wish, as long as we comply with the restriction (2.7). (We might for example throw dice).

Referring to the change in the cofactor a_s during a step, we usually call an algorithm *subtractive* if it always takes $b(i) = 1$; a *division* algorithm always chooses the maximal $b(i) = \left[\frac{a_s(i)}{a_t(i)} \right]$ (and has to avoid that $a_t(i) = 0$).

Applying these remarks to the case $n = 1$ we immediately obtain:

THEOREM 2.3. *The only 1-dimensional continued fraction algorithm is the ordinary continued fraction algorithm.*

PROOF. Because $s,t \in \{0,1\}$, (2.8) defines s and t uniquely, unless $a_0(i) = a_1(i)$. With $b(i) = \left[\frac{a_s(i)}{a_t(i)} \right]$ we have the ordinary division algorithm, with other choices of b(i) we merely make additional stops at some of the intermediate convergents (or all, if $b(i) = 1$ throughout): Note that

$$a_{s(i)}(i+1) = a_{s(i)}(i) - b a_{t(i)}(i) > a_{t(i)}(i) = a_{t(i)}(i+1)$$

[1]) For formal reasons we shall sometimes allow empty steps, i.e. $b = 0$.

if $b < \left[\frac{a_s(i)}{a_t(i)}\right]$, so that $s(i+1) = s(i)$ and $t(i+1) = t(i)$; in other words, the main convergents are never missed. If $a_0(i) = a_1(i)$, then ℓ is a rational line and the twin choice of s,t pairs corresponds to a rational number having two continued fraction expansions. □

2C. Convergence.

In this section we explain what we shall mean by the convergence of an expansion. We define the normalized height function h on $\mathbb{R}^{n+1}$ as the linear function given by $h(\ell_0) = |\ell_0|$ and $h(X) = 0$ for all $X \in \ell^*$. This gives

$$|h(X)| = d(X, \ell^*) \quad \text{for all } X \in \mathbb{R}^{n+1}.$$

DEFINITION 2.5. (a) An expansion $A(i)$, $i = 0,1,2,\ldots\ldots$ along a line ℓ is *strongly convergent* if for every $\varepsilon > 0$ there is an integer i_0 such that

$$(2.10) \qquad \max_{0 \le k \le n} d(A_k(i_0), \ell) < \varepsilon.$$

(b) It is *weakly convergent* if for every $\varepsilon > 0$ there is an integer i_0 such that

$$(2.11) \qquad \max_{0 \le k \le n} \frac{d(A_k(i_0), \ell)}{|h(A_k(i_0))|} < \varepsilon.$$

The latter concept generalizes the convergence of $|\alpha - \frac{p}{q}|$ to zero in the 1-dimensional case of chapter 1, the former that of $|q\alpha - p|$. By convergent we shall always mean strongly convergent, unless otherwise stated.

Not all lines can have strongly convergent expansions. Indeed, an immediate consequence of theorem 2.2 is:

LEMMA 2.4. *If ℓ admits a strongly convergent expansion, then it is an independent line.*

In section 6B we prove the converse of this lemma by explicit construction of a convergent expansion when ℓ is independent. Quantitative versions of lemma 2.4 may help to decide whether a given line is independent (see section 4A).

It usually depends on the algorithm whether we obtain a weakly convergent expansion along a dependent line. Consider for example the

algorithm of Brun (section 3H): When $n = 2$ only independent lines have weakly convergent expansions, whereas for $n \geq 3$ a dependent line may have one, too. On the other hand, most algorithms (including Brun's) have the property that they always choose t such that $a_t(i) > 0$. Then $a_k(i_0) = 0$ for some k and i_0 implies $A_k(i) = A_k(i_0)$ for all $i \geq i_0$ and the expansion is not weakly convergent. Thus only those dependent lines may have weakly convergent expansions, for which the algorithm does not prove the dependence by exhibiting a base with a cofactor equal to zero.

In view of these considerations we give:

DEFINITION 2.6. An n-dimensional continued fraction algorithm is strongly (weakly) convergent if it generates a strongly (weakly) convergent expansion of a lattice base A along an independent line ℓ, whenever A and ℓ are compatible with the initial constraints of the algorithm.

Anticipating what follows, we remark that it is relatively easy to prove weak convergence of a given algorithm, if it does converge weakly. But strong convergence is (when $n > 1$) much more difficult, if it exists at all. In fact, until very recently convergence nearly always meant weak convergence.

2D. The linear Diophantine equation.

As an easy application of multi-dimensional continued fraction algorithms we shall solve the equation

$$c_0x_0 + c_1x_1 + \dots + c_nx_n = 1 \tag{2.12}$$

in integers $x_0, x_1, \dots, x_n$, when $c_0, c_1, \dots, c_n$ are given positive integers without common divisor greater than one. Let C and X denote column vectors with respective entries $c_0, c_1, \dots, c_n$ and $x_0, x_1, \dots, x_n$, and let e_k be the column vector whose k-th component is 1 and whose other components are zero $(0 \leq k \leq n)$.

LEMMA 2.5. *Let D be an integral matrix with $|\det D| = 1$, such that $DC = e_k$. Then the general solution of (2.12) is*

$$X = D^T Y \tag{2.13}$$

where $Y = (y_0, y_1, \dots, y_n)$ is an arbitrary integral vector with $y_k = 1$, and

the ^{T} *denotes transposition.*

PROOF. Clearly (2.13) solves (2.12), since

$$C^TX = C^T\mathcal{D}^TY = (\mathcal{D}C)^TY = (e_k)^TY = y_k = 1.$$

For the converse, let X be a solution of (2.12), and determine the integral vector Y from $X = \mathcal{D}^TY$. This is possible because det $\mathcal{D} = \pm 1$. But we have

$$1 = C^TX = C^T\mathcal{D}^TY = (\mathcal{D}C)^TY = y_k,$$

as before. □

So the solution of (2.12) is complete if we can find $\mathcal{D}$ satisfying $\mathcal{D}C = e_k$ for some k. For this we use an n-dimensional continued fraction algorithm as follows.

THEOREM 2.6. *Let* $A(0) = I$ *be the standard base of* $\mathbb{Z}^{n+1}$, *consisting of* $e_0, e_1, \ldots, e_n$, *and put* $\ell_0 = a(0) = C$. *If an n-dimensional continued fraction algorithm has the property that it chooses t such that* $a_t > 0$, *then there is an index* i_0 *such that, for some k,* $a(i_0) = (A(i_0))^{-1}C = e_k$.

PROOF. From $A(0) = I$ and $\ell_0 = a(0) = C$, the identity $A(i)a(i) = C$ follows for all i, in view of (2.2). Furthermore, for all i the sum of the entries of a(i+1) (which are non-negative integers) is strictly less than that of a(i) by the nature of the algorithm. Thus the expansion ends for some i_0 with the impossibility of $a_t(i_0) > 0$, i.e. $a(i_0) = e_k$ for some k ($a(i_0)$ cannot be a multiple of e_k since the greatest common divisor of $c_0, c_1, \ldots, c_n$ is one). □

Not much seems to be known about two questions that theorem 2.6 raises: Is there a specific algorithm which finds the matrix $\mathcal{D} = (A(i_0))^{-1}$ in as few steps as possible? And is there an algorithm that gives as the particular solution $X = \mathcal{D}^Te_k$ (that is, $Y = e_k$ in (2.13)) a vector of least possible length?

The latter question was solved for $n \le 3$ by ROSSER[2], who gave a counterexample to his method with $n = 4$. His method is not a true continued fraction algorithm, but bears strong resemblance to it. Several authors studied the effect of some simple algorithms with regard to these

questions. JACOBI[1] noted that the obvious algorithm,

$$\begin{array}{ll} \text{first solve} & c_0x_0 + c_1x_1 = d_1 = \text{g.c.d.}(c_0,c_1), \\ \text{then} & d_1x_1' + c_2x_2 = d_2 = \text{g.c.d.}(d_1,c_2), \\ \ldots\ldots\ldots & \\ \text{finally} & d_{n-1}x_{n-1}' + c_nx_n = 1, \end{array}$$

may lead to rather big numbers which form the particular solution $X = \mathcal{D}^T e_k$ of (2.12), especially when the number 1 does not feature early in the sequence $d_1, d_2, \ldots, d_{n-1}, 1$. His own algorithm (section 3D) and an algorithm of EULER[1] usually give much smaller solutions. Also D.N. LEHMER[2] and BERNSTEIN[13] considered the particular solution of (2.12) obtained by Jacobi's algorithm. ROSSER[1] and D.H. LEHMER[1] worked with algorithms of Brun (section 3H) and Güting (section 3G) respectively.

2E. Outline.

A variety of problems from number theory and algebraic number theory can be formulated in terms of the approximation of a suitable line through the origin by non-zero points of some lattice. Among these problems, those concerning simultaneous Diophantine approximation, linear dependence or the units of algebraic number fields are straightforward multi-dimensional generalizations of the problems summarized at the end of chapter 1, for which the continued fraction algorithm provides a solution.

It is the purpose of this book to investigate whether multi-dimensional continued fraction algorithms can be used to solve such generalized approximation problems. The interest in this question grew about 1970 with the publication of an algorithm of Szekeres. In chapter 3 we give first a brief survey of the history of multi-dimensional continued fractions until that time, because no such survey is available elsewhere. The reader can follow most of the chapters 4 to 7 without knowledge of chapter 3.

CHAPTER 3. VECTORIAL ALGORITHMS

3A. Introduction.

In section 2B we saw that an n-dimensional continued fraction algorithm essentially is a rule or procedure to select at each step the indices s,t ($s \neq t$) and an integer $b \geq 1$ in accordance with (2.7), then to replace A_t by $A_t + bA_s$ (and simultaneously a_s by $a_s - ba_t$). In principle we impose no restrictions on this rule, apart from (2.7). But it must, of course, be relatively simple if it is to have any practical value. For instance, it should not look back to see what was done in more than the one or two preceding steps. We notice that there are two more or less opposed approaches. Firstly, we can examine what a given rule can achieve, and secondly, we can look into a problem and try to find a rule that will solve it. The first approach has an instructive value since it helps us to see the difficulties that will arise in the second approach. Also we may wish to know the further properties of a rule designed for a particular problem. The first approach is the older one, too, as this chapter will make clear.

The most extensively studied class of multi-dimensional continued fraction algorithms is that of the vectorial algorithms (also known as Jacobi-type algorithms), which we define as follows.

DEFINITION 3.1. A multi-dimensional continued fraction algorithm is called *vectorial* if its rule of selecting s, t and b depends only on the vectorial relation between the line ℓ and the lattice Ω (i.e., the vector of cofactors).

The important feature of such an algorithm is that it runs the same for any line and lattice combination having the same initial vector of cofactors, and in particular that it does not take into account the geometrical relation between ℓ and Ω.

Nearly all work done in the field up to 1970 comes in this class, as does a lot of later work. This chapter intends to give a brief survey of that work. Except for a previously unknown result (section 3I) and the general periodicity theorem (which applies also to other than vectorial algorithms), we refer to the literature for all proofs.

3B. Division or subtraction?

In the 1-dimensional case, the ordinary continued fraction algorithm can be used, as we saw in section 1C, either in division form with $b = \left[\frac{a_s}{a_t}\right]$ (as did Euler), or with smaller b, notably in subtraction form with $b = 1$ (as did Euclid) to obtain the intermediate convergents. The division form is merely a way of running faster through the subtraction form, stopping only at the principal convergents (but at all of them).

This is no longer true when $n \geq 2$, for then there are subtractive algorithms which cannot be accelerated in division form.

EXAMPLE. SELMER[2] defined the following algorithm: Let $s = s(i)$, $t = t(i)$ be given by

$$a_s(i) = \max_{0 \leq k \leq n} a_k(i)$$

and

$$a_t(i) = \min\{a_k(i) \mid a_k(i) > 0,\ k \neq s\},$$

and take $b(i) = 1$. Applied to a line and lattice base for which the initial vector of cofactors is $a(0) = (5, 4, 2)$, it gives

$$\begin{array}{ll} a(1) = (3, 4, 2) & a(5) = (1, 1, 1) \\ a(2) = (3, 2, 2) & a(6) = (0, 1, 1) \\ a(3) = (1, 2, 2) & a(7) = (0, 0, 1); \\ a(4) = (1, 1, 2) & \end{array}$$

whereas the only possibilities for a division algorithm are those in the scheme

$$\begin{array}{lccc} a(0) = & (5, 4, 2) & & \\ & \downarrow & \searrow & \\ a(1) = & (1, 4, 2) & & (5, 0, 2) \\ & \downarrow & \searrow & \downarrow \\ a(2) = & (1, 4, 0) & & (1, 0, 2) \\ & & \searrow & \downarrow \\ a(3) = & & & (1, 0, 0) \end{array}$$

as one easily recognizes.

On the other hand it is easy to convert one division into b(i) subtractions. The dilemma of subtraction and division is an old one that turns up repeatedly in the literature (but mostly with vectorial algorithms). The best-known division algorithm is that of Jacobi-Perron. Brun, Pipping, and more recently Szekeres, are among those who favour subtractions. So is, to some extent, the present author.

We notice, however, that purely subtractive algorithms show the defect of being essentially slower than division algorithms. With the ordinary continued fraction algorithm we find the (j+1)-th convergent p_j/q_j to $\alpha = [b_0;b_1,b_2,\ldots\ldots]$ after j+1 divisions, or after $b_0 + b_1 + \ldots + b_j$ subtractions. That is, in the former case we find on average 1 convergent at each step, while in the latter this average is zero for almost all α (in Lebesgue sense) because for almost all α,

$$\lim_{j\to\infty} \frac{1}{j} \sum_{k=0}^{j} b_k = \infty.$$

(see KHINTCHINE[2], p. 101).

3C. Periodicity of vectorial algorithms.

Particular attention has always been paid to the periodicity properties of vectorial algorithms in the following sense.

DEFINITION 3.2. The expansion of a lattice base $A(0)$ along a line ℓ by a vectorial algorithm is *periodic* if there exist integers $k_0 \geq 0$, $m > 0$, a real number λ and a permutation matrix P such that

$$(3.1) \qquad a(k+m) = \lambda P a(k) \quad \text{for } k \geq k_0.$$

Some authors restrict the permutation matrix to special types, e.g. cyclic permutations, but this makes a difference in the length m of the period only: When q is the order of P we have

$$(3.2) \qquad a(k+qm) = \lambda^q a(k) \quad \text{for } k \geq k_0$$

without any permutation.

The importance of periodicity lies in the next result.

THEOREM 3.1. *Suppose the expansion of* $A(0)$ *along* ℓ *by a vectorial algorithm is periodic. Then*

(a) *the number* λ *in (3.1) is an algebraic integer of degree* $\leq n+1$ *and norm* ± 1;

(b) *if the degree of* λ *equals* $n+1$, *then the numbers* $\frac{a_j(0)}{a_0(0)}$, $j=0,1,\ldots,n$, *constitute a base of the number field* $\mathbb{Q}(\lambda)$.

PROOF. By (2.6) we have

$$a(k+m) = I^b_{st}(k+m-1)\ldots\ldots I^b_{st}(k+1)I^b_{st}(k)\ a(k),$$

that is, $a(k+m) = Ja(k)$ for some $SL(n+1, \mathbb{Z})$ matrix J. Now (3.1) gives

$$P^{-1}Ja(k) = \lambda a(k) \tag{3.3}$$

so that λ is an eigenvalue of $P^{-1}J$. Since the characteristic polynomial of $P^{-1}J$ (also called the characteristic polynomial of the expansion) is a polynomial of degree n+1 with integral coefficients, leading coefficient $(-1)^{n+1}$ and constant term $\det(P^{-1}J) = \pm 1$, part (a) follows.

Now assume that λ is of degree n+1, and write

$$P^{-1}J = \begin{bmatrix} j_{00} & \cdots\cdots & j_{0n} \\ \vdots & & \vdots \\ j_{n0} & \cdots\cdots & j_{nn} \end{bmatrix}$$

At least one of the numbers $a_j(k)$, $j=0,1,\ldots,n$, is not zero. Suppose that $a_0(k) \neq 0$. Then the last n lines of (3.3) take the shape

$$\begin{bmatrix} (j_{11}-\lambda)a_1(k) + & \ldots\ldots + & j_{1n}a_n(k) = -j_{10}a_0(k) \\ \vdots & & \vdots \\ j_{n1}a_1(k) + & \ldots\ldots + & (j_{nn}-\lambda)a_n(k) = -j_{n0}a_0(k) \end{bmatrix} \tag{3.4}$$

The coefficient determinant Δ of (3.4) is not zero, otherwise λ were of degree $\leq n$, hence it follows by Cramer's rule that

(3.5) $\Delta a_j(k) = \alpha_j a_0(k), \quad j = 1,\ldots\ldots,n,$

for some numbers $\alpha_1,\ldots\ldots,\alpha_n \in \mathbb{Q}(\lambda)$ which in fact are polynomial expressions in λ of degree at most n-1 and with integral coefficients. Thus any non-trivial linear relation

$$\sum_{j=0}^{n} c_j a_j(k) = 0, \quad c_j \in \mathbb{Z},$$

would yield an n-th degree polynomial equation for λ, namely

$$c_0\Delta + \sum_{j=1}^{n} c_j\alpha_j = 0 .$$

This is another contradiction to the fact that λ is of degree n+1. Therefore, and by (3.5), $\frac{a_j(k)}{a_0(k)}$, $j = 0,1,\ldots.,n$, are n+1 rationally independent numbers in $\mathbb{Q}(\lambda)$, which proves part (b) since the $\frac{a_j(0)}{a_0(0)}$ are connected with them through the unimodular transformation $I^b_{st}(k-1)\ldots\ldots.I^b_{st}(1)I^b_{st}(0)$. □

<u>REMARK (1).</u> Whether or not the numbers $\frac{a_j(0)}{a_0(0)}$ belong to $\mathbb{Q}(\lambda)$ if λ is of degree less than n+1, cannot be decided without additional information on the algorithm. Consider, for example, the following 3-dimensional algorithm: Let e,f,g,h be a permutation of 0,1,2,3 such that $a_e \geq a_f \geq a_g \geq a_h$. Then put $s = e$, $t = g$. This algorithm is periodic (length 2 with suitable permutation) with eigenvalue $\frac{1}{2}(1+\sqrt{5})$ for

$$a(0) = \left(\tfrac{1}{2}\pi(1+\sqrt{5}),\ \tfrac{3}{2}(1+\sqrt{5}),\ \pi,\ 3\right).$$

On the other hand, PERRON[1] proved that the $\frac{a_j(0)}{a_0(0)}$ always belong to $\mathbb{Q}(\lambda)$ if the algorithm is that of Jacobi-Perron, regardless of the degree of λ.

<u>REMARK (2).</u> Theorem 3.1 enables us to obtain algebraic units from periodic expansions by means of a vectorial algorithm: When ω is an algebraic integer of degree n+1, and such an algorithm becomes periodic when applied to a line and lattice for which $a(0) = (1, \omega, \ldots., \omega^n)$ (or any other base of $\mathbb{Q}(\omega)$), then the eigenvalue is a unit of $\mathbb{Q}(\omega)$. Quite another approach is, to use the fact that if the entries of a(0) are algebraic integers of

$\mathbb{Q}(\omega)$, then so are all subsequent cofactors and nothing prevents one of them from being a unit.

EXAMPLE. Apply to $a(0) = (1, \omega, \omega^2)$ with $\omega = \sqrt[3]{2}$ the algorithm of Brun (section 3H), which takes s and t from

$$a_s = \max_{0 \le k \le n} a_k \ , \quad a_t = \max_{0 \le k \le n, k \ne s} a_k$$

and $b = 1$. The calculation is as follows (using the approximate values $\sqrt[3]{2} = 1.259921$, $\sqrt[3]{4} = 1.587401$):

$$\begin{aligned}
a(0) &= (\ 1 \qquad\qquad , \omega \quad , \omega^2 \qquad\quad) \\
a(1) &= (\ 1 \qquad\qquad , \omega \quad , \omega^2 - \omega \quad) \\
a(2) &= (\ 1 \qquad\qquad , \omega - 1, \omega^2 - \omega \quad) \\
a(3) &= (-\omega^2 + \omega + 1, \omega - 1, \omega^2 - \omega \quad) \\
a(4) &= (-2\omega^2 + 2\omega + 1, \omega - 1, \omega^2 - \omega \quad) \\
a(5) &= (-3\omega^2 + 3\omega + 1, \omega - 1, \omega^2 - \omega \quad) \\
a(6) &= (-3\omega^2 + 3\omega + 1, \omega - 1, \omega^2 - 2\omega + 1) \\
&\cdots\cdots \\
a(13) &= (-3\omega^2 + 3\omega + 1, 6\omega^2 - 2\omega - 7, 4\omega^2 - 9\omega + 5) \\
&\cdots\cdots \\
a(19) &= (-8\omega^2 - 5\omega + 19, -5\omega^2 + 19\omega - 16, 24\omega^2 - 35\omega + 6) = \\
&\quad = (-8\omega^2 - 5\omega + 19)a(1).
\end{aligned}$$

The fundamental unit $\varepsilon = \omega - 1$ appears in $a(2)$. Its powers $\varepsilon^2 = \omega^2 - 2\omega + 1$, $\varepsilon^3 = -3\omega^2 + 3\omega + 1$, $\varepsilon^4 = 6\omega^2 - 2\omega - 7$, $\varepsilon^5 = -8\omega^2 - 5\omega + 19$ appear in $a(6)$, $a(5)$, $a(13)$, $a(19)$ respectively. The eigenvalue of the period is ε^5, its length is 18.

3D. The Jacobi-Perron algorithm.

The earliest and most extensively studied of all multi-dimensional continued fraction algorithms is the Jacobi-Perron algorithm (J.P.A.). It was first published for two dimensions in 1868 in a post mortem paper of JACOBI[2], who investigated the matter at least as early as 1839. After the publication of some minor contributions to the theory of the algorithm in the nineteenth century (see KOKSMA[1], p. 50, for these early references),

it was PERRON[1] who gave the algorithm a rigorous foundation in all dimensions in his thesis (1907). Jacobi remarked casually that his method would work in any dimension, although he explained it for $n = 2$ only.

The most common definition in homogeneous form is as follows. If $\hat{a}(0) = (\hat{a}_0(0), \hat{a}_1(0), \ldots, \hat{a}_n(0))$ is any vector with positive entries, then the sequence of vectors $\hat{a}(i)$ given by

$$(3.6) \qquad \hat{a}(i+1) = (\hat{a}_1(i) - \hat{b}_1\hat{a}_0(i), \ldots, \hat{a}_n(i) - \hat{b}_n\hat{a}_0(i), \hat{a}_0(i))$$

with $\hat{b}_j = \left[\dfrac{\hat{a}_j(i)}{\hat{a}_0(i)}\right]$, is its Jacobi-Perron expansion.

If we convert this to the notations of chapter 2 by eliminating the cyclic rotation of indices we get:

<u>DEFINITION 3.3.</u> The algorithm

(1) $i := t(0) := 0;\ s(0) := 1;$

(2) $A(i+1) := A(i)I^b_{st}(i)$ with $b = b(i) = \left[\dfrac{a_s(i)}{a_t(i)}\right]$;

(3) <u>if</u> $s(i) + 1 = t(i)$ <u>then</u> $t(i+1) := t(i) + 1,\ s(i+1) := s(i) + 3;$ <u>else</u> $s(i+1) := s(i) + 1,\ t(i+1) := t(i);$

(4) $i := i + 1;$ <u>goto</u> (2);

is called the *Jacobi-Perron algorithm* (J.P.A.).

The indices s and t are, of course, to be read modulo n+1. In this definition we allow $b(i) = 0$ and some of the steps may be empty. Note that the n steps between two occurrences of the event $s(i) + 1 = t(i)$ correspond to the one big step in (3.6): In formula expression

$$\hat{a}_k(i) = a_{k+i}(ni) \quad \text{for all } k,i$$

where again the indices should be read modulo n+1.

We give two small examples, in which the cofactor holding the t-index for the next n (=2) steps is set in italics.

<u>EXAMPLE 1.</u> $a(0) = (\ \mathit{93},\ 167,\ 118)$

$a(2) = (\ 93,\ \mathit{74},\ 25)$	$b(0) = 1,\ b(1) = 1$
$a(4) = (\ 19,\ 74,\ \mathit{25})$	$b(2) = 0,\ b(3) = 1$
$a(6) = (\ \mathit{19},\ 24,\ 25)$	$b(4) = 0,\ b(5) = 2$
$a(8) = (\ 19,\ \mathit{5},\ 6)$	$b(6) = 1,\ b(7) = 1$

$$a(10) = (\ 4,\quad 5,\quad 1)$$
$$a(12) = (\ 0,\quad 0,\quad 1)$$

$$b(8) = 1,\quad b(9) = 3$$
$$b(10) = 4,\quad b(11) = 5$$

EXAMPLE 2. Let us take $\omega = \sqrt[3]{2}$ and $a(0) = (1,\ \omega,\ \omega^2)$.

$$a(0) = (\ 1 \qquad ,\ \omega \qquad ,\ \omega^2 \qquad)$$
$$a(2) = (\ 1 \qquad ,\ \omega - 1 \qquad ,\ \omega^2 - 1 \qquad)$$
$$a(4) = (\ 4 - 3\omega \qquad ,\ \omega - 1 \qquad ,\ \omega^2 - 2\omega + 1)$$
$$a(6) = (-3\omega^2 + 3\omega + 1,\ -3\omega^2 + 7\omega - 4,\ \omega^2 - 2\omega + 1)$$

$$b(0) = 1,\ b(1) = 1$$
$$b(2) = 2,\ b(3) = 3$$
$$b(4) = 3,\ b(5) = 3$$

Since $a_{k+1}(6) = (\omega - 1)a_k(4)$ for $k = 0,1,2$, the expansion is periodic from $a(4)$ on with eigenvalue $\omega - 1$ and length 2.

REMARK (1). Because of the choice of the $\hat{b}_j$ the last entry of the right hand side of (3.6) is the largest. This implies that

$$\hat{b}_n \geq 1 \quad \text{and} \quad \hat{b}_n \geq \hat{b}_k \geq 0 \text{ for } k = 1,\ldots,n-1, \tag{3.7}$$

at least when $i \geq 1$. One also recognizes that if $\hat{b}_n = \hat{b}_k$, then we have $\hat{b}_{n-1} \geq \hat{b}_{k-1}$ in the next step; if this is an equality also, then we must have $\hat{b}_{n-2} \geq \hat{b}_{k-2}$ in the step thereafter, and so on, where $\hat{b}_0$ is to be interpreted as $\hat{b}_0 = 1$. In our notation this means that

$$\begin{cases} b_{nj-1} \geq 1 & \text{for } j \geq 2 \\ b_{nj-1} \geq b_{nj-k} \geq 0 & \text{for } k = 2,\ldots,n,\ j \geq 2 \end{cases} \tag{3.8}$$

and

$$b_{nj-1} = b_{nj-k} \Rightarrow b_{(j+1)n-2} \geq b_{(j+1)n-k-1} \quad \text{for } k = 2,\ldots,n, \text{ etc.} \tag{3.9}$$

Relation (3.8) shows that at least one of the n steps into which (3.6) falls apart in definition 3.3, is not empty, namely the one with $s = t - 1$ (modulo n+1).

REMARK (2). Both JACOBI[1] and PERRON[1] preferred the greater regularity of the cyclic algorithm to the idea of using in each big step (3.6) the smallest non-zero cofactor as the divisor $\hat{a}_t$. The latter idea had already been used by EULER[1], who sought coefficients to make a linear form in

three or more real numbers small in absolute value. The cyclic algorithm has the disadvantage that it requires certain modifications (namely, transition to the (n-1)-dimensional case) when the divisor $\hat{a}_0(i)$ happens to be zero. By definition, such difficulties can occur only when the line ℓ is not independent; otherwise the gained regularity, resulting in convenient recurrent formulae akin to (1.2), (1.6), (1.8) was considered a more important advantage, particularly in proofs by induction. In the sequel we assume to deal only with expansions where no divisor happens to be equal to zero.

REMARK (3). The expansion of a given base by the J.P.A. depends usually on the way in which the elements of the base are ordered.

REMARK (4). PERRON[1], theorem IV, proved that two different vectors a(0) can never produce the same sequence of integers b(i) throughout, unless they differ by a factor only. He also determined under what conditions on a given sequence of integers c(i) there exists a vector a(0) such that $b(i) = c(i)$ for all $i \geq 0$; namely, the conditions which are necessary from (3.8), (3.9) are at the same time sufficient.

3E. Convergence of the J.P.A.

PERRON[1], §4, proved algebraically (see also BERNSTEIN[15]):

THEOREM 3.2. *The J.P.A. is weakly convergent in any dimension.*

MAUNSELL[1], correcting an earlier paper of BATEMAN[1], proved in 1930 for $n = 2$ by means of a geometric argument that

$$\max_{k=0,1,2} \frac{d(A_k(i),\ell)}{|h(A_k(i))|} < c\left(\frac{2}{3}\right)^{[i/3]}$$

where c is a constant depending on $A(0)$ and ℓ. VAUGHAN[1] gave an analogous result with $\frac{1}{2}$ instead of $\frac{2}{3}$, and DAVID[4], p. 36, proved that, for $n = 2$, $d(A_k(i),\ell)$ is bounded uniformly in i, whereas $\lim_{i\to\infty} |h(A_k(i))| = \infty$. Recently FISCHER[1] estimated for arbitrary $n \geq 1$

$$\max_{0\leq k\leq n} \frac{d(A_k(i),\ell)}{|h(A_k(i))|} < c\left(\sqrt[n]{1 - \left(\frac{1}{n+1}\right)^n}\right)^{[i/n]} .$$

However, we find in PERRON[1], p. 67, an example with $n \geq 3$ indicating

that strong convergence is not a generic property of the J.P.A. It seems likely that there exist 2-dimensional non-convergent expansions, too.

In 1875 already BACHMANN[1] proved: If the expansion of $a(0) = (1, \sqrt[3]{d}, \sqrt[3]{d^2})$ (d is a cube-free integer greater than 1) by the J.P.A. turns out to be periodic, then each of the points $A_k(i)$, $k = 0,1,2$, $i \geq 0$, satisfies

$$(3.10) \qquad d(A_k(i),\ell) < c|h(A_k(i))|^{-\frac{1}{2}}$$

where the constant c is independent of i and k. This is a special case of the following general theorem (PERRON[1], p. 67, also DUBOIS & PAYSANT-LE-ROUX[1]):

<u>THEOREM 3.3.</u> *A periodic Jacobi-Perron expansion is strongly convergent if and only if its characteristic polynomial is irreducible and has a P.V.-number for one of its roots.* [1])

The exponent $-\frac{1}{2}$ in (3.10) represents the fastest possible order of convergence. (In general, this exponent would be $-1/n$. For this and other results in the theory of Diophantine approximation we refer to CASSELS[2], KHINTCHINE[1], KOKSMA[1]). However, PERRON[1], theorem XVII, proved that in general a periodic Jacobi-Perron expansion can have the order of convergence $|h(A_k(i))|^{-1/n}$ only when $n = 1$, or when $n = 2$ and $1, \frac{a_1(0)}{a_0(0)}, \frac{a_2(0)}{a_0(0)}$ form a base of a cubic field with negative discriminant.

Other, computational, material presented by JURKAT, KRATZ & PEYERIMHOFF[1] and VAN DE LUNE & TE RIELE[1] also shows that the approximations furnished by the J.P.A. are in general rather bad. It is not surprising that more attention has been paid to the periodicity properties of the algorithm.

SCHMIDT[1] (for $n = 2$) and SCHWEIGER[9,10,11,12] (for $n \geq 3$) gave exact theorems on the speed with which the cofactors tend to zero during a Jacobi-Perron expansion. These theorems generalize the Hurwitz-Borel theorem 1.3 and involve, instead of $\sqrt{5}$, the number $\xi^n + n\xi^{-1}$, where

---

[1]) A P.V.-number, named after Pisot and Vijayaraghavan, is a real algebraic integer greater than one, all of whose other (real and complex) conjugates lie inside the unit circle.

$\xi^{n+1} = 1 + \xi^n$. SCHWEIGER[1-8] also studied the ergodic properties of the J.P.A.

3F. Periodicity of the J.P.A.

In his original article, JACOBI[2] shows that the expansions by J.P.A. of $a(0) = (1, \sqrt[3]{d}, \sqrt[3]{d^2})$ are periodic for $d = 2,3,5$ (compare example 2 of section 3D, p.32). This has stimulated intensive research into the periodicity properties of the algorithm. Usually one defines periodicity of a Jacobi-Perron expansion by the existence of integers $k_0 \geq 0$, $m > 0$ and a real number λ such that, in the notation of (3.6),

$$(3.11) \qquad \hat{a}(k+m) = \lambda \hat{a}(k) \quad \text{for } k \geq k_0,$$

that is, one restricts P in definition 3.2 to a cyclic permutation. When m and k_0 are the smallest possible, we call them the length of the period and pre-period, respectively. In what follows we shall mean by period length the minimal number m in (3.11).

PERRON[1], theorem XII, proved in addition to theorem 3.1, that for any $a(0)$ whose Jacobi-Perron expansion is periodic, the numbers $\frac{a_j(0)}{a_0(0)}$, $j = 0,1,\ldots,n$, all belong to $\mathbb{Q}(\lambda)$, regardless of the degree of λ. On the other hand, an important observation is that the characteristic polynomial of a periodic expansion may indeed be reducible, at least when $n \geq 3$ (PERRON[1], p. 61. Its irreducibility for $n = 2$ was proved in 1930 by COLEMAN[1,2][1])). With this observation, we have an example where the J.P.A. may fail to notice a dependence relation (because it never finds a base with a cofactor equal to zero) -- again this is not valid for $n = 2$, since DAVID[4], p. 46, showed that in the 2-dimensional case the J.P.A. will always detect a dependence relation if one exists.

Whether the converse, i.e. the equivalent of Lagrange's theorem 1.5, is true, is still an open question: Is the J.P.A.-expansion of any $a(0)$ whose components are a base of some algebraic number field ultimately periodic? In spite of many efforts, the answer cannot yet be reasonably conjectured. The case of periodic expansions of length 1 (in (3.11)) is relatively easy:

---

[1]) and independently by PERRON[3] in 1935.

THEOREM 3.4. (PERRON[1], theorem VII) *Let ξ be the largest real root of the polynomial $x^{n+1} - d_n x^n - \ldots - d_1 x - 1$, and suppose all other roots have absolute values less than ξ. Then the vector $a(0) = (a_0(0), \ldots, a_n(0))$ with*

$$a_k(0) = \xi^{-k}(1 + d_1\xi + \ldots + d_k\xi^k), \quad k = 0, \ldots, n,$$

has a purely periodic Jacobi-Perron expansion of length 1.

By theorem 3.3 the convergence of that expansion is strong only if ξ is a P.V.-number. Theorem 3.4 also appears in DUBOIS & PAYSANT-LE-ROUX[1] and HUA KOO LENG & WANG YUAN[1].

BERNSTEIN[1,2,3,4,6,7,8,14,15,17,19,21], LEVESQUE[1,2,3], DUBOIS & PAYSANT-LE-ROUX[2,3] discovered many classes of algebraic numbers admitting a periodic Jacobi-Perron expansion (where later results sometimes comprise earlier ones as special cases). The proofs are mainly by direct calculation; GREITER[1] gave shorter proofs for some of Bernstein's results. Two examples are:

EXAMPLE 1. (BERNSTEIN[2]) Suppose $D \in \mathbb{N}$, $D > 1$, and let $\omega = \sqrt[3]{D^3 + 3D}$. The Jacobi-Perron expansion of $a(0) = (1, \omega, \omega^2)$ is periodic with lengths $k_0 = m = 4$ in (3.11).

EXAMPLE 2. (BERNSTEIN[3]) Let $D, d, n \in \mathbb{N}$ be such that $n \geq 2$, $d|D$, $n \leq d \leq D/n$. If $\omega = \sqrt[n+1]{D^{n+1} + dD}$, then the Jacobi-Perron expansion of $a(0) = (a_0(0), \ldots, a_n(0))$ with

$$a_k(0) = d^{-k} \sum_{j=0}^{k} \binom{n+j-k}{j} D^j \omega^{k-j}$$

is purely periodic with $m = n+1$ (or $m = 1$ if $D = d^n$).

Also D.N. LEHMER[1,3] studied periodicity. He started with the integers $b(i)$ and studied the $a(0)$ corresponding to them, but his results are not really new, compared with PERRON[1]. The fact that the Jacobi-Perron expansion of some $a(0)$ usually changes when the components are permuted -- remark (3) of section 3D, p. 33 -- received the attention of ELSNER & HASSE[1]. For some cubic roots ω where a short period for $a(0) = (1, \omega, \omega^2)$ is known from Bernstein's work, they calculated the expansions of its permutations and did not always obtain periodicity

within computer precision (about 100 steps). BERNSTEIN[10,11] also recognized the importance of the ordering and found a few periodic expansions with other permutations.

3G. Variations of the J.P.A.

Scepticism about the performance of the J.P.A. was expressed already in the nineteenth century by HERMITE[1] in a letter to Stieltjes. Several authors studied variations in definition 3.3, hoping to eliminate one shortcoming or another.

In the twenties DAUS[1,2] was looking for periodic expansions of bases of cubic fields. In definition 3.3 he usually takes the maximal value $b(i) = \left[\frac{a_s(i)}{a_t(i)}\right]$, but sometimes chooses a smaller value. Without obtaining or suggesting a general rule how to do this, he succeeds, apparently by trial and error, to find many periodic expansions, also in cases where the J.P.A. is not known to exhibit periodicity.

EXAMPLE. Let $\omega = \sqrt[3]{13} \approx 2.351335$. The successive vectors of cofactors $a(i)$ and quotient integers $b(i)$ are (with the cofactor that holds the t-index in italics):

$$a(0) = (\mathit{1}, \omega, \omega^2) \qquad b(0) = 2,\ b(1) = 5$$

$$a(2) = (1, \omega - \mathit{2}, \omega^2 - 5) \qquad b(2) = 1,\ b(3) = 2$$

$$a(4) = (-2\omega + 5, \omega - 2, \omega^2 - \omega - \mathit{3}) \qquad b(4) = 1,\ b(5) = 1$$

$$a(6) = (-\omega^2 - \omega + \mathit{8}, -\omega^2 + 2\omega + 1, \omega^2 - \omega - 3) \qquad b(6) = 1,\ b(7) = 1$$

$$a(8) = (-\omega^2 - \omega + 8, \mathit{3}\omega - \mathit{7}, 2\omega^2 - 11) \qquad b(8) = 1,\ b(9) = 2$$

$$a(10) = (-\omega^2 - 7\omega + 22, 3\omega - 7, \mathit{2}\omega^2 - \mathit{3}\omega - \mathit{4}) \qquad b(10) = 3,\ b(11) = 14$$

$$a(12) = (-\mathit{7}\omega^2 + \mathit{2}\omega + \mathit{34}, -28\omega^2 + 45\omega + 49, 2\omega^2 - 3\omega - 4) \qquad b(12) = 3,\ b(13) = 2$$

$$a(14) = (-7\omega^2 + 2\omega + 34, -\mathit{7}\omega^2 + \mathit{39}\omega - \mathit{53}, 16\omega^2 - 7\omega - 72)$$

Since $a_k(14) = \lambda a_{k+1}(4)$ for $k = 0,1,2$, with $\lambda = 2\omega^2 - 3\omega - 4$, the expansion

becomes periodic. All b(i) are greatest integer values except for $b(11) = 14$ where $\left[\frac{a_s(11)}{a_t(11)}\right] = 15$. The J.P.A. exhibits no periodicity for $(1, \omega, \omega^2)$ within 16-digit computer precision (BERNSTEIN[14]).

In 1975 GÜTING[1,3] revived an old algorithm of SANG[1]. One obtains this algorithm by interchanging the cyclic rotation of s and t in definition 3.3 of the J.P.A.:

(1) $i := s(0) := 0;\ t(0) := 1;$

(2) $A(i+1) := A(i)I^b_{st}(i)$ with $b = b(i) = \left[\frac{a_s(i)}{a_t(i)}\right];$

(3) <u>if</u> $t(i) + 1 = s(i)$ <u>then</u> $s(i+1) := s(i) + 1,\ t(i+1) := s(i+1) + 1;$ <u>else</u> $s(i+1) := s(i),\ t(i+1) := t(i);$

(4) $i := i+1;$ <u>goto</u> (2);

(with suitable modifications if the divisor $a_t(i)$ would be zero).

PERRON[1], p. 3, and D.H. LEHMER[1] had also touched upon this idea, but had not given it further attention. Güting intended his algorithm to be applied to a base with initial cofactors $a_0(0), a_1(0), \ldots, a_n(0)$ ordered by

$$(3.12) \qquad a_0(0) \geq a_1(0) \geq \ldots\ldots \geq a_n(0).$$

His main goal was to find integers $x_0, x_1, \ldots, x_n$ such that $\left|\sum_{j=0}^{n} x_j a_j(0)\right|$ is very small.

In fact, however, Güting's algorithm is obtained -- as far as the cofactors are concerned -- from Jacobi's merely by doing the steps in a different order. If for the moment we give cofactors and quotients b(i) from Güting's algorithm a double dash to distinguish them from those of Jacobi's, then the translation of a result of SCHWEIGER[13] to our notation is that

$$(3.13) \qquad b''_k = b_{f(k)} \quad \text{for } k \geq 0$$

where

$$f(k) = 2n - 1 + k + h(n-2), \text{ with } 0 \leq h \leq n-1,\ h \equiv k \pmod{n}.$$

The proof of (3.13) is by induction. In fact one has

$$(3.14)\qquad \begin{cases} a''_s(k) = a_s(f(k)) \\ a''_t(k) = a_t(f(k)) \end{cases}$$

for all $k \geq 0$. Here (3.14) not only expresses the numerical equality of the cofactors, but also that of the indices s and t chosen by Güting's and Jacobi's algorithm for the (k+1)-th and (f(k)+1)-th step, respectively. The function f(k) satisfies the recurrence formulae

$$\begin{cases} f(k+1) = f(k) + n - 1 & \text{if } k \not\equiv -1 \pmod{n} \\ f(k+1) = f(k) + 1 - (n-1)(n-2) & \text{if } k \equiv -1 \pmod{n}, \end{cases}$$

and covers all natural numbers exactly once, except for leaving out $\frac{1}{2}n(n+1)$ values $\leq n^2 - n$ (and for $j \notin f(\mathbb{N})$, the J.P.A. takes $b(j) = 0$ because of the ordering (3.12)). Therefore Güting's algorithm gives in its cofactors no information not obtainable from the J.P.A., and actually Güting gets even less information, since he takes into consideration the a"(ni) only.

Other authors left the frame of definition 2.3 and considered the J.P.A. as an algebraic operation, allowing negative cofactors or fractional b(i).

The nearest integer J.P.A., with $b(i) = \left[\frac{1}{2} + \frac{a_s(i)}{a_t(i)}\right]$, received the attention of CHABAUTY & PISOT[1] and of DAVID[1,2,4], who both proved its convergence when $n = 2$. It was extended to higher dimensions by TÖRNQVIST[1], but he left the requirement that the $A(i)$ be bases of Ω. For $n = 2,3$ his work was compared with other work by PIPPING[7,8]. GÜTING[2] considered his own algorithm also with nearest integer choice of b(i).

The difficulty of proving or disproving a general periodicity criterion for the J.P.A. led BERNSTEIN[5,9,15] to consider definition 3.3 purely algebraically. He noted that the algorithm might work as well with other values of b(i) than $\left[\frac{a_s(i)}{a_t(i)}\right]$, for instance with rational values, and he defined the "modified Jacobi-Perron algorithm" (M.J.P.A.) as follows. With all b(i) rational, all entries of a(i) belong to some $\mathbb{Q}(\omega)$ if those of a(0) do. Hence they are polynomial expressions $a_k(i)(\omega)$ of degree at most n in ω with rational coefficients. In the M.J.P.A. one chooses $b(i) = \frac{a_s(i)([\omega])}{a_t(i)([\omega])}$.

One result of the M.J.P.A. is the following theorem (BERNSTEIN[5]): Suppose $d \in \mathbb{N}$, not divisible by an $(n+1)$-th power greater than 1, and let $\omega = \sqrt[n+1]{d}$. Then the M.J.P.A.-expansion of $a(0) = (a_0(0), a_1(0), \ldots, a_n(0))$, where

$$a_k(0) = \sum_{j=0}^{k} \binom{n+j-k}{j} \omega^{k-j}[\omega]^j, \quad k = 0, \ldots, n,$$

is a purely periodic expansion of length $m = n+1$ in (3.11) (or $m = 1$ if $d-1$ is a perfect $(n+1)$-th power). This theorem contains results such as example 2 of section 3F, p. 36, as special cases, namely where the M.J.P.A. happens to coincide with the J.P.A.

<u>EXAMPLE.</u> Let $\omega = \sqrt[3]{17}$, $[\omega] = 2$. On applying the above theorem we get

$$a(0) = (1, \quad \omega + 4, \quad \omega^2 + 2\omega + 4) \qquad b(0) = 6, \ b(1) = 12$$

$$a(2) = (1, \quad \omega - 2, \quad \omega^2 + 2\omega - 8) \qquad b(2) = \tfrac{4}{3}, \ b(3) = 6$$

$$a(4) = (-\tfrac{4}{3}\omega + \tfrac{11}{3}, \quad \omega - 2, \quad \omega^2 - 4\omega + 4) \qquad b(4) = \tfrac{2}{3}, \ b(5) = \tfrac{4}{3}$$

$$a(6) = (-\tfrac{2}{3}\omega^2 + \tfrac{4}{3}\omega + 1, \ -\tfrac{4}{3}\omega^2 + \tfrac{19}{3}\omega - \tfrac{22}{3}, \ \omega^2 - 4\omega + 4)$$

and $a(6) = \lambda a(0)$ with $\lambda = -\frac{2}{3}\omega^2 + \frac{4}{3}\omega + 1$. The norm of λ is 1, but it is not an algebraic integer. The ordinary J.P.A. for $a(0) = (1, \omega, \omega^2)$ is periodic with $m = 61$, $k_0 = 32$ in (3.11) (BERNSTEIN[14]).

Bernstein's "B-algorithm" (BERNSTEIN[10]), of complicated formulation akin to that of the M.J.P.A., has integral $b(i)$, but it is by no means clear why they should be non-negative.

3H. Brun's algorithm.

In 1884 POINCARÉ[1] proposed the following algorithm for two dimensions. Let f,g,h be a permutation of $0,1,2$ such that $a_f \geq a_g \geq a_h$. Then do two subtractive steps, the first with $s = f$, $t = g$, and the second with $s = g$, $t = h$. Now redefine f,g,h and start anew. This algorithm describes an attractive geometric idea, namely it is equivalent to dividing the triangle with vertices A_0, A_1, A_2 into six triangles using the medians, and seeing

through which of these six triangles ℓ passes.

But the next example shows a defect of Poincaré's algorithm. Take $a(0) = (\alpha,\ 1,\ \xi^{-1})$ where $\xi = \frac{1}{2}(1+\sqrt{5})$ and $\alpha > 1+\xi^2$. An easy calculation shows that always $f=0$, while g equals 1 or 2, alternatingly. In fact, for all $k \geq 0$

$$a(2k) = (\alpha - \sum_{j=0}^{2k-1} \xi^{-j},\ \xi^{-2k},\ \xi^{-2k-1})$$

and

$$a(2k+1) = (\alpha - \sum_{j=0}^{2k} \xi^{-j},\ \xi^{-2k-2},\ \xi^{-2k-1}),$$

and we always have $a_0(k) > \alpha - \sum_{j=0}^{\infty} \xi^{-j} = \alpha - \xi^2 > 1$, so that the cofactors do not even tend to zero. The point $A_0(0)$ remains unchanged during the whole expansion.

Attempting to remove the defect, BRUN[1,2] proposed in 1919 the following simple subtractive algorithm first for $n=2$, later for general n.

<u>DEFINITION 3.4.</u> The algorithm

(1) $i := 0$;

(2) choose $s = s(i)$ and $t = t(i)$ from

$$(3.15) \qquad a_s(i) = \max_{0\leq k\leq n} a_k(i), \qquad a_t(i) = \max_{0\leq k\leq n,\, k\neq s} a_k(i);$$

(3) $A(i+1) := A(i)I_{st}^{b}(i)$ with $b=1$;

(4) $i := i+1$; <u>goto</u> (2);

is called *Brun's algorithm.*

That is, a_s is the largest, a_t the second largest cofactor. When $a_s(i) - a_t(i) \geq a_t(i)$, the pair s,t will remain unchanged for the next step, so that there is a division form of the algorithm; when $a_s(i) - a_t(i) < a_t(i)$ the previous t will become the next s. If (3.15) does not define s and t uniquely, one takes just any of the compatible choices.

Brun[2] studied the 2-dimensional case in detail. He called the $(i+1)$-th step of type

α if $a_s(i) - a_t(i) \geq a_t(i)$,

β if $a_t(i) > a_s(i) - a_t(i) \geq a_{3-t-s}(i)$,

$$\gamma \text{ if } a_{3-t-s}(i) > a_s(i) - a_t(i);$$

that is, if the new cofactor $a_s(i) - a_t(i)$ would still be the largest cofactor, or have become the middle or smallest cofactor, respectively. The expansion of a given a(0) can thus be described by a string of Greek symbols α, β, γ.

<u>EXAMPLE.</u> The expansion of $(1, \sqrt[3]{2}, \sqrt[3]{4})$ given in section 3C, p. 30, yields the periodic string

$$\gamma\ \gamma\alpha\alpha\gamma\beta\alpha\alpha\beta\gamma\alpha\alpha\gamma\beta\gamma\gamma\gamma\gamma\beta\ \gamma\alpha\alpha\gamma\beta\alpha\alpha\beta\gamma\alpha\alpha\gamma\beta\gamma\gamma\gamma\gamma\beta\ \gamma\alpha\alpha\gamma\beta\alpha\alpha\beta\gamma\alpha\alpha\gamma\beta\gamma\gamma\gamma\gamma\beta\ \ldots.$$

where the blocks of 18 characters repeat indefinitely.

Conversely Brun showed that any given string consisting of α's, β's and γ's (without restrictions, in contrast to the J.P.A.'s restrictions (3.8) and (3.9)) corresponds to a unique a(0), up to scalar multiplication and permutation of the indices. He also showed that ℓ is independent if and only if the string contains infinitely many γ's; that is, if and only if the algorithm never finds a cofactor equal to zero.

It follows from the boundedness of the distances to ℓ, in fact

$$\max_{k=0,1,2} d(A_k(i+1), \ell) \le \max_{k=0,1,2} d(A_k(i), \ell) \text{ for } i \ge 0, \tag{3.16}$$

that the 2-dimensional algorithm is weakly convergent (the proof of FOKKER[1] is incorrect). But, contrary to what FERGUSON & FORCADE[1] believe, it is not strongly convergent, as a counterexample shows (section 3I).

About periodicity properties of Brun's algorithm not much is known. SELMER[2] for instance, could not find a periodic expansion of $(1, \sqrt[3]{3}, \sqrt[3]{9})$ within 24-digit precision; the expansion of $(1, \sqrt[3]{4}, \sqrt[3]{16})$ is periodic with length 45, according to BERGMAN[1].

Much later, Brun[4] studied his algorithm in the 3-dimensional case, $n = 3$. With now obvious meaning for the symbols α, β, γ, δ he proved that an expansion containing infinitely many δ's but not infinitely many groups $\delta\gamma$, $\gamma\delta$ or $\delta\delta$, is always weakly convergent (the proof is analogous to (3.16)), and can belong only to an independent line.

<u>EXAMPLE.</u> The string $\delta\alpha\delta\alpha\alpha\delta\alpha\alpha\alpha\alpha\delta\alpha\alpha\alpha\alpha\alpha\delta\alpha\alpha\alpha\alpha\alpha\alpha\delta\ldots\ldots$ (with the number of

intermittent α's equal to 1,2,4,5,6,8,9,...., that is, all natural numbers not $\equiv 3 \pmod 4$) corresponds to

$$a(0) = (\cosh 1 + \cos 1,\ \sinh 1 + \sin 1,\ \cosh 1 - \cos 1,\ \sinh 1 - \sin 1).$$

Therefore these entries are rationally independent (BRUN[5]).

But it is well possible that the 3-dimensional algorithm fails to detect dependence:

EXAMPLE. The periodical string $\beta\delta\gamma\ \beta\delta\gamma\ \beta\delta\gamma$ ..... results from

$$a(0) = (2 - \lambda^2,\ 1,\ 1 - \lambda,\ \lambda)$$

when $\lambda \approx 0.3473$ is a root of $\lambda^3 - 3\lambda + 1 = 0$. But ℓ is contained in the plane with equation $x_2 = x_1 - x_3$ (with respect to the base $A(0)$), hence not independent (BRUN[4]).

Neither does the analogue of (3.16) hold when $n \geq 3$, as the example of $a(0) = (3\lambda^2,\ 4\lambda^2 + 1,\ \lambda^3 + \lambda,\ 3\lambda^3)$ with $\lambda^4 - 4\lambda^3 - 6\lambda^2 - \lambda - 1 = 0$ shows (BERGMAN[1]).

In 1977 GREITER[1] brought the state of knowledge about Brun's algorithm to the level of that of the J.P.A. when he proved:

THEOREM 3.5. *Brun's algorithm is weakly convergent in any dimension.* (GREITER[1], p. 27).

He also established the one-to-one correspondence between a(0) and its associated "string of Greek symbols" for all dimensions. Together with some other arguments this led him to enounce that Brun's algorithm were a more natural generalization of continued fractions than the J.P.A.

Since the original articles of Brun were written in Norwegian, it was to be expected that someone else would rediscover the simple algorithm, and actually this was done at least five times, namely by PALEY & URSELL[1], ROSSER[1], BARBOUR[1], VAUGHAN[2] and BERGMAN[1]. They all arrived at the division form of Brun's algorithm.

VAUGHAN[2] proved geometrically that, when $n = 2$,

$$\max_{k=0,1,2} \frac{d(A_k(i),\ell)}{|h(A_k(i))|} < c\left(\frac{2}{3}\right)^{[i/2]}$$

for some constant c, but a better estimate can be obtained directly from formula (3.16), by which Brun had proved weak convergence.

BARBOUR[1] studied a problem in music theory, famous since the days of Huygens, namely that of simultaneous rational approximations with small denominators to the so-called tuning ratios $\log\frac{3}{2} : \log 2$ and $\log\frac{5}{4} : \log 2$ of the fifth and major third respectively [1]). After applying both the J.P.A. and the inverted J.P.A. (where s and t move around cyclically in the direction opposite to that of definition 3.3) he concluded that a certain mixture of these gives more approximations with small integers, and should be preferred. This mixture is exactly Brun's algorithm. (An editorial note added to the paper of ROSSER[3] shows that Brun himself was aware of this).

DAVID[4], p. 44, also considered Brun's algorithm as a mixture of the two rotational directions of the J.P.A. Of course, this interpretation of Brun's algorithm is valid when $n = 2$ only.

In a slightly confusing paper, where many proofs are incomplete or absent, PALEY & URSELL[1] make some comments upon division algorithms in general, then they specialize to $n = 2$. Their arguments in favour of the second largest cofactor as the divisor $a_t(i)$ (i.e., Brun's algorithm) are, however, not very convincing.

BERGMAN[1] arrived at Brun's algorithm in the following manner. Suppose that we have some numbers whose ratio is an approximation of the ratio of small integers which are to be reconstructed. If we consider each of the given numbers as containing a small error term, we might expect to get the least build-up of errors by subtracting the second largest number from the largest.

3I. A non-convergent 2-dimensional Brun expansion.

We give a 2-dimensional Brun expansion which is not strongly convergent. We take $\Omega = \mathbb{Z}^3$ and $\ell_0 = (1, \xi, \eta)$ where $1 > \xi > \eta$ are such that the Brun expansion corresponds to the string of Greek symbols

$$\alpha^{\lambda_1-1}\beta\alpha^{\mu_1-1}\beta\gamma\beta\ \alpha^{\lambda_2-1}\beta\alpha^{\mu_2-1}\beta\gamma\beta\ \alpha^{\lambda_3-1}\beta\alpha^{\mu_3-1}\beta\gamma\beta\ \ldots\ldots\ldots,$$

---

[1]) The problem was also taken as an example by BRUN[6], PIPPING[8], ROSSER[3] and SANG[1].

where the λ_i, μ_i are large integers to be determined later on. The plane ℓ^* has equation $x + \xi y + \eta z = 0$. A simple calculation shows that with $A_0(0) = (1, 0, 0)$, $A_1(0) = (0, 1, 0)$ and $A_2(0) = (0, 0, 1)$ we have

$$(3.17) \qquad |\underline{A}_0(0)| : |\underline{A}_1(0)| : |\underline{A}_2(0)| = \sqrt{\xi^2 + \eta^2} : \sqrt{1 + \eta^2} : \sqrt{1 + \xi^2},$$

in which the underlined points are the projections, parallel to ℓ, of the non-underlined points on ℓ^*.

Now choose a positive $\varepsilon_0 < \frac{1}{10}$. We put

$$(3.18) \qquad M_0 = |\underline{A}_2(0)|.$$

Let us consider the transformation corresponding to the block $\alpha^{\lambda_1 - 1}\beta\alpha^{\mu_1 - 1}\beta\gamma\beta$ of our string. It consists of four steps with successively

(1) $s = 0$, $t = 1$, $b = \lambda_1$;
(2) $s = 1$, $t = 0$, $b = \mu_1$;
(3) $s = 0$, $t = 1$, $b = 1$;
(4) $s = 1$, $t = 2$, $b = 1$.

From (1) we conclude, using (2.7), that $a_0(0) \geq \lambda_1 a_1(0)$, or $1 \geq \lambda_1 \xi$. With (3.17) this gives

$$(3.19) \qquad |\underline{A}_0(0)| < \varepsilon_0 M_0$$

if λ_1 is large enough (e.g. $\lambda_1 > \varepsilon_0^{-1}/2$), which we assume to be the case.

The algorithm finds successively (see figure 3.1, p. 46):

$$A_0(0) = A_0(1);\ A_0(2) = A_0(3) = A_0(4) = A_0(1) + \mu_1 A_1(1);$$
$$A_1(1) = A_1(2) = A_1(0) + \lambda_1 A_0(0);\ A_1(3) = A_1(4) = A_1(2) + A_0(2);$$
$$A_2(0) = A_2(1) = A_2(2) = A_2(3);\ A_2(4) = A_2(3) + A_1(3).$$

We derive the following (in-)equalities from the table of four steps and the definition of the algorithm of Brun:

$$a_0(0) > a_1(0) > a_2(0),$$
$$a_1(1) > a_0(1) > a_2(1),$$

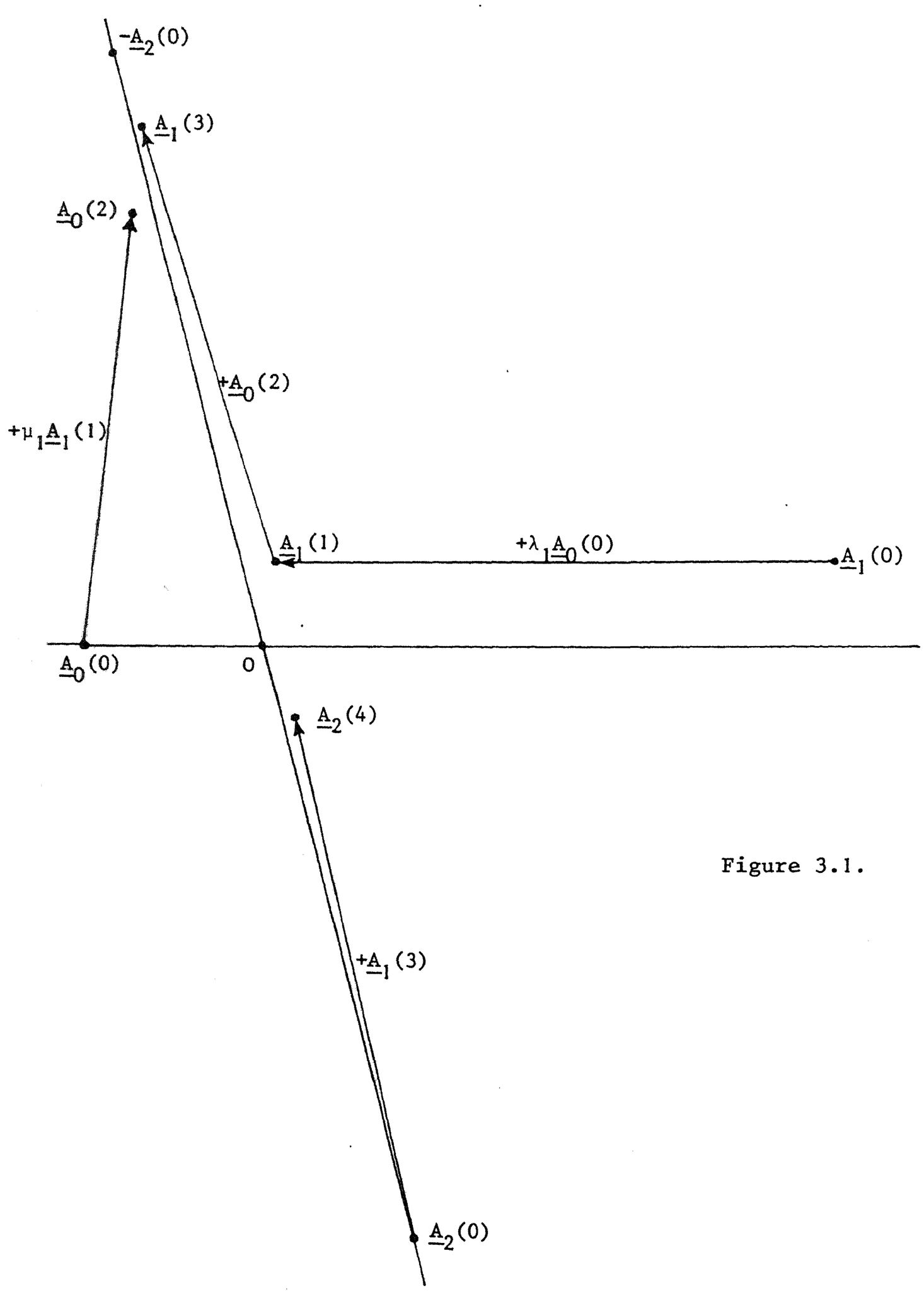

$-\underline{A}_2(0)$
$\underline{A}_1(3)$
$\underline{A}_0(2)$
$+\underline{A}_0(2)$
$+\mu_1\underline{A}_1(1)$
$\underline{A}_1(1)$
$+\lambda_1\underline{A}_0(0)$
$\underline{A}_1(0)$
$\underline{A}_0(0)$
0
$\underline{A}_2(4)$
$+\underline{A}_1(3)$
$\underline{A}_2(0)$

Figure 3.1.

$$a_0(2) > a_1(2) > a_2(2),$$
$$a_1(3) > a_2(3) > a_0(3),$$
$$a_2(4) > a_1(4) > a_0(4);$$

$$\begin{array}{lll} a_0(1) = a_0(0) - \lambda_1 a_1(0), & a_1(1) = a_1(0), & a_2(1) = a_2(0), \\ a_0(2) = a_0(1), & a_1(2) = a_1(1) - \mu_1 a_0(1), & a_2(2) = a_2(1), \\ a_0(3) = a_0(2) - a_1(2), & a_1(3) = a_1(2), & a_2(3) = a_2(2), \\ a_0(4) = a_0(3), & a_1(4) = a_1(3) - a_2(3), & a_2(4) = a_2(3). \end{array}$$

Since $\mu_1 = \left[\frac{a_1(1)}{a_0(1)}\right]$ we have $\frac{a_0(1)}{a_1(1)} > \frac{1}{1+\mu_1}$. Because of $a_2(1) = a_2(2) < a_1(2) = a_1(1) - \mu_1 a_0(1)$ this gives

$$\frac{a_2(1)}{a_1(1)} < \frac{1}{1+\mu_1}. \tag{3.20}$$

The fifth step being with $s = 2$, $t = 1$, $b = \lambda_2$ we see that $\lambda_2 a_1(4) < a_2(4)$. This gives

$$a_1(2) = a_1(3) = a_1(4) + a_2(3) = a_1(4) + a_2(4) < \left(1 + \frac{1}{\lambda_2}\right) a_2(4)$$

and also

$$a_0(2) - a_1(2) = a_0(3) = a_0(4) < a_1(4) < \frac{1}{\lambda_2}\, a_2(4).$$

Combining, we get

$$a_0(2) < a_1(2) + \frac{1}{\lambda_2} a_2(4) < \left(1 + \frac{2}{\lambda_2}\right) a_2(4) = \left(1 + \frac{2}{\lambda_2}\right) a_2(2),$$

whence

$$\frac{a_2(2)}{a_0(2)} > 1 - \frac{2}{\lambda_2}. \tag{3.21}$$

Now we find, using $a_1(1) > a_0(1)$, (3.19) and (3.20), that

$$|\underline{A}_1(1)| = \left| -\frac{a_0(1)}{a_1(1)} \underline{A}_0(1) - \frac{a_2(1)}{a_1(1)} \underline{A}_2(1) \right| < \left(\varepsilon_0 + \frac{1}{1+\mu_1}\right) M_0.$$

This gives, with (3.21),

$$|\underline{A}_0(4)| = |\underline{A}_0(2)| = \left| -\frac{a_2(2)}{a_0(2)}\,\underline{A}_2(2) - \frac{a_1(2)}{a_0(2)}\,\underline{A}_1(2)\right|$$

$$> \left(1 - \frac{2}{\lambda_2} - \varepsilon_0 - \frac{1}{1+\mu_1}\right) M_0 .$$

Also we find

$$\underline{A}_2(4) = \underline{A}_2(2) + \underline{A}_1(2) + \underline{A}_0(2) =$$

$$= \underline{A}_2(2) + \underline{A}_1(2) - \frac{a_2(2)}{a_0(2)}\,\underline{A}_2(2) - \frac{a_1(2)}{a_0(2)}\,\underline{A}_1(2),$$

so that (from (3.21) and $a_1(2) > a_2(2)$)

$$|\underline{A}_2(4)| < \frac{2}{\lambda_2}|\underline{A}_2(2)| + \frac{2}{\lambda_2}|\underline{A}_1(2)| < \frac{2}{\lambda_2}\left(1 + \varepsilon_0 + \frac{1}{1+\mu_1}\right) M_0 .$$

Putting $M_1 = |\underline{A}_0(4)|$ we have

$$|\underline{A}_2(4)| < \varepsilon_1 M_1$$

with

$$\varepsilon_1 = \frac{\frac{2}{\lambda_2}\left(1 + \varepsilon_0 + \frac{1}{1+\mu_1}\right)}{1 - \frac{2}{\lambda_2} - \varepsilon_0 - \frac{1}{1+\mu_1}} ,$$

and

$$M_1 > (1 - \eta_0) M_0$$

with $\eta_0 = \frac{2}{\lambda_2} + \varepsilon_0 + \frac{1}{1+\mu_1}$. Now we assume that μ_1 and λ_2 are so large that $\varepsilon_1 < \varepsilon_0^2$ and $\eta_0 < 3\varepsilon_0$. The base $A(4)$ has essentially the same properties as $A(0)$ (with the indices 0 and 2 interchanged), namely to (3.18) there corresponds

$$M_1 = |\underline{A}_0(4)|,$$

to (3.19)

$$|\underline{A}_2(4)| < \varepsilon_1 M_1 .$$

Therefore the effect of the next block of Greek symbols, $\alpha^{\lambda_2-1}\beta\alpha^{\mu_2-1}\beta\gamma\beta$, is identical if we choose μ_2 and λ_3 large enough, and so on. Continuing we obtain sequences of real numbers $M_0, M_1, M_2, \ldots$ and $\varepsilon_0, \varepsilon_1, \varepsilon_2, \ldots$ such that

$$\begin{cases} |\underline{A}_0(4k)| = M_k, & |\underline{A}_2(4k)| < \varepsilon_k M_k \quad (k \text{ odd}) \\ |\underline{A}_2(4k)| = M_k, & |\underline{A}_0(4k)| < \varepsilon_k M_k \quad (k \text{ even}) \end{cases}$$

and that

$$\varepsilon_{k+1} < \varepsilon_k^2, \quad M_{k+1} > M_k(1 - \eta_k)$$

for $k \geq 0$, where $\eta_k < 3\varepsilon_k < 3\varepsilon_0^{2^k} < (3\varepsilon_0)^{2^k}$, all this being true under the condition that the λ_j, μ_j are sufficiently large. Since $3\varepsilon_0 < \frac{1}{3}$ we conclude that

$$\begin{aligned} M_k &\geq M_0(1-\eta_0)(1-\eta_1)\ldots\ldots(1-\eta_k) \\ &> M_0(1-\tfrac{1}{3})(1-(\tfrac{1}{3})^2)(1-(\tfrac{1}{3})^4)\ldots\ldots(1-(\tfrac{1}{3})^{2^k}) \\ &> M_0(1-\tfrac{1}{3}-(\tfrac{1}{3})^2-(\tfrac{1}{3})^3-\ldots\ldots) = \tfrac{1}{2}M_0, \end{aligned}$$

from which we see that our expansion is not convergent.

3J. Some other algorithms.

In connection with Brun's algorithm, SELMER[2] studied the subtractive variation where one lets $a_s(i)$ be the largest cofactor of $A(i)$, $a_t(i)$ the smallest non-zero cofactor (instead of the second largest). It was hoped that this algorithm would yield more approximations because the cofactors tend to zero more slowly. Selmer showed that, from some moment on, the new cofactor $a_s(i) - a_t(i)$ will be the smallest or next-smallest cofactor of $A(i+1)$, so that a string of Greek symbols associated to the expansion in the same way as was done for Brun's algorithm, will contain thereafter only the last two symbols employed. Though this might be expected to give better periodicity results than

Brun's algorithm, there seems to be no essential improvement.

PIPPING[1,2,3] studied the following subtractive *branching* algorithm. From the base $A(0)$ one obtains n new bases $A(1)$ by taking for $a_s(0)$ the largest cofactor of $A(0)$ and for $a_t(0)$ successively all other cofactors. Proceeding in the same way, one produces n bases $A(2)$ from each $A(1)$, so that there are n^2 bases $A(2)$, and so on. Pipping proved: The line ℓ is dependent if and only if for some k one of the n^k bases $A(k)$ has two equal cofactors. In PIPPING[6] he improved the algorithm slightly by showing that the same result holds when one takes only n−1 new bases, omitting the choice of the smallest cofactor as a_t. In particular this establishes again the fact that Brun's algorithm for $n = 2$ will detect any dependence relation. TÖRNQVIST[1] studied further improvements when $n = 3$. For general n the expanding calculations would soon become prohibitive, and Pipping's work has as its principal merit that one can prove dependence by choosing the right path through the branching tree. His proofs are constructive if a dependence relation were known beforehand. In later papers PIPPING[7,8] also tried, in the case $n = 3$, to give criteria for the choice of a path in the tree when one has no previous knowledge of a dependence relation.

Finally we mention the strictly 2-dimensional subtractive algorithm of KOLLROS[1,2], who proposed a 2-dimensional analogue of the function $y = ?(x)$ of MINKOWSKI[3]:

(1) $i := 0;\ t(-1) := 0;$

(2) choose $s = s(i)$ and $t = t(i)$ such that $s \neq t(i-1)$, $t \neq t(i-1)$ and $a_s(i) \geq a_t(i)$;

(3) $A(i+1) := A(i)I_{st}(i);\ i := i+1;$ <u>goto</u> (2).

<u>EXAMPLE.</u> Take $a(0) = (1, \omega, \omega^2)$ with $\omega = \sqrt[3]{2}$. The cofactors of the expansion are

$$
\begin{array}{ll}
 & t(-1) = 0 \\
a(0) = (1 \quad , \omega \quad , \omega^2 \quad) & s(0) = 2,\ t(0) = 1 \\
a(1) = (1 \quad , \omega \quad , \omega^2 - \omega) & s(1) = 0,\ t(1) = 2 \\
a(2) = (1 + \omega - \omega^2, \omega \quad , \omega^2 - \omega) & s(2) = 1,\ t(2) = 0 \\
a(3) = (1 + \omega - \omega^2, \omega^2 - 1, \omega^2 - \omega) & s(3) = 1,\ t(3) = 2 \\
a(4) = (1 + \omega - \omega^2, \omega - 1 \ , \omega^2 - \omega) & s(4) = 0,\ t(4) = 1 \\
a(5) = (2 - \omega^2 \quad , \omega - 1 \ , \omega^2 - \omega). &
\end{array}
$$

We find $a(5) = (\omega - 1)(\omega^2, 1, \omega)$, so that the expansion becomes periodic. The algorithm can also be periodic for bases of cubic fields with positive discriminant. The following example shows that the algorithm can fail to detect a dependence relation (though one can prove that it is weakly convergent):

$$t(-1) = 0$$

$$\begin{aligned}
a(0) &= (1 \quad , 2-\sqrt{2} \ , \sqrt{2}-1) & s(0) &= 1, \ t(0) = 2 \\
a(1) &= (1 \quad , 3-2\sqrt{2}, \sqrt{2}-1) & s(1) &= 0, \ t(1) = 1 \\
a(2) &= (2\sqrt{2}-2, 3-2\sqrt{2}, \sqrt{2}-1) & s(2) &= 0, \ t(2) = 2 \\
a(3) &= (\sqrt{2}-1 \ , 3-2\sqrt{2}, \sqrt{2}-1) & s(3) &= 0, \ t(3) = 1 \\
a(4) &= (3\sqrt{2}-4, 3-2\sqrt{2}, \sqrt{2}-1).
\end{aligned}$$

Since $a(4) = (\sqrt{2}-1)(2-\sqrt{2}, \sqrt{2}-1, 1)$ the expansion becomes periodic and no cofactor will ever be zero.

CHAPTER 4. INTRODUCTION TO SOME APPROXIMATION PROBLEMS

4A. Linear dependence as an approximation problem.

Theorem 2.2 says that all points of the lattice Ω that are sufficiently close to the line ℓ lie in a certain $(n-r+1)$-dimensional subspace, where r is the dependence rank of ℓ on Ω. This explains the connection between dependence and the approximation of ℓ by lattice points. In theory, a strongly convergent algorithm can also be used to prove that a given line is independent (lemma 2.4). But in practice, apart from special cases like periodic expansions, it is impossible to know a strongly convergent expansion in a finite number of steps. Therefore we shall derive a quantitative result concerning the possibility of dependence when just one base of Ω is given. This we do for the particular case where it is asked whether given positive real numbers $1, \xi_1, \ldots\ldots, \xi_n$ are rationally independent.

We take $\ell_0 = (1, \xi_1, \ldots\ldots, \xi_n)$ and $\Omega = \mathbb{Z}^{n+1}$. Any non-zero vector $c = (c_0, c_1, \ldots\ldots, c_n) \in \mathbb{Z}^{n+1}$ is called a *$\mathbb{Z}$-relation* for ℓ_0 if $c.\ell_0 = 0$, that is

$$c_0 + \sum_{j=1}^{n} c_j \xi_j = 0.$$

THEOREM 4.1. *If $A = \{A_0, A_1, \ldots\ldots, A_n\}$ is a base of $\mathbb{Z}^{n+1}$ and*

$$\max_{0 \le k \le n} d(A_k, \ell) < \varepsilon, \tag{4.1}$$

then any $\mathbb{Z}$-relation c for ℓ_0 satisfies $|c| = \sqrt{c_0^2 + c_1^2 + \ldots\ldots + c_n^2} > 1/\varepsilon$.

PROOF. For at least one A_k we have $c.A_k \neq 0$, hence $|c.A_k| \ge 1$ because c and A_k are integral vectors. On the other hand, by $c.\ell_0 = 0$ we have

$$|c.A_k| = |c.\underline{A}_k| \le |c|\,|\underline{A}_k| < |c|\varepsilon,$$

when $\underline{A}_k$ is the projection of A_k on ℓ^* parallel to ℓ. Combining, we get $|c| > 1/\varepsilon$. □

In section 6B we define a strongly convergent algorithm with the property that it always finds a $\mathbb{Z}$-relation if one exists. In combination with this algorithm, theorem 4.1 allows us to give lower bounds on the lengths of possible $\mathbb{Z}$-relations. It also allows us to solve dependence questions in which an upper bound is known in the sense that there is a number M such that either a $\mathbb{Z}$-relation c with $|c| \leq M$ or no $\mathbb{Z}$-relation at all exists.

An example of the latter type is the following. Let $f(x) \in \mathbb{Z}[x]$ be a polynomial of degree n+1 with at least one real root α. We wish to decide whether f(x) is irreducible, and if not, find a factor g(x) of smaller degree. Such a factor is a $\mathbb{Z}$-relation for $\ell_0 = (1, \alpha, \ldots, \alpha^n)$ with $\Omega = \mathbb{Z}^{n+1}$. A rough upper bound M on the relation follows from:

<u>LEMMA 4.2.</u> *Let $f(x)$, $g(x) \in \mathbb{Z}[x]$ be monic polynomials with $f(0) \neq 0$ and $1 \leq \deg(g) < \deg(f) = n+1$, such that g divides f. Then*

$$\text{height}(g) < 2^n \cdot \max_{1 \leq j \leq n+1} |\alpha_j|^n < 2^n(1 + \text{height}(f))^n, \tag{4.5}$$

where the α_j are the (real and complex) roots of $f(x) = 0$.

<u>PROOF.</u> Suppose the α_j are numbered such that $f(x) = \prod_{j=1}^{n+1} (x - \alpha_j)$, $g(x) = \prod_{j=1}^{m} (x - \alpha_j)$ $(1 \leq m < n+1)$. If $g(x) = \sum_{k=0}^{m} g_k x^k$, then g_k can be expressed as the (m-k)-th symmetric function of $\alpha_1, \ldots, \alpha_m$ and we have for $k = 0, 1, \ldots, m$:

$$|g_k| \leq \binom{m}{k} \max_{1 \leq j \leq m} |\alpha_j|^{m-k} < 2^m \max_{1 \leq j \leq m} |\alpha_j|^m, \tag{4.6}$$

from which the first inequality in (4.5) follows. The second inequality holds because one always has $|\alpha_j| < 1 + \text{height}(f)$. □

The bound (4.5) can be reduced, e.g by observing that $g_0 \mid \prod_{j=1}^{n+1} \alpha_j$, or by using specific information about f(x). This does not, however, concern us now.

4B. Best approximations.

In the preceding section any convergent algorithm could be used to tackle a dependence problem, and only practical arguments would favour a

fast rate of convergence. We now turn our attention to problems where our algorithms must have a good rate of convergence. Before doing so, we recall the well-known lattice point theorem of Minkowski:

THEOREM 4.3. *The volume vol(K) of a convex body K in* $\mathbb{R}^{n+1}$, *which is symmetric about 0 and contains no other lattice point, satisfies*

$$(4.7) \qquad vol(K) \le 2^{n+1}.\Delta,$$

where $\Delta = vol(\{\sum_{j=0}^{n} x_j A_j \mid 0 \le x_j \le 1\})$ *is the volume of a base* $\{A_0, A_1, \ldots, A_n\}$ *of* Ω.

On $\mathbb{R}^{n+1}$, let the *height function* h be a linear function with $h(\ell_0) > 0$, and such that $h = 0$ on the orthogonal subspace ℓ^*. Usually we shall normalize h by $h(\ell_0) = |\ell_0|$, in which case we have

$$d(X, \ell^*) = |h(X)| \quad \text{for all } X \in \mathbb{R}^{n+1}.$$

The *radius function* ρ measures the distance to ℓ in some norm. It must of course satisfy

$$(4.8) \qquad \begin{cases} \text{(a)} & \rho(X) \ge 0, \text{ and } \rho(X) = 0 \Leftrightarrow X \in \ell, \text{ for } X \in \mathbb{R}^{n+1}, \\ \text{(b)} & \rho(\lambda X) = |\lambda| \rho(X), \text{ for } X \in \mathbb{R}^{n+1}, \lambda \in \mathbb{R}, \\ \text{(c)} & \rho(X+Y) \le \rho(X) + \rho(Y), \text{ for } X, Y \in \mathbb{R}^{n+1}, \\ \text{(d)} & \rho(X + \lambda \ell_0) = \rho(X), \text{ for } X \in \mathbb{R}^{n+1}, \lambda \in \mathbb{R}, \end{cases}$$

but no other properties are required yet. The restriction of ρ to ℓ^* is, of course, a norm on ℓ^* and every norm on ℓ^* induces a radius function by (4.8)(d). Notice that (4.8)(d) is an easy consequence of (a), (b) and (c), and is added merely to elucidate the concept.

Since we are dealing with finite-dimensional real spaces, we know that every two radius functions ρ_1, ρ are absolutely continuous with respect to each other; that is, there exist positive constants c_1, c_2 depending on ρ_1, ρ only, such that

$$(4.9) \qquad c_1 \rho(X) \le \rho_1(X) \le c_2 \rho(X) \quad \text{for all } X \in \mathbb{R}^{n+1}.$$

DEFINITION 4.1. A lattice point $B \neq 0$ is a *best approximation* of Ω to ℓ

(with respect to ρ and h) if there is no lattice point $P \neq 0$ for which

$$(4.10) \qquad \begin{cases} \rho(P) \leq \rho(B) \text{ and } |h(P)| < |h(B)| \\ \text{or } \rho(P) < \rho(B) \text{ and } |h(P)| = |h(B)|. \end{cases}$$

This generalizes the 1-dimensional concept of section 1B (formula (1.10)), which results from (4.10) if we put $h(x,y) = x$, $\rho(x,y) = |x\alpha - y|$, and take the inner product such that ℓ_0 and $(0,1)$ are orthogonal.

Note that in the case $n = 1$ the radius function is unique up to multiplication by a positive constant, since on ℓ^* (which now is a line) essentially one metric is possible. So if $n = 1$, the addition "with respect to ρ" in definition 4.1 is superfluous. This changes drastically when $n \geq 2$.

We define the ρ-body $K_\rho(P)$ of a point P to be the set

$$K_\rho(P) = \{X \in \mathbb{R}^{n+1} \mid \rho(X) \leq \rho(P),\ |h(X)| \leq |h(P)|\}.$$

This body is convex and symmetric about the origin, and its volume is given by

$$(4.11) \qquad \mathrm{vol}(K_\rho(P)) = c_{\rho h}(\rho(P))^n |h(P)|$$

because of (4.8)(b), where the constant $c_{\rho h}$ is

$$c_{\rho h} = \mathrm{vol}(\{X \in \mathbb{R}^{n+1} \mid \rho(X) \leq 1,\ |h(X)| \leq 1\}).$$

<u>EXAMPLE.</u> If ρ is the Euclidean radius in $\mathbb{R}^3$, $\rho(X) = d(X,\ell)$, and $|h(X)| = d(X,\ell^*)$, then $c_{\rho h} = 2\pi$.

If B is a best approximation, then every lattice point $P \neq 0$ in $K_\rho(B)$ satisfies $\rho(P) = \rho(B)$ and $h(P) = \pm h(B)$; usually the only such P will be B and $-B$. It follows from theorem 4.3 that

$$(4.12) \qquad \mathrm{vol}(K_\rho(B)) \leq 2^{n+1}.\Delta.$$

Now suppose we increase the height of $K_\rho(B)$, keeping its radius fixed at $\rho(B)$. By Minkowski's theorem we must hit a new lattice point B_1, not on

the edge (i.e., $\rho(B_1) < \rho(B)$), before the volume of the expanding body exceeds $2^{n+1}.\Delta$. That is,

$$(4.13) \qquad |h(B_1)| < \frac{2^{n+1}.\Delta}{c_{\rho h}(\rho(B))^n} .$$

Let B_1 be the lowest such point (= least positive height) and if there are several points of equal height, let B_1 be one among them of least radius. It is clear then that B_1 is a best approximation and also that there is no best approximation B' with $|h(B)| < |h(B')| < |h(B_1)|$, for we should have hit B' before B_1 in the expanding process. We call B_1 the (or a, if there is still a choice) *successor* of B. In similar fashion the *predecessor* of B is found by increasing the radius of $K_\rho(B)$ while keeping its height constant.

The set of all best approximations with positive height can thus be ordered by increasing height to form the *sequence of best approximations*, in which each point is the successor of the preceding one. This sequence is infinite in the direction of increasing height if and only if ℓ is not a rational line (cf. (4.13)). It is infinite in the direction of decreasing height if and only if ℓ^* contains no lattice point except the origin.

With δ and U as in theorem 2.2, p. 17, we have:

LEMMA 4.4. *When ℓ is a dependent line, the best approximation B satisfies $B \in U$ if*

$$|h(B)| > \frac{2^{n+1}.\Delta}{c_{\rho h}\delta^n c_1^n} ,$$

where $c_1 > 0$ is such that $\rho(X) \geq c_1 d(X,\ell)$ for all $X \in \mathbb{R}^{n+1}$ (cf. (4.9)).

PROOF. Immediate from (4.7), (4.11) and theorem 2.2. □ (A coordinate version of this lemma is found in MACK[1]).

Consequently best approximations will mainly be of interest when ℓ is independent, for otherwise we could diminish the dimension by considering ℓ with respect to the lattice $\Omega \cap U$.

For a more detailed discussion of the sequence of best approximations with respect to arbitrary ρ we refer to LAGARIAS[1,2,3].

4C. Periodicity of best approximations.

In this section we assume that ℓ is independent. Also we assume the height function h to be normalized by $h(\ell_0) = |\ell_0|$.

DEFINITION 4.2. The sequence $\{B_i \mid -\infty < i < \infty\}$ of best approximations of positive height is periodic if there is a unimodular linear transformation T of Ω onto itself, such that for all i

$$(4.14) \qquad TB_i = B_{i+m}$$

for some $m > 0$.

In an elementary way we shall derive a necessary condition for the existence of periodic sequences of best approximations. We then invoke a result of Minkowski and find that the condition can be satisfied only in very special cases when $n = 1$ or $n = 2$.

LEMMA 4.5. *If the transformation T has property (4.14), then ℓ_0 is an eigenvector of T with positive eigenvalue $\tau_0 > 1$.*

PROOF. First we observe that:

(1) for any k_0 there exist n+1 linearly independent vectors among the B_k $(k \geq k_0)$, because ℓ is independent;

(2) it follows from theorem 4.3 and the discreteness of Ω that for any k_0

$$\liminf_{k \geq k_0} \; (h(B_{k+1}) - h(B_k)) = \infty.$$

Now put $\|T\| = \max_{\rho(X)=1} \rho(TX)$ and let C_k be the projection, parallel to ℓ^*, of B_k on ℓ, so that $h(B_k) = h(C_k)$. Write $C_k = \lambda_k \ell_0$ and put $\delta = \rho(T\ell_0)$. We shall show that $\delta = 0$. Suppose to the contrary that $\delta > 0$ and select k such that

$$\rho(B_k) < \delta(1 + \|T\|)^{-1},$$

and $\lambda_k > 1$. Then

$$\begin{aligned} \delta < \lambda_k \delta = \rho(TC_k) &\leq \rho(TC_k - TB_k) + \rho(B_{k+m}) \leq \|T\|\rho(B_k) + \rho(B_{k+m}) \\ &\leq (1 + \|T\|)\rho(B_k) < \delta, \end{aligned}$$

a contradiction. Therefore $\delta = 0$ and ℓ_0 is an eigenvector of T. Let τ_0 be the corresponding eigenvalue. Define

$$\overline{|T|} = \max_{h(X)=0,\rho(X)=1} |h(TX)|.$$

Select k such that $h(B_{k+m}) - h(B_k) > \overline{|T|}\rho(B_k)$, which is possible by observation (2). Now $\tau_0 \le 1$ would give

$$h(B_{k+m}) - h(B_k) = h(TB_k) - h(C_k) \le h(TB_k) - \tau_0 h(C_k)$$
$$= h(TB_k) - h(TC_k) = h(T(B_k - C_k)) \le \overline{|T|}\rho(B_k),$$

a contradiction. Hence $\tau_0 > 1$. □

Now let $f(\tau) = 0$ be the characteristic equation of T. Let its real and complex roots be $\tau_0, \tau_1, \ldots, \tau_n$ with τ_0 as in lemma 4.5, so that $\tau_0 \in \mathbb{R}$.

<u>LEMMA 4.6.</u> *If* $\tau_j \ne \tau_0$ *then* $|\tau_j| \le \tau_o^{-1/n}$.

<u>PROOF.</u> Choose $0 < \varepsilon < \frac{1}{2}(\tau_0 - 1)$. Let $\mathcal{B} = \{B_{i_0}, B_{i_1}, \ldots, B_{i_n}\}$ be an independent set of best approximations with $0 < h(B_{i_0}) < \ldots < h(B_{i_n})$ and with B_{i_0} already so high that

$$\overline{|T|}\rho(B_{i_0}) < \varepsilon h(B_{i_0}).$$

Then

$$\overline{|T|}\rho(B_i) < \varepsilon h(B_i)$$

holds for all best approximations B_i with $i \ge i_0$. We calculate

$$|h(TB_i) - \tau_0 h(B_i)| = |h(TB_i) - \tau_o h(C_i)| = |h(TB_i - TC_i)|$$
$$\le \overline{|T|}\rho(B_i) < \varepsilon h(B_i),$$

for $i \ge i_0$, which gives

$$\frac{h(TB_i)}{h(B_i)} > \tau_0 - \varepsilon \quad (i \ge i_0). \tag{4.15}$$

From (4.11), (4.12) we infer that for $0 \le k \le n$ and for all $\nu > 0$

$$\rho(T^\nu B_{i_k}) \le \sqrt[n]{\frac{2^{n+1}.\Delta}{c_{\rho h} h(T^\nu B_{i_k})}} < \sqrt[n]{\frac{2^{n+1}.\Delta}{c_{\rho h}(\tau_0 - \varepsilon)^\nu h(B_{i_k})}} .$$

Let X be an arbitrary vector. Writing $X = \sum_{k=0}^{n} \xi_k B_{i_k}$ we have, with $\Xi = \max_{\rho(X)=1} \sum_{k=0}^{n} |\xi_k|$,

$$\rho(T^\nu X) \le \Xi \sqrt[n]{\frac{2^{n+1}.\Delta}{c_{\rho h}(\tau_0 - \varepsilon)^\nu h(B_{i_0})}} \cdot \rho(X),$$

and therefore

$$\rho(T^\nu X) < (\tau_0 - 2\varepsilon)^{-\nu/n}.\rho(X) \tag{4.16}$$

when ν is large enough, say $\nu \ge \nu_0 = \nu_0(\varepsilon)$.

In particular, let X be an eigenvector of T with real eigenvalue $\tau_j \ne \tau_0$. Then $\rho(X) \ne 0$, whence (4.16) yields $|\tau_j|^\nu < (\tau_0 - 2\varepsilon)^{-\nu/n}$, or

$$|\tau_j| < (\tau_0 - 2\varepsilon)^{-1/n}. \tag{4.17}$$

If τ_j is a complex eigenvalue, write $\tau_j = |\tau_j|(\cos\phi + i\sin\phi)$. Then there exist vectors Y and Z such that

$$T^\nu Y = |\tau_j|^\nu (Y\cos\nu\phi - Z\sin\nu\phi)$$
$$T^\nu Z = |\tau_j|^\nu (Y\sin\nu\phi + Z\cos\nu\phi).$$

Suppose that $\rho(Y) \ge \rho(Z)$ (the other case is similar) and choose ν such that not only $\nu \ge \nu_0$ but also $\cos\nu\phi > 1 - \frac{1}{2}\varepsilon$, and hence $|\sin\nu\phi| < \sqrt{\varepsilon}$. Then

$$\rho(Y\cos\nu\phi - Z\sin\nu\phi) \ge (1 - \tfrac{1}{2}\varepsilon - \sqrt{\varepsilon})\rho(Y)$$

and (4.16) applied to Y yields $|\tau_j|^\nu < \dfrac{(\tau_0 - 2\varepsilon)^{-\nu/n}}{1 - \frac{1}{2}\varepsilon - \sqrt{\varepsilon}}$,

whence

$$|\tau_j| < \frac{(\tau_0 - 2\varepsilon)^{-1/n}}{1 - \frac{1}{2}\varepsilon - \sqrt{\varepsilon}} . \tag{4.18}$$

Now (4.17) and (4.18) give the desired result if we let $\varepsilon \to 0$. □

Taking (4.16) with $\varepsilon < \frac{1}{2}(\tau_0 - 1)$ we see that for $\nu \geq \nu_0(\varepsilon)$

(4.19) $\qquad \rho(T^\nu X) < \rho(X) \quad$ for all $X \in \mathbb{R}^{n+1}$.

Let B be a best approximation such that

$$\overline{|T^\nu|}\rho(B) < \varepsilon h(B).$$

Define D to be the region $D = \{X \in \mathbb{R}^{n+1} \mid \rho(X) \leq \rho(B),\ h(X) \geq h(B)\}$. Then we have

$$\overline{|T^\nu|}\rho(X) < \varepsilon h(X) \quad \text{for all } X \in D.$$

For $X \in D$, let Y be the projection (parallel to ℓ^*) of X on ℓ. Then

$$\begin{aligned} |h(T^\nu X) - \tau_0^\nu h(X)| &= |h(T^\nu X) - h(\tau_0^\nu Y)| = |h(T^\nu(X-Y))| \\ &\leq \overline{|T^\nu|}\rho(X) < \varepsilon h(X), \end{aligned}$$

from which

(4.20) $\qquad h(T^\nu X) > (\tau_0^\nu - \varepsilon)h(X) \quad$ for $X \in D$.

By (4.19), (4.20) we have $T^\nu D \subset D$. Let $\mathcal{X} = \{X_0, X_1, \ldots, X_n\}$ be a base of $\mathbb{R}^{n+1}$, not necessarily consisting of lattice points, but such that $X_k \in D$, $k = 0,1,\ldots,n$, and that

$$\theta := \inf\ \{\rho(X) \mid h(X) = 1,\ X \notin \{\mathcal{X}\}^+\} > 0$$

(that is, $\mathcal{X}$ has strictly positive cofactors). Put $H = \max\ (h(B),\ \rho(B)/\theta)$ and $D' = \{X \in D \mid h(X) > H\}$. Then $D' \subset D \cap \{\mathcal{X}\}^+$.
Choosing k such that $(\tau_0^\nu - \varepsilon)^k h(B) > H$ we also find $T^{\nu k} D \subset D'$, so that $T^{\nu k} D \subset \{\mathcal{X}\}^+$. But this means that the matrix of $T^{\nu k}$ with respect to the base $\mathcal{X}$ has only positive entries (but not necessarily integral). By a theorem of PERRON[1], §14, (see also FROBENIUS[1]) we know then, that the largest eigenvalue of $T^{\nu k}$, which is $\tau_0^{\nu k}$ by lemma 4.6, is simple. But then τ_0 is a simple root of $f(\tau) = 0$, and since all other roots satisfy

$$|\tau_j| \le \tau_0^{-1/n} < 1,$$

$f(\tau)$ is irreducible. Moreover, if we would have $|\tau_j| < \tau_0^{-1/n}$ for any j, we would get

$$|\det T| = 1 = \prod_{0\le k\le n} |\tau_k| < \tau_0(\tau_0^{-1/n})^n = 1,$$

and thus it must be that

$$(4.21) \qquad |\tau_1| = |\tau_2| = \ldots\ldots = |\tau_n| = \tau_0^{-1/n}.$$

We have proved

<u>THEOREM 4.7.</u> *Let $\{B_i \mid -\infty < i < \infty\}$ be the sequence of best approximations ordered such that $h(B_{i+1}) > h(B_i)$ for all i. Suppose it is periodic, such that for some $m > 0$ and all i we have $TB_i = B_{i+m}$. Then the characteristic equation of T, $f(\tau) = 0$, is irreducible and has one root $\tau_0 > 1$. The other roots are all equal in absolute value. If $a_0, a_1, \ldots, a_n$ are the cofactors of a lattice base $\{A_0, A_1, \ldots, A_n\}$, then the numbers $1, \frac{a_1}{a_0}, \ldots\ldots, \frac{a_n}{a_0}$ form a base of the field $\mathbb{Q}(\tau_0)$.*

The last statement is proved along the same lines as part (b) of theorem 3.1, p. 28.

The condition, necessary for the existence of a periodic sequence of best approximations, is a very strong one. For MINKOWSKI[2] proved that the only fields containing a unit τ_0 with the property (4.21) are the trivial ones:

When $n = 1$, the real quadratic fields (where only one conjugate exists).

When $n = 2$, the real cubic fields of negative discriminant (where τ_1, τ_2 are complex conjugate numbers).

The necessity of this condition is complemented by sufficiency in the following sense. When $n = 1$, the best approximations are indeed periodic if $1, \frac{a_1}{a_0}$ form a base of a real quadratic field (theorem 1.4). When $n = 2$ and the numbers $1, \frac{a_1}{a_0}, \frac{a_2}{a_0}$ form a base of a cubic field of negative discriminant, then there is a choice of the radius function ρ and the

height function h such that the best approximations indeed become periodic. This is further dealt with in section 5F.

4D. Good approximations.

Roughly speaking, the successor B_1 of a best approximation B is the lowest point (i.e., least positive height) for which $\rho(B_1) < \rho(B)$. However, it may well be that a slight decrease in radius requires a large jump in height. If so, we would want to say that B is an approximation of good quality. In view of theorem 4.3 it seems natural to take the product $|h(B)|(\rho(B))^n$ as a measure of this quality. We therefore define the *quality* of a point A as an approximation to ℓ (with respect to ρ and h) to be the number

$$(4.22) \qquad qu(A) = \frac{c_{\rho h}}{2^{n+1}.\Delta} |h(A)|(\rho(A))^n.$$

A point A with quality $qu(A) \leq 1$ will then be called a *good approximation*. By (4.11) and (4.12) every best approximation is also a good approximation (but not conversely).

There are many problems in number theory where one is not interested in the complete sequence of best approximations, but only in a few approximations of sufficiently good (= small) quality and sufficiently large height (say 10^{40} or 10^{100}). Such problems arise for instance in connection with Baker's method in Diophantine equations -- see AGRAWAL, COATES, HUNT & VAN DER POORTEN[1] for a good example -- or in connection with the Riemann zeta-function, as in JURKAT & PEYERIMHOFF[1], or VAN DE LUNE & TE RIELE[1]. It is evident that, whenever possible, fast algorithms should be employed to find such good approximations, rather than the slower algorithms one would use in calculating the complete sequence of best approximations. Multi-dimensional continued fraction algorithms for these purposes are the subject of chapters 6 and 5, respectively.

Let us consider in some detail the case $n = 2$ when the radius function is Euclidean, that is $\rho(X) = d(X,\ell)$ for all $X \in \mathbb{R}^3$, and the height function has been normalized by $h(\ell_0) = |\ell_0|$. Denoting the projection of a point A on ℓ^* by underlining ($\underline{A}$) we have $\rho(A) = |\underline{A}|$.

<u>LEMMA 4.8.</u> *If* $A = \{A_0, A_1, A_2\}$ *is a base of* Ω, $\Delta = |\det A|$, $\ell_0 \in \{A\}^+$, *then*

$$(4.23) \qquad \min_{0\le k\le 2} |\underline{A}_k|^2 \le \frac{a_0\Delta/\sin\phi}{h(a_0A_0 + a_1A_1 + a_2A_2)}$$

where $\phi = \measuredangle(\underline{A}_1, \underline{A}_2)$ *is the angle between the projections* $\underline{A}_1$ *and* $\underline{A}_2$.

<u>PROOF.</u> From the cofactor relation $\ell_0 = a_0A_0 + a_1A_1 + a_2A_2$ it follows that

$$a_0\Delta = |\det(a_0A_0,\ A_1,\ A_2)| = |\det(\ell_0,\ A_1,\ A_2)| = |\det(\ell_0,\ \underline{A}_1,\ \underline{A}_2)|$$
$$= |\ell_0|\,|\underline{A}_1|\,|\underline{A}_2|\sin\phi$$

and this gives (4.23) if we observe that $|\ell_0| = h(\ell_0) = h(a_0A_0 + a_1A_1 + a_2A_2)$ and that $\min_{0\le k\le 2} |\underline{A}_k|^2 \le |\underline{A}_1|\,|\underline{A}_2|$. □

Formula (4.23) tells us, for example, that we have $qu(A_0) \le \frac{\pi}{2\sqrt{3}} < 1$ if $|\underline{A}_0| = \min_{0\le k\le 2} |\underline{A}_k|$ and $\frac{\pi}{3} \le \phi \le \frac{2\pi}{3}$. In general it suggests that we consider the following points when we wish to construct an expansion containing good (or best) approximations:

-- For any M an index j should exist such that $\min_{0\le k\le 2} h(A_k(j)) > M$.

-- There should be infinitely many bases $A(i)$ in the expansion for which $\min_{f\ne g} \measuredangle(\underline{A}_f(i), \underline{A}_g(i))$ exceeds, say, $\frac{\pi}{3}$. That is, the triangles with vertices $\underline{A}_0(i)$, $\underline{A}_1(i)$, $\underline{A}_2(i)$ in ℓ^* should not degenerate into needle-like triangles with two angles near zero and one near π.

Needless to say that these principles trivially extend to higher dimensions.

4E. Simultaneous Diophantine approximation.

In the usual theory of Diophantine approximations the lattice Ω consists of all points with integral coordinates, with respect to an orthonormal system, and ℓ_0 is given by $\ell_0 = (1, \xi_1, \ldots\ldots, \xi_n)$, where $\xi_1, \ldots\ldots, \xi_n$ are positive real numbers.

We come across two notions of best approximation, which happen to coincide in the 1-dimensional case. One distinguishes in fact between approximating ℓ by lattice points and approximating the orthogonal subspace ℓ^* (with equation $x_0 + \xi_1x_1 + \ldots\ldots + \xi_nx_n = 0$) by lattice points.

(1) In the first case it is said that $X = (x_0, x_1, \ldots\ldots, x_n)$ with $x_0 > 0$ is a best approximation if

$$(4.24) \qquad \max_{1 \le k \le n} |y_0 \xi_k - y_k| \ge \max_{1 \le k \le n} |x_0 \xi_k - x_k|$$

whenever $0 < y_0 \le x_0$. That is, the radius function

$$(4.25) \qquad \rho_1(X) = \max_{1 \le k \le n} |x_0 \xi_k - x_k|$$

and the height function $h_1(X) = x_0$ are used.

(2) In the second case it is said that $X = (x_0, x_1, \ldots\ldots, x_n) \neq 0$ is a "best approximation" if

$$(4.26) \qquad |y_0 + y_1 \xi_1 + \ldots\ldots + y_n \xi_n| \ge |x_0 + x_1 \xi_1 + \ldots\ldots + x_n \xi_n|$$

whenever $0 < \max_{1 \le k \le n} |y_k| \le \max_{1 \le k \le n} |x_k|$. This definition cannot be brought to terms with definition 4.1 for the simple reason that $\max_{1 \le k \le n} |x_k|$ is not a radius function (its value in ℓ_0 is not zero).

But for small heights it is, of course, nearly a radius function. Define ρ_2 by

$$(4.27) \qquad \rho_2(X) = \max_{1 \le k \le n} |x_k| \text{ for } X \in \ell^*$$

and extend by $\rho_2(X + \lambda \ell_0) = \rho_2(X)$ for all λ; put $h_2(X) = X.\ell_0$.

It can easily be shown that all points Z with $|h_2(Z)| < \frac{1}{2}|\ell_0|^2$ which are "best approximations" by the standard of (4.26) are also best approximations with respect to ρ_2 and h_2. Conversely, if Z is a best approximation with respect to ρ_2 and h_2, and $|h_2(Z)| < \frac{1}{2}|\ell_0|^2$, then either it is a "best approximation" by the standard of (4.26), or there is a "best approximation" $Y \neq Z$ in (4.26) with $\max_{1 \le k \le n} |y_k| = \max_{1 \le k \le n} |z_k|$.

<u>REMARK.</u> When $n \ge 2$, the radius functions ρ_1 and ρ_2 do not necessarily give the same points as best approximations. Take for example $n = 2$. The inequality $\rho_2(X) \le 1$ defines on ℓ^* a parallelogram Π_2. The inequality $\rho_1(X) \le |\ell_0|$ defines on ℓ^* a parallelogram Π_1 which, though congruent to

Π_2, is rotated in ℓ^* over $\frac{1}{2}\pi$ with respect to Π_2.

Most authors writing on best approximations adhere to one of the definitions (4.24), (4.26), particularly when discussing multi-dimensional continued fractions. See for example CUSICK[1-4,6], FURTWÄNGLER[1], SZEKERES[1]. The Euclidean radius function was considered repeatedly in relation to cubic number fields with negative discriminant, e.g. by VORONOI[1], BERWICK[1], DUBOIS[2,4,6]. This relation will be explained in section 5F. The first to consider the Euclidean radius as a general alternative to the sup norm functions ρ_1, ρ_2 were JURKAT, KRATZ & PEYERIMHOFF[1]. From a non-algorithmic point of view LAGARIAS[1] seems to be the first who studied properties of best approximations with respect to arbitrary radius functions.

CHAPTER 5. THE 2-DIMENSIONAL CASE

5A. The computation of best approximations.

In this chapter we consider only the 2-dimensional case, $n = 2$.

Given a best approximation B one can determine its successor B_1 in the following way. From the discussion in section 4B, notably formula (4.13), we know that

$$B_1 \in K = \{X \in \mathbb{R}^3 \mid \rho(X) < \rho(B),\ |h(X)| < \frac{8\Delta}{c_{\rho h}(\rho(B))^2}\}.$$

So if we first find all finitely many lattice points in K, then the lowest of them (least positive height) will be B_1. In case there are several points of lowest height we choose B_1 among them such that it has the smallest radius.

Various authors have given detailed versions of the method crudely outlined above, with special emphasis on reducing the number of "candidates" for B_1 as far as possible. We mention algorithms of MINKOWSKI[1] and FURTWÄNGLER[1] for the square norm (4.25), and work of VORONOI[1], BERWICK[1] and DUBOIS[4,5] for the Euclidean radius function. Some of these methods calculate predecessors instead of successors. All these algorithms further use the fact that every two consecutive best approximations can be completed to a base of the lattice with a third point ("auxiliary approximation") [1]). Each of them reduces the number of candidates for B_1 to somewhere between five and ten. But none of these algorithms possesses the simple additive structure that is characteristic of multi-dimensional continued fraction algorithms.

It is not even clear, a priori, whether there exists for given ℓ_0 and Ω and given radius and height functions ρ and h, an expansion which contains all best approximations with respect to that ρ and h. The main

---

[1]) Though *three* consecutive best approximations do not always constitute a lattice base: Their determinant may be zero or arbitrarily large (LAGARIAS[1,3]).

purpose of this chapter is to answer that question in the affirmative for the special case where $n = 2$ and ρ is the Euclidean radius function, and we do so by explicit construction of an algorithm.

5B. Preliminary results.

We shall always denote the projection on ℓ^*, parallel to ℓ, of a point A by underlining, that is by $\underline{A}$. The Euclidean radius function is now simply given by

(5.1) $$d(X,\ell) = |\underline{X}| \quad \text{for all } X \in \mathbb{R}^3.$$

If $\mathcal{A} = \{A_0, A_1, A_2\}$ is a base of Ω such that $\ell_0 \in \{\mathcal{A}\}^+$, then the projection of the cofactor relation (2.1) on ℓ^*,

(5.2) $$a_0 \underline{A}_0 + a_1 \underline{A}_1 + a_2 \underline{A}_2 = 0 ,$$

shows that 0 is in the convex hull of $\underline{A}_0$, $\underline{A}_1$, $\underline{A}_2$. This implies

(5.3) $$\measuredangle(\underline{A}_0, \underline{A}_1) + \measuredangle(\underline{A}_1, \underline{A}_2) + \measuredangle(\underline{A}_2, \underline{A}_0) = 2\pi,$$

and in particular the sum of any two of these angles must be at least π. Conversely, if the sum of any two of these angles is at least π, then $\ell_0 \in \{\mathcal{A}\}^+$.

We introduce the symbolic notation

(5.4) $$h(\mathcal{A}) > 0$$

to express that $\min\limits_{0 \le k \le 2} h(A_k) > 0$. It is easy to see that (5.4) holds for all bases of an expansion if it holds for the initial base $\mathcal{A}(0)$.

We now formulate the two lemmata which form the basis of our algorithm in the next section. The underlying idea is to study those lattice points P which are contained in $\{\mathcal{A}(i)\}^+$ but not in $\{\mathcal{A}(i+1)\}^+$. This is a natural idea since $P \notin \{\mathcal{A}(i)\}^+$ implies $P \notin \mathcal{A}(j)$ for all $j \ge i$ because of the additive structure of expansions.

LEMMA 5.1. *Let $\mathcal{A} = \{A_0, A_1, A_2\}$ be a base of Ω such that $h(\mathcal{A}) > 0$ and $\ell_0 \in \{\mathcal{A}\}^+$. Suppose s, t, u form a permutation of $0, 1, 2$ such that $\ell_0 \in \{\mathcal{A}I_{st}\}^+$*

(i.e., $a_s \geq a_t$*), and suppose*

(5.5) $$\measuredangle(\underline{A}_t, \underline{A}_u) \leq \tfrac{2}{3}\pi \quad \textit{and} \quad \measuredangle(\underline{A}_t, \underline{A}_t + \underline{A}_s) \leq \tfrac{2}{3}\pi .$$

If $P \neq A_t$ *is a lattice point in* $\{A\}^+$ *but not in* $\{AI_{st}\}^+$*, then P is not a best approximation.*

LEMMA 5.2. *Lemma 5.1 holds when (5.5) is replaced by*

(5.6) $$|\underline{A}_t + k\underline{A}_u| \geq |\underline{A}_u| \quad \textit{and} \quad |\underline{A}_t + k(\underline{A}_t + \underline{A}_s)| \geq |\underline{A}_t + \underline{A}_s| \quad \textit{for all } k \in \mathbb{N}\ ^{1)}.$$

PROOFS. The reader can follow these proofs in the accompanying figures 5.1 and 5.2, though we shall not rely on them.

Suppose then, that $P = pA_s + qA_t + rA_u \neq A_t$ is a lattice point in $\{A\}^+ - \{AI_{st}\}^+$. If we want to prove that P is not a best approximation, it is by definition sufficient to find a point R with

(5.7) $$0 < h(R) < h(P), \quad |\underline{R}| \leq |\underline{P}| .$$

Since $P = (p-q)A_s + q(A_t + A_s) + rA_u$, our assumption on P implies $q > p \geq 0$, $r \geq 0$ with $(p,q,r) \neq (0,1,0)$. We write P as

(5.8) $$P = p(A_t + A_s) + (q-p)A_t + rA_u,$$

which in particular implies

(5.9) $$h(P) > h(A_t)$$

since $P \neq A_t$ and $h(A) > 0$. Using (5.8) and the cofactor relation

$$a_s\underline{A}_s + a_t\underline{A}_t + a_u\underline{A}_u = 0$$

we have

(5.10) $$\underline{P} = \left(q - p + \frac{r(a_s - a_t)}{a_u}\right)\underline{A}_t + \left(p - \frac{ra_s}{a_u}\right)(\underline{A}_t + \underline{A}_s)$$

---

1) The natural numbers $\mathbb{N}$ start with 1.

and

$$(5.11)\qquad \underline{P} = \left(q - p + \frac{p(a_s - a_t)}{a_s}\right)\underline{A}_t + \left(r - \frac{pa_u}{a_s}\right)\underline{A}_u .$$

The coefficients of $\underline{A}_t + \underline{A}_s$ in (5.10) and $\underline{A}_u$ in (5.11) are of opposite sign. Writing $D = A_u$ if $ra_s > pa_u$, $D = A_t + A_s$ if $ra_s < pa_u$ we obtain

$$(5.12)\qquad \underline{P} = \tau\underline{A}_t + \delta\underline{D}$$

for some $\tau \geq 1$, $\delta \geq 0$ (we have $\tau \geq 1$ because of $q - p \geq 1$ and $a_s \geq a_t$).

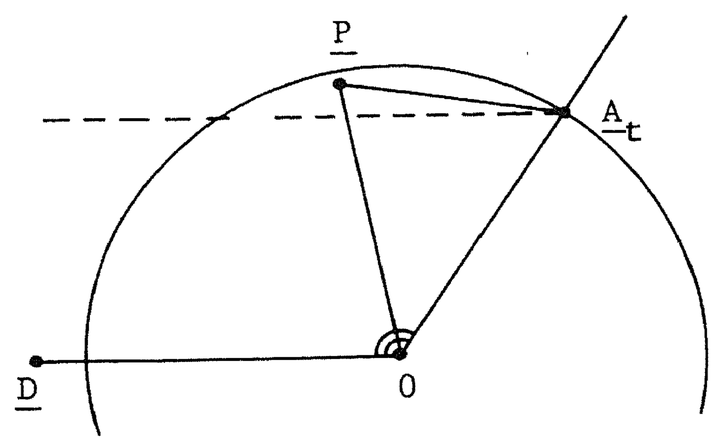


Figure 5.1: Lemma 5.1.

Proof of lemma 5.1. (Figure 5.1). If $|\underline{P}| \geq |\underline{A}_t|$ we immediately have (5.7) with $R = A_t$ by (5.9). Suppose now that $|\underline{P}| < |\underline{A}_t|$. We shall show that (5.7) holds for $R = P - A_t$. From $|\underline{P}| < |\underline{A}_t|$ we obtain, by (5.12),

$$\tau^2\underline{A}_t.\underline{A}_t + 2\delta\tau\underline{A}_t.\underline{D} + \delta^2\underline{D}.\underline{D} < \underline{A}_t.\underline{A}_t,$$

or

$$(5.13)\qquad 2\delta\tau\underline{A}_t.\underline{D} < |\underline{A}_t|^2(1 - \tau^2) - \delta^2|\underline{D}|^2.$$

On the other hand, (5.5) gives $\measuredangle(\underline{A}_t,\underline{D}) \leq \frac{2}{3}\pi$, whence

$$(5.14)\qquad \underline{A}_t.\underline{D} \geq -\tfrac{1}{2}|\underline{A}_t|\,|\underline{D}|$$

by the cosine formula. Combining (5.13) and (5.14) we get

$$(\tau^2 - 1)|\underline{A}_t|^2 - \delta\tau|\underline{A}_t|\,|\underline{D}| + \delta^2|\underline{D}|^2 < 0,$$

from which

$$\delta|\underline{D}| < \frac{1}{2}|\underline{A}_t|(\tau + \sqrt{4 - 3\tau^2}).$$

Since $\tau \geq 1$, this implies

(5.15) $\qquad \delta|\underline{D}| < \tau|\underline{A}_t|.$

Using (5.14), (5.15) and $\tau \geq 1$ one sees that

$$(1 - 2\tau)|\underline{A}_t|^2 - 2\delta\underline{A}_t.\underline{D} < 0$$

is true, which by direct calculation is equivalent to $|\underline{P} - \underline{A}_t|^2 < |\underline{P}|^2$.

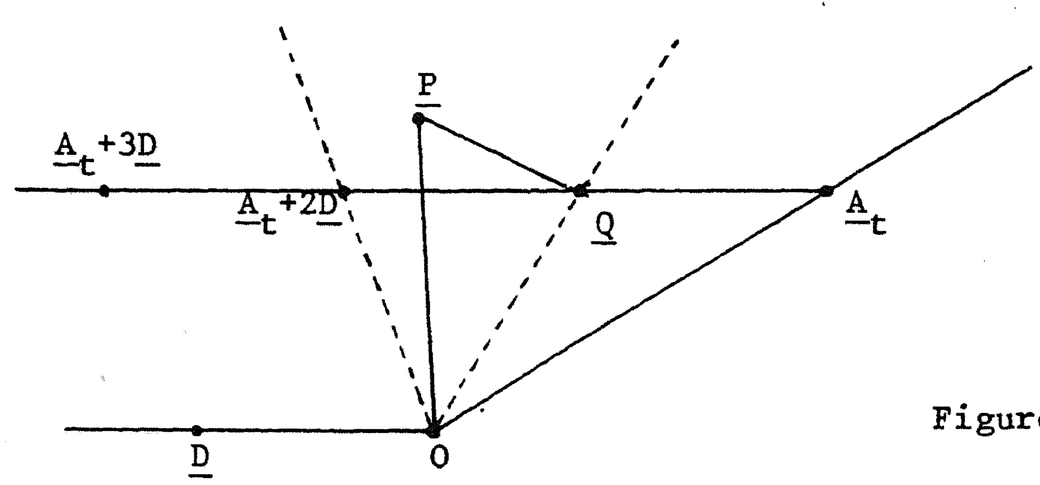


Figure 5.2: Lemma 5.2.

<u>Proof of lemma 5.2.</u> (Figure 5.2). If $P = A_t + kD$ for some $k \in \mathbb{N}$, then we have (5.7) with $R = D$ because of (5.6). So suppose P is not of the form $A_t + kD$. With δ, τ as in (5.12), define

(5.16) $\qquad m = \left[\frac{\delta}{\tau}\right] \geq 0.$

We shall prove (5.7) with $R = P - Q$, where $Q = A_t + mD$. By supposition $P \neq Q$, so that $P - Q \neq 0$. In case $D = A_u$,

$$r = \delta + p\frac{a_u}{a_s} \geq \delta \geq m\tau \geq m$$

so that $P - Q \in \{A\}^+$, whence

(5.17) $\qquad 0 < h(P - Q) < h(P).$

Similarly (5.17) holds if $D = A_t + A_s$ because

$$p = \delta + r\frac{a_s}{a_u} \geq \delta \geq m.$$

Define

$$\theta = \frac{\delta}{\tau} - m,$$

so that $0 \leq \theta < 1$ by (5.16); and by (5.12)

$$\underline{P} = \tau\underline{Q} + \tau\theta\underline{D},$$

$$\underline{P} - \underline{Q} = (\tau - 1)\underline{Q} + \tau\theta\underline{D}.$$

Now $|\underline{P} - \underline{Q}| \leq |\underline{P}|$, which we are going to prove, is equivalent to

(5.18) $\qquad (1 - 2\tau)|\underline{Q}|^2 - 2\tau\theta\underline{Q}.\underline{D} \leq 0.$

By (5.6), $|\underline{Q} + \underline{D}| = |\underline{A}_t + (m+1)\underline{D}| \geq |\underline{D}|$, so that $2\underline{Q}.\underline{D} \geq -|\underline{Q}|^2$, and this yields

$$(1 - 2\tau)|\underline{Q}|^2 - 2\tau\theta\underline{Q}.\underline{D} \leq (1 - 2\tau)|\underline{Q}|^2 + \tau\theta|\underline{Q}|^2 = |\underline{Q}|^2(1 - \tau(2 - \theta)) \leq 0$$

because $\tau \geq 1$, $\theta < 1$; which implies (5.18). □

We conclude with a key of access to the condition (5.6) of lemma 5.2:

<u>LEMMA 5.3.</u> $|\underline{A}_t + k\underline{D}| \geq |\underline{D}|$ *holds*

(i) *for* $k \geq 1$, *when* $\measuredangle(\underline{A}_t, \underline{D}) \leq \frac{1}{2}\pi$;

(ii) *for* $k \geq 1$, *when* $\measuredangle(\underline{A}_t, \underline{D}) \leq \frac{2}{3}\pi$ *and* $|\underline{A}_t| \geq |\underline{D}|$;

(iii) *for* $k \geq 2$, *when* $\measuredangle(\underline{A}_t, \underline{D}) \leq \frac{5}{6}\pi$.

<u>PROOF.</u> Straightforward from the cosine formula,

$$|\underline{A}_t + k\underline{D}|^2 = |\underline{A}_t|^2 + 2k|\underline{A}_t||\underline{D}|\cos\measuredangle(\underline{A}_t, \underline{D}) + k^2|\underline{D}|^2. \ \square$$

<u>5C. The 2-dimensional best approximation algorithm.</u>

We shall say that a lattice base $A = \{A_0, A_1, A_2\}$ with $\ell_0 \in \{A\}^+$ is of the *first kind* if

(5.19) $\qquad \measuredangle(\underline{A}_i, \underline{A}_j) \geq \frac{1}{3}\pi \quad$ when $i \neq j$.

It is of the *second kind* if for some permutation f,g,h of 0,1,2

$$(5.20)\qquad \begin{cases} \measuredangle(\underline{A}_g, \underline{A}_h) < \frac{1}{3}\pi \\ |\underline{A}_g - \underline{A}_h + k\underline{A}_f| \geq |\underline{A}_f| \quad \text{for all } k \in \mathbb{Z}. \end{cases}$$

It is possible that A is neither of the first nor of the second kind.

Now let the base $A(0) = \{A_0(0), A_1(0), A_2(0)\}$ be of the first or second kind and such that $h(A(0)) > 0$ and $\ell_0 \in \{A(0)\}^+$. We define a sequence of bases $A(i)$, $i = 1,2,\ldots\ldots$, from $A(0)$ iteratively by $A(i+1) = A(i)I_{st}$ according to the following algorithm for the determination of s and t.

<u>I.</u> If $A(i)$ is of the first kind, we specify the permutation f,g,h of 0,1,2 by

$$(5.21)\qquad |\underline{A}_f(i)| \leq \min\,(|\underline{A}_g(i)|, |\underline{A}_h(i)|), \quad a_h(i) \geq a_g(i).$$

<u>Ia.</u> If

$$(5.22)\qquad a_g(i) = 0$$

then $s = f$, $t = h$; else

<u>Ib.</u> if

$$(5.23)\qquad |\underline{A}_g(i) + k\underline{A}_f(i)| \geq |\underline{A}_f(i)| \quad \text{for all } k \in \mathbb{N}$$

then $s = h$, $t = g$; else

<u>Ic.</u> if

$$(5.24)\qquad \measuredangle(\underline{A}_g(i), \underline{A}_h(i)) \leq \measuredangle(\underline{A}_g(i), \underline{A}_f(i))$$

then $s = f$, $t = g$; else

<u>Id.</u> $s = g$, $t = f$.

<u>II.</u> If $A(i)$ is not of the first kind (whether or not of the second) we specify f,g,h by

$$(5.25)\qquad \measuredangle(\underline{A}_g(i), \underline{A}_h(i)) < \frac{1}{3}\pi, \quad |\underline{A}_g(i)| \geq |\underline{A}_h(i)|$$

and take $s = f$, $t = g$.

REMARKS.

1.- If (5.21) or (5.25) does not define f,g,h uniquely we choose any of the compatible permutations.

2.- If one of the points of $A(i)$ lies on ℓ (that is, the other two cofactors are zero), it is projected into the origin and the necessary angles do not exist. In this case we agree that the sequence of bases *terminates* with $A(i)$.

3.- For practical purposes we notice that (5.23) holds for all $k \in \mathbb{N}$ if it holds for

$$k = \max\left(1, \left[\frac{1}{2} - \frac{\underline{A}_g(i).\underline{A}_f(i)}{|\underline{A}_f(i)|^2}\right]\right).$$

The verification of the conditions therefore presents no difficulties.

4.- If ℓ is known to be independent, then case Ia may be disregarded.

THEOREM 5.4. *Let the base $A(0)$ with $\ell_0 \in \{A(0)\}^+$ and $h(A(0)) > 0$ be either of the first or of the second kind. Each base $A(i)$ of the sequence defined by the above algorithm has the following properties:*

(i) $\ell_0 \in \{A(i)\}^+$;

(ii) if $P \notin A(i)$ is a lattice point in $\{A(i)\}^+$ but not in $\{A(i+1)\}^+$, then P is not a best approximation.

Furthermore, the number of consecutive bases not of the first kind is always finite.

Statement (i) means that the sequence of bases the algorithm defines, is indeed an expansion in the sense of definition 2.3. Once the theorem has been proved it will be easy to show that the expansion actually contains all best approximations in $\{A(0)\}^+$ (theorem 5.7). Before giving a proof of theorem 5.4 we formulate two arguments we need more than once.

LEMMA 5.5. *Let $A = \{A_0, A_1, A_2\}$ be a base of Ω, $h(A) > 0$, $\ell_0 \in \{A\}^+$. Suppose that for some permutation s,t,u of $0,1,2$*

$$|\underline{A}_s| \le |\underline{A}_t| \tag{5.26}$$

and

(5.27) $\measuredangle(\underline{A}_s,\underline{A}_t) \geq \frac{2}{3}\pi$ *and* $\measuredangle(\underline{A}_s,\underline{A}_u) \geq \frac{2}{3}\pi$.

Then $\ell_0 \in \{AI_{st}\}^+$ *and* $\measuredangle(\underline{A}_s,\underline{A}_t+\underline{A}_s) \geq \frac{1}{3}\pi$. *Moreover, if* $P \neq A_t$ *is a lattice point in* $\{A\}^+$ *but not in* $\{AI_{st}\}^+$, *then* P *is not a best approximation.*

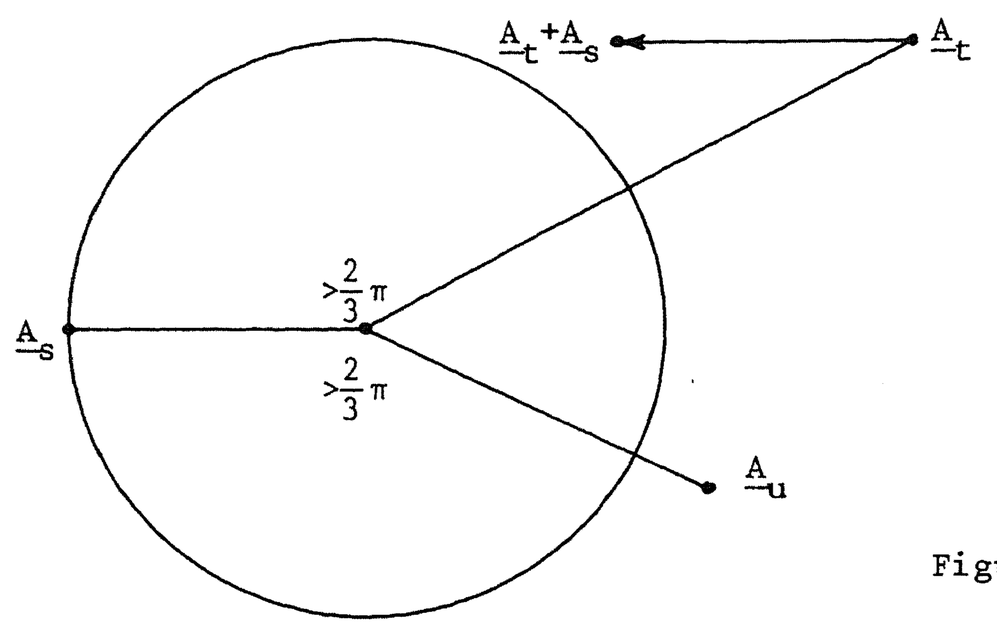


Figure 5.3: Lemma 5.5.

PROOF. (See figure 5.3). From (5.26) and (5.27) it easily follows that

(5.28) $\measuredangle(\underline{A}_s,\underline{A}_t+\underline{A}_s) \geq \frac{1}{2}\measuredangle(\underline{A}_s,\underline{A}_t) \geq \frac{1}{3}\pi$.

Now

$$\measuredangle(\underline{A}_u,\underline{A}_s) + \measuredangle(\underline{A}_s,\underline{A}_t+\underline{A}_s) \geq \frac{2}{3}\pi + \frac{1}{3}\pi = \pi,$$

from which we conclude that $\ell_0 \in \{AI_{st}\}^+$. The last statement of the lemma follows from lemma 5.1 if we observe that

$$\measuredangle(\underline{A}_u,\underline{A}_t) \leq \frac{2}{3}\pi$$

by (5.27), and

$$\measuredangle(\underline{A}_t,\underline{A}_t+\underline{A}_s) = \measuredangle(\underline{A}_s,\underline{A}_t) - \measuredangle(\underline{A}_s,\underline{A}_t+\underline{A}_s) \leq \pi - \frac{1}{3}\pi = \frac{2}{3}\pi$$

by (5.28). □

LEMMA 5.6. *Let* $A = \{A_0,A_1,A_2\}$ *be a lattice base,* $\ell_0 \in \{A\}^+$. *Suppose that for some permutation* s,t,u *of* $0,1,2$, $a_u \geq a_t$ *and* $\measuredangle(\underline{A}_s,\underline{A}_t) \geq \measuredangle(\underline{A}_u,\underline{A}_t)$. *Then*

(5.29) $$a_s|\underline{A}_s| \geq a_u|\underline{A}_u|$$

and

(5.30) $$|\underline{A}_t + \underline{A}_u| \geq |\underline{A}_u|.$$

PROOF. If $a_t = 0$ we have $\measuredangle(\underline{A}_s,\underline{A}_u) = \pi$ and the lemma is obvious, with equality in (5.29). Now suppose $a_t > 0$. Inserting the cofactor relation into the given angle inequality, we obtain, using the cosine formula,

$$\frac{-a_s|\underline{A}_s|^2 - a_u\underline{A}_s.\underline{A}_u}{|\underline{A}_s|} \leq \frac{-a_s\underline{A}_s.\underline{A}_u - a_u|\underline{A}_u|^2}{|\underline{A}_u|}$$

whence

(5.31) $$|\underline{A}_s||\underline{A}_u|(a_s|\underline{A}_s| - a_u|\underline{A}_u|) \geq \underline{A}_s.\underline{A}_u(a_s|\underline{A}_s| - a_u|\underline{A}_u|).$$

By the Cauchy-Schwarz inequality, (5.31) is possible only if the bracket factor is non-negative (of course, $|\underline{A}_s||\underline{A}_u| = \underline{A}_s.\underline{A}_u$ would imply $\measuredangle(\underline{A}_s,\underline{A}_u) = 0$ which is impossible), and this proves (5.29).
Furthermore, (5.30) is equivalent to

$$a_s^2|\underline{A}_s|^2 + 2a_s(a_u - a_t)\underline{A}_s.\underline{A}_u + (a_u - a_t)^2|\underline{A}_u|^2 \geq a_t^2|\underline{A}_u|^2$$

upon elimination of A_t (by means of the cofactor relation); and this we rewrite as

(5.32) $$x^2 + 2x|\underline{A}_u|(a_u + (a_u - a_t)\cos\measuredangle(\underline{A}_s,\underline{A}_u)) + 2a_u(a_u - a_t)|\underline{A}_u|^2(1 + \cos\measuredangle(\underline{A}_s,\underline{A}_u)) \geq 0$$

with $x = a_s|\underline{A}_s| - a_u|\underline{A}_u| \geq 0$. But (5.32) is clearly true for all $x \geq 0$ because $a_u \geq a_t > 0$. □

PROOF OF THEOREM 5.4. In most cases we shall show that, if $\mathcal{A}(i)$ is of the first or second kind, then so is $\mathcal{A}(i+1)$; in one case (Id) we may have to consider $\mathcal{A}(i+2)$. For the statement on best approximations we always refer to the lemmata 5.1, 5.2 and 5.5. Throughout the proof we omit the step indicator i, since no confusion can arise.

Suppose, then, that $\mathcal{A} = \{A_0,A_1,A_2\}$ is a base of Ω with non-negative

cofactors a_0, a_1, a_2 and with $h(A) > 0$. First let A be of the second kind with f,g,h specified by (5.25). From (5.25), and (5.20) with $k = 0$, it follows easily that

$$|\underline{A}_g| > |\underline{A}_g - \underline{A}_h| \geq |\underline{A}_f|.$$

Also,

$$\measuredangle(\underline{A}_f, \underline{A}_g) \geq \pi - \measuredangle(\underline{A}_g, \underline{A}_h) \geq \frac{2}{3}\pi$$

and similarly

$$\measuredangle(\underline{A}_f, \underline{A}_h) \geq \frac{2}{3}\pi.$$

Therefore we can apply lemma 5.5 with $s = f$, $t = g$, $u = h$, and it follows that (i) and (ii) of theorem 5.4 hold in this case. Also we have $\measuredangle(\underline{A}_f, \underline{A}_g + \underline{A}_f) \geq \frac{1}{3}\pi$, so if $AI_{fg} = \{A_f, A_g + A_f, A_h\}$ is not of the first kind, then we must have

$$\measuredangle(\underline{A}_g + \underline{A}_f, \underline{A}_h) < \frac{1}{3}\pi.$$

Since for all $k \in \mathbb{Z}$

$$|(\underline{A}_g + \underline{A}_f) - \underline{A}_h + k\underline{A}_f| = |\underline{A}_g - \underline{A}_h + (k+1)\underline{A}_f| \geq |\underline{A}_f|$$

by (5.20), we see that AI_{fg} then is of the second kind again. Observe that

$$1 + \left[\frac{-\underline{A}_g \cdot \underline{A}_f}{|\underline{A}_f|^2}\right] + \left[\frac{-\underline{A}_h \cdot \underline{A}_f}{|\underline{A}_f|^2}\right]$$

is a trivial bound on the maximal number of steps during which the base can remain of the second kind.

We now assume A to be of the first kind with f,g,h specified by (5.21). First suppose $a_g = 0$ (case Ia). This means that $0 = a_f \underline{A}_f + a_h \underline{A}_h$, whence

$$\measuredangle(\underline{A}_f, \underline{A}_h) = \pi$$

and

$$a_f = a_h \frac{|\underline{A}_h|}{|\underline{A}_f|} \geq a_h.$$

Thus $\mathcal{A}I_{fh} = \{A_f, A_g, A_h + A_f\}$ has non-negative cofactors. Furthermore, if $|\underline{A}_h| > |\underline{A}_f|$, then

$$\measuredangle(\underline{A}_h, \underline{A}_h + \underline{A}_f) = 0 < \frac{2}{3}\pi,$$

and

$$\measuredangle(\underline{A}_h, \underline{A}_g) = \pi - \measuredangle(\underline{A}_f, \underline{A}_g) \leq \frac{2}{3}\pi$$

since $\mathcal{A}$ is of the first kind. Therefore lemma 5.1 applies with $s = f$, $t = h$ to give us (i) and (ii) of theorem 5.4, and the new base is trivially of the first kind. But if $|\underline{A}_h| = |\underline{A}_f|$, that is $\underline{A}_f + \underline{A}_h = 0$, we argue directly: Let $P \neq A_h$ be a lattice point in $\{\mathcal{A}\}^+ - \{\mathcal{A}I_{fh}\}^+$, then $P = pA_h + q(A_h + A_f) + rA_g$ for some integers $p \geq 1$, $q, r \geq 0$. Notice that $h(P) > h(A_h)$ since $P \neq A_h$. Using $\underline{A}_f + \underline{A}_h = 0$ and

$$\measuredangle(\underline{A}_g, \underline{A}_h) = \pi - \measuredangle(\underline{A}_f, \underline{A}_g) \leq \frac{2}{3}\pi$$

we obtain

$$|\underline{P}| = |p\underline{A}_h + r\underline{A}_g| \geq |\underline{A}_h|,$$

which implies that P is not a best approximation.

From now on, $a_g > 0$. First suppose that condition (5.23) holds (case Ib; figure 5.4, p. 78). From the specification (5.21) the cofactors $a_f, a_g, a_h - a_g$ of $\{A_f, A_g + A_h, A_h\}$ are non-negative, which proves (i) of the theorem. Also we have

$$\measuredangle(\underline{A}_f, \underline{A}_g) + \measuredangle(\underline{A}_g, \underline{A}_g + \underline{A}_h) = \measuredangle(\underline{A}_f, \underline{A}_g + \underline{A}_h) \leq \pi. \tag{5.33}$$

But $\measuredangle(\underline{A}_f, \underline{A}_g) \geq \frac{1}{3}\pi$ from (5.19), hence

$$\measuredangle(\underline{A}_g, \underline{A}_g + \underline{A}_h) \leq \frac{2}{3}\pi.$$

If also $\measuredangle(\underline{A}_g, \underline{A}_f) \leq \frac{2}{3}\pi$ we apply lemma 5.1 with $s = h$, $t = g$ to obtain the desired result. And if $\measuredangle(\underline{A}_g, \underline{A}_f) > \frac{2}{3}\pi$ then

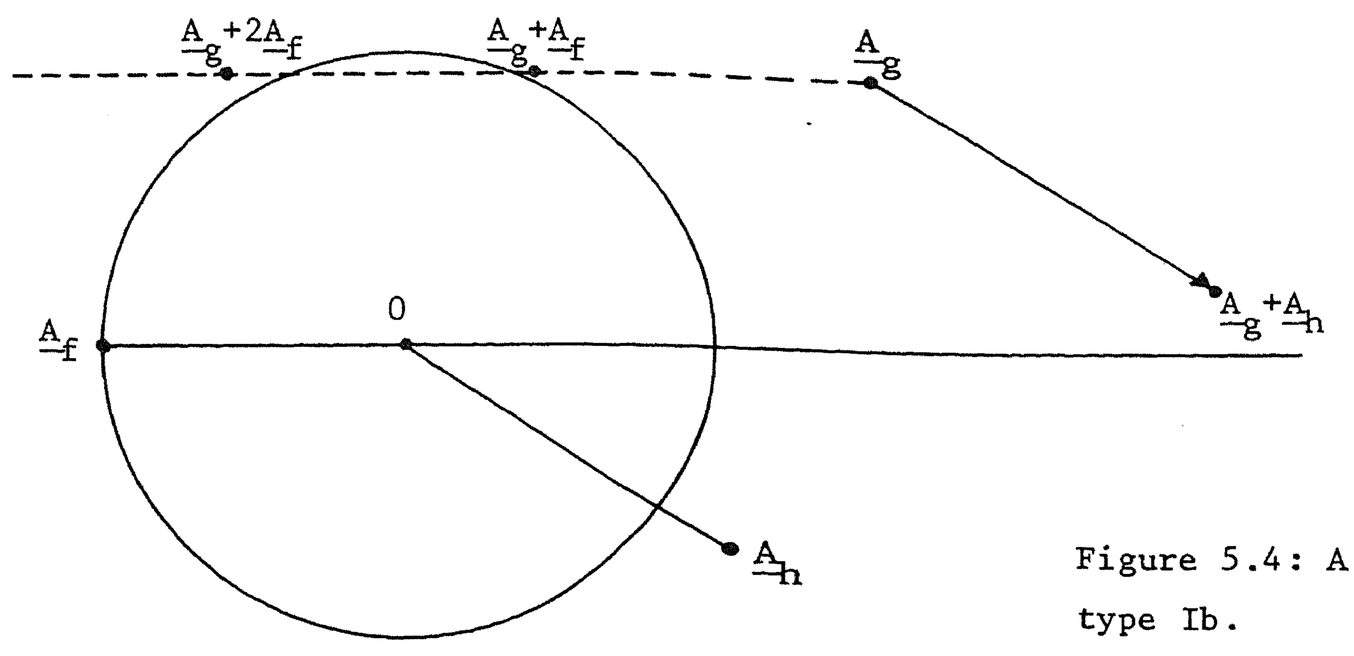


Figure 5.4: A step of type Ib.

(5.34) $\qquad \measuredangle(\underline{A}_g, \underline{A}_g + \underline{A}_h) < \frac{1}{3}\pi$

by (5.33). Hence the conditions of lemma 5.2 (with $s = h$, $t = g$) are satisfied by (5.23) and by lemma 5.3(i) in combination with (5.34). Moreover, if the new base $\{A_f, A_g + A_h, A_h\}$ is not of the first kind, then we must have

$$\measuredangle(\underline{A}_g + \underline{A}_h, \underline{A}_h) < \frac{1}{3}\pi.$$

Now

$$|(\underline{A}_g + \underline{A}_h) - \underline{A}_h + k\underline{A}_f| = |\underline{A}_g + k\underline{A}_f| \geq |\underline{A}_f|$$

follows from the specification (5.21) when $k = 0$, from the condition (5.23) when $k \geq 1$, and from lemma 5.3(ii) (with $A_t = A_g$, $D = -A_f$) when $k \leq -1$, in view of $\measuredangle(\underline{A}_g, \underline{A}_f) \geq \frac{1}{3}\pi$. Thus the new base is of the second kind.

We now assume that (5.23) does not hold, but that (5.24) is true (case Ic; figure 5.5). By lemma 5.3(ii) with $A_t = A_g$, $D = A_f$, (5.23) can be false only if

(5.35) $\qquad \measuredangle(\underline{A}_f, \underline{A}_g) > \frac{2}{3}\pi.$

Thus lemma 5.5 applies (with $s = f$, $t = g$) to obtain the desired result if also $\measuredangle(\underline{A}_f, \underline{A}_h) \geq \frac{2}{3}\pi$. Therefore suppose now that

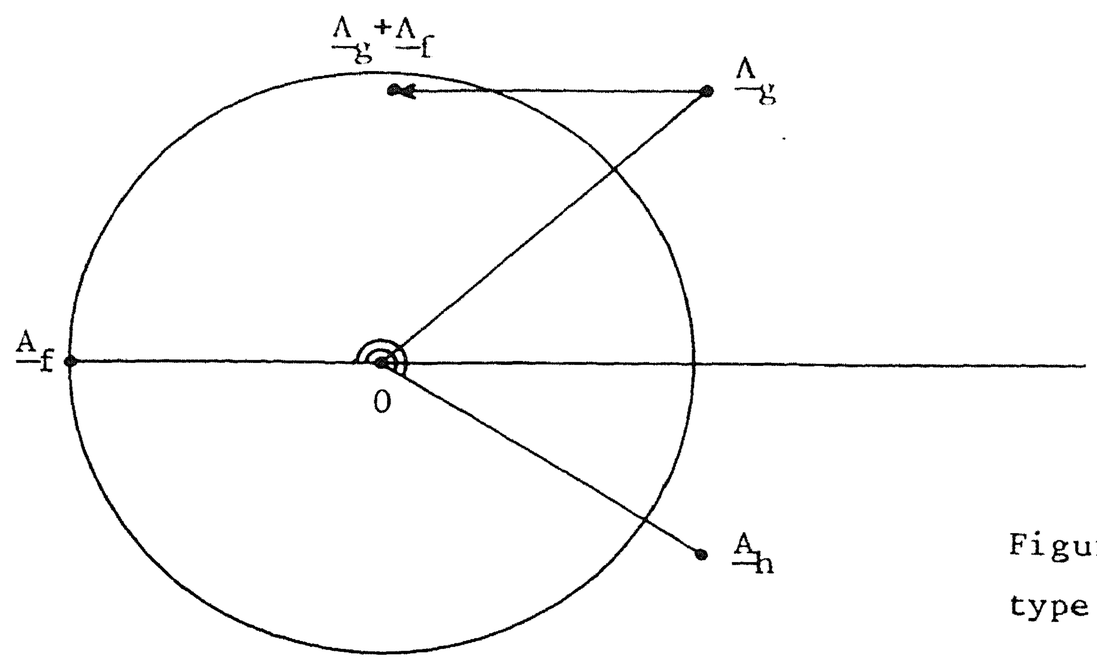


Figure 5.5: A step of type Ic.

(5.36) $\quad \frac{1}{3}\pi \le \measuredangle(\underline{A}_f,\underline{A}_h) < \frac{2}{3}\pi,$

and note that (5.24) together with the lower bound in (5.36) implies that

(5.37) $\quad \measuredangle(\underline{A}_g,\underline{A}_h) \le \frac{5}{6}\pi.$

Because of (5.21) and (5.24), lemma 5.6 applies with $s=f$, $t=g$, $u=h$, so that

$$a_f|\underline{A}_f| \ge a_h|\underline{A}_h|.$$

But $|\underline{A}_f| \le |\underline{A}_h|$ by specification, hence

(5.38) $\quad a_f \ge a_h \ge a_g$

in view of (5.21). Again by lemma 5.6,

(5.39) $\quad |\underline{A}_g + \underline{A}_h| \ge |\underline{A}_h|.$

Furthermore

(5.40) $\quad |\underline{A}_g + k\underline{A}_h| \ge |\underline{A}_h|, \quad k \ge 2,$

follows from (5.37) and lemma 5.3(iii). Lastly, using the trivial estimate $\underline{A}_f.\underline{A}_g \ge -|\underline{A}_f|\,|\underline{A}_g|$ one obtains easily

(5.41) $\quad |\underline{A}_g + k(\underline{A}_g + \underline{A}_f)| \geq |\underline{A}_g + \underline{A}_f|, \quad k \geq 1.$

Now (5.38), (5.39), (5.40) and (5.41) allow application of lemma 5.2 with $s = f$, $t = g$. Lastly, by (5.35) and $|\underline{A}_g| \geq |\underline{A}_f|$ we have

$$\measuredangle(\underline{A}_f, \underline{A}_g + \underline{A}_f) \geq \frac{1}{2}\measuredangle(\underline{A}_f, \underline{A}_g) > \frac{1}{3}\pi$$

so that the new base $\{A_f, A_g + A_f, A_h\}$ is again of the first kind.

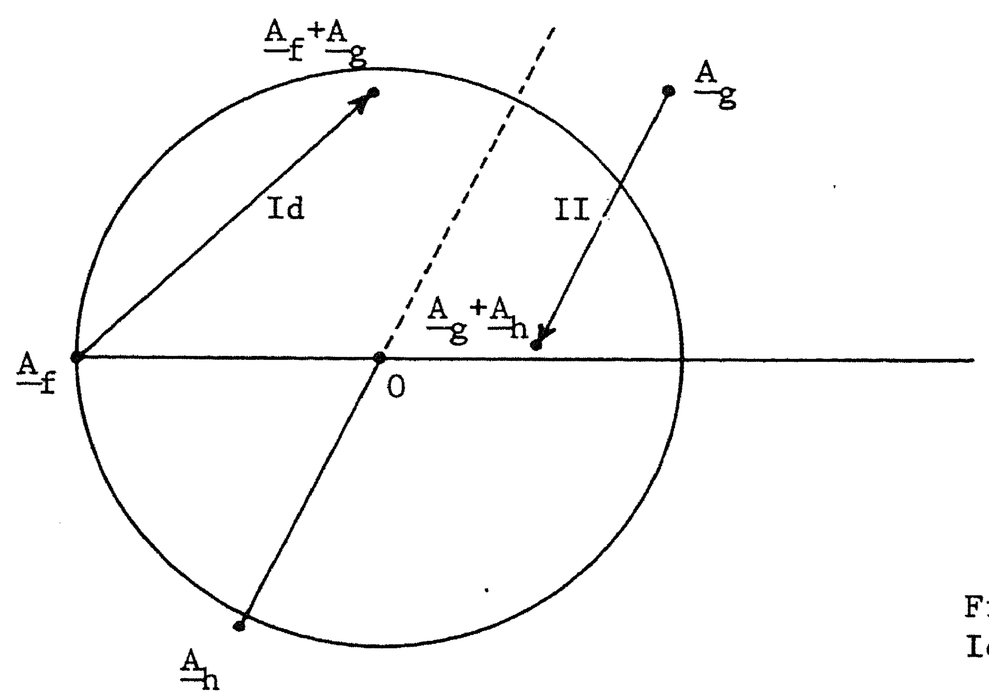


Figure 5.6: A step of type Id followed by one of type II.

It remains to consider the case Id (figure 5.6), where neither (5.23) nor (5.24) is true. That means

(5.42) $\quad |\underline{A}_g + k\underline{A}_f| < |\underline{A}_f| \quad$ for at least one $k \in \mathbb{N}$

and

(5.43) $\quad \measuredangle(\underline{A}_f, \underline{A}_g) < \measuredangle(\underline{A}_g, \underline{A}_h).$

Since A is of the first kind, (5.43) implies $\measuredangle(\underline{A}_f, \underline{A}_g) \leq \frac{5}{6}\pi$, whence

$$|\underline{A}_g + k\underline{A}_f| \geq |\underline{A}_f| \quad \text{for } k \geq 2$$

by lemma 5.3(iii). Therefore (5.42) gives

(5.44) $\quad |\underline{A}_g + \underline{A}_f| < |\underline{A}_f|.$

Now if we would have $a_f \geq a_g$, then lemma 5.6 (with $s = h$, $t = g$, $u = f$) would give $|\underline{A}_f + \underline{A}_g| \geq |\underline{A}_f|$; therefore (5.44) implies

$$(5.45) \qquad a_h \geq a_g > a_f$$

so that $AI_{gf} = \{A_f + A_g, A_g, A_h\}$ has non-negative cofactors $a_f, a_g - a_f, a_h$. Combining (5.35) and (5.43) we see that

$$(5.46) \qquad \frac{1}{3}\pi \leq \measuredangle(\underline{A}_f, \underline{A}_h) < \frac{2}{3}\pi.$$

Also

$$(5.47) \qquad \measuredangle(\underline{A}_f, \underline{A}_f + \underline{A}_g) = \measuredangle(\underline{A}_h, \underline{A}_f + \underline{A}_g) - \measuredangle(\underline{A}_f, \underline{A}_h) \leq \pi - \frac{1}{3}\pi = \frac{2}{3}\pi.$$

The application of lemma 5.1 with $s = g$, $t = f$ is now allowed because of (5.45), (5.46) and (5.47), and gives us (i) and (ii) of the theorem.

If $\measuredangle(\underline{A}_f + \underline{A}_g, \underline{A}_g) \geq \frac{1}{3}\pi$, then the new base AI_{gf} is of the first kind, but this is not necessarily so (see figure 5.6). If not, then the next step, according to the definition of the algorithm, will be to replace A_g by $A_g + A_h$ (case II; notice that $|\underline{A}_g| \geq |\underline{A}_f| > |\underline{A}_g + \underline{A}_h|$ by (5.21) and (5.44)), and we obtain the base

$$AI_{gf}I_{hg} = \{A_f + A_g, A_g + A_h, A_h\},$$

whose cofactors $a_f, a_g - a_f, a_h - (a_g - a_f)$ are indeed non-negative because of (5.45). Lemma 5.1 (with $s = h$, $t = g$) tells us that, if $P \neq A_g$ is a lattice point in $\{AI_{gf}\}^+ - \{AI_{gf}I_{hg}\}^+$, then P is not a best approximation: Indeed, we assumed that $\measuredangle(\underline{A}_f + \underline{A}_g, \underline{A}_g) < \frac{1}{3}\pi$; and it follows from $a_h \geq a_g$, (5.35) and (5.46) that

$$(5.48) \qquad |\underline{A}_g| \geq |\underline{A}_h|,$$

whence

$$\measuredangle(\underline{A}_g, \underline{A}_g + \underline{A}_h) \leq \frac{1}{2}\measuredangle(\underline{A}_g, \underline{A}_h) \leq \frac{1}{2}\pi.$$

It remains to show that $\{A_f + A_g, A_g + A_h, A_h\}$ is of the first kind. Notice that $\measuredangle(\underline{A}_f + \underline{A}_g, \underline{A}_h) > \measuredangle(\underline{A}_f, \underline{A}_h) \geq \frac{1}{3}\pi$. Also we have

$$\measuredangle(\underline{A}_g + \underline{A}_h, \underline{A}_h) \geq \frac{1}{2}\measuredangle(\underline{A}_g, \underline{A}_h) > \frac{1}{3}\pi$$

because of (5.48), and (5.35), (5.43). The inequality

$$\measuredangle(\underline{A}_f + \underline{A}_g, \underline{A}_g + \underline{A}_h) \geq \frac{1}{3}\pi$$

is easy if we observe that, in the triangle with vertices 0, $\underline{A}_f + \underline{A}_g$ and $\underline{A}_g + \underline{A}_h$, the edge opposite 0 is the longest. To see this, we first use (5.46) and $|\underline{A}_h| \geq |\underline{A}_f|$, and (5.44), to establish that

$$|\underline{A}_f - \underline{A}_h| \geq |\underline{A}_f| > |\underline{A}_f + \underline{A}_g|.$$

It now suffices to prove that

$$(5.49) \qquad |\underline{A}_g + \underline{A}_h| \leq |\underline{A}_f + \underline{A}_g|.$$

We prove (5.49) as follows: Write $\underline{A}_g = -\sigma\underline{A}_f - \tau\underline{A}_h$ with $\sigma = a_f/a_g$ and $\tau = a_h/a_g$, so that

$$(5.50) \qquad 0 \leq \sigma < 1 \leq \tau$$

by (5.45). From (5.44) we obtain by substitution

$$\sigma^2|\underline{A}_f|^2 - 2\sigma(|\underline{A}_f|^2 - \tau\underline{A}_f\cdot\underline{A}_h) + (\tau^2|\underline{A}_h|^2 - 2\tau\underline{A}_f\cdot\underline{A}_h) < 0.$$

In this formula the term without σ is non-negative because of $|\underline{A}_h| \geq |\underline{A}_f|$, $\tau \geq 1$ and (5.46). Therefore $\sigma \geq 0$ implies

$$|\underline{A}_f|^2 - \tau\underline{A}_f\cdot\underline{A}_h \geq 0$$

whence

$$(5.51) \qquad |\underline{A}_f|^2 - \underline{A}_f\cdot\underline{A}_h \geq 0$$

by $\tau \geq 1$. Now (5.49) is equivalent to

$$|\underline{A}_f|^2 - |\underline{A}_h|^2 - 2\sigma(|\underline{A}_f|^2 - \underline{A}_f\cdot\underline{A}_h) + 2\tau(|\underline{A}_h|^2 - \underline{A}_f\cdot\underline{A}_h) \geq 0.$$

To check its truth it suffices to insert the minimal τ and maximal σ of the range (5.50) because the coefficients of τ and $-\sigma$ are both non-negative by (5.51) and $|\underline{A}_h| \geq |\underline{A}_f|$. With $\sigma = \tau = 1$ we get

$$|\underline{A}_h|^2 - |\underline{A}_f|^2 \geq 0,$$

which is true. This completes the proof of theorem 5.4. □

REMARKS. For future use we collect some properties of the algorithm that turned up during the preceding proof:

1.- If A is of the second kind with f,g,h as in (5.25), then $|\underline{A}_g| > |\underline{A}_f|$.

2.- If case Id applies to A with f,g,h as in (5.21), then $\measuredangle(\underline{A}_g, \underline{A}_h) \neq \pi$. This follows from (5.35) and the fact that A is of the first kind.

3.- If some base $A(i)$ in the expansion is neither of the first nor of the second kind, then the preceding and next bases, $A(i-1)$ and $A(i+1)$, are of the first kind and the i-th step was of type Id.

4.- In the proof for case Ib no use was made of the fact that $\measuredangle(\underline{A}_g, \underline{A}_h) \geq \frac{1}{3}\pi$. This means that we may also apply step Ib instead of step II to a base of the second kind if it satisfies (with the same specification of f, g and h), in addition to the first inequality of (5.25), also (5.21), (5.23) and $a_g > 0$.

5.- When a step of type II, Ia, Ic, Id is applied to $A(i)$ we have

$$|\underline{A}_t(i) + \underline{A}_s(i)| \leq |\underline{A}_t(i)|.$$

For case Id this follows from (5.44), for the other cases from $\measuredangle(\underline{A}_s, \underline{A}_t) \geq \frac{2}{3}\pi$ and $|\underline{A}_t| \geq |\underline{A}_s|$.

6.- If the algorithm chooses t such that

$$|\underline{A}_t| < \min_{k \neq t} |\underline{A}_k|,$$

then the step is of type Id and in that case, $|\underline{A}_t + \underline{A}_s| < |\underline{A}_t|$ by (5.44).

We now formulate the main theorem on the algorithm.

THEOREM 5.7. *Let $A(0)$ be as in theorem 5.4 and construct its expansion $A(i)$, $i = 1, 2, \ldots\ldots$, by the above algorithm (p. 72). If $P \in \{A(0)\}^+$ is a best approximation, then $P \in A(j)$ for some $j \geq 0$.*

PROOF. First suppose the expansion terminates, say, with the step $A_0(j) + A_1(j) = A_0(j+1) \in \ell$. That requires $\measuredangle(\underline{A}_0(j), \underline{A}_1(j)) = \pi$ and $|\underline{A}_0(j)| = |\underline{A}_1(j)|$. Therefore $A(j)$ is of the first kind: Indeed, an appeal to case II could be made only when $A(j)$ were of the second kind (but then $|\underline{A}_0(j)| = |\underline{A}_1(j)|$ contradicts with remark 1 of p. 83), or when the preceding base $A(j-1)$ were of the first kind and the j-th step of type Id (but then $\measuredangle(\underline{A}_0(j), \underline{A}_1(j)) = \pi$ contradicts with remark 2 of p. 83).

Now any best approximation in $\{A(0)\}^+$ we might not yet have found, would still be in $\{A(j+1)\}^+$ by theorem 5.4. Since it also should be lower than $A_0(j+1)$, it is necessarily of the form $pA_1(j+1) + qA_2(j+1) = pA_1(j) + qA_2(j)$ for some $p,q \geq 1$. But since $A(j)$ is of the first kind, we have

$$\measuredangle(\underline{A}_1(j), \underline{A}_2(j)) = \pi - \measuredangle(\underline{A}_2(j), \underline{A}_0(j)) \leq \frac{2}{3}\pi,$$

and therefore

$$|p\underline{A}_1(j) + q\underline{A}_2(j)| \geq \min\ (|\underline{A}_1(j)|,\ |\underline{A}_2(j)|),$$

so that $pA_1(j) + qA_2(j)$ is not a best approximation.

Now suppose no termination occurs. Let P be a best approximation in $\{A(0)\}^+$ and assume there is an index i_0 such that, for some permutation f,g,h of 0,1,2,

$$h(A_f(i_0)) > h(P),\ h(A_g(i_0)) > h(P).$$

Then the only lattice points in $\{A(i_0)\}^+$ not higher than P might be $A_h(i_0)$ and some of its multiples. Thus either $P = A_h(i_0)$ or $P \notin \{A(i_0)\}^+$. In the latter case we have $P \in A(j)$ for some $j < i_0$ by theorem 5.4(ii). So it remains to show the existence of such an i_0. If there were none, there would exist an i_1 such that, for some permutation f,g,h,

$$A_f(i) = A_f(i_1),\quad A_g(i) = A_g(i_1) \quad \text{for all } i \geq i_1, \tag{5.52}$$

and, say, $|\underline{A}_g(i_1)| \geq |\underline{A}_f(i_1)|$. That is, A_f and A_g remain fixed after the i_1-th step, and $t = h$ when $i \geq i_1$. If we would have $a_h(i_1) > 0$ the whole expansion would terminate after at most

$$i_1 + \left[\frac{a_f(i_1)}{a_h(i_1)}\right] + \left[\frac{a_g(i_1)}{a_h(i_1)}\right]$$

steps, contrary to the assumption of non-termination. Therefore $a_h(i_1) = 0$. By theorem 5.4 $A(i_2)$ is of the first kind for some $i_2 \geq i_1$. The algorithm then decides to take the step

$$A_g(i_2+1) = A_g(i_2) + A_f(i_2)$$

either as a type Ia step (if $|\underline{A}_f(i_2)| \leq |\underline{A}_h(i_2)|$) or as a type Ib step (if $|\underline{A}_f(i_2)| > |\underline{A}_h(i_2)|$). But this contradicts with (5.52). □

5D. Further discussion of the algorithm and an example.

We give an example of an expansion by the 2-dimensional best approximation algorithm of the preceding section.

Take the lattice $\Omega = \mathbb{Z}^3$ in $\mathbb{R}^3$, and let ℓ be given by $\ell_0 = (1, \sqrt[3]{5}, \sqrt[3]{25})$. Let the inner product on $\mathbb{R}^3$ be defined by the quadratic form in the coordinates x,y,z,

$$x^2 + (y - x\sqrt[3]{5})^2 + (z - x\sqrt[3]{25})^2.$$

With this inner product, the orthogonal plane ℓ^* is the yz-plane and the height function $h(x,y,z) = x$ is normalized. The radius function is

$$\rho(x,y,z) = \sqrt{(y - x\sqrt[3]{5})^2 + (z - x\sqrt[3]{25})^2},$$

or $\rho(0,y,z) = \sqrt{y^2 + z^2}$ on ℓ^*. We wish to find the best approximations with respect to this radius function. We start with the standard base of $\mathbb{Z}^3$, consisting of

$$\begin{aligned} A_0(0) &= (1,\ 0,\ 0), \quad \text{cofactor } a_0(0) = 1 \\ A_1(0) &= (0,\ 1,\ 0), \quad \text{cofactor } a_1(0) = \sqrt[3]{5} \\ A_2(0) &= (0,\ 0,\ 1), \quad \text{cofactor } a_2(0) = \sqrt[3]{25}. \end{aligned}$$

This base has the property that $\underline{A}_j(0) \cdot \underline{A}_k(0) \leq 0$ when $j \neq k$, which allows us to use

<u>LEMMA 5.8.</u> *Suppose* $A = \{A_0, A_1, A_2\}$ *is a lattice base with non-negative*

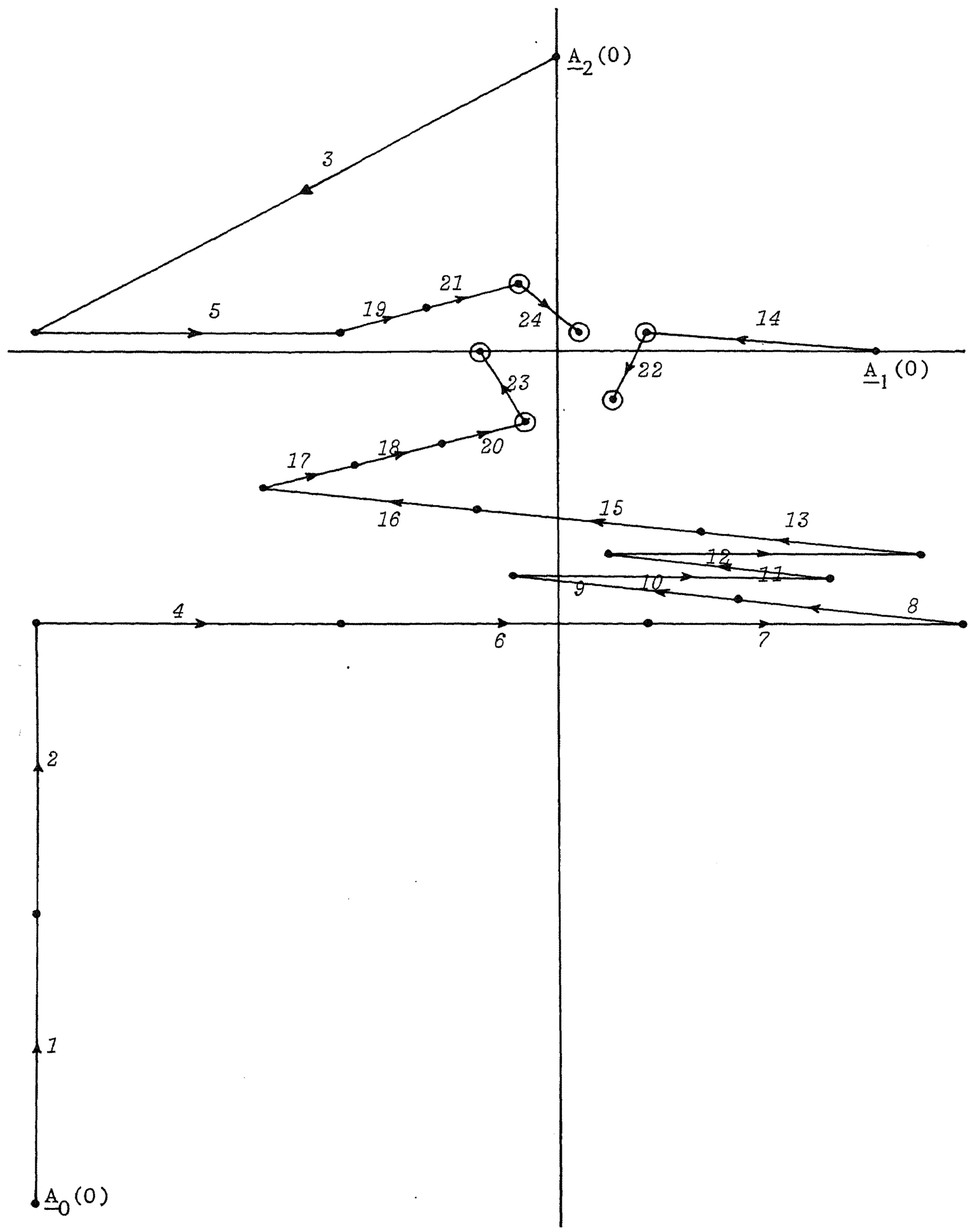


Figure 5.7. Best approximations are encircled. Step numbers in italics.

TABLE 5.1. (All figures rounded to 3 decimal places).

step	s	a_s	t	A_t	$\underline{A}_t$	$\lvert\underline{A}_t\rvert$	case
begin	0	1	0	(1, 0, 0)	(-1.710, -2.924)	3.387	
	1	1.710	1	(0, 1, 0)	(1 , 0)	1	
	2	2.924	2	(0, 0, 1)	(0 , 1)	1	Ib
1	2	1.924	0	(1, 0, 1)	(-1.710, -1.924)	2.574	Ib
2	2	0.924	0	(1, 0, 2)	(-1.710, -0.924)	1.944	Ib
3	0	0.076	2	(1, 0, 3)	(-1.710, 0.076)	1.712	II
4	1	1.634	0	(1, 1, 2)	(-0.710, -0.924)	1.165	II
5	1	0.710	2	(1, 1, 3)	(-0.710, 0.076)	0.7140	II
6	1	0.634	0	(1, 2, 2)	(0.290, -0.924)	0.969	Ib
7	1	0.558	0	(1, 3, 2)	(1.290, -0.924)	1.587	II
8	2	0.848	0	(2, 4, 5)	(0.580, -0.848)	1.027	II
9	2	0.772	0	(3, 5, 8)	(-0.130, -0.772)	0.783	Ib
10	1	0.482	0	(3, 6, 8)	(0.870, -0.772)	1.163	II
11	2	0.696	0	(4, 7, 11)	(0.160, -0.696)	0.7141	Ib
12	1	0.406	0	(4, 8, 11)	(1.160, -0.696)	1.353	II
13	2	0.620	0	(5, 9, 14)	(0.450, -0.620)	0.766	II
14	2	0.214	1	(1, 2, 3)	(0.290, 0.076)	0.300	Ib
15	2	0.138	0	(6, 10, 17)	(-0.260, -0.544)	0.603	Ib
16	2	0.062	0	(7, 11, 20)	(-0.970, -0.468)	1.077	II
17	1	0.330	0	(8, 13, 23)	(-0.680, -0.392)	0.785	II
18	1	0.254	0	(9, 15, 26)	(-0.390, -0.316)	0.502	II
19	1	0.192	2	(2, 3, 6)	(-0.420, 0.152)	0.447	II
20	1	0.116	0	(10, 17, 29)	(-0.100, -0.240)	0.260	Ib
21	1	0.054	2	(3, 5, 9)	(-0.130, 0.228)	0.262	Ic
22	0	0.022	1	(11, 19, 32)	(0.190, -0.164)	0.251	Ib
23	2	0.040	0	(13, 22, 38)	(-0.230, -0.012)	0.230	Ib
24	1	0.014	2	(14, 24, 41)	(0.060, 0.064)	0.088	

cofactors and $h(A) > 0$, such that $\underline{A}_j \cdot \underline{A}_k \le 0$ whenever $j \ne k$. If a lattice point P with $h(P) > 0$ satisfies $|\underline{P}| < \min(|\underline{A}_0|, |\underline{A}_1|, |\underline{A}_2|)$, then $P \in \{A\}^+$.

PROOF. If $P \notin \{A\}^+$, we may suppose $P = pA_0 + qA_1 + rA_2$ for some $p,q \ge 0$, $r \le -1$. Then $|\underline{P}|^2 = |p\underline{A}_0 + q\underline{A}_1|^2 + r^2|\underline{A}_2|^2 + 2pr\underline{A}_0 \cdot \underline{A}_2 + 2qr\underline{A}_1 \cdot \underline{A}_2 \ge r^2|\underline{A}_2|^2$,

contradicting $|\underline{P}| < |\underline{A}_2|$. □

Therefore we shall find all best approximations B with $|\underline{B}| < 1$ in the expansion of $A(0)$ [1]).

The expansion is given in table 5.1 and figure 5.7. The table is to be read as follows. Each step is represented by one line, containing only the new cofactor a_s and the new point $A_t = (x,y,z)$ with its projection $\underline{A}_t = (y - x\sqrt[3]{5},\ z - x\sqrt[3]{25})$. If one wants to know $A(k)$, one looks up the last entries of each of the symbols 0,1,2 in the first k steps in the column headed "t". The cofactors are found at the last entries of each of 0,1,2 in the column "s" during the first k steps. In the last column on the k-th line we find the classification of $A(k)$ by the algorithm (p. 72). The values of s and t then chosen are found on the next line. We notice that for $i = 0$, 1 and 2, we have $|\underline{A}_1(i)| = |\underline{A}_2(i)| = 1$. This implies a possibility of choice in determining f from (5.21); we have chosen $f = 1$.

Our last base is $A(24)$, consisting of $A_0(24) = (13, 22, 38)$, $A_1(24) = (11, 19, 32)$ and $A_2(24) = (14, 24, 41)$. Let B be a best approximation. If we have not yet found B, and $|\underline{B}| < 1$, then $B \in \{A(24)\}^+$ by theorem 5.4. In particular this gives $h(B) \geq 11 + 13 = 24$. So all best approximations with height up to 24 have been found. Rearranging the points in the table according to increasing height (= x-coordinate) we easily find these best approximations to be

(1, 2, 3),	radius 0.300
(3, 5, 9),	radius 0.262
(10, 17, 29),	radius 0.260
(11, 19, 32),	radius 0.251
(13, 22, 38),	radius 0.230
(14, 24, 41),	radius 0.088.

Further examples will be found in section 5F. We now turn to discuss the convergence of the algorithm.

---

[1]) Note: $h(A_1(0)) = h(A_2(0)) = 0$, so that $A(0)$ does not meet the requirement $h(A) > 0$. However, an inspection of the proofs of lemmata 5.1 and 5.2 (p. 68) shows that if $h(A(0)) \geq 0$, a best approximation B can be missed only if there is another lattice point B' with $h(B') = h(B)$ and $|\underline{B}'| = |\underline{B}|$. In our case this is impossible, since ℓ is independent.

THEOREM 5.9. *Applied to a line of dependence rank r the 2-dimensional best approximation algorithm of section 5C will yield a base with 3-r points arbitrarily close to that line.*

COROLLARY. *The algorithm is strongly convergent.*

PROOF OF THEOREM 5.9. Let ℓ be rational $(r = 2)$. When $B \neq 0$ is a lattice point on ℓ with minimal value of $h(B) > 0$, it is a best approximation and so $B \in A(j)$ for some $j \geq 0$ by theorem 5.7.

Now suppose $r \leq 1$ and choose $\varepsilon > 0$. Let j_0 be such that

$$\min_{0 \leq k \leq 2} |\underline{A}_k(j_0)| < \tfrac{1}{2}\varepsilon$$

(which certainly occurs when a best approximation B with $|\underline{B}| < \frac{1}{2}\varepsilon$ is found). Let $A_0(j_0)$ realize the minimum, and let $A(j_1)$, $j_1 > j_0$, be the first base to contain a point $A_k(j_1)$ with $|\underline{A}_k(j_1)| < |\underline{A}_0(j_0)|$. By remark 6 of p. 83 we have $A_0(j_0) = A_0(j_1 - 1)$, and if $k \neq 0$ also $A_0(j_0) = A_0(j_1)$. In case $k \neq 0$ we have

$$\max\,(|\underline{A}_0(j_1)|, |\underline{A}_k(j_1)|) < \tfrac{1}{2}\varepsilon.$$

In case $k = 0$ the j_1-th step was, say, $A_0(j_1) = A_0(j_1-1) + A_g(j_1-1)$ $(g \in \{1,2\})$ and it follows that

$$|\underline{A}_g(j_1-1)| \leq |\underline{A}_0(j_1-1)| + |\underline{A}_0(j_1)| < (\tfrac{1}{2} + \tfrac{1}{2})\varepsilon = \varepsilon,$$

so

$$\max\,(|A_g(j_1-1)|, |\underline{A}_0(j_1-1)|) < \varepsilon.$$

This proves the theorem when $r = 1$. It remains to consider the case of an independent line, $r = 0$. If the theorem were false, there was a $\delta > 0$ such that

$$\max_{0 \leq k \leq 2} |\underline{A}_k(j)| > \delta \quad \text{for all } j.$$

Let j_2 be such that, say, $|\underline{A}_1(j_2)| < \delta/10$ and $|\underline{A}_2(j_2)| < \delta/10$, hence $|\underline{A}_0(j_2)| > \delta$. Suppose for $j \geq j_2$ there will never be a step of Ib type with $s = 0$. Then we prove by induction that

(5.53) $\max\,(|\underline{A}_1(j)|,|\underline{A}_2(j)|) < \frac{\delta}{10}, \qquad |\underline{A}_0(j)| > \delta \quad \text{for } j \geq j_2.$

Indeed, suppose (5.53) holds for a particular value $j_3 \geq j_2$. If the (j_3+1)-th step is not of type Ib, then (5.53) holds for $j = j_3+1$ by remark 5 of p. 83. If the (j_3+1)-th step is of type Ib, then $t = 0$ so that $A_k(j_3+1) = A_k(j_3)$ for $k = 1,2$, and again (5.53) holds for $j = j_3+1$ (observe that a step of type Ib always involves the point of largest radius).

From (5.53) and remark 5 of p. 83 we conclude that there will be no step at all with $s = 0$ after the j_2-th step. By the independence of ℓ and again remark 5 of p. 83, the number of steps with $\{s,t\} = \{1,2\}$ will be finite; and so will, by the independence of ℓ, the number of steps with $t = 0$. Therefore the whole expansion is finite. From this contradiction we conclude that there will be after the j_2-th step a first step of type Ib of the form $A_g(j_4) = A_g(j_4-1) + A_0(j_4-1)$ for some $j_4 \geq j_2$ and $g \in \{1,2\}$. A simple estimate shows that

$$|\underline{A}_0(j_4)| > \delta, \quad |\underline{A}_g(j_4)| > \frac{9}{10}\delta, \quad |\underline{A}_0(j_4) - \underline{A}_g(j_4)| < \frac{1}{10}\delta,$$

and the algorithm continues with type II steps at least until

$$\max\,(|\underline{A}_0(j)|,|\underline{A}_g(j)|) < \frac{2}{10}\delta \quad \text{for some } j > j_4,$$

contradicting our assumption. □

In the present form our algorithm has the disadvantage of being subtractive and thus it may be too slow (compare section 3B). We shall make it faster:

<u>LEMMA 5.10.</u> *Let $A(i) = A = \{A_0, A_1, A_2\}$ be a lattice base with $h(A) > 0$, $\ell_0 \in \{A\}^+$. Suppose (5.23) holds when f,g,h are specified by (5.21), also suppose that $a_g > 0$ and that A is of the first kind or that $\measuredangle(\underline{A}_g, \underline{A}_h) < \frac{1}{3}\pi$. Define*

$$b = \left[\frac{a_h}{a_g}\right] - \left[\frac{|\underline{A}_f|}{d(\underline{A}_h, m)}\right]$$

where m is the line in ℓ^* through 0 and $\underline{A}_f$. If $b \geq 2$ and $h(A_g + A_h) \geq h(A_f)$, then theorem 5.4 remains valid if we replace the Ib-step $A_g(i+1) = A_g(i) + A_h(i)$ by $A_g(i+1) = A_g(i) + bA_h(i)$.

PROOF. By definition of b we have

$$(5.54)\qquad |\underline{A}_g + b'\underline{A}_h + k\underline{A}_f| \geq d(\underline{A}_g + b'\underline{A}_h, m) \geq |\underline{A}_f|$$

for $0 \leq b' \leq b-1$ and $k \in \mathbb{Z}$. Therefore and by remark 4 of p. 83 the succession of Ib type steps $A_g \to A_g + A_h \to \ldots\ldots \to A_g + bA_h$ can be done simultaneously. We only have to see why $A_g + A_h$, ..., $A_g + (b-1)A_h$ are not best approximations, and this follows from (5.54) (with $k = 0$) and the height condition $h(A_g + A_h) > h(A_f)$. □

LEMMA 5.11. *Let $A(i) = A = \{A_0, A_1, A_2\}$ be a base of the second kind with f, g, h specified by (5.25). Define*

$$b = \left[-\frac{|\underline{A}_g|}{|\underline{A}_f|}\left(\gamma + \frac{1}{3}\sqrt{3 - 3\gamma^2}\right)\right], \quad \gamma = \cos \measuredangle(\underline{A}_f, \underline{A}_g).$$

If $b \geq 2$ then theorem 5.4 remains true if we replace the type II step $A_g(i+1) = A_g(i) + A_f(i)$ by $A_g(i+1) = A_g(i) + bA_f(i)$.

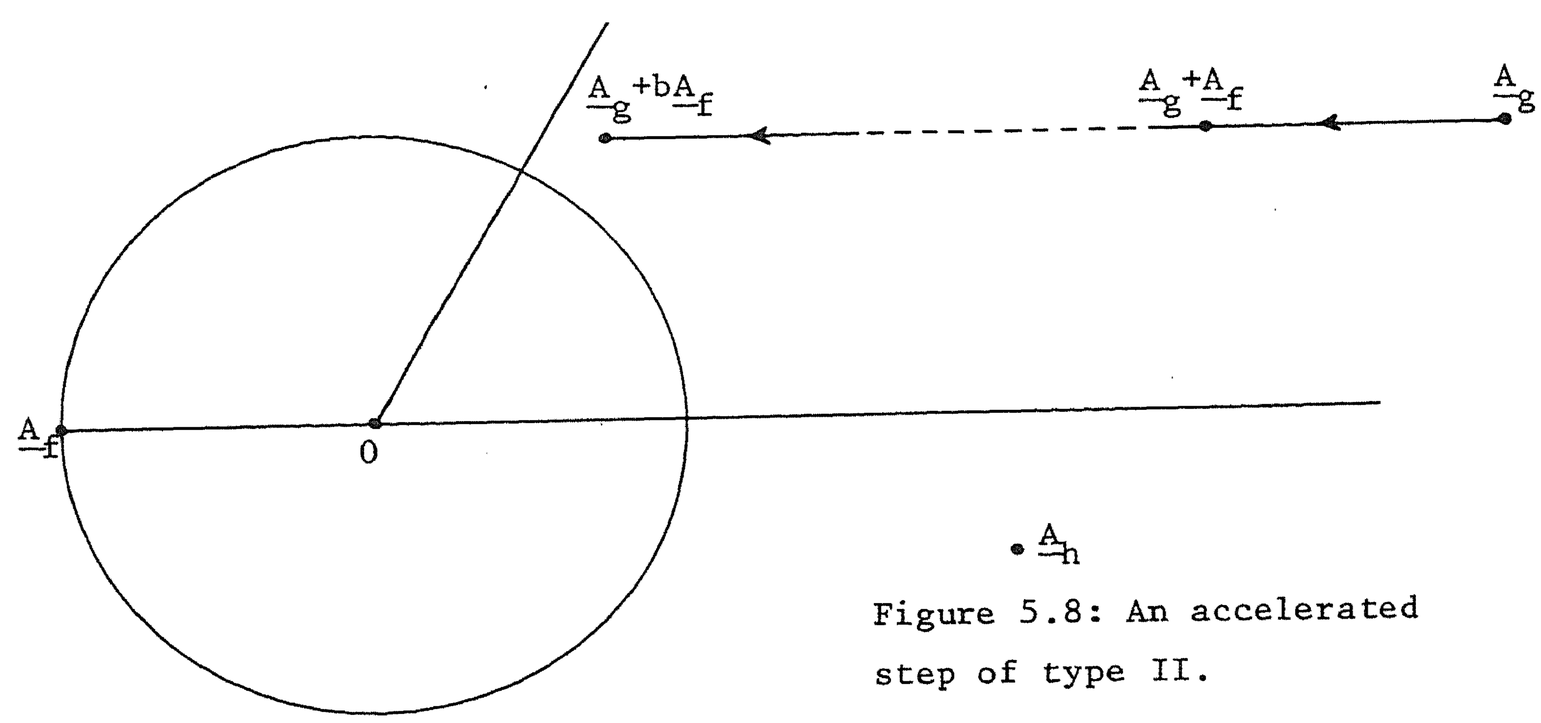


Figure 5.8: An accelerated step of type II.

PROOF. By the choice of b we have (see figure 5.8)

$$(5.55)\qquad \measuredangle(\underline{A}_f, \underline{A}_g + b'\underline{A}_f) \geq \frac{2}{3}\pi \quad \text{for } 0 \leq b' \leq b,$$

so that to all steps $A_g \to A_g + A_f \to \ldots\ldots \to A_g + bA_f$ lemma 5.5 applies. The fact that none of the points $A_g + b'A_f$, $1 \leq b' \leq b-1$, is a best approximation

follows trivially from $h(A_g + b'A_f) > h(A_f)$ and

$$|\underline{A}_g + b'\underline{A}_f| = |\underline{A}_g + b\underline{A}_f - (b - b')\underline{A}_f| > |\underline{A}_f|,$$

where the last inequality is a consequence of (5.55). □

REMARK. A similar acceleration could be applied to a step of type Ic. In practice, however, it is found that most steps will be of types Ib and II. Type Ic occurs occasionally and type Id hardly ever (but sometimes).

EXAMPLE. We reconsider the example with which we started this section. Table 5.2 gives the expansion by the accelerated algorithm. The acceleration has more substantial effect in examples 4 and 5 of section 5F.

TABLE 5.2. (All figures rounded to 3 decimal places).

step	s	a_s	t	A_t	$\underline{A}_t$	$\lvert\underline{A}_t\rvert$	case	b
begin	0	1	0	(1, 0, 0)	(-1.710, -2.924)	3.387		
	1	1.710	1	(0, 1, 0)	(1 , 0)	1		
	2	2.924	2	(0, 0, 1)	(0 , 1)	1	Ib	2
1	2	0.924	0	(1, 0, 2)	(-1.710, -0.924)	1.944	Ib	1
2	0	0.076	2	(1, 0, 3)	(-1.710, 0.076)	1.712	II	1
3	1	1.634	0	(1, 1, 2)	(-0.710, -0.924)	1.165	II	1
4	1	0.710	2	(1, 1, 3)	(-0.710, 0.076)	0.7140	II	1
5	1	0.634	0	(1, 2, 2)	(0.290, -0.924)	0.969	Ib	2
6	1	0.482	0	(1, 4, 2)	(2.290, -0.924)	2.469	II	3
7	2	0.696	0	(4, 7, 11)	(0.160, -0.696)	0.7141	Ib	1
8	1	0.406	0	(4, 8, 11)	(1.160, -0.696)	1.353	II	1
9	2	0.620	0	(5, 9, 14)	(0.450, -0.620)	0.766	II	1
10	2	0.214	1	(1, 2, 3)	(0.290, 0.076)	0.300	Ib	1
11	2	0.138	0	(6, 10, 17)	(-0.260, -0.544)	0.603	Ib	1
12	2	0.062	0	(7, 11, 20)	(-0.970, -0.468)	1.077	II	3
13	1	0.178	0	(10, 17, 29)	(-0.100, -0.240)	0.260	Ib	2
14	1	0.054	2	(3, 5, 9)	(-0.130, 0.228)	0.262	Ic	1
15	0	0.022	1	(11, 19, 32)	(0.190, -0.164)	0.251	Ib	1
16	2	0.040	0	(13, 22, 38)	(-0.230, -0.012)	0.230	Ib	1
17	1	0.014	2	(14, 24, 41)	(0.060, 0.064)	0.088		

We conclude with a simple technique for *inhomogeneous approximations* based upon our algorithm. Given a (non-zero) vector $Y \in \ell^*$ the problem of inhomogeneous approximation is that of finding lattice points arbitrarily close to the line $Y + \ell$. It is known from Kronecker's theorem that such lattice points always exist when ℓ is an independent line. The technique is as follows.

Let a point $P \in \Omega$ be given such that $|Y - \underline{P}| = \delta > 0$. To find a point $P_1 \in \Omega$ closer to $Y + \ell$ we use the algorithm to obtain a lattice base $A = \{A_0, A_1, A_2\}$ with

$$\max_{0 \le k \le 2} |\underline{A}_k| < \delta.$$

Write $Y - \underline{P} = \eta_f \underline{A}_f + \eta_g \underline{A}_g$ where f,g are chosen such that $\eta_f \ge 0$ and $\eta_g \ge 0$. To see that this is possible, first note that any two of $\underline{A}_0, \underline{A}_1, \underline{A}_2$ are independent vectors in ℓ^* because ℓ is independent. Secondly, if $Y - \underline{P} = \eta_1 \underline{A}_1 + \eta_2 \underline{A}_2$ with $\min(\eta_1, \eta_2) < 0$, put

$$\sigma = \min\ (\eta_1/a_1\ ,\ \eta_2/a_2).$$

Then $\sigma < 0$, and

$$Y - \underline{P} = Y - \underline{P} - \sigma O = -\sigma a_0 \underline{A}_0 + (\eta_1 - \sigma a_1)\underline{A}_1 + (\eta_2 - \sigma a_2)\underline{A}_2$$

is of the required form. Now take $P_1 = P + [\frac{1}{2} + \eta_f]A_f + [\frac{1}{2} + \eta_g]A_g$. Then

$$|Y - \underline{P}_1| = |(\eta_f - [\tfrac{1}{2} + \eta_f])\underline{A}_f + (\eta_g - [\tfrac{1}{2} + \eta_g])\underline{A}_g| < \tfrac{1}{2}\delta + \tfrac{1}{2}\delta = \delta.$$

Repeating the technique we find lattice points arbitrarily close to $Y + \ell$.

<u>EXAMPLE.</u> Consider the (academic) problem of determining integers x,y,z which make

$$\max\ (|y - x\sqrt[3]{5} + \sqrt{2}|\ ,\ |z - x\sqrt[3]{25} + \sqrt{3}|)$$

small. We take $\Omega = \mathbb{Z}^3$ and $\ell_0 = (1,\ \sqrt[3]{5},\ \sqrt[3]{25})$, and $Y = (0,\ -\sqrt{2},\ -\sqrt{3})$. We start with $P = (1,\ 0,\ 1)$, so that $\underline{P} = (-1.710,\ -1.924)$ and $Y - \underline{P} = (0.296,\ 0.192)$ in 3 decimal figures. Since $|Y - \underline{P}| = 0.352$ we look in table 5.2 for the first base $A(i)$ with

$$\max_{0\le k\le 2} |\underline{A}_k(i)| < 0.352,$$

and this is $A(14)$, consisting of $A_0(14) = (10,\ 17,\ 29)$, $A_1(14) = (1,\ 2,\ 3)$ and $A_2(14) = (3,\ 5,\ 9)$. We find approximately $Y - \underline{P} = 1.22\,\underline{A}_1(14) + 0.44\,\underline{A}_2(14)$, therefore we take $P_1 = P + A_1(14) = (2,\ 2,\ 4)$. We have $|Y - \underline{P}_1| = 0.112$.

5E. Other convergent algorithms.

One might ask whether a 2-dimensional continued fraction algorithm to give all best Euclidean norm approximations must necessarily distinguish between several cases, as in section 5C. Let us therefore investigate two simpler selection rules. We shall see that they do not always find all best approximations.

First we consider the (subtractive) *greatest angle algorithm*, where one chooses s and t such that

$$(5.56)\qquad \measuredangle(\underline{A}_s,\underline{A}_t) = \max_{j\ne k} \measuredangle(\underline{A}_j,\underline{A}_k), \quad a_s \ge a_t.$$

When (5.56) does not define s and t uniquely we take any compatible choice. Notice that the algorithm, like the best approximation algorithm, has to terminate when one of the points A_0, A_1, A_2 lies on ℓ.

LEMMA 5.12. *Let $A = \{A_0, A_1 . A_2\}$ be a lattice base with $\ell_0 \in \{A\}^+$. If s,t,u form a permutation of $0,1,2$ such that (5.56) holds, then $|\underline{A}_t + \underline{A}_s| \le |\underline{A}_t|$.*

COROLLARY. *With the greatest angle algorithm we have $|\underline{A}_k(i+1)| \le |\underline{A}_k(i)|$ for $k = 0,1,2$ and all $i \ge 0$.*

PROOF OF LEMMA 5.12. From $\measuredangle(\underline{A}_s,\underline{A}_t) \ge \measuredangle(\underline{A}_s,\underline{A}_u)$ we deduce

$$(5.57)\qquad a_t|\underline{A}_t| \ge a_u|\underline{A}_u|$$

as in lemma 5.6. Now from $a_s \ge a_t$, (5.57) and the cofactor relation we obtain the following series of inequalities:

$$a_s^2(|\underline{A}_s + \underline{A}_t|^2 - |\underline{A}_t|^2)$$
$$= (a_s - a_t)^2|\underline{A}_t|^2 - 2(a_s - a_t)a_u\underline{A}_t.\underline{A}_u + a_u^2|\underline{A}_u|^2 - a_s^2|\underline{A}_t|^2$$

$$\leq (a_s - a_t)\{(a_s - a_t)|\underline{A}_t|^2 - 2a_u \underline{A}_t \cdot \underline{A}_u - (a_s + a_t)|\underline{A}_t|^2\}$$

$$= 2(a_s - a_t)(-a_t|\underline{A}_t|^2 - a_u \underline{A}_t \cdot \underline{A}_u)$$

$$\leq 2(a_s - a_t)(-a_t|\underline{A}_t|^2 + a_u|\underline{A}_t||\underline{A}_u|)$$

$$= 2|\underline{A}_t|(a_s - a_t)(a_u|\underline{A}_u| - a_t|\underline{A}_t|) \leq 0,$$

whence $|\underline{A}_s + \underline{A}_t| \leq |\underline{A}_t|$. □

In order to prove convergence of the greatest angle algorithm we establish the following auxiliary inequalities.

LEMMA 5.13. *Let X and Y be non-zero vectors with $\measuredangle(X,Y) \geq \frac{2}{3}\pi$ and $|X| \leq |Y|$.*

(a) If $|Y| \geq \frac{5}{4}|X|$ then $|X+Y| \leq |Y| - \frac{1}{10}|X|$.

(b) If $|Y| < \frac{5}{4}|X|$ and $\measuredangle(X,Y) \geq \frac{3}{4}\pi$, then $|X+Y| \leq \frac{9}{10}|X| \leq |Y| - \frac{1}{10}|X|$.

(c) If $|Y| < \frac{5}{4}|X|$ and $\measuredangle(X,Y) < \frac{3}{4}\pi$, then $\measuredangle(Y,X+Y) \leq \measuredangle(X,X+Y) < \frac{1}{2}\pi$.

PROOF. By direct calculation. □

THEOREM 5.14. *The 2-dimensional greatest angle algorithm is strongly convergent.*

PROOF. 1. Let ℓ be independent. Suppose $\delta = \inf\{|\underline{A}_k(i)| \mid 0 \leq k \leq 2,\ i \geq 0\} > 0$. We shall show that for any given i_0 there is a $j > i_0$ such that

$$\sum_{k=0}^{2} |\underline{A}_k(j)| \leq \sum_{k=0}^{2} |\underline{A}_k(i_0)| - \frac{1}{10}\delta, \tag{5.58}$$

from which clearly a contradiction results.

Put $X = \underline{A}_{s(i_0)}(i_0)$ and $Y = \underline{A}_{t(i_0)}(i_0)$. If $|Y| \geq \frac{5}{4}|X|$ or if $|Y| \geq |X|$ and $\measuredangle(X,Y) \geq \frac{3}{4}\pi$, then we have (5.58) with $j = i_0+1$ (see lemma 5.13). Otherwise we have

$$\measuredangle(\underline{A}_{t(i_0)}(i_0+1),\ \underline{A}_{s(i_0)}(i_0+1)) < \frac{1}{2}\pi,$$

by lemma 5.13(c) if $|Y| \geq |X|$ and trivially if $|Y| < |X|$, so that the greatest angle of $A(i_0+1)$ exceeds $\frac{3}{4}\pi$. Suppose there is a least index $j_0 \geq i_0+1$ for which $|\underline{A}_{t(j_0)}(j_0)| \geq |\underline{A}_{s(j_0)}(j_0)|$. Since $|\underline{A}_t| < |\underline{A}_s|$ trivially

implies $\measuredangle(\underline{A}_t + \underline{A}_s, \underline{A}_s) \le \frac{1}{2}\pi$, the greatest angle of $A(j_0)$ still exceeds $\frac{3}{4}\pi$. Thus lemma 5.13(a) or (b) applies to give (5.58) with $j = j_0+1$. Now suppose there is no such j_0, which means

$$(5.59) \qquad |\underline{A}_{t(i)}(i)| < |\underline{A}_{s(i)}(i)| \quad \text{for all } i > i_0$$

(and thus the greatest angle will always exceed $\frac{3}{4}\pi$). If the index h is such that

$$|\underline{A}_h(i_0+1)| = \max_{0\le k\le 2} |\underline{A}_k(i_0+1)|,$$

then (5.59) means $t(i) \ne h$ for $i > i_0$ and $A_h(i) = A_h(i_0+1)$ for $i > i_0$. Let $j_1 \ge i_0+1$ be the first step index for which $s(j_1) \ne h$. The existence of j_1 follows from the independence of ℓ, in fact

$$j_1 \le i_0 + \sum_{k=0}^{2} \left[\frac{a_h(i_0+1)}{a_k(i_0+1)}\right] .$$

Define the indices f,g by

$$|\underline{A}_f(j_1)| \le |\underline{A}_g(j_1)| \le |\underline{A}_h(j_1)|,$$

so that $s(j_1) = g$, $t(j_1) = f$ by (5.59). By lemma 5.13(b) with $X = \underline{A}_f(j_1)$ and $Y = \underline{A}_g(j_1)$, we would have (5.58) with $j = j_1+1$ unless

$$|\underline{A}_f(j_1)| < \frac{4}{5}\, |\underline{A}_g(j_1)|,$$

which we now assume. Suppose for some $j_2 > j_1$ we have $|\underline{A}_g(j_2)| \le |\underline{A}_f(j_2)|$. Then we have, using the corollary to lemma 5.12,

$$|\underline{A}_g(j_2)| \le |\underline{A}_f(j_2)| \le |\underline{A}_f(j_1)| < \frac{4}{5}\, |\underline{A}_g(j_1)|,$$

whence (5.58) with $j = j_2$. Otherwise we have for all $i \ge j_1$

$$|\underline{A}_f(i)| < |\underline{A}_g(i)| < |\underline{A}_h(i)|$$

and by (5.59) the expansion can contain at most

$$j_1 + \left[\frac{a_h(j_1)}{a_g(j_1)}\right] + \left[\frac{a_h(j_1)}{a_f(j_1)}\right] + \left[\frac{a_g(j_1)}{a_f(j_1)}\right]$$

steps, contradicting independence.

2. We have concluded that $\delta = 0$, that is, for any $\varepsilon > 0$ there is an i_0 such that

$$\min_{0 \le k \le 2} |\underline{A}_k(i_0)| < \varepsilon.$$

Let $\varepsilon > 0$ and $|\underline{A}_0(i_0)| < \varepsilon$. Let $A(i_1)$, $i_1 > i_0$, be the first base which contains a point $A_k(i_1)$ with $|\underline{A}_k(i_1)| < |\underline{A}_0(i_0)|$. By lemma 5.12 that means $|\underline{A}_0(i_1-1)| = |\underline{A}_0(i_0)|$, and if $k \neq 0$ also $|\underline{A}_0(i_1)| = |\underline{A}_0(i_0)|$. So if $k \neq 0$ we have

$$\max\,(|\underline{A}_0(i_1)|, |\underline{A}_k(i_1)|) < \varepsilon,$$

and if $k = 0$, the i_1-th step was, say, $A_0(i_1) = A_0(i_1-1) + A_g(i_1-1)$ $(g = 1$ or $2)$ so that

$$\max\,(|\underline{A}_0(i_1-1)|, |\underline{A}_g(i_1-1)|) < 2\varepsilon.$$

Thus we shall always find in a finite number of steps a base with at least two points arbitrarily close to ℓ.

3. Suppose that $\max_{0 \le k \le 2} |\underline{A}_k(i)| \ge \delta > 0$ for all i. Choose i_0 such that, say, $\max(|\underline{A}_0(i_0)|, |\underline{A}_1(i_0)|) < \frac{1}{2}\delta$ so that $|\underline{A}_2(i_0)| > \delta$. By lemma 5.12 we have for all $i \ge i_0$ that $|\underline{A}_0(i)| < \frac{1}{2}\delta$ and $|\underline{A}_1(i)| < \frac{1}{2}\delta$, so that $|\underline{A}_2(i)| > \delta$.

Again from lemma 5.12 it now follows that for $i \ge i_0$, (s,t) can only be one of the ordered pairs $(0,1)$, $(1,0)$, $(0,2)$, $(1,2)$. The last two will appear at most

$$\left[\frac{a_0(i_0)}{a_2(i_0)}\right] + \left[\frac{a_1(i_0)}{a_2(i_0)}\right]$$

times, so $\{s,t\} = \{0,1\}$ for $i \ge i_1$, say. But with lemma 5.12 and the independence of ℓ this leads to a clear contradiction. Therefore $\delta = 0$, and the theorem has been proved. □

Since rough worst-case arguments allowed us to prove the convergence of the greatest angle algorithm it may be expected that in practice the

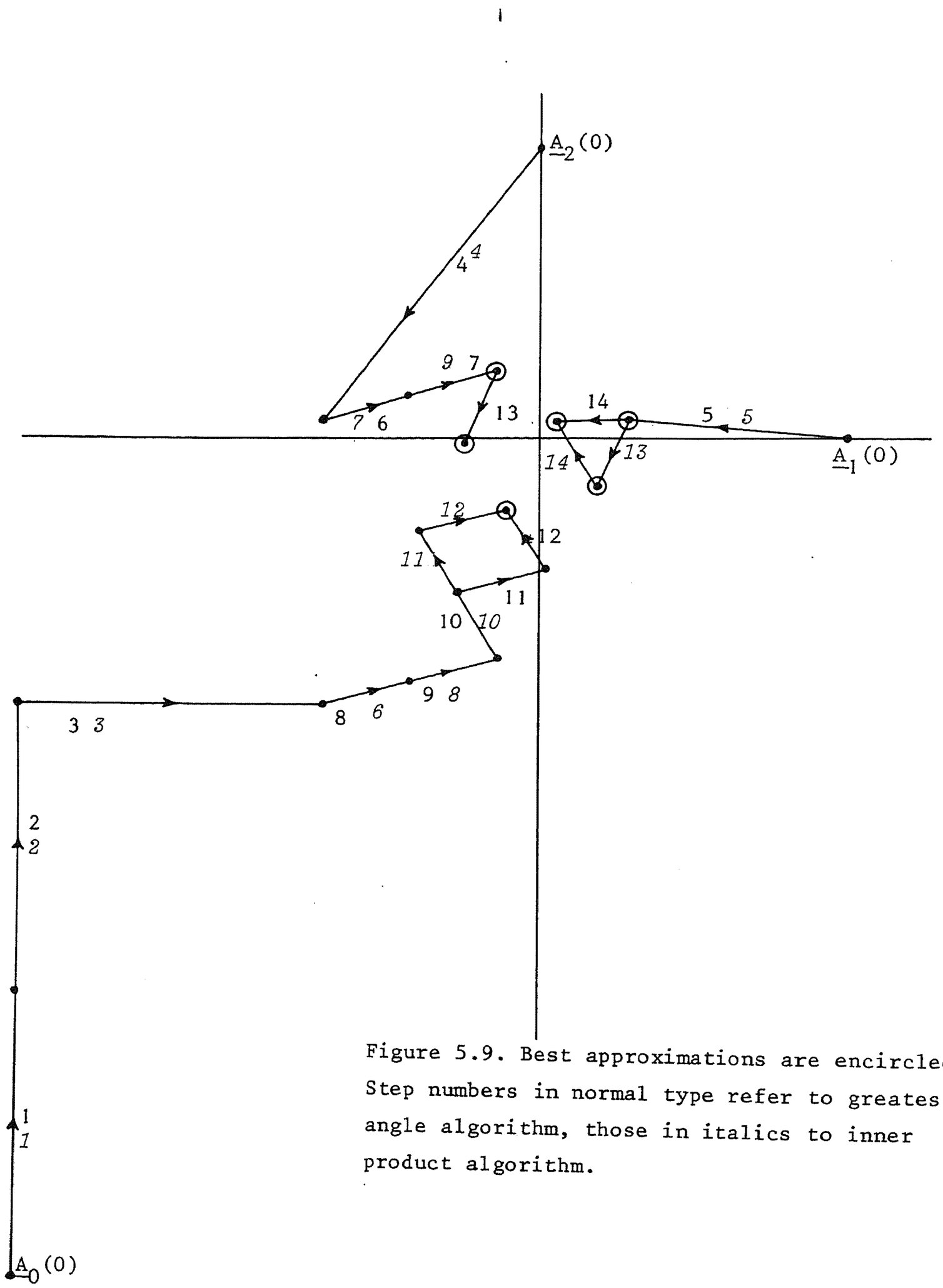


Figure 5.9. Best approximations are encircled. Step numbers in normal type refer to greatest angle algorithm, those in italics to inner product algorithm.

TABLE 5.3. (All figures rounded to 3 decimal places).

step	s	a_s	t	A_t	$\underline{A}_t$	$\lvert\underline{A}_t\rvert$	$\measuredangle(\underline{A}_0,\underline{A}_1)$	$\measuredangle(\underline{A}_1,\underline{A}_2)$	$\measuredangle(\underline{A}_2,\underline{A}_0)$
							in degrees		
begin	0	1	0	(1, 0, 0)	(-1.710, -2.924)	3.387			
	1	1.710	1	(0, 1, 0)	(1 , 0)	1			
	2	2.924	2	(0, 0, 1)	(0 , 1)	1	120	90	150
1	2	1.924	0	(1, 0, 1)	(-1.710, -1.924)	2.574	132	90	138
2	2	0.924	0	(1, 0, 2)	(-1.710, -0.924)	1.944	152	90	118
3	1	0.710	0	(1, 1, 2)	(-0.710, -0.924)	1.165	127	90	143
4	0	0.076	2	(1, 1, 3)	(-0.710, 0.076)	0.714	127	174	59
5	2	0.214	1	(1, 2, 3)	(0.290, 0.076)	0.300	142	159	59
6	1	0.496	2	(2, 3, 6)	(-0.420, 0.152)	0.447	142	151	67
7	1	0.282	2	(3, 5, 9)	(-0.130, 0.228)	0.262	142	105	113
8	1	0.206	0	(2, 3, 5)	(-0.420, -0.848)	0.946	131	105	124
9	1	0.130	0	(3, 5, 8)	(-0.130, -0.772)	0.783	115	105	140
10	2	0.138	0	(6, 10, 17)	(-0.260, -0.544)	0.603	130	105	125
11	1	0.054	0	(7, 12, 20)	(0.030, -0.468)	0.469	101	105	154
12	2	0.062	0	(10, 17, 29)	(-0.100, -0.240)	0.260	127	105	128
13	0	0.014	2	(13, 22, 38)	(-0.230, -0.012)	0.230	127	168	65
14	2	0.008	1	(14, 24, 41)	(0.060, 0.064)	0.088			

algorithm has very good approximation properties. The practical experience is that usually most of the best approximations are found, though occasionally one of them is missed.

EXAMPLE. Let us take the same example as in section 5D. That is, $\Omega = \mathbb{Z}^3$, $\ell_0 = (1, \sqrt[3]{5}, \sqrt[3]{25})$, $h(x,y,z) = x$, $\rho^2(x,y,z) = (y - x\sqrt[3]{5})^2 + (z - x\sqrt[3]{25})^2$, ℓ^* is the yz-plane. The first steps of the expansion are given in table 5.3 and in figure 5.9. The table explains itself. On comparing with the previous results (p. 88) we see that the best approximation (11, 19, 32) is missing.

Another simple selection rule for s and t is the *inner product algorithm*. Here we choose s such that

$$(5.60) \qquad a_s = \max_{0 \le k \le 2} a_k$$

and t such that

(5.61) $\quad \underline{A}_t \cdot \underline{A}_s = \min_{k \neq s} \underline{A}_k \cdot \underline{A}_s .$

LEMMA 5.15. *Formulae (5.60) and (5.61) imply* $|\underline{A}_t + \underline{A}_s| \le |\underline{A}_t|$.

COROLLARY. *With the inner product algorithm we have* $|\underline{A}_k(i+1)| \le |\underline{A}_k(i)|$ *for* $k = 0,1,2$ *and all* $i \ge 0$.

PROOF OF LEMMA 5.15. Write $u = 3-t-s$. It follows from the cofactor relation $a_s \underline{A}_s + a_t \underline{A}_t + a_u \underline{A}_u = 0$ that

(5.62) $\quad a_s |\underline{A}_s|^2 + a_t \underline{A}_t \cdot \underline{A}_s + a_u \underline{A}_u \cdot \underline{A}_s = 0.$

Suppose that $\underline{A}_t \cdot \underline{A}_s > -\frac{1}{2}|\underline{A}_s|^2$. Then $\underline{A}_u \cdot \underline{A}_s > -\frac{1}{2}|\underline{A}_s|^2$ by (5.61), and thus

$$a_s |\underline{A}_s|^2 + a_t \underline{A}_t \cdot \underline{A}_s + a_u \underline{A}_u \cdot \underline{A}_s > |\underline{A}_s|^2 (a_s - \tfrac{1}{2}a_t - \tfrac{1}{2}a_u) \ge 0$$

by (5.60), contradicting (5.62). Therefore $\underline{A}_t \cdot \underline{A}_s \le -\frac{1}{2}|\underline{A}_s|^2$, whence $|\underline{A}_t + \underline{A}_s| \le |\underline{A}_t|$. □

THEOREM 5.16. *The 2-dimensional inner product algorithm is strongly convergent.*

PROOF. Let ℓ be independent and suppose $\delta = \inf |\underline{A}_k(i)| > 0$, where the inf is taken over $k = 0,1,2$ and all $i \ge 0$. From (5.60) it is clear that

$$\lim_{i \to \infty} a_k(i) = 0 \quad \text{for } k = 0,1,2.$$

Now we have, for $g \neq f$, $h = 3-f-g$,

$$\begin{aligned} |\underline{A}_g(i)|\,|\underline{A}_f(i)| \sin \measuredangle(\underline{A}_g(i), \underline{A}_f(i)) &= |\det\,(\underline{A}_g(i), \underline{A}_f(i))| \\ &= \frac{1}{|\ell_0|} |\det\,(\ell_0, A_g(i), A_f(i))| \\ &= \frac{1}{|\ell_0|} |\det\,(a_h(i) A_h(i), A_g(i), A_f(i))| \\ &= \frac{a_h(i)}{|\ell_0|} |\det A(i)|, \end{aligned}$$

so that

$$\sin \measuredangle(\underline{A}_g(i), \underline{A}_f(i)) < \frac{a_h(i)\,|\det A(i)|}{|\ell_0|\,\delta^2}.$$

Therefore $\lim_{i\to\infty} \sin \measuredangle(\underline{A}_g(i),\underline{A}_f(i)) = 0$. Thus, for i large enough two of the angles $\measuredangle(\underline{A}_g(i),\underline{A}_f(i))$, $g \neq f$, exceed $\frac{3}{4}\pi$ while the third is less than $\frac{1}{2}\pi$. The proof now proceeds along exactly the same lines as that of theorem 5.14 (notice that (5.61) implies that $\measuredangle(\underline{A}_s,\underline{A}_t) > \frac{1}{2}\pi$). □

Like the greatest angle algorithm the inner product algorithm usually finds most of the best approximations, but it sometimes misses one.

EXAMPLE. In our previous example, the inner product expansion (table 5.4 and figure 5.9) does not include the best approximation (13, 22, 38).

The speed of the inner product algorithm can be increased somewhat if we replace the subtractive step $A_t(i+1) = A_t(i) + A_s(i)$ by $A_t(i+1) = A_t(i) + b\,A_s(i)$ with

TABLE 5.4. (All figures rounded to 3 decimal places).

step	s	a_s	t	A_t	$\underline{A}_t$	$\underline{A}_0\cdot\underline{A}_1$	$\underline{A}_1\cdot\underline{A}_2$	$\underline{A}_2\cdot\underline{A}_0$
begin	0	1	0	(1, 0, 0)	(-1.710, -2.924)			
	1	1.710	1	(0, 1, 0)	(1 , 0)			
	2	2.924	2	(0, 0, 1)	(0 , 1)	-1.710	0	-2.924
1	2	1.924	0	(1, 0, 1)	(-1.710, -1.924)	-1.710	0	-1.924
2	2	0.924	0	(1, 0, 2)	(-1.710, -0.924)	-1.710	0	-0.924
3	1	0.710	0	(1, 1, 2)	(-0.710, -0.924)	-0.710	0	-0.924
4	0	0.076	2	(1, 1, 3)	(-0.710, 0.076)	-0.710	-0.710	0.434
5	2	0.214	1	(1, 2, 3)	(0.290, 0.076)	-0.276	-0.200	0.434
6	1	0.634	0	(2, 3, 5)	(-0.420, -0.848)	-0.186	-0.200	0.234
7	1	0.420	2	(2, 3, 6)	(-0.420, 0.152)	-0.186	-0.110	0.048
8	1	0.344	0	(3, 5, 8)	(-0.130, -0.772)	-0.096	-0.110	-0.063
9	1	0.130	2	(3, 5, 9)	(-0.130, 0.228)	-0.096	-0.020	-0.159
10	2	0.138	0	(6, 10, 17)	(-0.260, -0.544)	-0.117	-0.020	-0.091
11	2	0.062	0	(9, 15, 26)	(-0.390, -0.316)	-0.137	-0.020	-0.022
12	1	0.054	0	(10, 17, 29)	(-0.100, -0.240)	-0.047	-0.020	-0.042
13	0	0.022	1	(11, 19, 32)	(0.190, -0.164)	0.021	-0.062	-0.042
14	2	0.008	1	(14, 24, 41)	(0.060, 0.064)	-0.021	0.007	-0.042
15	1	0.032	0	(24, 41, 70)	(-0.040, -0.176)			

TABLE 5.5. (All figures rounded to 3 decimal places).

step	s	a_s	t	A_t	$\underline{A}_t$	$\lvert\underline{A}_t\rvert$	
begin	0	1	0	(1, 0, 0)	(-1.710, -2.924)	3.387	
	1	1.710	1	(0, 1, 0)	(1 , 0)	1	
	2	2.924	2	(0, 0, 1)	(0 , 1)	1	b:=2
1	2	0.924	0	(1, 0, 2)	(-1.710, -0.924)	1.944	b:=1
2	1	0.710	0	(1, 1, 2)	(-0.710, -0.924)	1.165	b:=1
3	0	0.076	2	(1, 1, 3)	(-0.710, 0.076)	0.714	b:=1
4	2	0.214	1	(1, 2, 3)	(0.290, 0.076)	0.300	b:=3
5	1	0.482	0	(4, 7, 11)	(0.160, -0.696)	0.714	b:=2
6	1	0.054	2	(3, 5, 9)	(-0.130, 0.228)	0.262	b:=2
7	2	0.062	0	(10, 17, 29)	(-0.100, -0.240)	0.260	b:=1
8	0	0.022	1	(11, 19, 32)	(0.190, -0.164)	0.251	b:=1
9	2	0.008	1	(14, 24, 41)	(0.060, 0.064)	0.088	b:=2
10	1	0.010	0	(38, 65,111)	(0.020, -0.112)	0.115	

$$b = \min\left(\left[\frac{a_s}{a_t}\right], \left[\frac{1}{2} - \frac{\underline{A}_s \cdot \underline{A}_t}{|\underline{A}_s|^2}\right]\right).$$

Notice that $b \geq 1$, since $a_s \geq a_t$ and $\underline{A}_s \cdot \underline{A}_t \leq -\frac{1}{2}|\underline{A}_s|^2$ (cf. lemma 5.15). The accelerated algorithm is also strongly convergent, because we have $|\underline{A}_t + b\underline{A}_s| \leq |\underline{A}_t|$ as an easy equivalent to lemma 5.15, and this was the relevant property in the proof of theorem 5.16.

EXAMPLE. Applied to our standard example the accelerated inner product algorithm (table 5.5) still misses the best approximation (13, 22, 38). It now runs for only 10 steps.

In 1979 JURKAT, KRATZ & PEYERIMHOFF[1] proposed a 2-dimensional algorithm also with the intention to obtain best approximations to the Euclidean radius function. Their algorithm was the first one for which something was proved about its relation to best approximations, though it is not true that an expansion by their algorithm always contains all best approximations. We give a brief sketch of the algorithm (which we shall call the JKP-algorithm).

The central notion is that of a *regulated* lattice base: Let

$A = \{A_0, A_1, A_2\}$ be such that $\ell_0 \in \{A\}^+$ and $h(A) > 0$, and specify the permutation f,g,h of 0,1,2 by $a_h \geq a_g \geq a_f$. Then A is regulated if

$$\frac{1}{3}\pi \leq \measuredangle(\underline{A}_f, \underline{A}_g) \leq \frac{2}{3}\pi$$

and

$$h(A_h) + \min\ (h(A_f), h(A_g)) \geq \max\ (h(A_f), h(A_g)).$$

The idea is that a regulated base contains at least one approximation of good quality (see lemma 4.8).

Now suppose that $A(i)$ is regulated, with $f = f(i)$, $g = g(i)$, $h = h(i)$ such that

$$(5.63) \qquad a_h(i) \geq a_g(i) \geq a_f(i).$$

Let c_1, c_2, c be integers satisfying $c_1 = \left[\frac{a_h(i)}{a_g(i)}\right]$,

$$\measuredangle(\underline{A}_h, \underline{A}_g + c_2\underline{A}_h) \geq \frac{2}{3}\pi > \measuredangle(\underline{A}_h, \underline{A}_g + (c_2+1)\underline{A}_h),$$

and $c = \min\ (c_1, c_2)$. In the next regulated base $A(i+1)$, write $f' = f(i+1)$, $g' = g(i+1)$, $h' = h(i+1)$ such that

$$(5.64) \qquad a_{h'}(i+1) \geq a_{g'}(i+1) \geq a_{f'}(i+1).$$

Then $\{A_{g'}(i+1), A_{f'}(i+1)\}$ are, without regard to ordering,

(i) $\{A_g(i),\ A_h(i)\}$ <u>if</u> $\frac{1}{3}\pi \leq \measuredangle(\underline{A}_g(i), \underline{A}_h(i)) \leq \frac{2}{3}\pi$; <u>else</u>

(ii) $\{A_f(i),\ A_h(i)\}$ <u>if</u> $\frac{1}{3}\pi \leq \measuredangle(\underline{A}_f(i), \underline{A}_h(i)) \leq \frac{2}{3}\pi$; <u>else</u>

(iii) $\{A_f(i),\ A_g(i) + (c+1)A_h(i)\}$ <u>if</u> $c = c_1 < c_2$; <u>else</u>

(iv) $\{A_h(i),\ A_g(i) + (c+1)A_h(i)\}$ <u>if</u> $\frac{1}{3}\pi \leq \measuredangle(\underline{A}_h(i), \underline{A}_g(i) + (c+1)\underline{A}_h(i)) \leq \frac{2}{3}\pi$; <u>else</u>

(v) $\{A_g(i) + cA_h(i),\ A_g(i) + (c+1)A_h(i)\}$.

In all cases the third point of $A(i+1)$, $A_{h'}(i+1)$, is the point determined uniquely by (5.64) and $\ell_0 \in \{A(i+1)\}^+$. Notice that by (5.63) $c_1 \geq 1$, so if $c < 0$ we have $c_2 < 0$ which means that (i) applies. Thus, if (iii), (iv) or (v) applies, then $c \geq 0$ and $A(i+1) \neq A(i)$. The algorithm has the purely

additive structure postulated in definition 2.3. In fact, the transition from $A(i)$ to $A(i+1)$ consists of the following (sequence of successive) step(s) (we give s, t and b; "b = max" means $b = \left[\frac{a_s}{a_t}\right]$):

(i) $s = h$, $t = f$, $b = \max$; then $s = g$, $t = f$, $b = \max$.
(ii) $s = h$, $t = g$, $b = \max$.
(iii) $s = h$, $t = g$, $b = c+1$; then $s = g$, $t = h$, $b = \max$; then $s = f$, $t = h$, $b = \max$.
(iv) $s = h$, $t = g$, $b = c+1$; then $s = g$, $t = f$, $b = \max$; then $s = h$, $t = f$, $b = \max$.
(v) $s = h$, $t = g$, $b = c$; then $s = h$, $t = f$, $b = \max$; then $s = g$, $t = h$, $b = 1$; then $s = g$, $t = f$, $b = \max$.

Now Jurkat, Kratz and Peyerimhoff associate with each base $A(i)$ in an expansion of the JKP-algorithm a best approximation $B(i)$. It is the non-zero point with

$$|\underline{B}(i)| \leq \min_{0 \leq k \leq 2} |\underline{A}_k(i)|$$

for which $|h(B(i))|$ is minimal. Often $B(i) \in A(i)$. At the cost of some long and intricate proofs [1]), all in planar geometry, it turns out that for any i there are at most four best approximations B' with

$$h(B(i)) < h(B') < h(B(i+1)),$$

and these are found among a well-defined set of nine points. The speed of the JKP-algorithm is comparable to that of the accelerated best approximation algorithm (lemmata 5.10 and 5.11), but it does not guarantee that all best approximations occur in the expansion itself.

EXAMPLE. In our standard example, the expansion by the algorithm of Jurkat, Kratz and Peyerimhoff (see table 5.6) does not include the best approximation (11, 19, 32). Notice that the standard base of $\mathbb{Z}^3$ is not regulated. Therefore we start with another base, which is regulated.

---

[1]) Unfortunately there seems to be a flaw in the proof of theorem 4, case 4.

TABLE 5.6. (All figures rounded to 3 decimal places).

step	s	a_s	t	A_t	$\underline{A}_t$	f g h	case
begin	0	1	0	(1, 1, 2)	(-0.710, -0.924)		
	1	0.710	1	(0, 1, 0)	(1 , 0)		
	2	0.924	2	(0, 0, 1)	(0 , 1)	1 2 0	(v) c=0 b=0,1,1,0
1	0	0.290	1	(1, 1, 2)	(0.290, -0.924)		
	2	0.634	0	(1, 1, 3)	(-0.710, 0.076)	0 2 1	(ii) b=1
2	1	0.076	2	(1, 2, 3)	(0.290, 0.076)	1 0 2	(ii) b=2
3	2	0.054	0	(3, 5, 9)	(-0.130, 0.228)	2 1 0	(ii) b=3
4	0	0.062	1	(10, 17, 29)	(-0.100, -0.240)	2 0 1	(iv) c=0 b=1,1,0
5	1	0.014	0	(13, 22, 38)	(-0.230, -0.012)		
	0	0.008	2	(14, 24, 41)	(0.060, 0.064)		

5F. Application: Units in cubic number fields.

Let $f(x) \in \mathbb{Z}[x]$ be an irreducible polynomial of the third degree and with negative discriminant. Then it has one real root θ and two complex conjugate roots θ', θ''. The number θ defines a real cubic number field $\mathbb{Q}(\theta)$. For $\alpha \in \mathbb{Q}(\theta)$ we denote its conjugates by α', α''. Let $\mathcal{O}(\theta)$ be the maximal ring of integers in $\mathbb{Q}(\theta)$, $E(\theta)$ its unit group. From the unit theorem of Dirichlet we know that there is a so-called fundamental unit ε_0 such that $E(\theta) = \{\pm\varepsilon_0^m \mid m \in \mathbb{Z}\}$. With ε_0 also $-\varepsilon_0$ and $1/\varepsilon_0$ are fundamental; therefore we fix $\varepsilon_0 > 1$. Dirichlet's theorem does not tell us how to find ε_0 in a given field $\mathbb{Q}(\theta)$. We shall see that ε_0 can be determined with the 2-dimensional best approximation algorithm.

With $\mathcal{O}(\theta)$ we associate the lattice

$$\Omega(\theta) = \{(\alpha, \operatorname{Re}\alpha', \operatorname{Im}\alpha') \mid \alpha \in \mathcal{O}(\theta)\}$$

in $\mathbb{R}^3$. The point $(\alpha, \operatorname{Re}\alpha', \operatorname{Im}\alpha')$ will also be denoted by α. We choose the height function

(5.65) $\quad h(x,y,z) = x$

and the Euclidean radius function

(5.66) $\quad \rho(x,y,z) = \sqrt{y^2 + z^2}.$

This gives

(5.67) $$\begin{cases} h(\alpha) = \alpha \\ \rho(\alpha) = |\alpha'| \end{cases}$$

for $\alpha \in \Omega(\theta)$. Notice that $|h(\alpha)| = |h(\beta)|$ implies $\alpha = \pm\beta$.

LEMMA 5.17. *If $\varepsilon \in E(\theta)$, then ε is a best approximation of $\Omega(\theta)$ to the x-axis with respect to the height and distance functions (5.65), (5.66).*

PROOF. Suppose some $\alpha \in \Omega(\theta)$ satisfies simultaneously $\rho(\alpha) \le \rho(\varepsilon)$ and $|h(\alpha)| < |h(\varepsilon)|$. By (5.67) this implies $|\alpha'| \le |\varepsilon'|$ and $|\alpha| < |\varepsilon|$. If we denote the norm of α by $N(\alpha) = \alpha\alpha'\alpha'' = \alpha|\alpha'|^2$, this gives $|N(\alpha)| < |N(\varepsilon)| = 1$, whence $N(\alpha) = 0$, or $\alpha = 0$. □

LEMMA 5.18. *If β is a best approximation, ε a unit, then $\beta\varepsilon$ is a best approximation.*

PROOF. Suppose some α satisfies simultaneously $\rho(\alpha) \le \rho(\beta\varepsilon)$ and $|h(\alpha)| < |h(\beta\varepsilon)|$. By (5.67) this means $|\alpha'| \le |\beta'\varepsilon'|$ and $|\alpha| < |\beta\varepsilon|$, whence $\left|\frac{\alpha'}{\varepsilon'}\right| \le |\beta'|$ and $\left|\frac{\alpha}{\varepsilon}\right| < |\beta|$. Therefore $\rho\left(\frac{\alpha}{\varepsilon}\right) \le \rho(\beta)$ and $\left|h\left(\frac{\alpha}{\varepsilon}\right)\right| < |h(\beta)|$, so that $\alpha = 0$ because β is a best approximation. □

COROLLARY. *If $\varepsilon_0 > 1$ is the fundamental unit of $\mathbb{Q}(\theta)$ and $1 < \beta_1 < \ldots < \beta_j = \varepsilon_0$ are the best approximations with heights between 1 and ε_0, then every best approximation is of the form $\pm\beta_k\varepsilon_0^m$ for some $m \in \mathbb{Z}$ and $k \in \{1,\ldots,j\}$.*

PROOF. If β is a best approximation, then so is $|\beta|\varepsilon_0^{-m}$ by lemma 5.18, where $m \in \mathbb{Z}$ may be chosen such that $\log|\beta| > m.\log\varepsilon_0 \ge \log|\beta| - \log\varepsilon_0$. But we have $1 < |\beta|\varepsilon_0^{-m} \le \varepsilon_0$, hence $|\beta|\varepsilon_0^{-m} = \beta_k$ for some $k \in \{1,\ldots,j\}$. □

REMARK. This is the periodicity result we promised at the end of section 4C. Multiplication by ε_0 represents the periodicity transformation T in (4.14).

Lemma 5.17 and the corollary to lemma 5.18 show that all best approximations, including the fundamental unit, can be determined by computing a finite part of the expansion of some suitably chosen base of $\Omega(\theta)$ by means of the 2-dimensional best approximation algorithm.

The question arises whether the sequence of best approximations with respect to another radius function than ρ in (5.66) can be periodic. This indeed is possible. Take, for example, a cubic field $\mathbb{Q}(\theta)$ of negative discriminant in which the only best approximations with respect to $\rho(x,y,z) = \sqrt{y^2 + z^2}$ are the units. Let the integer $t \geq 2$ be such that $|N(\alpha)| \geq t$ for all $\alpha \in \mathcal{O}(\theta)$, $\alpha \notin E(\theta) \cup \{0\}$, and suppose $t > \varepsilon_0$. Such fields exist (see example 2 below). Take $u > 1$ to be a real number satisfying $u^4 < \min(\varepsilon_0, t/\varepsilon_0)$, and let the radius function ρ_1 be such that

$$\frac{1}{u}\rho(\alpha) \leq \rho_1(\alpha) \leq u\rho(\alpha) \quad \text{for all } \alpha \in \Omega(\theta)$$

(e.g. $\rho_1(x,y,z) = \sqrt{y^2 + u^2z^2}$). First we show: Every unit $\varepsilon = \varepsilon_0^m$ is a best approximation with respect to ρ_1. Indeed, if ε were not, then for some $\alpha \in \mathcal{O}(\theta)$, $\alpha < \varepsilon$, we would have $\rho_1(\alpha) \leq \rho_1(\varepsilon)$, whence

$$|\alpha'| = \rho(\alpha) \leq u\rho_1(\alpha) \leq u\rho_1(\varepsilon) \leq u^2\rho(\varepsilon) = u^2|\varepsilon'|.$$

Now if $\alpha \notin E(\theta)$ this gives $|N(\alpha)| = |\alpha||\alpha'|^2 \leq u^4\varepsilon|\varepsilon'|^2 = u^4 < t/\varepsilon_0 < t$ by choice of u, whence $\alpha = 0$. And if $\alpha \in E(\theta)$, then $\alpha < \varepsilon = \varepsilon_0^m$ implies $\alpha \leq \varepsilon_0^{m-1}$, whence $|N(\alpha)| \leq \varepsilon_0^{m-1}u^4|\varepsilon_0'|^{2m} = u^4/\varepsilon_0 < 1$, a contradiction.

Secondly, we show: If $\alpha \notin E(\theta)$, then α is not a ρ_1-best approximation. Choose $m \in \mathbb{Z}$ such that $\varepsilon_0^m < \alpha < \varepsilon_0^{m+1}$ and suppose $\rho_1(\alpha) \leq \rho_1(\varepsilon_0^m)$. Then $|\alpha'| \leq u^2|\varepsilon_0'|^m$, whence $|N(\alpha)| \leq \varepsilon_0^{m+1}u^4|\varepsilon_0'|^{2m} = u^4\varepsilon_0 < t$, hence $\alpha = 0$. Therefore $\alpha \notin E(\theta) \cup \{0\}$ implies $\rho_1(\alpha) \geq \rho_1(\varepsilon_0^m)$, so that α is not a best approximation with respect to ρ_1.

We shall now give a few examples.

<u>EXAMPLE 1.</u> Let θ be the real root of

$$x^3 + 3x - 2 = 0.$$

By standard methods (DELONE & FADDEEV[1], §17) we find that $1, \theta, \theta^2$ form

an integral base of $\mathcal{O}(\theta)$. We start our expansion with $\mathcal{A}(0)$ consisting of $A_0(0) = 1$, $A_1(0) = \theta$, $A_2(0) = 3 + \theta^2 = 2/\theta$. The three angles

$$\begin{cases} \sphericalangle(\underline{A}_0(0), \underline{A}_1(0)) = \sphericalangle(\underline{A}_0(0), \underline{A}_2(0)) = \arg\ \theta' \approx 99^\circ \\ \sphericalangle(\underline{A}_1(0), \underline{A}_2(0)) \approx 162^\circ \end{cases}$$

all exceed $\frac{1}{2}\pi$. Thus by lemma 5.8 all points α with $|\alpha'| < 1$, $\alpha > 0$, are in $\{\mathcal{A}(0)\}^+$, in particular so all best approximations higher than 1, including ε_0. Table 5.7 gives the beginning of the expansion. The last base in the table, $\mathcal{A}(11)$, consists of

$$\begin{aligned} A_0(11) &= 17 + 3\theta + 5\theta^2, \\ A_1(11) &= 10 + 2\theta + 3\theta^2, \\ A_2(11) &= 40 + 7\theta + 12\theta^2. \end{aligned}$$

Except for $10 + 2\theta + 3\theta^2$ any non-zero lattice point in $\{\mathcal{A}(11)\}^+$ has at least the height of $17 + 3\theta + 5\theta^2$. Therefore all best approximations between 1 and $17 + 3\theta + 5\theta^2$ have been found. We list them in increasing order of height:

TABLE 5.7. (All figures rounded to 3 decimal places).

step	s	a_s	t	$A_t = \alpha$	Reα'	Imα'	$\lvert\alpha'\rvert^2$	case
begin	0	0.643	0	1	(1 ,	0)	1	
	1	1.077	1	θ	(−0.298,	1.807)	3.355	
	2	1.807	2	$3 + \theta^2$	(−0.178,	−1.077)	1.192	Ib
1	2	0.730	1	$3 + \theta + \theta^2$	(−0.476,	0.730)	0.759	Ib
2	2	0.087	0	$4 + \theta^2$	(0.822,	−1.077)	1.836	II
3	1	0.434	0	$7 + \theta + 2\theta^2$	(0.346,	−0.347)	0.241	II
4	1	0.347	2	$6 + \theta + 2\theta^2$	(−0.654,	−0.347)	0.547	Ib
5	1	0.260	2	$9 + 2\theta + 3\theta^2$	(−1.130,	0.383)	1.424	II
6	0	0.556	2	$16 + 3\theta + 5\theta^2$	(−0.784,	0.036)	0.616	II
7	0	0.296	1	$10 + 2\theta + 3\theta^2$	(−0.130,	0.383)	0.163	Ib
8	0	0.209	2	$23 + 4\theta + 7\theta^2$	(−0.438,	−0.311)	0.289	Ib
9	0	0.122	2	$30 + 5\theta + 9\theta^2$	(−0.092,	−0.658)	0.441	II
10	1	0.173	2	$40 + 7\theta + 12\theta^2$	(−0.222,	−0.275)	0.125	Ib
11	1	0.051	0	$17 + 3\theta + 5\theta^2$	(0.216,	0.036)	0.048	

$$3 + \theta + \theta^2, \quad \text{norm } 3$$
$$6 + \theta + 2\theta^2, \quad \text{norm } 4$$
$$7 + \theta + 2\theta^2, \quad \text{norm } 2$$
$$10 + 2\theta + 3\theta^2, \quad \text{norm } 2$$
$$17 + 3\theta + 5\theta^2, \quad \text{norm } 1.$$

Since $17 + 3\theta + 5\theta^2$ is the first unit that occurs, it is fundamental.

EXAMPLE 2. Let $\theta \approx 1.325$ be the real root of

$$x^3 - x - 1 = 0.$$

Starting with $A(0) = \{1, \theta, \theta^2\}$, which forms a base of $\Omega(\theta)$, the first step of the algorithm is of type Ib and takes 1 into $1 + \theta = \theta^3$, so that $A(1) = \{\theta^3, \theta, \theta^2\}$. We see that the projection of $A(1)$ on the yz-plane is obtained from that of $A(0)$ by rotation over arg θ' and multiplication with $|\theta'|$. Therefore the algorithm gives a periodic expansion with a period length of one step, and it follows that all (and only) positive powers of θ will be found: In every step the algorithm adds the second-lowest to the lowest point and obtains a new best approximation.

CUSICK[6] determined the best approximations (with $\Omega = \mathbb{Z}^3$ and $\ell_0 = (1, \theta, \theta^2)$) in the square norm sense of Diophantine approximation (section 4E): All of them are units, but not all units occur as best approximations. There is in fact an "almost periodic" pattern in the sequence of those $m \in \mathbb{Z}$ for which θ^m is a square norm best approximation; the lengths of the "almost periods" are integers j such that j.arg θ' is approximately an integral multiple of π. This example shows that a square norm is not always natural.

EXAMPLE 3. Let D be a cube-free integer greater than 1, and write $D = fg^2$ where f and g are square-free integers. Put $\theta = \sqrt[3]{fg^2}$, $\phi = \sqrt[3]{f^2g}$. Since $\mathbb{Q}(\theta) = \mathbb{Q}(\phi)$ we may suppose that $\phi > \theta$ (or $f > g$).

According to DEDEKIND[1] an integral base of $\mathcal{O}(\theta)$ is

$$(5.68) \qquad 1, \theta, \phi \quad \text{if } D \not\equiv \pm 1 \pmod 9$$

$$(5.69) \qquad 1, \theta, \gamma = \frac{u + v\theta + \phi}{3} \text{ if } D \equiv \pm 1 \pmod 9, \begin{cases} u \equiv g \pmod 3 \\ v \equiv fg \pmod 3. \end{cases}$$

It is never difficult to find a good initial base $A(0)$ for our expansions, because we have:

LEMMA 5.19. *(a) The base (5.68) is of the first kind.*
(b) The base (5.69) is of the first kind if we take $v \in \{-1,1\}$, $u \in \{1,2\}$.

PROOF. (a) We have $\theta' = \omega\theta$ and $\phi' = \omega^2\phi$ with $\omega = e^{2\pi i/3}$. Thus the three angles of the base (5.68) are all exactly $\frac{2}{3}\pi$.
(b) We have $\measuredangle(\underline{1},\underline{\theta}) = \frac{2}{3}\pi$ as above, and $\measuredangle(\underline{\theta},\underline{\gamma}) > \frac{1}{3}\pi$ because of

$$\operatorname{Im}\gamma' = \frac{1}{6}\sqrt{3}(v\theta - \phi) \le \frac{1}{6}\sqrt{3}(\theta - \phi) < 0$$

(by $\phi > \theta$ and $v \le 1$). It remains to see that $\measuredangle(\underline{1},\underline{\gamma}) \ge \frac{1}{3}\pi$, for which it suffices to show that $\operatorname{Re}\gamma' \le \frac{-1}{\sqrt{3}}\operatorname{Im}\gamma'$. Indeed we find

$$\operatorname{Re}(3\gamma') = u - \frac{1}{2}v\theta - \frac{1}{2}\phi \le \frac{1}{2}\phi - \frac{1}{2}v\theta = \frac{-1}{\sqrt{3}}\operatorname{Im}(3\gamma')$$

if $\phi \ge 2$. However, if $\phi < 2$, then also $\theta < 2$, so that $fg < 4$, and D can be 2 or 3 only, hence not $\pm 1 \pmod 9$. □

EXAMPLE 4. Consider the field $\mathbb{Q}(\theta)$ with $\theta = \sqrt[3]{29}$. An integral base is $1, \theta, \theta^2$ by (5.68). After 64 steps of the accelerated 2-dimensional best approximation algorithm (235 in the slow version) we have obtained the following best approximations:

β_1	= (	9,	3,	1)	of norm 4
β_2	= (	123,	40,	13)	of norm 24
β_3	= (	132,	43,	14)	of norm 7
β_4	= (	387,	126,	41)	of norm 14
β_5	= (	1699,	553,	180)	of norm 12
β_6	= (	5352,	1742,	567)	of norm 7
β_7	= (	23494,	7647,	2489)	of norm 6
β_8	= (	148015,	48177,	15681)	of norm 28
β_9	= (	324876,	105743,	34418)	of norm 3
β_{10}	= (	2445641,	796025,	259096)	of norm 22
β_{11}	= (	6539156,	2128412,	692771)	of norm 35
β_{12}	= (	8984797,	2924437,	951867)	of norm 12
β_{13}	= (	33818423,	11007466,	3582790)	of norm 11

$\beta_{14} = (128734536,\ 41901452,\ 13638689)$ of norm 21
$\beta_{15} = (214665852,\ 69871001,\ 22742121)$ of norm 34
$\beta_{16} = (343400388,\ 111772453,\ 36380510)$ of norm 25.

This information suffices to prove that $\varepsilon_0 = \beta_9^3/3 = \beta_9^3/N(\beta_9)$, if we use the following argument of WILLIAMS[2]: Write $\psi = \beta_9$. Using $N(\psi) = 3$ we easily verify that $\psi^3/3 \in \mathcal{O}(\theta)$ and $N(\psi^3/3) = 1$. Now $|\psi'| < 1$, whence $\psi > N(\psi) = 3$ and $\psi^3/3 > 1$. Also $\psi < \varepsilon_0$ since otherwise our table would have contained ε_0 before $\beta_9 = \psi$. Thus $\psi^3/3 < \varepsilon_0^3$. We conclude that $\psi^3/3$ equals ε_0 or ε_0^2. Suppose $\psi^3/3 = \varepsilon_0^2$. Because $\psi < 10^6$ we then have $\varepsilon_0 < 10^9/\sqrt{3}$ and ε_0 would have been found before β_{16} (which is greater than 10^9). This not being the case we conclude that $\psi^3/3 = \varepsilon_0 \approx 3.086\times10^{17}$.

A table of units for the fields $\mathbb{Q}(\sqrt[3]{D})$, $2 \le D \le 199$, D cube-free, was given by SVED[1]. These units were calculated by means of an algorithm of Szekeres (to be treated in section 6A), which has about the same approximation qualities as the algorithms of section 5E (in particular it does not guarantee that one will find all best approximations). Sved did not stop after the first unit had been found, but she tried to find several units: If all are powers of the first one it is reasonable to suppose the first one to be fundamental. Using a multi-precision computer program of the 2-dimensional best approximation algorithm, Te Riele checked some values of D where Sved had large units, namely

D = 41, 59, 69, 77, 83, 85, 87, 97, 101, 103, 111, 113, 119, 131, 137, 146, 149, 155, 157, 159, 167, 173, 177, 178, 183, 187, 191, 193, 194.

In all instances it was shown that Sved's unit is indeed fundamental. In the largest case, D = 167, a 154-digit precision was used to find the unit, which is about 6×10^{95}; here Sved had used 360 digits without finding a second unit [1]).

---

[1]) Recently, JEANS & HENDY[1] gave a method to decide whether a given unit in a pure cubic field $\mathbb{Q}(\sqrt[3]{D})$ is fundamental. Essentially they use the fact that a unit > 1 is a P.V.-number to give a more sophisticated version of checking all prime roots to see if they are algebraic integers. They also decided that Sved's units for D = 167, 177 are fundamental.

REMARK. WILLIAMS, CORMACK & SEAH[1] have shown how the use of multi-precision arithmetic can largely be avoided and replaced by integer arithmetic with integers not exceeding D^6 (for large D). They did so for the method of Voronoi mentioned in section 5A, but their technique is equally suited for our algorithm. For details on computer implementation of multi-dimensional continued fraction algorithms we refer to the appendix.

EXAMPLE 5. Consider the field $\mathbb{Q}(\theta)$, $\theta = \sqrt[3]{D}$, where $D = d^3 - 1$ ($d \in \mathbb{Z}$, $d \geq 2$). We calculate the expansion of $A(0) = \{1, \theta, \theta^2\}$ (which is not necessarily a base of $\mathcal{O}(\theta)$, but it is a base of a $\mathbb{Z}$-module in $\mathcal{O}(\theta)$). As an aid in the verification of certain inequalities we put $\delta = d - \theta = (d^2 + d\theta + \theta^2)^{-1}$. Since $d \geq 2$ we have $\delta < \frac{1}{4}$, and hence

$$(5.70) \qquad d - \frac{1}{4} < \theta < d.$$

We estimate $0 < 3d\delta = \dfrac{3d}{d^2 + d\theta + \theta^2} < \dfrac{3d}{3\theta^2} < \dfrac{d}{(d - \frac{1}{4})^2} < \frac{3}{4}$, using $d \geq 2$, hence

$$(5.71) \qquad 0 < d\delta < \frac{1}{4}.$$

1. We start with an accelerated step of type Ib, using lemma 5.10 and $[\theta] = d - 1$. This gives $A(1) = \{1, \theta, \alpha_1\}$ with $\alpha_1 = \theta^2 + (d-1)\theta$. We estimate

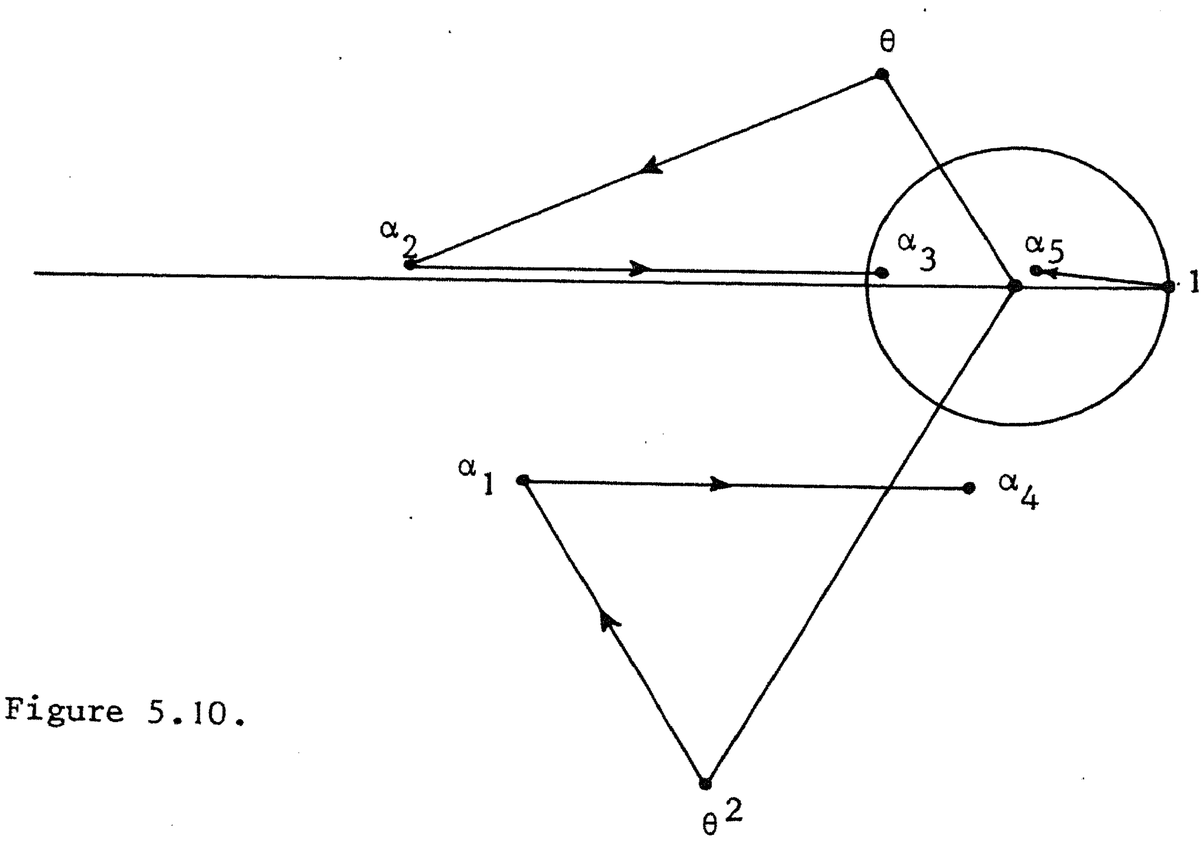


Figure 5.10.

$$\mathrm{Re}(\alpha_1') = -\tfrac{1}{2}\theta^2 - \tfrac{1}{2}\theta(d-1) = -\tfrac{1}{2}(d-\delta)(2d-\delta-1) \begin{cases} > -d^2 + \frac{1}{2}d & (\text{by } \delta > 0) \\ < -d^2 + \frac{1}{2}d + \frac{1}{2} & (\text{by } (5.71)) \end{cases}$$

$$\mathrm{Im}(\alpha_1') = \tfrac{1}{2}\theta\sqrt{3}(d-1-\theta) = \tfrac{1}{2}\sqrt{3}(d-\delta)(\delta-1) \begin{cases} < -\frac{7}{16}(d-\frac{1}{8})\sqrt{3} < -1 & (\text{by } \delta > \frac{1}{8}) \\ > -\frac{1}{2}d\sqrt{3} & (\text{by } \delta > 0) \end{cases}$$

and easily verify

$$(5.72) \qquad \tfrac{5}{6}\pi < \measuredangle(\underline{1}, \underline{\alpha}_1) < \pi.$$

2. The second step is also of type Ib: It gives $A(2) = \{1, \alpha_2, \alpha_1\}$ with $\alpha_2 = \theta + \alpha_1 = \theta^2 + d\theta$. We estimate

$$\mathrm{Re}(\alpha_2') = -\tfrac{1}{2}\theta^2 - \tfrac{1}{2}d\theta = -\tfrac{1}{2}(d-\delta)(2d-\delta) \begin{cases} > -d^2 & (\text{using } \delta > 0) \\ < -d^2 + \frac{1}{2} & (\text{using } (5.71)) \end{cases}$$

$$0 < \mathrm{Im}(\alpha_2') = \tfrac{1}{2}\theta\sqrt{3}(d-\theta) = \tfrac{1}{2}(d-\delta)\delta\sqrt{3} < \tfrac{1}{8}\sqrt{3} \quad (\text{using } (5.71)).$$

3. For the third step we take $\alpha_3 = \alpha_2 + d^2 - 1 = d^2 - 1 + d\theta + \theta^2$, so that $A(3) = \{1, \alpha_3, \alpha_1\}$. This is an accelerated step of type II (lemma 5.11). From the estimates for α_2' we have

$$\begin{cases} -1 < \mathrm{Re}(\alpha_3') < -\frac{1}{2} \\ 0 < \mathrm{Im}(\alpha_3') < \frac{1}{8}\sqrt{3} \end{cases}$$

so that

$$(5.73) \qquad \measuredangle(\underline{1}, \underline{\alpha}_3) > \tfrac{5}{6}\pi.$$

Furthermore we find $N(\alpha_3) = 3d^2 - 3d$ from the norm form,

$$N(x + y\theta + z\theta^2) = x^3 + Dy^3 + D^2z^3 - 3Dxyz.$$

Therefore $|\alpha_3'|^2 = \dfrac{3d^2 - 3d}{d^2 - 1 + d\theta + \theta^2} < 1$ (using (5.71)).

4. The point of $A(3)$ with minimal radius being α_3 we continue with a sequence of type Ib steps to obtain $A(4) = \{1, \alpha_3, \alpha_4\}$ with $\alpha_4 = \alpha_1 + d^2 - [\frac{1}{2}d] - 1$ (note: this is not the maximal value of b in lemma 5.10). From $\mathrm{Im}(\alpha_1') < -1$ we know that the points $\alpha_1 + c$, $0 < c \le d^2 - [\frac{1}{2}d] - 1$,

are not best approximations. We have

$$\mathrm{Re}(\alpha_4') = \mathrm{Re}(\alpha_1') + d^2 - [\tfrac{1}{2}d] - 1 \begin{cases} < \frac{1}{2}d - [\tfrac{1}{2}d] - \frac{1}{2} \le 0 \\ > \frac{1}{2}d - [\tfrac{1}{2}d] - 1 \ge -1 \end{cases}$$

from the estimates of $\mathrm{Re}(\alpha_1')$. Also

$$\mathrm{Im}(\alpha_4') = \mathrm{Im}(\alpha_1') < -\frac{7}{16}(d - \tfrac{1}{8})\sqrt{3} < -\sqrt{3} \qquad \text{for } d \ge 3$$

and, when $d = 2$,

$$\begin{cases} \mathrm{Im}(\alpha_4') = \frac{1}{2}\sqrt{3}\sqrt[3]{7}(1 - \sqrt[3]{7}) \approx 0.8732\sqrt{3} \\ \mathrm{Re}(\alpha_4') = 2 - \frac{13}{2}\sqrt[3]{7}(1 + \sqrt[3]{7}) \approx -0.7861, \end{cases}$$

so that

$$|\mathrm{Im}(\alpha_4')| > \sqrt{3}|\mathrm{Re}(\alpha_4')|$$

in either case, whence

$$\text{(5.74)} \qquad \tfrac{1}{3}\pi < \measuredangle(\underline{1}, \underline{\alpha}_4) < \tfrac{2}{3}\pi.$$

5. At this point we interrupt the algorithm to do the step $\alpha_5 = 1 + \alpha_3 = d^2 + d\theta + \theta^2$, which makes $A(5) = \{\alpha_5, \alpha_3, \alpha_4\}$. We verify that

$$\mathrm{Re}(\alpha_5') = \tfrac{3}{2}d\delta - \tfrac{1}{2}\delta^2 \ge \tfrac{3}{2}(d - \delta)\delta = \sqrt{3}\,\mathrm{Im}(\alpha_5'),$$

so that

$$\measuredangle(\underline{1}, \underline{\alpha}_5) < \tfrac{1}{6}\pi.$$

With (5.74) this allows application of lemma 5.1: There is no best approximation in $\{A(4)\}^+ - \{A(5)\}^+$ except for 1. It also proves that $\ell_0 \in \{A(5)\}^+$.

We have $N(\alpha_5) = 1$ and the lowest point in $\{A(5)\}^+$, not (a multiple of) α_3, α_4 or α_5, is $\alpha_3 + \alpha_4$. Since the height of $\alpha_3 + \alpha_4$ exceeds that of α_5, all best approximations between 1 and the unit $\alpha_5 = \varepsilon_0$ have been found: $\alpha_3 = \varepsilon_0 - 1$ is the only one, and the sequence of best

approximations is

$$\ldots\ldots,\ 1,\ \alpha_3,\ \varepsilon_0,\ \alpha_3\varepsilon_0,\ \varepsilon_0^2,\ \alpha_3\varepsilon_0^2,\ \ldots\ldots\ldots$$

HISTORICAL REMARKS. In the past, some small and not always entirely correct tables of fundamental units were computed by hand, chiefly by trial and error methods. The oldest seems to be given by MARKOV[1] in 1891 for cubic fields $\mathbb{Q}(\sqrt[3]{D})$, $2 \le D \le 70$ (reproduced in DELONE & FADDEEV[1], p. 304). Other such tables exist from CASSELS[1] and SELMER[1] (for $D \le 50$, $D \le 100$ respectively). WOLFE[1] calculated the fundamental unit of the module with base $\{1, \sqrt[3]{D}, \sqrt[3]{D^2}\}$ for $2 \le D \le 100$. This module is not always $\mathcal{O}(\sqrt[3]{D})$; in fact, its fundamental unit may be any power of the fundamental unit ε_0 of $\mathcal{O}(\sqrt[3]{D})$, as NAGELL[1], note I, pointed out. REID[1,2] gave an incomplete table of bases, discriminants, units and class numbers for many fields $\mathbb{Q}(\theta)$ with θ defined by $\theta^3 - p\theta - q = 0$, $|p|, |q| \le 9$. The method of USPENSKY[1] depends on successive minima of certain quadratic forms associated to $\mathbb{Q}(\theta)$.

It was recognized early that the theory of 2-dimensional continued fractions provided a more systematic approach: The problem has been an example of the application of nearly all algorithms that have been proposed. The oldest methods in this respect are those mentioned in section 3C, remark 2 (p. 29): To obtain units as eigenvalues of periodic expansions, or to find a cofactor which happens to be a unit. Typical results in this directions are the tables of BERNSTEIN[14] and GÜTING[2], who both tried to find units of $\mathbb{Q}(\sqrt[3]{D})$, $2 \le D \le 998$, with the Jacobi-Perron algorithm and Güting's algorithm, respectively. Both authors used a 16-digit precision, not enough for most values of D; however they also missed units which should be obtainable with that precision, e.g. for $D = 34$, 51. Other papers looking in this direction are BERNSTEIN[15,16,18,20,21]. STENDER[1], WILLIAMS[1] and RUDMAN[1] considered the question whether some units of Bernstein are fundamental. DAUS[2] gave a completer version of Reid's table, using his variation of the Jacobi-Perron algorithm (section 3G). A better result was obtained by SVED[1], who gave a unit table for $\mathbb{Q}(\sqrt[3]{D})$, $2 \le D \le 199$, calculated by Szekeres' algorithm (see example 4 of this section). A slightly larger table was given by WADA[1] (for $2 \le D \le 250$). It is not clear which method he used.

The characterization of units as best approximations in an appropriate setting was first found in the work of VORONOI[1] and BERWICK[1]; later

DUBOIS[2,3,5,6] employed it, too, but none of these authors did his calculations with a continued fraction algorithm. Instead, they resorted to the method outlined in section 5A. Voronoi's method is explained in detail in DELONE[1] and DELONE & FADDEEV[1]; a computer adaptation will be found in WILLIAMS & ZARNKE[2].

At present, the largest unit tables have been calculated by Voronoi's method, mainly by Williams and his colleagues. A table for $\mathbb{Q}(\sqrt[3]{D})$, $2 \le D \le 998$, was given by BEACH, WILLIAMS & ZARNKE[1]; the largest unit in this table is of order 10^{330}, for $D = 951$. A table for $\mathbb{Q}(\theta)$, with $\theta^3 - p\theta - q = 0$, $|p|, |q| \le 15$, is found in WILLIAMS & ZARNKE[2,3]. ANGELL[1] listed all cubic fields with discriminant between -20000 and 0 (there are 3169 such fields), and gave the fundamental units as calculated by the method of Voronoi.

WILLIAMS & ZARNKE[1] and WILLIAMS & HOLTE[1] asked for which D (≤ 15000, 50000 respectively) $x^3 - Dy^3 = 1$ has a solution in integers x,y with $y \ne 0$. As is known (NAGELL[1]) this is the case if and only if $x - y\sqrt[3]{D}$ is the fundamental unit ε_0^{-1} of $\mathbb{Q}(\sqrt[3]{D})$ (unless $D = 19$, 20 or 28). To determine if for given D there exists a solution, it is necessary to calculate ε_0, unless there is an easy argument (e.g., congruence considerations) to prove that there is no solution.

In another project involving $\mathbb{Q}(\sqrt[3]{D})$ for D large, WILLIAMS, CORMACK & SEAH[1] and WILLIAMS[2] tried to find such fields with D ($\le 100{,}000$ and 200,000 respectively) a prime $\equiv 2 \pmod 3$, such that the field has class number one. To compute the class number one needs the regulator $R = \log \varepsilon_0$ of the field. Since only $\log \varepsilon_0$, not ε_0 itself, is needed, they managed to avoid multi-length arithmetic by implementing special techniques in Voronoi's method. The largest regulator obtained was $\log \varepsilon_0 \approx 455713.750$ for $D = 199109$, where ε_0 is the 134645-th best approximation greater than 1.

CHAPTER 6. THE MULTI-DIMENSIONAL CASE

6A. The Szekeres and inner product algorithms.

In this chapter we survey the scarce work that has been done for dimensions higher than two.

SZEKERES[1] published in 1970 the first multi-dimensional continued fraction algorithm where the principal goal was to find good or best approximations. His article started with a list of properties that an ideal generalization of the continued fraction algorithm would have to possess: It should find all best approximations (and preferably only good ones), and the cofactors, which are integral linear combinations of $a_0(0), a_1(0), \ldots, a_n(0)$, should be good approximations to zero compared with the size of the coefficients of those combinations. Furthermore, it should be able to determine the dependence rank of the line ℓ, and to find units of an algebraic number field when applied to an integral base of that field. Last but not least, the algorithm should be reasonably simple. The 2-dimensional best approximation algorithm satisfies several of these requirements.

Szekeres then defined the following subtractive algorithm, ρ and h being a radius and height function as in chapter 4.

(1) $i := s := 0$;

(2) determine $t \neq s$ such that

$$\rho\left(\frac{A_t(i)}{h(A_t(i))} - \frac{A_s(i)}{h(A_s(i))}\right) \tag{6.1}$$

is maximal;

(3) if $a_t(i) > a_s(i)$ then interchange s and t;

(4) $A(i+1) = A(i)I_{st}$; $i := i+1$; goto (2).

Geometrically, one considers the central projection toward the origin of $A(i)$ on the plane $h = 1$, and takes $A_t(i)$ such that its central projection is farthest from that of $A_s(i)$ (farthest as measured by ρ; originally Szekeres works with $\Omega = \mathbf{Z}^{n+1}$ and the square radius function (4.25)). Szekeres does not make clear why he keeps the index s fixed until a t is chosen such that $a_t > a_s$. One argument is, doubtless, that we have to

TABLE 6.1. (All figures rounded to 3 decimal places).

step	s	a_s	t	A_t	corr. for $(1, \sqrt[3]{5}, \sqrt[3]{25})$
begin	0	0.076	0	(1, 0, 0)	(1, 1, 2)
	1	0.214	1	(1, 1, 0)	(1, 1, 3)
	2	0.710	2	(1, 1, 1)	(1, 2, 3)
1	2	0.634	0	(2, 1, 1)	(2, 3, 5)
2	2	0.420	1	(2, 2, 1)	(2, 3, 6)
3	2	0.344	0	(3, 2, 2)	(3, 5, 8)
4	2	0.130	1	(3, 3, 2)	(3, 5, 9)
5	2	0.054	0	(4, 3, 3)	(4, 7, 11)
6	0	0.022	2	(5, 4, 4)	(5, 9, 14)
7	1	0.192	0	(7, 6, 5)	(7, 12, 20)
8	1	0.138	2	(8, 7, 6)	(8, 14, 23)
9	1	0.116	0	(10, 9, 7)	(10, 17, 29)
10	1	0.062	2	(11, 10, 8)	(11, 19, 32)
11	1	0.008	2	(14, 13, 10)	(14, 24, 41)
12	0	0.014	1	(13, 12, 9)	(13, 22, 38)

evaluate only n quantities in step (2); if also s were free and the maximum were to be taken over all s,t-pairs, then this would require $\frac{1}{2}n(n+1)$ evaluations.

Szekeres claims that "it is quite likely that best approximations [with respect to ρ and h] are never missed out by the algorithm" when applied to a base $A(0)$ with $\ell_0 \in \{A(0)\}^+$ and $h(A(0)) > 0$.

EXAMPLE. With $n = 2$ we run our favourite example of sections 5D,E. However, we must have $h(A(0)) > 0$ because of the divisions in (6.1), therefore we take, according to Szekeres, $\ell_0 = (1, \sqrt[3]{25} - 2, \sqrt[3]{5} - 1)$ and let $A(0)$ consist of $A_0(0) = (1, 0, 0)$, $A_1(0) = (1, 1, 0)$ and $A_2(0) = (1, 1, 1)$. We give the expansion with the Euclidean radius function $\rho(x,y,z) = \sqrt{y^2 + z^2}$ on ℓ^* and the height function $h(x,y,z) = x$ in table 6.1. The last column contains the approximations to $(1, \sqrt[3]{5}, \sqrt[3]{25})$ which correspond to the approximations A_t of ℓ_0. We see that all best approximations have indeed been found.

Nevertheless, there exist counterexamples to Szekeres' claim (see the example with $n = 6$ below; also see DUBOIS[7] for a counterexample in two

dimensions).

SZEKERES[1] proved that his algorithm is weakly convergent in all dimensions. Nevertheless, the nature of the selection rule for s and t, which depends on central rather than orthogonal projection, makes it very difficult to prove strong convergence, let alone something on the quality of the approximations. Apart from a paper by CUSICK[5], who verified a conjecture of Szekeres on a particular example, no theoretical results have been published on the algorithm. It has been compared experimentally with other n-dimensional algorithms, mainly that of Jacobi-Perron, and the results indicate that the approximation quality of Szekeres' algorithm usually is much better than that of the J.P.A. See JURKAT, KRATZ & PEYERIMHOFF[1] and VAN DE LUNE & TE RIELE[1].

FERGUSON & FORCADE[1] mention the following example: If Szekeres' algorithm (in its original version) is applied to the four numbers

$$\xi_1 = 0.5366654780$$
$$\xi_2 = 0.3595184257$$
$$\xi_3 = 0.2420244429$$
$$\xi_4 = \xi_1 + \xi_2 + \xi_3 - 1,$$

then it does not detect a dependence relation before exhausting single computer precision (about 150 steps), in spite of the existence of the relation $1 - \xi_1 - \xi_2 - \xi_3 + \xi_4 = 0$.

Prompted by such examples Szekeres has very recently modified his original selection rule for the index t (formula (6.1)). He now considers, according to MACK[private communication, 1981], the following selection rule of t, still using central projections: Let H_k $(k \neq s)$ be the (n-1)-dimensional hyperplane in ℓ^* passing through the point $\frac{\underline{A}_k + \underline{A}_s}{h(A_k + A_s)}$ and the n-1 points $\frac{\underline{A}_i}{h(A_i)}$ $(i \neq k,s)$. Then determine $t \neq s$ such that H_t has maximal perpendicular distance from $\frac{\underline{A}_s}{h(A_s)}$. As yet no results of this new rule are known.

We now present another algorithm which experimentally gives approximations of good quality, and which compares favourably with Szekeres' algorithm in several aspects: The *n-dimensional inner product algorithm* takes s such that

(6.2) $$a_s = \max_{0\le k\le n} a_k$$

and t ($\neq$ s) such that

(6.3) $$\underline{A}_t\cdot\underline{A}_s = \min_{k\neq s} \underline{A}_k\cdot\underline{A}_s .$$

It may be accelerated by choosing the step size

(6.4) $$b = \max\left(1,\ \min\left(\left[\frac{a_s}{a_t}\right],\ \left[\frac{1}{2}-\frac{\underline{A}_s\cdot\underline{A}_t}{|\underline{A}_s|^2}\right]\right)\right).$$

The number of evaluations required to determine s and t from (6.2), (6.3) is linear in n, as was the case with Szekeres' algorithm. If we would not fix s by (6.2) but take the minimum in (6.3) over all s,t-pairs, this would require $\frac{1}{2}n(n+1)$ evaluations of inner products.

From (6.2) and (6.3) and $a_0\underline{A}_0 + a_1\underline{A}_1 + \ldots + a_n\underline{A}_n = 0$ we easily derive

$$\underline{A}_t\cdot\underline{A}_s \le -\frac{1}{n}|\underline{A}_s|^2$$

as in lemma 5.15, but only in the case $n\le 2$ this does provide a simple proof of strong convergence (theorem 5.16). For $n\ge 3$ nothing has been proved yet.

<u>EXAMPLE.</u> In $\mathbb{R}^7$ we take $\Omega = \mathbb{Z}^7$ and $\ell_0 = (1,\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,\alpha_6)$ with $\alpha_1 = \frac{\log 13}{\log 17}$, $\alpha_2 = \frac{\log 11}{\log 17}$, $\alpha_3 = \frac{\log 7}{\log 17}$, $\alpha_4 = \frac{\log 5}{\log 17}$, $\alpha_5 = \frac{\log 3}{\log 17}$, $\alpha_6 = \frac{\log 2}{\log 17}$. We take the height function $h(x_0,x_1,\ldots,x_6) = x_0$ and the Euclidean radius function

$$\rho(x_0,x_1,\ldots,x_6) = \left(\sum_{k=1}^{6} (x_0\alpha_k - x_k)^2\right)^{\frac{1}{2}} ;$$

ℓ^* is the hyperplane $x_0 = 0$. In table 6.2 we give the height (x_0-coordinate) and radius of selected points B from the Szekeres expansion. The selection criterion is that the expansion contain no P with both $|h(P)| < |h(B)|$ and $\rho(P)\le\rho(B)$. For each selected B the number of the step in which the point was found is given. As a measure of the quality we add the number $h(B)(\rho(B))^6$, which differs from the formal quality as defined by (4.22) only by a constant factor: $qu(B) = \frac{\pi^3}{384}|h(B)|(\rho(B))^6$. In table 6.3 we do the

TABLE 6.2. (Szekeres' algorithm).

step	$h(B) = x_0$	$\rho(B)$	$h\rho^6$	
9	3	0.6941	0.34	=
22	13	0.6864	1.36	
26	23	0.6635	1.96	
28	25	0.5804	0.96	
38	54	0.5123	0.98	
40	83	0.4805	1.02	
42	118	0.3560	0.24	=
49	201	0.2896	0.12	=
69	1,153	0.2895	0.56	
73	1,341	0.2494	0.32	
84	3,327	0.2267	0.45	
86	6,156	0.2178	0.66	
91	10,055	0.1861	0.42	=
94	13,583	0.1727	0.36	=
99	28,747	0.1300	0.14	=
111	58,647	0.1073	0.090	=
117	111,326	0.0891	0.056	=
126	169,973	0.0868	0.073	=
130	249,225	0.0863	0.10	
143	1,083,345	0.0814	0.32	=
148	1,516,126	0.0775	0.33	=
154	3,008,119	0.0732	0.46	
157	4,524,245	0.0480	0.055	=
170	9,467,688	0.0423	0.054	=
176	14,241,158	0.0382	0.044	=
208	350,448,032	0.0318	0.36	
224	1,648,635,145	0.0285	0.88	
236	3,639,936,733	0.0205	0.27	
244	8,335,041,604	0.0110	0.015	=

TABLE 6.3. (Inner product algorithm).

step	$h(B) = x_0$	$\rho(B)$	$h\rho^6$
12	3	0.6941	0.34
16	7	0.5667	0.23
17	12	0.5094	0.21
18	13	0.4906	0.18
21	32	0.4858	0.42
23	44	0.4389	0.31
28	106	0.4282	0.65
29	118	0.3560	0.24
31	201	0.2896	0.12
32	294	0.2591	0.089
47	2,174	0.2547	0.59
45	3,033	0.2483	0.71
46	3,528	0.2220	0.42
52	8,714	0.2092	0.73
54	10,055	0.1861	0.42
55	13,583	0.1727	0.36
57	22,591	0.1605	0.39
61	28,747	0.1300	0.14
62	58,647	0.1073	0.090
65	111,326	0.0891	0.056
67	169,973	0.0868	0.073
77	1,083,345	0.0814	0.32
78	1,516,126	0.0775	0.33
83	4,524,245	0.0480	0.055
93	9,467,688	0.0423	0.054
91	14,241,158	0.0382	0.044
102	98,072,695	0.0302	0.074
105	319,068,923	0.0294	0.21
107	669,516,955	0.0167	0.015
116	2,970,419,778	0.0158	0.046
121	8,335,041,604	0.0110	0.015

same for the inner product expansion. Note that a point B can easily be recovered from its x_0-coordinate by $x_1 \log 17 \approx x_0 \log 13$, etc.

We find that both algorithms produce about the same quality of approximation. The inner product algorithm has the advantage of achieving this in about half as many steps as Szekeres' algorithm. This is not due to the acceleration (6.4), because as it happened, the inner product algorithm took $b = 1$ throughout the expansion. More likely it is a result of the preference of Szekeres' algorithm for adding low points. The reason of this tendency is, of course, the fact that in formula (6.1) the points are weighted by the inverse of their heights. (The new selection rule of Szekeres will probably show the same tendency).

In tables 6.2 and 6.3 we find examples where Szekeres' algorithm misses a best approximation. Consider, for example, the point $B = (294,\ldots.)$ from table 6.3, with $\rho(B) = 0.2591$. If it is a best approximation, then Szekeres' algorithm has missed it. If it is not a best approximation, then there exists a lattice point $B_1 \neq 0$ with $|h(B_1)| < 294$ and $\rho(B_1) < 0.2591$. However, table 6.2 contains no such B_1. By a similar argument, the inner product algorithm also misses best approximations (consider, e.g., the point $B = (1341,\ldots.)$ from table 6.2).

To solve *inhomogeneous approximation* problems we can use the same method as was explained at the end of section 5D (p. 93). Suppose that ℓ is an independent line and that we wish to find lattice points arbitrarily close to the line $Y + \ell$, where Y is a (non-zero) vector in ℓ^*.

Let a point $P \in \Omega$ be given such that $|Y - \underline{P}| = \delta > 0$. To find a point $P_1 \in \Omega$ closer to $Y + \ell$ we use a multi-dimensional continued fraction algorithm, hoping to obtain a lattice base $A = \{A_0, A_1, \ldots., A_n\}$ with

$$(6.5) \qquad \max_{0 \le k \le n} |\underline{A}_k| < \frac{2\delta}{n}.$$

Solve for the unknowns $\eta_1, \ldots\ldots, \eta_n$ the equation $Y - \underline{P} = \eta_1 \underline{A}_1 + \ldots\ldots + \eta_n \underline{A}_n$. This is possible since ℓ is independent. If $\min(\eta_1, \ldots., \eta_n) < 0$ put

$$\sigma = \min(\eta_1/a_1, \ldots\ldots, \eta_n/a_n).$$

Then $\sigma < 0$ and

$$Y - \underline{P} = Y - \underline{P} - \sigma O = -\sigma a_0 \underline{A}_0 + \sum_{k=1}^{n} (\eta_k - \sigma a_k)\underline{A}_k$$

is a linear combination with positive coefficients of (at most) n projections of points of the lattice base A. Thus $Y - \underline{P} = \sum_{k \neq f} \xi_k \underline{A}_k$ for some $f \in \{0,\ldots,n\}$ and $\xi_k \geq 0$.

If we now take $P_1 = P + \sum_{k \neq f} [\frac{1}{2} + \xi_k]A_k$, then

$$|Y - \underline{P}_1| = \left| \sum_{k \neq f} (\xi_k - [\tfrac{1}{2} + \xi_k])\underline{A}_k \right| < \frac{n}{2} \sum_{k \neq f} |\underline{A}_k| < \delta$$

by (6.5), so that P_1 is closer to $Y + \ell$ than P.

Contrary to the 2-dimensional case there is in general no warrant that we shall obtain a base A satisfying (6.5). In applications, however, it is sufficient if we do find such a base. The algorithm of the next section does guarantee that we shall obtain such a base A. Unfortunately, it is computationally less feasible than the algorithms of this section.

6B. Convergent algorithms by induction.

In 1979 FERGUSON & FORCADE[1] showed how to construct convergent continued fraction algorithms by induction on the dimension. It was then proved for the first time that convergent algorithms exist in any dimension.

Ferguson and Forcade agree that an expansion $A(i)$, $i = 0,1,2,\ldots,$ *terminates* as soon as a base $A(j)$ with a cofactor equal to zero is found. They call an n-dimensional continued fraction algorithm *splitting* if it yields for every line ℓ and lattice base $A(0)$ with $\ell_0 \in \{A(0)\}^+$ either a terminating or a strongly convergent expansion. A splitting algorithm is strongly convergent, since an independent line cannot have a terminating expansion. Of course, the ordinary continued fraction algorithm is splitting.

Now let $n \geq 2$, and suppose we have an (n-1)-dimensional splitting algorithm SplA(n-1) at our disposition. We construct an n-dimensional splitting algorithm SplA(n) as follows.

Let $A(i_0)$ be a lattice base with $\ell_0 \in \{A(i_0)\}^+$. By underlining we denote, as usual, the projection on ℓ^*, parallel to ℓ.

(1) If $A(i_0)$ has a cofactor equal to zero, then SplA(n) terminates by

definition.

(2) Otherwise we choose the index f such that

$$(6.6) \qquad |\underline{A}_f(i_0)| = \min_{0\le k\le n} |\underline{A}_k(i_0)|.$$

In the n-dimensional subspace ℓ^* the base $A^*(i_0) = \{\underline{A}_k(i_0) \mid k \neq f\}$ defines a lattice Ω^* of rank n. Indeed, if the $\underline{A}_k(i_0)$, $k \neq f$, were dependent, then there would exist a relation

$$\sum_{k\neq f} \lambda_k \underline{A}_k(i_0) = 0,$$

whence

$$\sum_{k\neq f} \lambda_k A_k(i_0) = \lambda \ell_0$$

for some $\lambda \neq 0$ (because the $A_k(i_0)$ are independent) and some λ_k not all zero. But comparing this with the cofactor relation (2.1) we see that $a_k(i_0) = \lambda_k/\lambda$ for $k \neq f$ and $a_f(i_0) = 0$, contradicting our assumption.

In ℓ^*, let m be the line through 0 and $\underline{A}_f(i_0)$. With $m_0 = -a_f(i_0)\underline{A}_f(i_0)$ it follows from the projected cofactor relation,

$$\sum_{k=0}^{n} a_k(i_0)\underline{A}_k(i_0) = 0,$$

that

$$\sum_{k\neq f} a_k(i_0)\underline{A}_k(i_0) = m_0,$$

i.e. the cofactors of $A^*(i_0)$ with respect to m_0 are precisely the $a_k(i_0)$, $k \neq f$. This makes the following construction possible.

(3) To $A^*(i_0)$ and m we apply in ℓ^* the algorithm SplA(n-1).

For each step $\underline{A}_t(i+1) = \underline{A}_t(i) + b\underline{A}_s(i)$ of SplA(n-1) (where $s,t \neq f$) the corresponding step of SplA(n) will be $A_t(i+1) = A_t(i) + bA_s(i)$, and $\underline{A}_t(i+1)$ is indeed the projection of $A_t(i+1)$ because projection commutes with the transformation I^b_{st}.

We continue this until either a termination occurs in SplA(n-1) and hence in SplA(n), or a base $A(i_1)$, $i_1 \ge i_0$, with $A_f(i_1) = A_f(i_0)$ is

obtained such that

$$\max_{k\neq f} d(\underline{A}_k(i_1),m) < \delta|\underline{A}_f(i_0)| \tag{6.7}$$

and

$$\underline{A}_k(i_1).\underline{A}_f(i_0) < 0 \text{ for all } k\neq f. \tag{6.8}$$

Here δ *is a number, fixed in advance, in the range* $0 < \delta < \frac{1}{2}\sqrt{3}$.
The fact that (6.7) will be obtained if no termination occurs follows from the induction hypothesis, i.e. the assumption that SplA(n-1) is splitting. The same is true for (6.8): Let ϕ be the linear function on ℓ^* defined by $\phi(\underline{A}_k(i_0)) = 1$ for $k\neq f$, which exists because the $\underline{A}_k(i_0)$, $k\neq f$, are linearly independent. Then all points found during the application of SplA(n-1) satisfy $\phi \geq 1$ in view of the additive nature of the algorithm. Now (6.8) follows if

$$\max_{k\neq f} d(\underline{A}_k(i_1),m) < \varepsilon := \inf \{d(X,m) \mid X\in \ell^*,\ X.\underline{A}_f(i_0) \geq 0,\ \phi(X) \geq 1\},$$

where trivially $\varepsilon > 0$.

(4) Next we go from $A(i_1)$ *to* $A(i_2)$, *where* $A_f(i_2) = A_f(i_0)$ *and for* $k\neq f$

$$A_k(i_2) = A_k(i_1) + b_k A_f(i_0) \tag{6.9}$$

with

$$b_k = \left[-\frac{\underline{A}_f(i_0).\underline{A}_k(i_1)}{|\underline{A}_f(i_0)|^2} \right].$$

By (6.8) we have $b_k \geq 0$, and in fact b_k is chosen such that

$$\underline{A}_f(i_0).\underline{A}_k(i_2) \leq 0 < \underline{A}_f(i_0).(\underline{A}_k(i_2) + \underline{A}_f(i_0)). \tag{6.10}$$

Also, the only cofactor that changes during the (at most) n steps (6.9), is that of $A_f(i_0)$. But we have

$$\sum_{k=0}^{n} a_k(i_2)\,\underline{A}_k(i_2) = 0\ ,$$

whence

$$\sum_{k=0}^{n} a_k(i_2)\underline{A}_k(i_2).\underline{A}_f(i_0) = 0$$

and this implies $a_f(i_2) \geq 0$ by (6.10). Therefore we still have $\ell_0 \in \{A(i_2)\}^+$. Moreover, when C_k denotes the orthogonal projection of $\underline{A}_k(i_2)$ on m $(k \neq f)$ we have $|C_k| < |\underline{A}_f(i_0)|$ by (6.10), and hence

$$(6.11) \qquad |\underline{A}_k(i_2)| = \sqrt{|C_k|^2 + (d(\underline{A}_k(i_2),m))^2} < |\underline{A}_f(i_0)|\sqrt{1+\delta^2}$$

for $k \neq f$ by (6.7).

(5) Now if for some $g \neq f$, $|C_g| \leq \frac{1}{2}|\underline{A}_f(i_0)|$ we even have

$$(6.12) \qquad |\underline{A}_g(i_2)| \leq |\underline{A}_f(i_0)|\sqrt{\tfrac{1}{4}+\delta^2}$$

instead of (6.11).

If, however, $\min_{k\neq f} |C_k| > \frac{1}{2}|\underline{A}_f(i_0)|$ *we choose any* $g \neq f$ *and replace* $A_g(i_2)$ *or* $A_f(i_2)$ *by the sum* $A_g(i_2) + A_f(i_2)$*, depending on whether* $a_g(i_2) \leq a_f(i_2)$ *or* $a_g(i_2) > a_f(i_2)$*, respectively.*

For this g we then have

$$(6.13) \qquad |\underline{A}_g(i_2) + \underline{A}_f(i_2)| \leq |\underline{A}_f(i_0)|\sqrt{\tfrac{1}{4}+\delta^2}.$$

During the cycle (1),(2),(3),(4),(5) -- which we shall call one *iteration* of SplA(n) -- the index f keeps the same value, defined by (6.6). After (5) we return to (1), and in (2) f may receive a new value. Starting the iteration with $A(j_0)$ $(j_0 = i_0)$ and ending with $A(j_1)$ $(j_1 = i_2$ or $i_2+1)$ we have from (6.6) and (6.12) or (6.13)

$$(6.14) \qquad \min_{0\leq k\leq n} |\underline{A}_k(j_1)| \leq \sqrt{\tfrac{1}{4}+\delta^2}.\ \min_{0\leq k\leq n} |\underline{A}_k(j_0)|.$$

Continuing this way we obtain a sequence of bases $A(j_0)$, $A(j_1)$, $A(j_2)$, ..., in our expansion, each resulting from its predecessor by an iteration of SplA(n), such that

$$(6.15) \qquad \min_{0\leq k\leq n} |\underline{A}_k(j_m)| \leq (\sqrt{\tfrac{1}{4}+\delta^2})^m.\ \min_{0\leq k\leq n} |\underline{A}_k(j_0)|$$

for $m \geq 0$. If no termination occurs this implies

$$\lim_{m\to\infty} \min_{0\leq k\leq n} |\underline{A}_k(j_m)| = 0$$

since $\delta < \frac{1}{2}\sqrt{3}$. The proof of convergence is now completed if we observe that after step (4) we have (6.11), i.e.

$$\max_{0\leq k\leq n} |\underline{A}_k(j_{m+1})| < \sqrt{1+\delta^2}.\ \min_{0\leq k\leq n} |\underline{A}_k(j_m)|$$

$$\leq \sqrt{1+\delta^2}(\sqrt{\tfrac{1}{4}+\delta^2})^m.\ \min_{0\leq k\leq n} |\underline{A}_k(j_0)|.$$

HISTORICAL REMARK. The original algorithm of FERGUSON & FORCADE[1] differs from the one presented here in the following respects: It allows negative cofactors and prefers nearest integer to greatest integer choices, and it measures distance by the maximum of the absolute values of a vector's coordinates instead of by the Euclidean metric. Moreover, its main purpose was to detect dependence relations rather than to solve approximation problems.

EXAMPLE. Using theorem 4.1 and their algorithm, FERGUSON & FORCADE[1] proved that the imaginary parts $\gamma_1,\ldots\ldots,\gamma_7$ of the first seven non-trivial zeros of Riemann's zeta function have no integral linear relation with coefficients less than 65.

We present a proof by means of the inner product algorithm of the preceding section. The standard base of $\mathbb{Z}^7$ was expanded along $\ell_0 = (\gamma_1,\ldots\ldots,\gamma_7)$ with the values of $\gamma_1,\ldots\ldots,\gamma_7$ taken correct to 24 decimal places from a table of TE RIELE[1]. After 249 steps a base $A = \{A_0,\ldots\ldots,A_6\}$ was obtained with

$$\max_{0\leq k\leq 6} d(A_k,\ell) < 0.001.$$

The first coordinates of the points of this base are

1,142,857,059,047,434,529
3,329,045,285,199,282,919
3,975,536,609,301,247,926
7,378,042,537,276,921,345

8,700,621,178,764,960,734
9,447,144,539,016,545,816
9,575,223,657,159,743,444.

By theorem 4.1 (with $\varepsilon = 0.001$) we conclude that there is no integral linear relation $c_1\gamma_1 + \ldots\ldots + c_7\gamma_7 = 0$ with

$$0 < \sqrt{c_1^2 + \ldots\ldots + c_7^2} \le \frac{1}{0.001} = 1000$$

or

$$0 < \max_{1\le k\le 7} |c_k| \le \frac{1000}{\sqrt{7}} \approx 377.9.$$

It is somewhat curious that Ferguson and Forcade write they obtained their bound of 65 "employing a multiple precision of only 750 decimal places".

In the case $n = 2$ the factor $\sqrt{\frac{1}{4} + \delta^2}$ in (6.14) can be improved to $\sqrt{\frac{1}{16} + \delta^2}$ without great cost. For this, we use the simple fact that for any real numbers x,y:

$$\min\,(\|x\|, \|y\|, \|x+y\|, \|2x+y\|) \le \frac{1}{4}, \tag{6.16}$$

when $\|x\| = |x - [x + \frac{1}{2}]|$ is the difference between x and the integer nearest to it.

Suppose that, after step (3) of an iteration, we have the lattice base $A(i_1) = \{A_f(i_1), A_g(i_1), A_h(i_1)\}$ where f,g,h are a permutation of 0,1,2 and f was defined in step (2) by (6.6). Let the orthogonal projections of $\underline{A}_g(i_1)$ and $\underline{A}_h(i_1)$ on the line m be $-x\underline{A}_f(i_1)$ and $-y\underline{A}_f(i_1)$ respectively. If $\|x\| \le \frac{1}{4}$ or $\|y\| \le \frac{1}{4}$ we do nothing extra.

Else, if $\|x+y\| \le \frac{1}{4}$ we replace $A_g(i_1)$ or $A_h(i_1)$ by the sum $A_g(i_1) + A_h(i_1)$ (depending on which has the smaller cofactor).

And if $\|x\|$, $\|y\|$ and $\|x+y\|$ all exceed $\frac{1}{4}$, then we first construct $A_g(i_1) + A_h(i_1)$ as above, then either $(A_g(i_1) + A_h(i_1)) + A_g(i_1)$ or $(A_g(i_1) + A_h(i_1)) + A_h(i_1)$, depending on the cofactors. By (6.16) we have both $\|x + 2y\| \le \frac{1}{4}$ and $\|2x + y\| \le \frac{1}{4}$.

After this we continue with step (4), which now has the result that for $j = g$ or h (or both), either $|C_j| \le \frac{1}{4}|\underline{A}_f(i_0)|$ or $|C_j| \ge \frac{3}{4}|\underline{A}_f(i_0)|$ is true.

In the latter case we replace $A_j(i_2)$ or $A_f(i_2)$ by their sum, as was done in step (5), and we obtain (6.14) with the factor $\sqrt{\frac{1}{16}+\delta^2}$ instead of $\sqrt{\frac{1}{4}+\delta^2}$.

We remark here that choosing δ very small will increase the number of steps which are needed to obtain (6.7) without substantially improving the multiplier $\sqrt{\frac{1}{16}+\delta^2}$. In practice one may choose, e.g., $\delta=\frac{1}{5}$ to gain in each iteration a factor $\sqrt{\frac{1}{16}+\frac{1}{25}} < \frac{1}{3}$.

MULTI-DIMENSIONAL REMARK. The properties (6.7) and (6.8) of $A(i_1)$ remain unchanged during the one or two extra steps that may be needed to obtain the multiplier $\sqrt{\frac{1}{16}+\delta^2}$ in the above discussion. For (6.8) this is trivial. For (6.7), however, it follows from the special property of the ordinary continued fraction algorithm SplA(1) that these extra steps only bring our points nearer to the line m. When $n>2$ the situation is therefore slightly more complicated. The same technique, based on (6.16), will now guarantee not the factor $\sqrt{\frac{1}{16}+\delta^2}$, but rather $\sqrt{\frac{1}{16}+9\delta^2}$. On the other hand, the "chance" (meant heuristically) that we already have $|C_j| \le \frac{1}{4}|\underline{A}_f(i_0)|$ or $|C_j| \ge \frac{3}{4}|\underline{A}_f(i_0)|$ for at least one $j \ne f$ after step (4), increases with the dimension. Even when $n=2$, the expected number of extra steps per iteration is only $\frac{5}{16}$, if we assume a uniform probability distribution of x and y.

CHAPTER 7. AN ALGORITHM FOR RELATIVE MINIMA

7A. Relative minima.

In chapters 4, 5 and 6 we studied best and good approximations with respect to a given metric (radius function). We now consider an approximation problem which is defined with respect to a system of coordinate axes. Therefore, let $\mathbb{R}^{n+1}$ be equipped with a system of coordinates to be denoted by $(x_0, x_1, \ldots\ldots, x_n)$. We fix an inner product such that these coordinates are orthonormal. Let Ω be a lattice in $\mathbb{R}^{n+1}$.

DEFINITION 7.1. A lattice point $R = (r_0, r_1, \ldots\ldots, r_n) \neq 0$ is a *relative minimum* of Ω with respect to the coordinate axes if there exists no lattice point $P = (p_0, p_1, \ldots\ldots, p_n) \neq 0$ such that

$$|p_k| \leq |r_k| \quad \text{for } k = 0, 1, \ldots, n$$

unless

$$|p_k| = |r_k| \quad \text{for all } k.$$

To avoid certain complications we assume throughout this chapter that Ω is *irreducible*, which means that, if $P = (p_0, p_1, \ldots\ldots, p_n) \in \Omega$ and $p_k = 0$ for some k, then $P = 0$. In other words, two different lattice points can never have equal k-th coordinates $(k = 0, 1, \ldots, n)$. In all applications the lattice will be irreducible.

Given a relative minimum $R = (r_0, r_1, \ldots, r_n)$ we define its *x_j-successor* $(R)_{x_j}$ to be the lattice point in the region

$$(7.1) \qquad |x_k| < |r_k| \quad \text{for } k \neq j$$

for which $|x_j|$ is as small as possible. If we require in addition $x_j > 0$, then $(R)_{x_j}$ is uniquely determined because Ω is irreducible. From Minkowski's theorem 4.3 we know that there is at least one point of Ω in (7.1) such that

$$(7.2) \qquad |x_j| \leq \frac{\Delta}{\prod_{k \neq j} |r_k|}$$

when Δ is the basic volume of Ω.

Continuing the process of constructing x_j-successors we obtain the *x_j-chain* $\langle R\rangle_{x_j} = R_0, R_1, R_2, \ldots$ of a relative minimum $R = R_0$. It is defined by

$$R_{i+1} = (R_i)_{x_j} \quad \text{for } i \geq 0.$$

Such chains of relative minima play a role in certain problems of algebraic number theory (see section 7D).

REMARK. Since a relative minimum can be the x_j-successor of several other relative minima, there is no uniquely determined "x_j-predecessor" (contrary to the case of best approximations).

The x_j-successor of a given relative minimum can in principle be found in the same way as the successor of a best approximation in section 5A. First one seeks all lattice points in the bounded region (7.1), (7.2), e.g. by a Gaussian elimination process. Then one determines the one with least positive value of x_j. This method was employed, e.g., by STEINER[1] and RUDMAN & STEINER[1]. For the case $n = 2$ VORONOI[1] gave a method which reduces the number of "candidates" for the x_j-successor to at most five. His method is explained in detail in DELONE & FADDEEV[1] and in WILLIAMS & ZARNKE[2].

In this chapter we study the case $n = 2$. We start from a lattice base $A(0)$ in $\mathbb{R}^3$, one of whose points is a relative minimum R, and construct from it a two-dimensional continued fraction expansion along the x_j-axis containing all points of the x_j-chain $\langle R\rangle_{x_j}$ ($j = 0, 1$ or 2) under a certain constraint on $A(0)$.

7B. Notations and easy cases.

In $\mathbb{R}^3$ we denote the coordinate system by (x,y,z). When A_k or $A_k(i)$ ($k = 0,1,2$, $i \geq 0$) is a lattice point, it will have as coordinates (x_k, y_k, z_k) or $(x_k(i), y_k(i), z_k(i))$. Let $A = \{A_0, A_1, A_2\}$ be a base of Ω, Δ the absolute value of its coordinate determinant. For ℓ_0 we take the vector $\ell_0 = (\Delta, 0, 0)$, so that ℓ is the x-axis and ℓ^* the yz-plane. As in the preceding chapters, underlining denotes projection on ℓ^*, thus $\underline{A} = (0, y, z)$ if $\underline{A} = (x, y, z)$. For the height function we take $h(x,y,z) = x$ and we write $h(A) > 0$ to express that $\min(x_0, x_1, x_2) > 0$. Finally, if $A = (x_A, y_A, z_A)$,

then $\Sigma(A)$ denotes the open rectangle

$$(7.3) \qquad |y| < |y_A|, \ |z| < |z_A|$$

in ℓ^* with $\underline{A}$ as a vertex. We now introduce two auxiliary notions, valid for bases A with $h(A) > 0$.

We say that A is *positive with respect to* a lattice point $A \neq 0$, if $P \in \{A\}^+$ whenever P is a lattice point with $\underline{P} \in \Sigma(A)$ and $h(P) > 0$ (in short, A is *A-positive*).

We say that A is *regular with respect to* one of its points, say A_k, if for every $j \neq k$ at least one of the inequalities $y_k y_j < 0$, $z_k z_j < 0$ is true (in short, A is *A_k-regular*). In other words, A is A_k-regular if for $j \neq k$ the points $\underline{A}_j$ and $\underline{A}_k$ do not lie in the same quadrant of the yz-plane. A_k-regularness does not exclude that the two points $\underline{A}_i, \underline{A}_j$ (where $\{i,j,k\} = \{0,1,2\}$) lie in the same quadrant. Notice that apart from 0 no lattice points with zero y- or z-coordinates exist, since Ω is irreducible. Therefore we need not bother about the axes when speaking of the quadrants of the yz-plane.

<u>LEMMA 7.1.</u> *Suppose A is positive with respect to a lattice point A. Then $\ell_0 \in \{A\}^+$.*

<u>PROOF.</u> Suppose, say, that the cofactor a_2 is negative, and choose $\frac{1}{2} > \varepsilon > 0$. By $a_2 < 0$, there is an integer M such that $a_2 < -1/M$. From Dirichlet's theorem we know that there exist integers q, p_0, p_1, p_2 such that

$$|qa_j - p_j| < \varepsilon \quad \text{for } j = 0,1,2$$

and

$$q > M.$$

Now consider the point $P = p_0A_0 + p_1A_1 + p_2A_2$. Note that $q > M$ and $a_2 < -1/M$ give $p_2 \leq -1$. But we have

$$d(P,\ell) \leq \varepsilon . \sum_{j=0}^{2} d(A_j,\ell),$$

so that $\underline{P} \in \Sigma(A)$ if ε is small enough. □

Our approach will now be as follows. Given a base A which contains a relative minimum R and which is positive and regular with respect to R, we shall construct a new base A' which contains the x-successor $(R)_x$ and which is $(R)_x$-positive and -regular. This construction will consist of purely additive steps, so that the transition from A to A' can be considered to form a part of an expansion as defined in section 2B. Our algorithm for the x-chain of R then is a mere repetition of this procedure, and the condition on the initial base is that it contain R and be R-positive and R-regular.

A difficulty arises because the points of the chain have to be found strictly in increasing order of $|x|$, which poses the problem that we must immediately decide whether a newly found point belongs to the chain. (This difficulty did not arise in the algorithm of section 5C, where it was quite possible that a best approximation was found before a lower one).

Throughout this section and the next we assume that $A = \{A_0, A_1, A_2\}$ is a lattice base with $h(A) > 0$, and the index $f \in \{0,1,2\}$ is such that A_f is the relative minimum whose x-successor we are seeking. (We say A_f is the minimum associated with A). The remaining indices g,h will be specified in a moment. We also assume A to be A_f-positive and A_f-regular, so that $\ell_0 \in \{A\}^+$ by lemma 7.1. To facilitate notation we apply a scalar change of coordinates in the yz-plane by defining

$$(7.4) \qquad \eta = -\frac{y}{y_f}, \quad \zeta = -\frac{z}{z_f}.$$

This gives $\underline{A}_f$ the convenient $\eta\zeta$-coordinates $(-1, -1)$ [1]) and makes $\Sigma(A_f)$ the square

$$|\eta| < 1, \quad |\zeta| < 1.$$

Furthermore we write in $\eta\zeta$-coordinates

$$(7.5) \qquad \begin{cases} \underline{A}_g = (\alpha, \beta) \\ \underline{A}_h = (\gamma, \delta), \end{cases}$$

---

1) A vector with two entries should be read as $\eta\zeta$-coordinates, one with three entries as xyz-coordinates.

where we have specified g and h such that

(7.6) $\alpha \geq \beta, \quad \gamma \leq \delta.$

The cofactors of $\mathcal{A}$ have the ratio

(7.7) $a_f : a_g : a_h = (\alpha\delta - \beta\gamma) : (\delta - \gamma) : (\alpha - \beta).$

Because $\mathcal{A}$ is A_f-regular we have

(7.8) $\alpha > 0, \quad \delta > 0.$

The following lemma is very useful to prove that a new base is still A_f-positive:

<u>LEMMA 7.2.</u> *Let $A \neq 0$ be a lattice point and $\mathcal{A}$ an A-positive lattice base. Let the permutation s,t,u of $0,1,2$ be such that $a_s \geq a_t$. Assume there is a linear form L on ℓ^* such that*

$$L(\underline{A}_t + \underline{A}_s) \geq 0, \quad L(\underline{A}_u) \geq 0$$

and

$$L(\underline{A}_t) > L(X) \quad \text{for all } X \in \Sigma(A).$$

Then $\underline{P} \in \Sigma(A)$ implies $P \in \{\mathcal{A}I_{st}\}^+$ whenever P is a lattice point with $h(P) > 0$; i.e., $\mathcal{A}I_{st}$ is A-positive.

<u>PROOF.</u> Let $P \in \Omega$ be such that $h(P) > 0$, $\underline{P} \in \Sigma(A)$. Since $\mathcal{A}$ is A-positive we may write $P = pA_s + qA_t + rA_u$ with $p,q,r \geq 0$. From the conditions on L we have

$$L(\underline{A}_t) > L(\underline{P}) = L(p(\underline{A}_s + \underline{A}_t) + (q-p)\underline{A}_t + r\underline{A}_u) \geq (q-p)L(\underline{A}_t),$$

whence $q - p \leq 0$ (note that $L(\underline{A}_t) > L(0) = 0$). Thus

$$P = (p-q)A_s + q(A_s + A_t) + rA_u \in \{\mathcal{A}I_{st}\}^+. \ \square$$

<u>NOTE:</u> On the closure of $\Sigma(A)$, L assumes its maximal value in one of the vertices, so that verification of the conditions is always easy.

Now let us see how to recognize the successor of A_f.

LEMMA 7.3. *Let* A *be a lattice base,* $h(\mathsf{A}) > 0$, f,g,h *a permutation of* $0,1,2$. *Suppose* A *is* A_f*-positive and* A_f*-regular,* A_f *being a relative minimum of* Ω.

(a) *If* $\underline{A}_g, \underline{A}_h \in \Sigma(A_f)$ *and* $x_h < x_g$, *then* $(A_f)_x = A_h$ *and* A *is positive and regular with respect to* $(A_f)_x$.

(b) *If* $\underline{A}_h \in \Sigma(A_f)$, $\underline{A}_g \notin \Sigma(A_f)$ *and* $\beta < 0$, *then* $(A_f)_x = A_h$ *and* A *is positive and regular with respect to* $(A_f)_x$.

(c) *If* $\underline{A}_h \in \Sigma(A_f)$, $\underline{A}_g \notin \Sigma(A_f)$, $\beta > 0$ *and* $a_f < a_g$, *then* $(A_f)_x$ *is either* A_h *or* $A_f + A_g$, *and in either case* $\mathsf{A}I_{gf}$ *is* $(A_f)_x$*-positive and -regular.*

PROOF. (a) By definition $(A_f)_x$ is the lowest non-zero point of Ω of which the projection lies inside $\Sigma(A_f)$. By A_f-positivity of A, only points of the form $p\underline{A}_f + q\underline{A}_g + r\underline{A}_h$ with $p,q,r \geq 0$ lie in $\Sigma(A_f)$, and among these A_h clearly has the least x-coordinate (because $x_h < x_g$). Therefore $A_h = (A_f)_x$. Furthermore, A is also A_h-positive because $\Sigma(A_h) \subset \Sigma(A_f)$. If $\underline{A}_g, \underline{A}_h$ were in the same quadrant of the $\eta\zeta$-plane we would have $\underline{A}_g - \underline{A}_h \in \Sigma(A_f)$ because of

$$\begin{cases} |\alpha - \gamma| \leq \max(|\alpha|, |\gamma|) < 1 \\ |\beta - \delta| \leq \max(|\beta|, |\delta|) < 1. \end{cases}$$

But this contradicts the fact that A is A_f-positive. Thus A is A_h-regular.

(b) Since $\underline{A}_h \in \Sigma(A_f)$ and A is A_f-positive we know that $(A_f)_x$ is either A_h or $pA_f + qA_g$ for some $p,q \geq 1$. But if $p \geq 1$, then the ζ-coordinate of $p\underline{A}_f + q\underline{A}_g$ satisfies

$$\zeta = -p + q\beta \leq -1$$

by $\beta < 0$, and thus $p\underline{A}_f + q\underline{A}_g \notin \Sigma(A_f)$ for all $p,q \geq 1$. We conclude that $(A_f)_x = A_h$. Since $\Sigma(A_h) \subset \Sigma(A_f)$, A is also A_h-positive. From $\underline{A}_f \in$ III [1]), $\underline{A}_g \in$ IV (because of (7.8) and $\beta < 0$) and $\underline{A}_h \in$ II $\cup$ I (from (7.8)) we see that A is also A_h-regular.

(c) (See figure 7.1, p. 136). We use the following lemma:

---

[1]) The roman figures denote the quadrants of the $\eta\zeta$-plane:

II	I
III	IV

LEMMA 7.4. *If* $(A_f)_x$ *is of the form* $pA_f + qA_g$, *then* $q = 1$.

PROOF. The fact that $p\underline{A}_f + q\underline{A}_g \in \Sigma(A_f)$ implies $|-p + q\alpha| < 1$, whence

$$\left| -\frac{p}{q} + \alpha \right| < \frac{1}{q}.$$

But this gives $\left| -\left[\frac{p}{q}\right] + \alpha \right| < \frac{1}{q} + \frac{q-1}{q} = 1$. Using the similar result for the ζ-coordinates we see that $\underline{A}_g + \left[\frac{p}{q}\right]\underline{A}_f \in \Sigma(A_f)$, and this is a point of smaller height than $pA_f + qA_g$ if $q \geq 2$. □

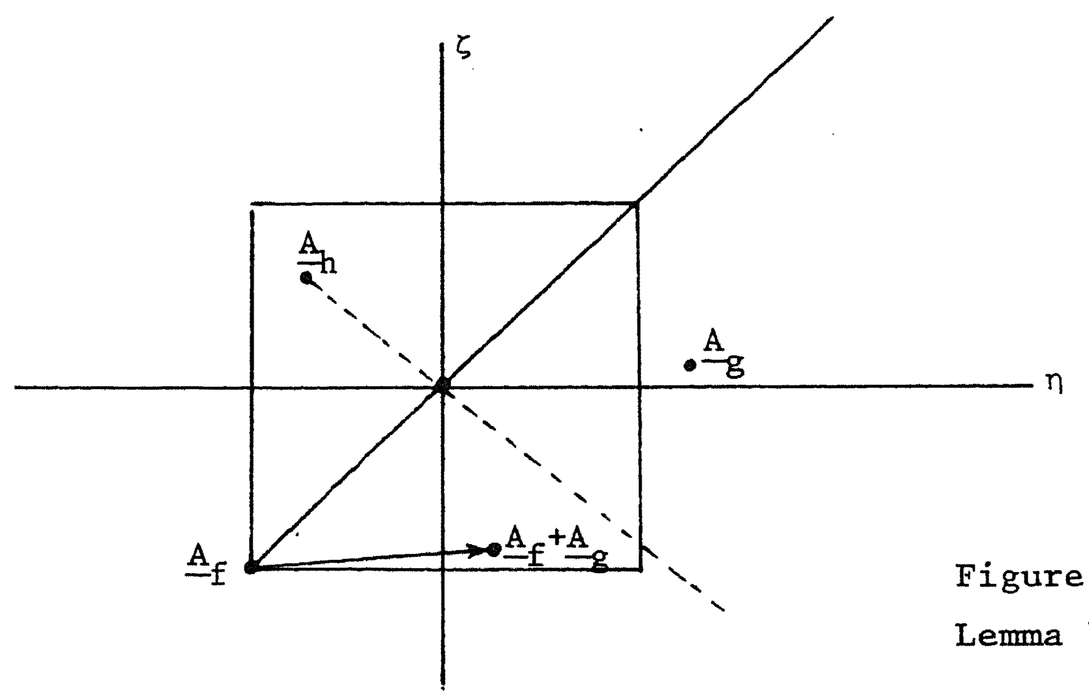


Figure 7.1: Lemma 7.3(c).

Proof of lemma 7.3(c): Since $\alpha \geq \beta$, $\beta > 1$ would give $\alpha\delta - \beta\gamma \geq \beta(\delta - \gamma) \geq \delta - \gamma$, contradicting $a_f < a_g$ (cf. (7.7)). Hence $0 < \beta < 1$.
Now by the fact that A is A_f-positive and that $\underline{A}_h \in \Sigma(A_f)$, we have either $(A_f)_x = A_h$ or $(A_f)_x = pA_f + qA_g$ for some $p, q \geq 1$. By lemma 7.4, we must have $q = 1$, and this gives $p = 1$, because for $p \geq 2$ the point $p\underline{A}_f + \underline{A}_g$ satisfies

$$\zeta = -p + \beta < -1,$$

hence lies outside $\Sigma(A_f)$. So $(A_f)_x = A_h$ or $(A_f)_x = A_f + A_g$ and in either case $AI_{gf} = \{A_f + A_g, A_g, A_h\}$ contains $(A_f)_x$.

Since $\underline{A}_g \notin \Sigma(A_f)$ and $0 < \beta < 1$ we have

$$\alpha = \max\,(|\alpha|, |\beta|) > 1$$

and this implies $\gamma < 0$ in view of $\alpha\delta - \beta\gamma < \delta - \gamma$. Therefore $\underline{A}_f + \underline{A}_g$, $\underline{A}_g$ and $\underline{A}_h$ lie in three different quadrants so that AI_{gf} is $(A_f)_x$-regular.

From lemma 7.2 with $L = \gamma\zeta - \delta\eta$ we see that AI_{gf} is also $(A_f)_x$-positive (and even A_f-positive): Indeed, we have

$$L(\underline{A}_f + \underline{A}_g) = (\delta - \gamma) - (\alpha\delta - \beta\gamma) > 0,$$
$$L(\underline{A}_h) = \gamma\delta - \delta\gamma = 0,$$
$$L(\underline{A}_f) = \delta - \gamma,$$

and for $X \in \Sigma(A_f)$, $L(X) < \delta - \gamma$ since $\gamma < 0 < \delta$. □

LEMMA 7.5. *Let A be a lattice base, $h(A) > 0$; let f,g,h be a permutation of $0,1,2$ and suppose A_f is a relative minimum with respect to which A is positive and regular. If in the notation of (7.4)-(7.6), $\delta - \gamma \geq \max(2, \alpha - \beta)$, then AI_{gh} is A_f-positive and A_f-regular.*

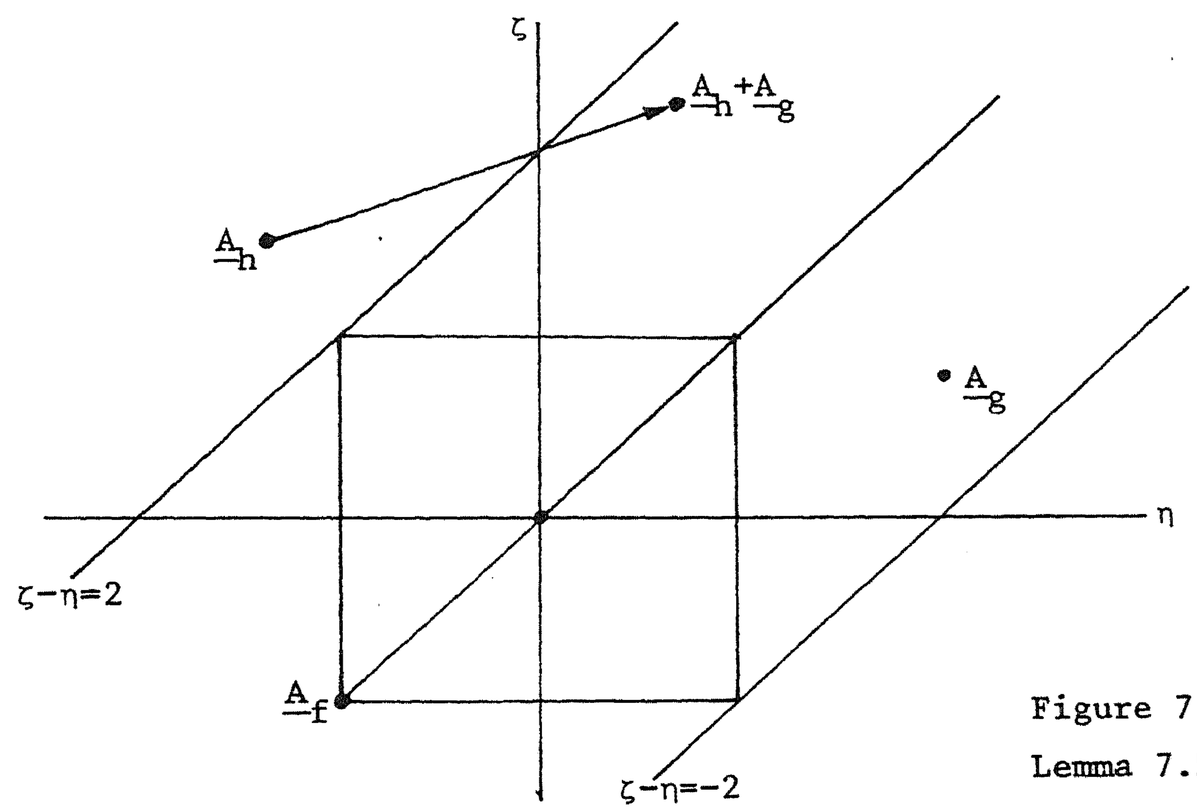


Figure 7.2: Lemma 7.5.

PROOF. Suppose $\beta + \delta < 0$. Then $\delta - \gamma \geq \alpha - \beta$ gives $\alpha + \gamma < 0$, hence $\alpha\delta - \beta\gamma < 0$. But this means $a_f < 0$, a contradiction. Therefore $\beta + \delta > 0$, so that AI_{gh} is A_f-regular. It follows from lemma 7.2 with $L = \zeta - \eta$ that AI_{gh} is also A_f-positive. In fact, $L(\underline{A}_f) = 0$, $L(\underline{A}_h + \underline{A}_g) = (\delta - \gamma) - (\alpha - \beta) \geq 0$, $L(\underline{A}_h) = \delta - \gamma \geq 2$ and $L(X) < 2$ for all $X \in \Sigma(A_f)$. □

LEMMA 7.6. *Let A be a lattice base, $h(A) > 0$; let f,g,h be a permutation of $0,1,2$, A_f a relative minimum. Suppose A is A_f-positive and A_f-regular. If*

$\alpha > 1$ and $a_f \geq a_g$, then AI_{fg} is A_f-positive and A_f-regular.

PROOF. Since $\alpha > 1$ we have $\alpha - 1 > 0$ so that $\underline{A}_f + \underline{A}_g \notin$ III. Thus $AI_{fg} = \{A_f, A_g + A_f, A_h\}$ is A_f-regular. Its A_f-positivity follows from lemma 7.2 with $L = \eta$ (if $\gamma > 0$) or $L = \delta\eta - \gamma\zeta$ (if $\gamma < 0$). □

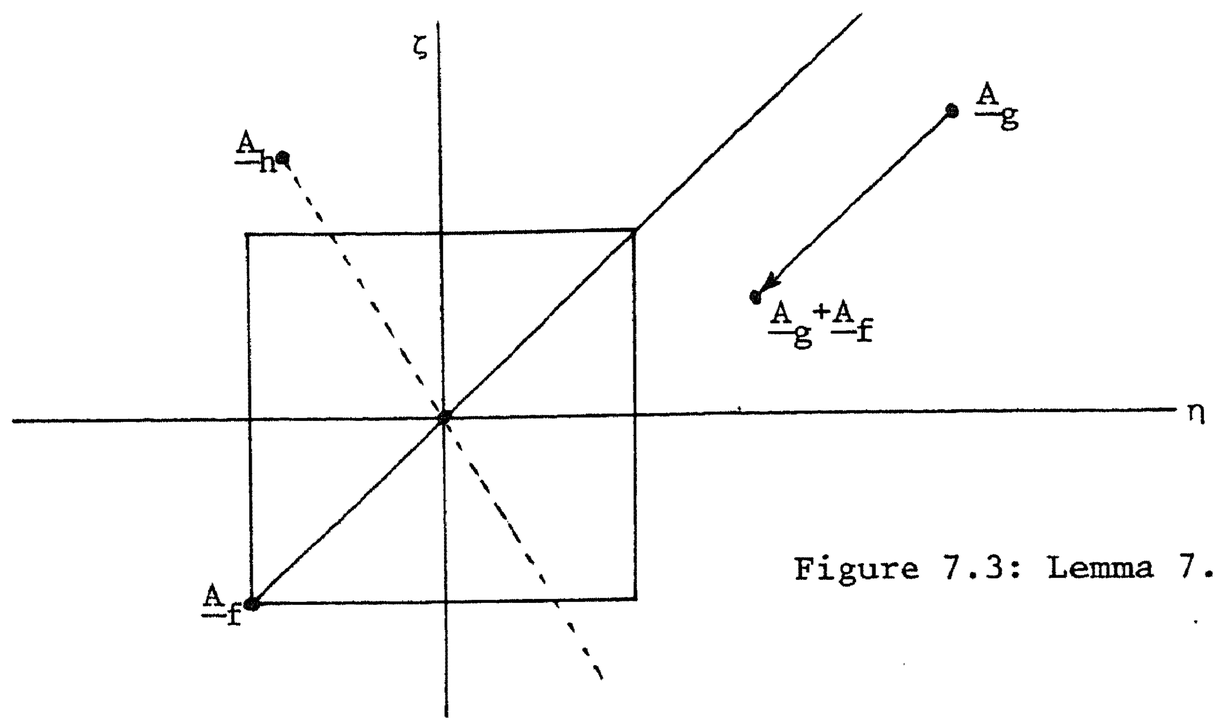


Figure 7.3: Lemma 7.6.

REMARK. The lemmata 7.3, 7.5 and 7.6 are symmetric in the coordinates η and ζ. They remain valid if the indices g and h and their corresponding cofactors and coordinates are interchanged.

7C. A continued fraction algorithm for relative minima.

THEOREM 7.7. *Let A be a lattice base, $h(A) > 0$, and assume that $A_f \in A$ is a relative minimum for some $f \in \{0,1,2\}$. Suppose A is A_f-positive and A_f-regular. Then*

(i) A contains $(A_f)_x$ and is positive and regular with respect to it,

or (ii) there exist indices s,t according to lemma 7.5, 7.6 or 7.3(c) such that AI_{st} is positive and regular with respect to A_f or $(A_f)_x$ (and in the latter case contains $(A_f)_x$),

or (iii) there exists a (sequence of) step(s) according to table 7.1 (p. 141) or its symmetric counterpart such that at the end the new base contains $(A_f)_x$ and is positive and regular with respect to it.

PROOF. Part 1. If one of the lemmata 7.3, 7.5 or 7.6 applies, then (i) or (ii) holds. So suppose that none of these lemmata or their symmetric

counterparts is applicable to A. That means, in the notations of the preceding section,

(7.9) $\underline{A}_g \notin \Sigma(A_f)$ and $\underline{A}_h \notin \Sigma(A_f)$

and

(7.10) $\max(\alpha - \beta, \delta - \gamma) < 2.$

Furthermore, if both $\alpha < 1$ and $\delta < 1$ then $\beta < -1$ and $\gamma < -1$ by (7.6) and (7.9) so that the cofactor a_f $(\sim\alpha\delta - \beta\gamma)$ is negative. But $a_f \geq 0$ and hence at least one of $\alpha > 1$, $\delta > 1$ is true. We assume

(7.11) $\alpha > 1.$

(The case $\delta > 1$ leads to the symmetric counterpart of table 7.1). Because lemma 7.6 is not applicable, (7.11) implies

(7.12) $a_f < a_g.$

By the same argument we have

(7.13) if $\delta > 1$ then $a_f < a_h$.

Now $\beta > 1$ would in view of $\alpha \geq \beta$ give $\alpha\delta - \beta\gamma \geq \beta(\delta - \gamma) \geq \delta - \gamma$, contradicting (7.12). Hence

(7.14) $\beta < 1,$

and from $\alpha > 1 > \beta$, $\delta \geq \gamma$, $\alpha\delta - \beta\gamma < \delta - \gamma$ we infer

(7.15) $\gamma < 0.$

We have now restricted the positions of $\underline{A}_h$ and $\underline{A}_g$ to the areas shown in figure 7.4 (p. 140).

Part 2. We now show, under the restrictions (7.9)-(7.15),

(7.16) if $\underline{A}_f + \underline{A}_g \in \Sigma(A_f)$ then $(A_f)_x = A_f + A_g$ or $(A_f)_x = A_g + A_h$.

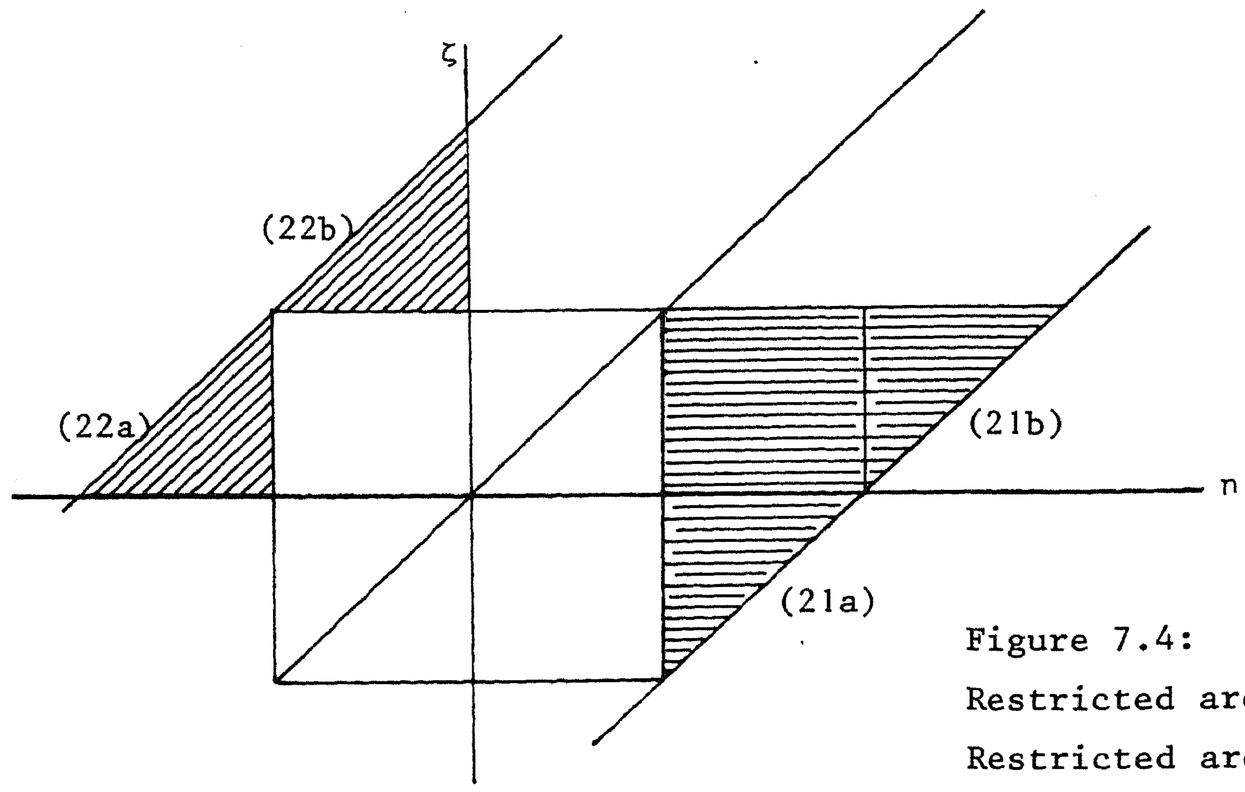


Figure 7.4:
Restricted area for $\underline{A}_h$:
Restricted area for $\underline{A}_g$:

Proof: By (7.9), (7.14), $\underline{A}_f + \underline{A}_g \in \Sigma(A_f)$ implies

(7.17) $\quad 0 < \beta < 1 < \alpha < 2.$

Now if $(A_f)_x \neq A_f + A_g$, then we must have $(A_f)_x = pA_f + rA_h$ $(p,r \geq 1)$ or $(A_f)_x = qA_g + rA_h$ $(q,r \geq 1)$ because A is A_f-positive and $h(A) > 0$.

However, $p\underline{A}_f + r\underline{A}_h$ satisfies $\eta = -p + r\gamma < -p \leq -1$ because of (7.15), hence lies outside $\Sigma(A_f)$. If $\beta + \delta > 1$ then $q\underline{A}_g + r\underline{A}_h$ satisfies $\zeta = q\beta + r\delta \geq \beta + \delta > 1$ because of $\beta > 0$ (see (7.17)), hence lies outside $\Sigma(A_f)$. But if $\beta + \delta < 1$ then we must have $0 < \beta < 1$ and $0 < \delta < 1$. The latter gives $-2 < \gamma < -1$ by (7.9) and (7.10), and thus $-1 < \alpha + \gamma < 1$ by (7.17), so that $\underline{A}_g + \underline{A}_h \in \Sigma(A_f)$.

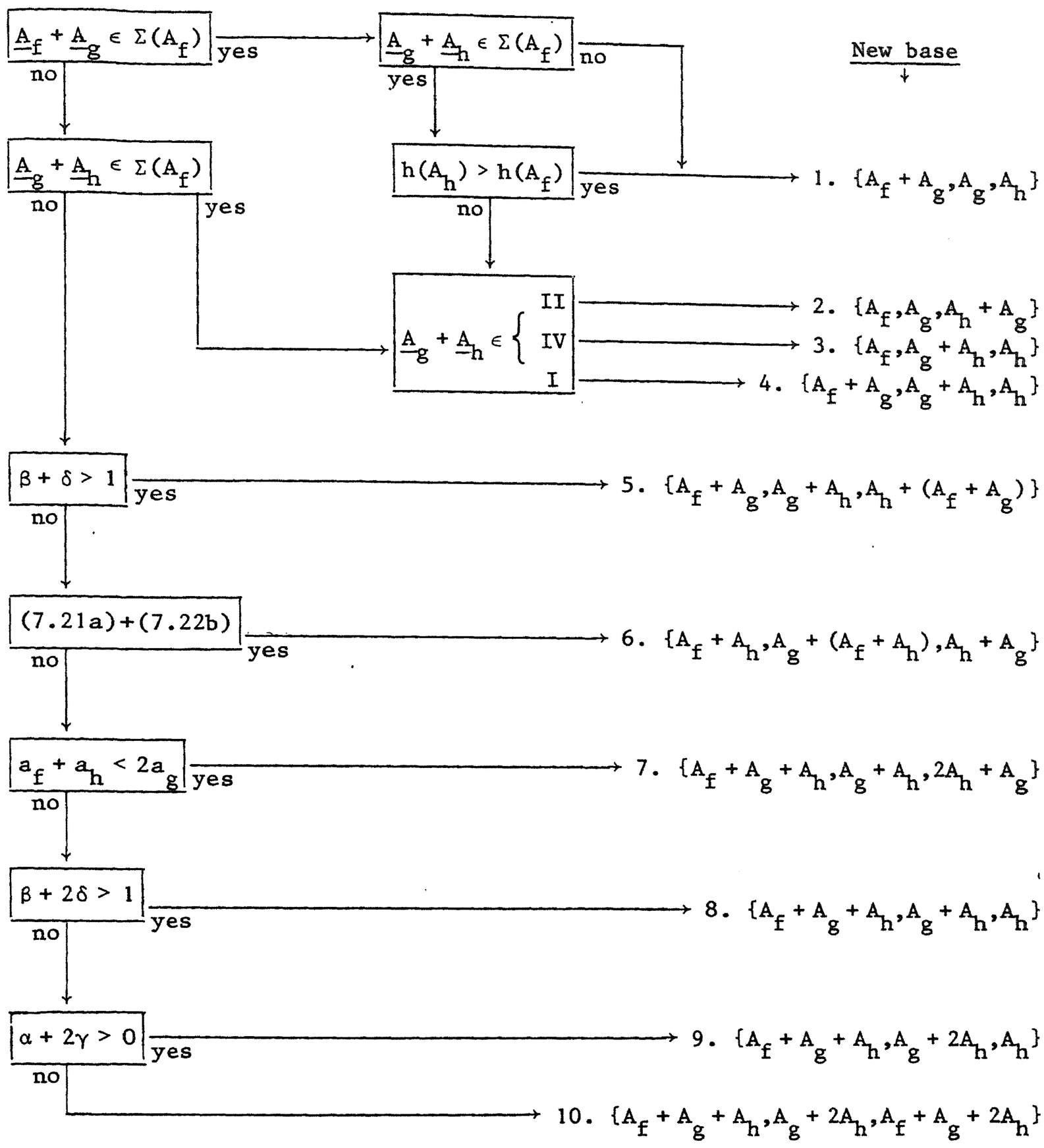
<u>Part 3.</u> Let us now assume

(7.18) $\quad \underline{A}_f + \underline{A}_g \notin \Sigma(A_f).$

We show: If $\underline{A}_g + \underline{A}_h \in \Sigma(A_f)$, then $(A_f)_x = A_g + A_h$.

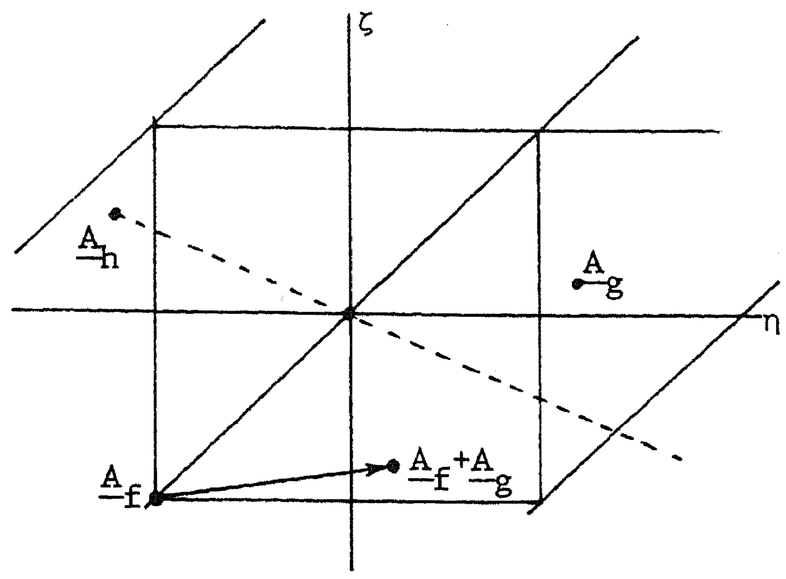
Proof: If not, then $(A_f)_x$ were $pA_f + qA_k$ for some $p,q \geq 1$ and $k = g$ or h. By lemma 7.4, $q = 1$. But if $k = h$, then $p\underline{A}_f + \underline{A}_h \in \Sigma(A_f)$ gives $\eta = -p + \gamma > -1$, whence $p < 1$ by (7.15). And if $k = g$, then $p\underline{A}_f + \underline{A}_g \in \Sigma(A_f)$ gives $\zeta = -p + \beta > -1$, whence $p < 2$ by (7.14). But $p = 1$ is impossible in view of the assumption (7.18).

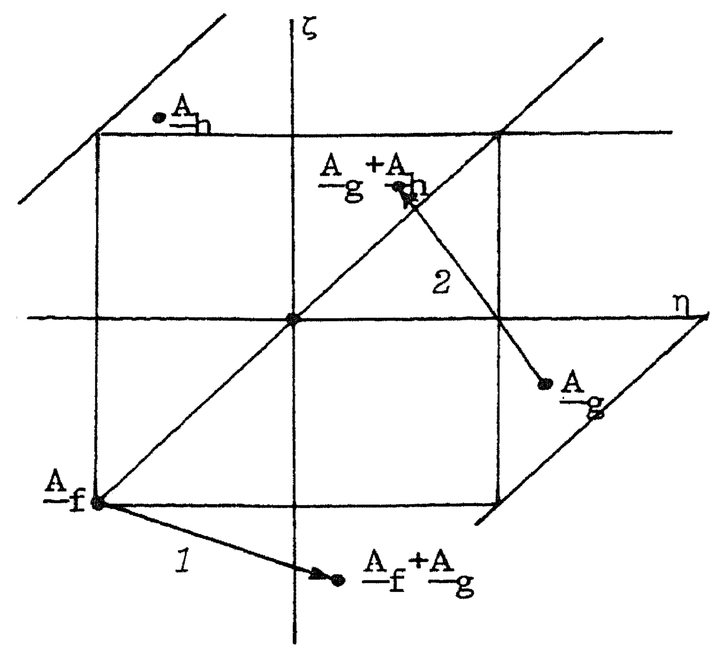
TABLE 7.1. (Case $\alpha > 1$).



Part 4. We can now verify statement (iii) of the theorem for the first four "new bases" of table 7.1.

New base 1: In this case, $(A_f)_x = A_f + A_g$ by (7.16). Apply lemma 7.2 with $L = \gamma\zeta - \delta\eta$ (like in the proof of lemma 7.3(c); see figure 7.5, p. 142). This gives us A_f-positivity, hence $(A_f)_x$-positivity





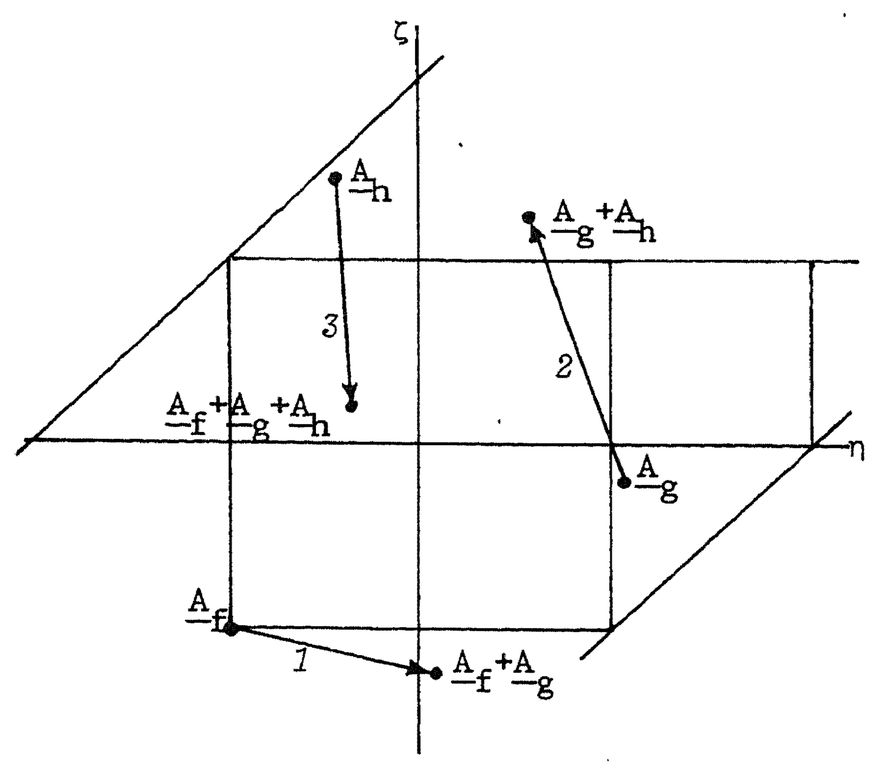
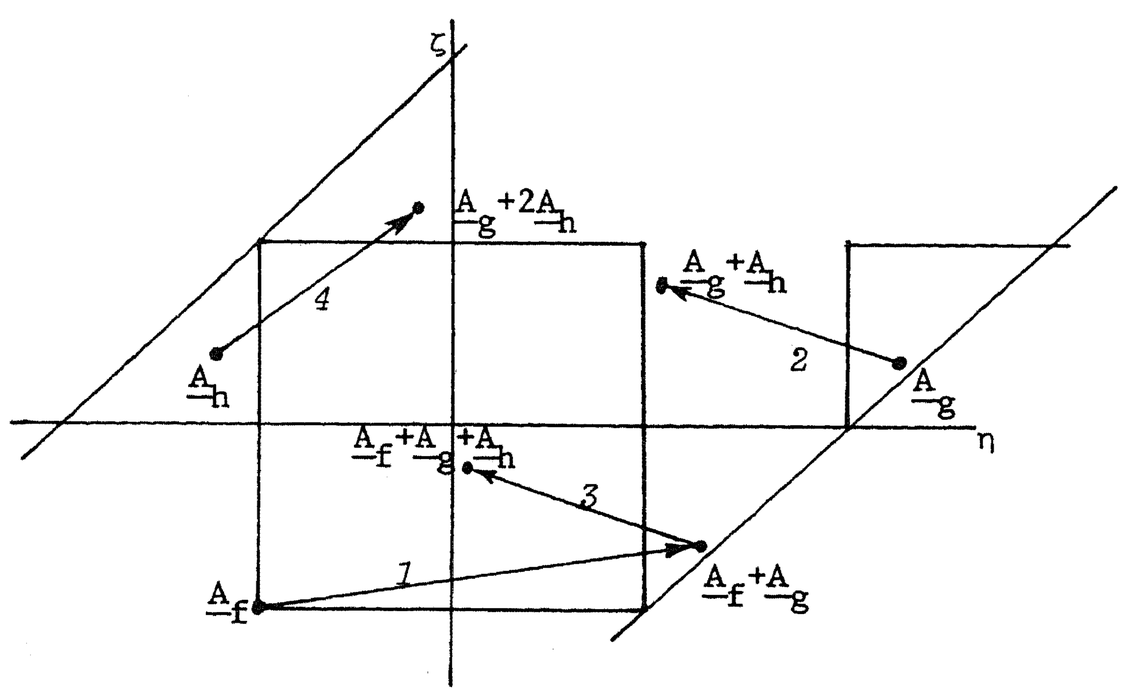


Figure 7.5 (top left):
Table 7.1, line 1.
Figure 7.6 (top right):
Table 7.1, line 4.
Figure 7.7 (left):
Table 7.1, line 5.
Figure 7.8 (bottom):
Table 7.1, line 7.



of AI_{gf}. We end with $\underline{A}_f + \underline{A}_g \in$ IV (by $\beta < 1$), $\underline{A}_g \in$ I, $\underline{A}_h \in$ II, and thus with a regular base.

New base 2: Here $(A_f)_x = A_g + A_h$. We apply lemma 7.2 with $s = g$, $t = h$ and $L = -\eta$ (note that $\alpha > 1$ and $\alpha + \gamma < 0$ imply $\gamma < -1$). We end with $\underline{A}_g + \underline{A}_h \in$ II, $\underline{A}_g \in$ I or IV and $\underline{A}_f \in$ III.

New base 3: Again $(A_f)_x = A_g + A_h$. From $\underline{A}_g + \underline{A}_h \in$ IV it follows that $\beta + \delta < 0$, hence $\beta < 0$. Suppose P is a lattice point in $\{A\}^+$ but not in $\{AI_{hg}\}^+$. Then $P = pA_g + q(A_g + A_h) + rA_f$ for some $p \geq 1$, $q, r \geq 0$. If $r = 0$ then $\underline{P}$ satisfies

$$\eta = p\alpha + q(\alpha + \gamma) \geq \alpha > 1$$

by (7.11). If $r \geq 1$ then $\underline{P}$ satisfies

$$\zeta = p\beta + q(\beta + \delta) - r < -1$$

because $\beta < \beta + \delta < 0$. In both cases $\underline{P} \notin \Sigma(A_f)$ and therefore AI_{hg} is A_f-positive. Its regularness follows from $\underline{A}_f \in$ III, $\underline{A}_h \in$ II and $\underline{A}_g + \underline{A}_h \in$ IV.

New base 4: (See figure 7.6). Again $(A_f)_x = A_g + A_h$. The base is obtained by two steps. The first, with $s = g$ and $t = f$, preserves A_f-positivity by lemma 7.2 with $L = \gamma\zeta - \delta\eta$; for the second step, with $s = h$ and $t = g$, we use $L = \eta$. At the end we have $\underline{A}_f + \underline{A}_g \in$ IV (because $\beta < 1 < \alpha$), $\underline{A}_h \in$ II and $\underline{A}_g + \underline{A}_h \in$ I, so that our new base is regular.

<u>Part 5.</u> From now on we may assume in addition to (7.9)-(7.15) also (7.18) and

$$\underline{A}_g + \underline{A}_h \notin \Sigma(A_f). \tag{7.19}$$

We show

$$\underline{A}_f + \underline{A}_g + \underline{A}_h \in \Sigma(A_f). \tag{7.20}$$

In fact, from (7.9)-(7.15) and (7.18) we have for $\underline{A}_g$ the following possibilities:

(7.21a) $\begin{cases} 1 < \alpha < 2 \\ -1 < \beta < 0 \end{cases}$, (7.21b) $\begin{cases} 2 < \alpha < 3 \\ 0 < \beta < 1. \end{cases}$

For $\underline{A}_h$ we have the possibilities

(7.22a) $\begin{cases} -2 < \gamma < -1 \\ 0 < \delta < 1 \end{cases}$, (7.22b) $\begin{cases} -1 < \gamma < 0 \\ 1 < \delta < 2. \end{cases}$

The combination of (7.21a) and (7.22a) gives $-1 < \alpha + \gamma < 1$ and $-1 < \beta + \delta < 1$, contradicting (7.19). And (7.21b) with (7.22b) gives $\alpha\delta - \beta\gamma > 2 > \delta - \gamma$ in view of (7.10), which contradicts (7.12). In the other two cases, (7.21a)+(7.22b) and (7.21b)+(7.22a), we get $-1 < \alpha + \gamma - 1 < 1$ and $-1 < \beta + \delta - 1 < 1$, which imply (7.20).

Part 6. Under the assumptions of part 5 we show that $(A_f)_x \neq A_f + A_g + A_h$ implies $(A_f)_x = A_g + 2A_h$.

In fact, if $(A_f)_x \neq A_f + A_g + A_h$, then $(A_f)_x = pA_f + qA_k$ for $k = g$ or h, or $(A_f)_x = pA_g + qA_h$ for some $p,q \geq 1$. By lemma 7.4, $(A_f)_x = pA_f + qA_k$ requires $q = 1$. However, (7.15) gives $p\underline{A}_f + \underline{A}_h \notin \Sigma(A_f)$ for $p \geq 1$, and (7.14) (for $p \geq 2$), (7.18) (for $p = 1$) give $p\underline{A}_f + \underline{A}_g \notin \Sigma(A_f)$.

There remains the possibility $(A_f)_x = pA_g + qA_h$, $p,q \geq 1$ and $p \neq q$ (the latter from (7.19)). In the case (7.21a)+(7.22b), $p > q$ implies that $p\underline{A}_g + q\underline{A}_h \notin \Sigma(A_f)$, since it satisfies

$$\eta = p\alpha + q\gamma = \alpha(p - q) + q(\alpha + \gamma) > p - q \geq 1,$$

whereas $p < q$ implies the same because

$$\zeta = p\beta + q\delta = p(\beta + \delta) + \delta(q - p) > q - p \geq 1.$$

In the case (7.21b)+(7.22a), $\beta + \delta > 1$ gives $p\underline{A}_g + q\underline{A}_h \notin \Sigma(A_f)$ from $\zeta = p\beta + q\delta \geq \beta + \delta > 1$, and $p > q$ gives the same because then $p\underline{A}_g + q\underline{A}_h$ satisfies

$$\eta = p\alpha + q\gamma = \alpha(p - q) + q(\alpha + \gamma) > 2(p - q) \geq 2.$$

But if $\beta + \delta < 1$ we have $1 < \alpha + \gamma < 2$ by (7.19), hence $-1 < \alpha + 2\gamma < 1$. Therefore we have $\underline{A}_g + 2\underline{A}_h \in \Sigma(A_f)$ if also $\beta + 2\delta < 1$ (but $p\underline{A}_g + q\underline{A}_h \notin \Sigma(A_f)$ if $\beta + 2\delta > 1$

and $q > p$ since $\zeta = p\beta + q\delta \geq \beta + 2\delta > 1$).

Part 7. We can now verify statement (iii) of the theorem for the last six "new bases" of table 7.1. According to part 6 we have $(A_f)_x = A_f + A_g + A_h$ in the cases leading to base 5, 6 or 8. In the other cases the successor may be either $A_f + A_g + A_h$ or $A_g + 2A_h$, and the new bases contain them both.

New base 5: (See figure 7.7). First we prove that $\beta + \delta > 1$ implies

(7.23) $\alpha + \gamma < 1$.

Suppose that $\alpha + \gamma > 1$. Then we get $(\alpha - 1)\delta - (\beta - 1)\gamma > -\delta\gamma + \delta\gamma = 0$, hence $\alpha\delta - \beta\gamma > \delta - \gamma$, which contradicts (7.12).
The new base results from three steps. For the first step, with $s = g$, $t = f$, we use lemma 7.2 with $L = \gamma\zeta - \delta\eta$; for the second step, with $s = h$ and $t = g$, we use $L = \eta$ (note that $\alpha + \gamma > 0$). The third step is with $s = f$ and $t = h$. In case $\delta > 1$ we use lemma 7.2 with $L = \zeta$ (recall that $\beta + \delta > 1$). In case $0 < \delta < 1$, let P be a lattice point in $\{AI_{gf}I_{hg}\}^+$ but not in $\{AI_{gf}I_{hg}I_{fh}\}^+$. Then we can write

$$P = pA_h + q(A_f + A_g + A_h) + r(A_g + A_h), \quad p \geq 1, \; q,r \geq 0.$$

Now if $r = 0$, then $\underline{P}$ satisfies $\eta = p\gamma + q(\alpha + \gamma - 1) \leq \gamma < -1$ by $\gamma < -1$ and (7.23). And if $r > 0$, then $\underline{P}$ satisfies $\zeta = p\beta + q(\beta + \delta - 1) + r(\beta + \delta) \geq \beta + 2\delta > 1$ by $\beta + \delta > 1$. Therefore $\underline{P} \notin \Sigma(A_f)$ and $AI_{gf}I_{hg}I_{fh}$ is positive with respect to A_f, hence to $A_f + A_g + A_h$ by (7.20).
Regularness follows since $\underline{A}_f + \underline{A}_g \in$ IV, $\underline{A}_g + \underline{A}_h \in$ I and $\underline{A}_f + \underline{A}_g + \underline{A}_h \in$ II (the latter because $\alpha + \gamma < 1 < \beta + \delta$).

New base 6: This case is symmetric to the preceding by interchange of η and ζ: Note that $\beta + \delta < 1$ implies $\alpha + \gamma > 1$ by (7.19).

New base 7: Apply lemma 7.2 with, respectively,
- $s = g$, $t = f$, $L = \gamma\zeta - \delta\eta$,
- $s = h$, $t = g$, $L = \eta$,
- $s = h$, $t = f$, $L = \eta$,
- $s = g$, $t = h$, $L = -(\alpha + \gamma - 1)\zeta + (\beta + \delta - 1)\eta$.

(See figure 7.8). In verifying the conditions for L, note that

$\alpha > 2$ and $\alpha + \gamma > 1$, and recall that $a_h + a_f < 2a_g$, which yields $(\alpha - \beta) + (\alpha\delta - \beta\gamma) < 2(\delta - \gamma)$.

At the end we have $\underline{A}_f + \underline{A}_g + \underline{A}_h \in$ IV (because $\beta + \delta < 1 < \alpha + \gamma$), $\underline{A}_g + \underline{A}_h \in$ I and $\underline{A}_g + 2\underline{A}_h \in$ II, the latter because $a_h + a_f < 2a_g$ implies $\alpha + 2\gamma < \beta(1 + \gamma) + \delta(2 - \alpha) < 0$.

New base 8: Apply lemma 7.2 with, successively,

- $s = g$, $t = f$, $L = \gamma\zeta - \delta\eta$,
- $s = h$, $t = g$, $L = \eta$,
- $s = h$, $t = f$, $L = \eta$.

Then we have obtained $\underline{A}_f + \underline{A}_g + \underline{A}_h \in$ IV, $\underline{A}_g + \underline{A}_h \in$ I and $\underline{A}_h \in$ II.

New base 9: We start with the same three steps as those of new base nr. 8. For the fourth step we take $s = h$, $t = g$, and we use lemma 7.2 with $L = \eta$. Then we still have $\underline{A}_f + \underline{A}_g + \underline{A}_h \in$ IV, $\underline{A}_h \in$ II, moreover $\underline{A}_g + 2\underline{A}_h \in$ I. For verification, recall that $\alpha + 2\gamma > 0$.

New base 10: We start with the same three steps as those of new base nr. 8. For the fourth step we take $s = h$, $t = g$, and we use lemma 7.2 with $L = (\alpha + \gamma - 1)\zeta - (\beta + \delta - 1)\eta$. For the fifth step we take $s = f$, $t = h$, and we use lemma 7.2 with $L = -\eta$. Note that $\alpha + 2\gamma < 0$. After these five steps (see figure 7.9) we have obtained a lattice base consisting of $\underline{A}_f + \underline{A}_g + \underline{A}_h \in$ IV, $\underline{A}_g + 2\underline{A}_h \in$ II and $\underline{A}_f + \underline{A}_g + 2\underline{A}_h \in$ III (because $\beta + 2\delta < 1$).

This completes the proof of theorem 7.7. □

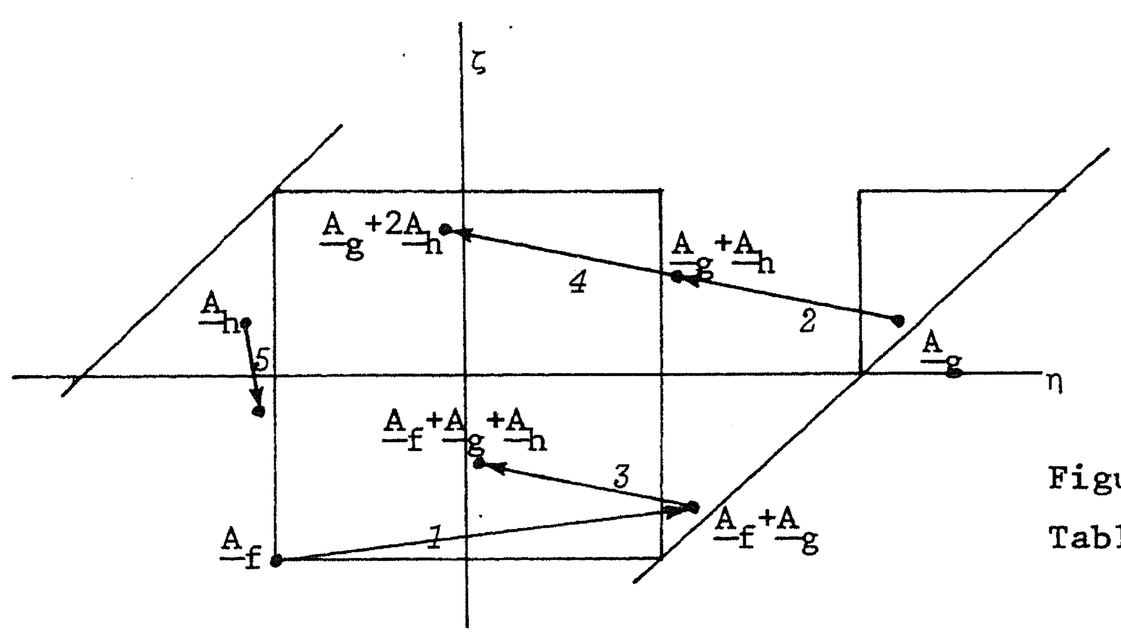


Figure 7.9:
Table 7.1, line 10.

From theorem 7.7 it is obvious how we can construct an expansion which contains all points of an x-chain $\langle R_0\rangle_x$, once we have an R_0-positive and R_0-regular base $A(0)$ with $R_0 \in A(0)$ and $h(A(0)) > 0$:

Let $A(i)$ be given, with associated minimum $A_f(i)$ (initially $A_f(0) = R_0$). To find the next base,

(i) try a step with $s \neq f$, $t \neq f$, using lemma 7.5 (figure 7.2). If this is not possible, then

(ii) try a step with $s = f$, using lemma 7.6 (figure 7.3). If the conditions of that lemma are not satisfied for $t = g$, nor for $t = h$, then

(iii) there is a base $A(i+k)$ $(0 \leq k \leq 5)$ according to lemma 7.3 or table 7.1 such that $A(i+k)$ contains $(A_f(i))_x$ and is positive and regular with respect to it. Do the corresponding k steps and let $(A_f(i))_x$ be the minimum associated with $A(i+k)$.

THEOREM 7.8. *Let $A(0)$ be as above and suppose ℓ is independent. The expansion by the algorithm just explained contains all points of the chain* $\langle R_0\rangle_x = R_0, R_1, R_2, \ldots\ldots$

PROOF. Let i_0 be such that $R_j \in A(i_0)$ for some $j \geq 0$. We show $R_{j+1} \in A(i_1)$ for some $i_1 \geq i_0$. By the independence of ℓ lemma 7.5 does not apply to $A(i_2)$ for some least $i_2 \geq i_0$. Also by the independence of ℓ lemma 7.6 does not apply infinitely often, hence not to, say, $A(i_3)$ for some least $i_3 \geq i_2$. Moreover, lemma 7.5 still does not apply to $A(i_3)$, for $\delta - \gamma$ $(\sim a_g)$ and $\alpha - \beta$ $(\sim a_h)$ have not changed since $A(i_2)$. Thus $R_{j+1} \in A(i_3+k)$ for some $k \leq 5$ according to lemma 7.3 or table 7.1. □

We conclude with an acceleration of those steps where lemma 7.5 or lemma 7.6 applies.

LEMMA 7.9. *(a) Suppose A satisfies the conditions of lemma 7.5, and put*

$$b = \left[\frac{\delta - \gamma}{\alpha - \beta}\right] - \left[\frac{2}{\alpha - \beta}\right].$$

If $b \geq 2$, then the lemma remains true with AI_{gh} replaced by AI^b_{gh}.
(b) Suppose A satisfies the conditions of lemma 7.6, and put

$$b = \min\left(\left[\frac{a_f}{a_g}\right],\ [\alpha]\right).$$

If $b \geq 2$, then the lemma remains true with AI_{fg} replaced by AI^b_{fg}.

PROOF. (a) By choice of b we have $(\delta - \gamma) - b'(\alpha - \beta) > \max(2, \alpha - \beta)$ for $0 \leq b' \leq b-1$, so that to each of the b steps lemma 7.5 applies.
(b) By choice of b we have $\alpha - b' > 1$ for $0 \leq b' \leq b-1$, so that to each of the b steps lemma 7.6 applies. □

7D. Application: Units in cubic number fields.

Let $f(x) \in \mathbb{Z}[x]$ be an irreducible polynomial of the third degree and with positive discriminant. Then it has three distinct real roots θ, θ' and θ''. The number θ defines a real cubic number field $\mathbb{Q}(\theta)$. We denote the maximal ring of integers in $\mathbb{Q}(\theta)$ by $\mathcal{O}(\theta)$, the unit group by $\mathcal{E}(\theta)$. From the unit theorem of Dirichlet we know the structure of $\mathcal{E}(\theta)$: There are two so-called fundamental units ε_1 and ε_2 such that

$$\mathcal{E}(\theta) = \{\pm\varepsilon_1^a\varepsilon_2^b \mid a,b \in \mathbb{Z}\},$$

and $\varepsilon_1^a\varepsilon_2^b = 1$ only if $a = b = 0$. When ε_1 and ε_2 are fundamental, then so are $\varepsilon_1^a\varepsilon_2^b$ and $\varepsilon_1^c\varepsilon_2^d$ if $ad - bc = \pm 1$.

However, Dirichlet's theorem does not tell us how to find a pair of fundamental units in a given field $\mathbb{Q}(\theta)$. In this section we describe a method involving chains of relative minima in $\mathbb{R}^3$. The method is due to VORONOI[1], but for the proofs we refer to the standard work of DELONE & FADDEEV[1], where a detailed stereometric treatment will be found in chapter IVA.

With $\mathcal{O}(\theta)$ we associate the irreducible lattice

$$\Omega(\theta) = \{(\alpha, \alpha', \alpha'') \mid \alpha \in \mathcal{O}(\theta)\},$$

where α' and α'' are the conjugates of α. The point $(\alpha, \alpha', \alpha'')$ will also be denoted by α.

LEMMA 7.10. *Let α be a relative minimum of $\Omega(\theta)$. Then its norm satisfies $|N(\alpha)| < \sqrt{D(\theta)}$, where $D(\theta)$ is the field discriminant of $\mathbb{Q}(\theta)$.*

PROOF. The basic volume of $\Omega(\theta)$ is $\sqrt{D(\theta)}$. From theorem 4.3 we find that $8|N(\alpha)| < 8\sqrt{D(\theta)}$, since the body $|x| < |\alpha|$, $|y| < |\alpha'|$, $|z| < |\alpha''|$ has volume $8|\alpha\alpha'\alpha''| = 8|N(\alpha)|$ and contains no lattice point except 0. □

LEMMA 7.11. *Let* α *be a relative minimum of* $\Omega(\theta)$. *Then so is* $\alpha\varepsilon$ *for every* $\varepsilon \in E(\theta)$. *If* α_1 *is the x-successor of* α, *then* $\alpha_1\varepsilon$ *is the x-successor of* $\alpha\varepsilon$.

PROOF. To prove both assertions at once, it suffices to show that there is no non-zero $\beta \in \Omega(\theta)$ which satisfies simultaneously $|\beta| < |\alpha_1\varepsilon|$, $|\beta'| < |\alpha'\varepsilon'|$ and $|\beta''| < |\alpha''\varepsilon''|$. Indeed, we would obtain $|\beta/\varepsilon| < |\alpha_1|$, $|\beta'/\varepsilon'| < |\alpha'|$ and $|\beta''/\varepsilon''| < |\alpha''|$. Since α is a relative minimum and α_1 its x-successor, this implies $\beta = 0$. □

Now take any relative minimum α_0 and construct its x-chain

$$\alpha_0 < \alpha_1 < \alpha_2 < \ldots\ldots\ldots \tag{7.24}$$

of relative minima (or its y- or z-chain). Since there are in $\Omega(\theta)$ only finitely many non-associated elements of a given norm and the norms of the points α_i are bounded (by lemma 7.10), this chain must contain two points $\alpha_k < \alpha_{k+m}$ $(m > 0)$ whose quotient $\varepsilon = \alpha_{k+m}/\alpha_k$ is a non-trivial unit in $E(\theta)$. By lemma 7.11 the chain (7.24) must be periodic at least from α_k on, so that it reads

$$\alpha_k < \ldots\ldots < \alpha_{k+m} = \alpha_k\varepsilon < \alpha_{k+1}\varepsilon < \ldots\ldots < \alpha_{k+m}\varepsilon = \alpha_k\varepsilon^2 < \ldots\ldots$$

But this chain can also be extended leftwards if we use lemma 7.11 with ε^{-1} instead of ε. We then obtain the *two-sided x-chain*

$$\ldots\ldots < \alpha_k\varepsilon^{-1} < \ldots\ldots < \alpha_k < \ldots\ldots < \alpha_k\varepsilon < \ldots\ldots < \alpha_k\varepsilon^2 < \ldots\ldots, \tag{7.25}$$

in which every element is the x-successor of the one immediately to its left. Note that (7.25) does not necessarily contain α_0 if $k > 0$.

LEMMA 7.12. *Two two-sided chains of minima with respect to the same coordinate axis have no or all elements in common.*

PROOF. Trivial. □

LEMMA 7.13. *Two two-sided chains of minima with respect to different coordinate axes always have a common element.*

PROOF. See DELONE & FADDEEV[1], p. 258, theorem 6. □

From lemma 7.12 and lemma 7.13 Voronoi derives the following method to obtain a pair of fundamental units $\varepsilon_1, \varepsilon_2$ of $E(\theta)$:

(a) Choose an arbitrary relative minimum α_0 (e.g. $\alpha_0 = 1$) and construct its x-chain $\alpha_0 < \alpha_1 < \alpha_2 < \ldots\ldots$

(b) Let ε_1 be the first automorphism of the chain, i.e.

$$\varepsilon_1 = \frac{\alpha_{k+m}}{\alpha_k}$$

where k and m are such that $m > 0$ and $\alpha_j/\alpha_i \notin E(\theta)$ when $k+m > j > i \geq 0$.

(c) Construct the y-chain (or z-chain) of α_k, say $\alpha_k = \beta_0 < \beta_1 < \beta_2 < \ldots\ldots$

(d) Determine the least $j \geq 1$ such that $\beta_j/\alpha_i \in E(\theta)$ for some i, $k \leq i < k+m$. Then

$$\varepsilon_2 = \frac{\beta_j}{\alpha_i}.$$

<u>THEOREM 7.14.</u> *When ε_1 and ε_2 are determined by the above method they constitute a fundamental pair.*

<u>PROOF.</u> See DELONE & FADDEEV[1], pp. 259-261. □

For the computation of the chains in (a) and (c) the algorithm of section 7C can be used. Therefore it is possible to find the unit group by means of a 2-dimensional continued fraction algorithm. Note that in (b) and (d) only elements of equal norm must be compared.

<u>EXAMPLE.</u> Let θ be the positive root of

$$f(x) = x^3 - 7x - 2. \tag{7.26}$$

Rounded to 3 decimal places we have $\theta = 2.778$, $\theta' = -0.289$, $\theta'' = -2.489$. By standard methods (DELONE & FADDEEV[1], §17) we find that 1, θ and $\frac{1}{2}(\theta^2 + \theta)$ form an integral base of $\mathcal{O}(\theta)$.

Let us first calculate the x-chain of the relative minimum $1 = (1, 1, 1)$. We start with the base $A(0) = \{A_0(0), A_1(0), A_2(0)\}$ consisting of $A_0(0) = 1$, $A_1(0) = \theta$ and $A_2(0) = \frac{1}{2}(\theta^2 + \theta)$. In $\eta\zeta$-coordinates we have $\underline{A}_0(0) = (-1, -1)$, $\underline{A}_1(0) = (0.289, 2.489)$ and $\underline{A}_2(0) = (0.103, -1.854)$. It can easily be verified that $A(0)$ is positive and regular with respect to the associated minimum 1.

TABLE 7.2. (All figures rounded to 3 decimal places).

step	s	a_s	t	A_t	$\underline{A}_t$ (η,ζ)	applicable is
begin	0	0.792	0	1	(−1 , −1)	
	1	1.957	1	θ	(0.289, 2.489)	
	2	2.200	2	$\frac{1}{2}\theta + \frac{1}{2}\theta^2$	(0.103, −1.854)	lemma 7.5 (b=1)
1	2	0.243	1	$\frac{3}{2}\theta + \frac{1}{2}\theta^2$	(0.392, 0.635)	lemma 7.3(b): $(1)_x = \frac{1}{2}(\theta^2 + 3\theta)$
2	redefine $\eta\zeta$-co-ordinates		0	1	(2.551, 1.575)	
			1	$\frac{3}{2}\theta + \frac{1}{2}\theta^2$	(−1 , −1)	
			2	$\frac{1}{2}\theta + \frac{1}{2}\theta^2$	(−0.263, 2.920)	lemma 7.5 (b=2)
3	0	0.306	2	$2 + \frac{1}{2}\theta + \frac{1}{2}\theta^2$	(4.839, 6.070)	lemma 7.6 (b=5)
4	1	0.742	2	$2 + 8\theta + 3\theta^2$	(−0.161, 1.070)	lemma 7.6 (b=2)
5	1	0.130	0	$1 + 3\theta + \theta^2$	(0.551, −0.425)	lemma 7.3(b): $(\frac{1}{2}(\theta^2+3\theta))_x = 1+3\theta+\theta^2$

We find that $(1)_x = \tau = \frac{1}{2}(\theta^2 + 3\theta)$ with $N(\tau) = 2$, and that $(\tau)_x = \varepsilon_1 = 1 + 3\theta + \theta^2$, a unit with $N(\varepsilon_1) = -1$. From now on the chain is periodic and it contains $1, \tau, \varepsilon_1, \tau\varepsilon_1, \varepsilon_1^2, \ldots\ldots$

To obtain a second unit we have to construct the y- or z-chain of the first element of the period of the x-chain, which is the point 1. Let us take the z-chain $\langle 1\rangle_z$. Consider the base $1, -\theta, \frac{1}{2}(\theta^2 + \theta)$, with the minus sign before θ because the heights, that is the z-coordinates (second conjugates), must be positive. In projection on the $\xi\eta$-plane we get $\underline{1} = (-1, -1)$, $\underline{-\theta} = (2.778, -0.289)$ and $\underline{\frac{1}{2}(\theta^2 + \theta)} = (-5.249, 0.103)$, if we round to 3 decimal places. However, the cofactors of this base are not all positive. Instead, we start with $A(0)$ consisting of

$$A_0(0) = 1,$$

$$A_1(0) = -\theta - 1,$$

$$A_2(0) = \tfrac{1}{2}(\theta^2 + \theta).$$

TABLE 7.3. (All figures rounded to 3 decimal places).

step	s	a_s	t	A_t	$\underline{A}_t\ (\xi,\eta)$	applicable is
begin	0	4.121	0	1	(−1 , −1)	
	1	5.352	1	$-1-\theta$	(3.778, 0.711)	
	2	3.067	2	$\frac{1}{2}\theta+\frac{1}{2}\theta^2$	(−5.249, 0.103)	lemma 7.5 (b=1)
1	1	2.285	2	$-1-\frac{1}{2}\theta+\frac{1}{2}\theta^2$	(−1.471, 0.814)	lemma 7.5 (b=1)
2	2	0.782	1	$-2-\frac{3}{2}\theta+\frac{1}{2}\theta^2$	(2.307, 1.525)	lemma 7.5 (b=1)
3	1	1.503	2	$-3-2\theta+\theta^2$	(0.836, 2.339)	lemma 7.6 (b=1)
4	0	3.339	2	$-2-2\theta+\theta^2$	(−0.164, 1.339)	lemma 7.6 (b=2)
5	0	0.333	1	$-\frac{3}{2}\theta+\frac{1}{2}\theta^2$	(0.307, −0.475)	lemma 7.3(b): $(1)_z=\frac{1}{2}(\theta^2-3\theta)$

The z-coordinate of $A_1(0)$ is $-\theta''-1=1.489>0$. It can now easily be verified that $A(0)$ is positive and regular with respect to 1. Its expansion is given in table 7.3.

The z-successor of 1 appears to be $\varepsilon_2=\frac{1}{2}(\theta^2-3\theta)$ with norm $N(\varepsilon_2)=-1$. Since ε_2 clearly is the first element of the z-chain associated to ε_1 or τ, we know from theorem 5.14 that $\varepsilon_1=1+3\theta+\theta^2$ and $\varepsilon_2=\frac{1}{2}(\theta^2-3\theta)$ are a pair of fundamental units.

HISTORICAL REMARKS. To find the successor of a given relative minimum Voronoi gave a technique where one has to choose that successor among at most five candidates. It can be found in geometric terms in DELONE & FADDEEV[1], §59, and in a version adapted to computer use in WILLIAMS & ZARNKE[2]. It was used for the calculations in (a) and (c) of Voronoi's method by WILLIAMS & ZARNKE[2,3], who gave a table of the fundamental units of all fields of positive discriminant with defining equation $x^3-px-q=0$, $|p|,|q|\le 15$, and by ANGELL[2], who compiled a unit table for all 4794 cubic fields $\mathbb{Q}(\theta)$ with discriminant $D(\theta)$ in the range $0<D(\theta)<10^5$. Using Voronoi's method, STENDER[2] disproved the fundamentality of a pair of units which BERNSTEIN[15] and BERNSTEIN & HASSE[1] obtained by means of the Jacobi-Perron algorithm for fields defined by

$$f(x) = x(x+d_1)(x+d_2) - d = 0,$$

where d_1, d_2 and d underly certain restrictions (notably $d|(d_1,d_2)$). DAUS[2] had some success at finding units by his variation of the Jacobi-Perron algorithm (section 3G). The theory of BERGMANN[1-5] is for a good part related to Voronoi's method and work of MINKOWSKI[1], but seems computationally less feasible.

BERWICK[2] proved in 1932 that any two of the units $\varepsilon_1,\varepsilon_2,\varepsilon_3$ defined by

$$\begin{array}{lll} |\varepsilon_1'| < 1, & |\varepsilon_1''| < 1, & \varepsilon_1 > 0 \text{ and minimal,} \\ |\varepsilon_2| < 1, & |\varepsilon_2''| < 1, & \varepsilon_2' > 0 \text{ and minimal,} \\ |\varepsilon_3| < 1, & |\varepsilon_3'| < 1, & \varepsilon_3'' > 0 \text{ and minimal} \end{array} \tag{7.26}$$

form a fundamental pair. The question remains how to compute these units. The algorithm proposed by BILLEVICH[1] (see STEINER & RUDMAN[1]) is efficient only if the units have small coefficients when expressed in terms of the lattice base: When 1, ω_1, ω_2 form such a base, he solves (for ε_1) the simultaneous inequalities

$$|p + q\omega_1' + r\omega_2'| < 1$$
$$|p + q\omega_1'' + r\omega_2''| < 1$$

in integers q,r for $p = 0,1,2,\ldots\ldots$ successively. We might also obtain ε_1 in (7.26) as a best approximation to the x-axis with respect to a certain Euclidean radius function if we have an idea of the magnitude of $|\varepsilon_1'/\varepsilon_1''|$:

<u>LEMMA 7.15.</u> *Let ε_1 be defined by (7.26). Suppose for all $\alpha \in O(\theta)$ with $\alpha \notin E(\theta) \cup \{0\}$ we have $|N(\alpha)| \geq t$. Also suppose there is a real number u such that $\varepsilon_1'^2 + u^2\varepsilon_1''^2 < \min(1,u^2)$ and*

$$(t - \sqrt{t^2-1})\left|\frac{\varepsilon_1'}{\varepsilon_1''}\right| \leq u \leq (t + \sqrt{t^2-1})\left|\frac{\varepsilon_1'}{\varepsilon_1''}\right|. \tag{7.27}$$

Then ε_1 is a best approximation to the x-axis for the Euclidean radius function $\rho(x,y,z) = \sqrt{y^2 + u^2z^2}$.

<u>PROOF.</u> Suppose there is a non-zero $\alpha \in O(\theta)$ with $\alpha < \varepsilon_1$ and $\rho(\alpha) < \rho(\varepsilon_1)$.

Since 1 is a relative minimum, and $\varepsilon_1'^2 + u^2\varepsilon_2''^2 < \min(1,u^2)$, we have $1 < \alpha < \varepsilon_1$. Under the restriction $\rho(\alpha) < \rho(\varepsilon_1)$, which yields $\alpha'^2 + u^2\alpha''^2 < \varepsilon_1'^2 + u^2\varepsilon_1''^2$, an upper bound for $|\alpha'\alpha''|$ is $(\varepsilon_1'^2 + u^2\varepsilon_1''^2)/2u$. From (7.27) we find $\varepsilon_1'^2 + u^2\varepsilon_1''^2 \leq 2tu|\varepsilon_1'\varepsilon_1''|$, hence $|\alpha'\alpha''| < t|\varepsilon_1'\varepsilon_1''|$. With $1 < \alpha < \varepsilon_1$ this gives $|N(\alpha)| < t$. Hence α is a unit, in contradiction with the minimality of ε_1. □

So given a suitable t and u, we can find ε_1 by the 2-dimensional best approximation algorithm. But usually one cannot know the magnitude of $|\varepsilon_1'/\varepsilon_1''|$ before ε_1 is known. (If an upper bound M for ε_1 is known, we have $M^{-1} < |\varepsilon_1'/\varepsilon_1''| < M$).

Another way to obtain units in cubic fields of positive discriminant was proposed in 1956 by GODWIN[1]. For $\varepsilon \in E(\theta)$ define $S(\theta) = (\varepsilon' - \varepsilon'')^2 + (\varepsilon'' - \varepsilon)^2 + (\varepsilon - \varepsilon')^2$. Let $\varepsilon_1 \in E(\theta) \setminus \{\pm 1\}$ be such that $S(\varepsilon_1)$ is minimal and let $\varepsilon_2 \in E(\theta) \setminus \{\pm\varepsilon_1^m \mid m \in \mathbb{Z}\}$ be such that $S(\varepsilon_2)$ is minimal. Godwin conjectured that ε_1, ε_2 form a fundamental pair except perhaps for finitely many exceptional cases. BRUNOTTE & HALTER-KOCH[1] proved that ε_1, ε_2 generate a subgroup of $E(\theta)$ of index at most 4.

Finally we mention work of STEINER[1] and RUDMAN & STEINER[1]. They take ε_1 and ε_2 to be the first autumorphisms of the x- and y-chain of the minimum 1. Though independent, these units do not always make a fundamental pair (STEINER[1]).

The special case of a cyclic field (where θ' and θ'' belong to $\mathbb{Q}(\theta)$, or equivalently, $D(\theta)$ is a square) was studied by HASSE[1], GRAS-MONTOUCHET[1], COHN & GORN[1].

APPENDIX. ON COMPUTER IMPLEMENTATION

The purpose of this appendix is to discuss some of the difficulties connected with computer implementation of multi-dimensional continued fraction algorithms. The principal problems are:

- How to deal with round-off errors.
- How to use multi-length arithmetic in the most efficient way.

Writing an efficient computer program is mainly a question of balance, which involves technical aspects, like the nature of the multi-length package, the available storage and the relative speed of multiplication and addition.

The first step in writing a program for a subtractive multi-dimensional continued fraction algorithm is: to set up arrays which must carry the information the algorithm needs.

The lattice base points $A_k(i)$.

If we wish to compute approximations, we have to include an array containing the $A_k(i)$. Usually it is convenient to express these points in integral coordinates $A_k(i) = (x_{k0}(i),\ldots\ldots,x_{kn}(i))$ with respect to the initial base $A(0)$. As was pointed out in the example on p. 120-122, it is sufficient to compute only one coordinate, say $x_{k0}(i)$, of each $A_k(i)$. If we do so, $A_0,\ldots\ldots,A_n$ can be represented in a 1-dimensional array of length n+1 with initial values (1, 0, ..., 0). Each step requires one addition in this array, that is

$$\text{(a.1)} \qquad \begin{cases} x_{t0}(i+1) = x_{t0}(i) + x_{s0}(i) \\ x_{k0}(i+1) = x_{k0}(i) \quad \text{for } k \neq t. \end{cases}$$

Moreover, the initial values and each addition being exact, there will be no round-off errors at all.

The cofactors.

The indispensable array (vector) $\hat{a}$ of cofactors will be loaded with initial values $\hat{a}_0(0),\ldots\ldots,\hat{a}_n(0)$ which are approximations to the true cofactors $a_0(0),\ldots\ldots,a_n(0)$, say correct to D decimal places, so that

(a.2) $$|\hat{a}_k(0) - a_k(0)| < 10^{-D} \quad \text{for } k = 0,\ldots\ldots,n.$$

Each step requires one subtraction, that is

(a.3) $$\begin{cases} \hat{a}_s(i+1) = \hat{a}_s(i) - \hat{a}_t(i) \\ \hat{a}_k(i+1) = \hat{a}_k(i) \quad \text{for } k \neq s. \end{cases}$$

Since subtraction produces no round-off error, any error $|\hat{a}_k(i) - a_k(i)|$ is the consequence of the round-off errors in the initial values. We give two ways to estimate this error.

1. Each cofactor $a_k(i)$ is a linear combination of $a_0(0),\ldots\ldots,a_n(0)$:

$$a_k(i) = \sum_{j=0}^{n} y_{kj}(i) a_j(0).$$

From (a.2) we have

(a.4) $$|\hat{a}_k(i) - a_k(i)| < 10^{-D} \cdot \sum_{j=0}^{n} |y_{kj}(i)|.$$

The computation of the $y_{kj}(i)$ requires an $(n+1)\times(n+1)$ array Y (initially $Y = I$) with n+1 subtractions $y_{sj}(i+1) = y_{sj}(i) - y_{tj}(i)$ in each step.

2. Instead of Y we define the cofactor error estimating array δ, with initial values $\delta_j(0) = 10^{-D}$ for $j = 0,\ldots\ldots,n$, and subject to the change

(a.5) $$\begin{cases} \delta_s(i+1) = \delta_s(i) + \delta_t(i) \\ \delta_k(i+1) = \delta_k(i) \quad \text{for } k \neq s \end{cases}$$

along with (a.3). By induction one shows that

(a.6) $$|\hat{a}_k(i) - a_k(i)| < \delta_k(i) \quad \text{for all } k \text{ and } i.$$

The estimate (a.6) seems much cheaper in calculation than (a.4), since only one addition is required instead of n+1 subtractions per step. However, the $\delta_k(i)$ become much larger than the $y_{kj}(i)$ (because of cancellation in the latter), hence a greater precision may be required. Therefore it depends on the problem at hand whether (a.6) or (a.4) is cheaper.

EXAMPLE. After the 249-th step in the example on p. 127 the highest observed value of $10^{-24}.\sum_{j=0}^{n}|y_{kj}(i)|$ was 5092.10^{-24} (for $i=248$, $k=6$). At that time the $\delta_k(i)$ had gone up to about 10^{-12}.

Using (a.4) or (a.6) the computer can *prove* a cofactor inequality $a_f \geq a_g$ by verifying that

$$\hat{a}_f - \hat{a}_g \geq 10^{-D}.\sum_{j=0}^{n}|y_{fj} - y_{gj}|$$

or

$$\hat{a}_f - \hat{a}_g \geq \delta_f + \delta_g$$

respectively. However, we must stop the expansion as soon as the algorithm requires us to compare two cofactors a_f and a_g for which the computer can prove neither $a_f \geq a_g$ nor $a_g \geq a_f$ in this way.

EXAMPLE. In the example of p. 127 this occurred after 249 steps when the estimate (a.4) was used, but after 181 steps already when (a.6) was used. To obtain the result of the example with (a.6), it would have been necessary to know the first seven non-trivial zeros of the zeta function correct to at least 32 decimal places.

The projected lattice points.

For a vectorial algorithm -- such as the J.P.A. or Brun's algorithm -- the above is all we need. For most of the other algorithms discussed in this book we need information on the projections $\underline{A}_0,\ldots\ldots,\underline{A}_n$ of the base $\{A_0,\ldots\ldots,A_n\}$.

In ℓ^* we choose an orthonormal coordinate system with respect to which we write $\underline{A}_k(i) = (z_{k1}(i),\ldots\ldots,z_{kn}(i))$. To carry the projections we feed n+1 points $\hat{\underline{A}}_k(0) = (\hat{z}_{k1}(0),\ldots\ldots,\hat{z}_{kn}(0))$ into an $(n+1)\times n$ array $\hat{\underline{A}}$. We assume that $\hat{z}_{kj}(0) \in 10^{-D}.\mathbb{Z}$, and that

$$(a.7) \qquad |z_{kj}(0) - \hat{z}_{kj}(0)| < 10^{-D}$$

for $k=0,\ldots\ldots,n$ and $j=1,\ldots\ldots,n$. Each step requires us to do n additions

in the array $\underline{\hat{A}}$, namely

$$\text{(a.8)} \qquad \begin{cases} \hat{z}_{tj}(i+1) = \hat{z}_{tj}(i) + \hat{z}_{sj}(i) \\ \hat{z}_{kj}(i+1) = \hat{z}_{kj}(i) \quad \text{for } k \neq t \end{cases} \qquad j = 1,\ldots,n.$$

If $A_k(i) = \sum_{q=0}^{n} x_{kq}(i)A_q(0)$, then (a.7) gives the estimate

$$\text{(a.9)} \qquad |\hat{z}_{kj}(i) - z_{kj}(i)| < 10^{-D}. \sum_{q=0}^{n} x_{kq}(i), \quad \text{all } k,j,i,$$

for the accumulated error in $\hat{z}_{kj}(i)$ (note that the addition (a.8) is exact and produces no round-off error). To apply (a.9) it is not necessary to know all $x_{kq}(i)$. In fact, (a.9) may be replaced by

$$\text{(a.10)} \qquad |\hat{z}_{kj}(i) - z_{kj}(i)| < 10^{-D}\left(\varepsilon + x_{k0}(i). \sum_{q=0}^{n} \frac{a_q(0)}{a_0(0)}\right),$$

so that we need only the value of $x_{k0}(i)$, which we find in the array of first coordinates. Here ε is a small quantity, a precise estimate of which depends on $A(0)$ and $|\underline{A}_k(i)|$.

<u>EXAMPLE.</u> If $\underline{A}_1(0),\ldots\ldots,\underline{A}_n(0)$ are orthonormal in ℓ^*, then

$$\begin{aligned} |\underline{A}_k(i)| &= \left|\sum_{q=0}^{n} x_{kq}(i)\underline{A}_q(0)\right| \\ &= \left|\sum_{q=0}^{n} x_{kq}(i)\underline{A}_q(0) - \sum_{q=0}^{n} \frac{x_{k0}(i)}{a_0(0)} a_q(0)\underline{A}_q(0)\right| \\ &= \left|\sum_{q=1}^{n} \left(x_{kq}(i) - \frac{x_{k0}(i)}{a_0(0)} a_q(0)\right)\underline{A}_q(0)\right| \\ &\geq \left|x_{kq'}(i) - \frac{x_{k0}(i)}{a_0(0)} a_{q'}(0)\right| \end{aligned}$$

for $q' = 1,\ldots,n$, so that we may take $\varepsilon = n.|\underline{A}_k(i)|$.

From the error bounds (a.9) or (a.10) for the coordinates $\hat{z}_{kj}(i)$ we can deduce estimates of the errors in various quantities related to these projections.

<u>EXAMPLE.</u> Put $\varepsilon_k(i) = \max_{1\leq j\leq n} |\hat{z}_{kj}(i) - z_{kj}(i)|$. Then we have for $f,g \in \{0,\ldots,n\}$,

omitting the step index i

(a.11) $$|\underline{A}_f \cdot \underline{A}_g - \hat{\underline{A}}_f \cdot \hat{\underline{A}}_g| \le \sqrt{n}(\varepsilon_f |\hat{\underline{A}}_g| + \varepsilon_g |\hat{\underline{A}}_f|) + n\varepsilon_f \varepsilon_g .$$

Proof: Write $v_f = \underline{A}_f - \hat{\underline{A}}_f$ and $v_g = \underline{A}_g - \hat{\underline{A}}_g$. Then

$$\begin{aligned} |\underline{A}_f \cdot \underline{A}_g - \hat{\underline{A}}_f \cdot \hat{\underline{A}}_g| &= |(\hat{\underline{A}}_f + v_f) \cdot (\hat{\underline{A}}_g + v_g) - \hat{\underline{A}}_f \cdot \hat{\underline{A}}_g| \\ &= |v_f \cdot \hat{\underline{A}}_g + v_g \cdot \hat{\underline{A}}_f + v_f \cdot v_g| \\ &\le |v_f| |\hat{\underline{A}}_g| + |v_g| |\hat{\underline{A}}_f| + |v_f| |v_g| \\ &\le \sqrt{n}(\varepsilon_f |\hat{\underline{A}}_g| + \varepsilon_g |\hat{\underline{A}}_f|) + n\varepsilon_f \varepsilon_g , \end{aligned}$$

where the last inequality holds because the coordinate system $(z_1, \ldots, z_n)$ in ℓ^* was assumed orthonormal.

<u>The inner products.</u>

Several algorithms are formulated entirely in convenient terms of the inner products $\underline{A}_f \cdot \underline{A}_g$, such as the 2-dimensional best approximation algorithm or the inner product algorithm itself.

We present two ways to compute the inner products.

<u>1.</u> The first way is to use the formula

(a.12) $$\hat{\underline{A}}_f(i) \cdot \hat{\underline{A}}_g(i) = \sum_{j=1}^{n} \hat{z}_{fj}(i) \hat{z}_{gj}(i)$$

to calculate after each step the n+1 inner products $\hat{\underline{A}}_t(i) \cdot \hat{\underline{A}}_k(i) = \hat{\underline{A}}_k(i) \cdot \hat{\underline{A}}_t(i)$ that have changed during the step. Since the inner products are only meant to be compared, this can be done in single precision from the first 10 or 12 significant digits of the $\hat{z}_{kj}(i)$, $\hat{z}_{tj}(i)$. Therefore this approach requires $n^2 + n$ single precision multiplications in each step.

<u>2.</u> The second way is to use an $(n+1) \times (n+1)$ array $\hat{J}$ for the inner products. Initially, its elements are

(a.13) $$\hat{J}_{fg}(0) = \hat{\underline{A}}_f(0) \cdot \hat{\underline{A}}_g(0) = \sum_{j=1}^{n} \hat{z}_{fj}(0) \hat{z}_{gj}(0), \quad f,g = 0, \ldots, n.$$

Note that $\hat{z}_{fj}(0)$, $\hat{z}_{gj}(0) \in 10^{-D}.\mathbb{Z}$, so that $\hat{J}_{fg}(0) \in 10^{-2D}.\mathbb{Z}$, and the entries of $\hat{J}$ require 2D decimal places. In each step we do the n+3 additions

$$\text{(a.14)} \qquad \begin{cases} \hat{J}_{kt}(i+1) = \hat{J}_{tk}(i+1) = \hat{J}_{tk}(i) + \hat{J}_{sk}(i) & \text{for all } k \neq t \\ \hat{J}_{tt}(i+1) = \hat{J}_{tt}(i) + 2\hat{J}_{st}(i) + \hat{J}_{ss}(i) & \\ \hat{J}_{fg}(i+1) = \hat{J}_{fg}(i) & \text{if } f \neq t \text{ and } g \neq t. \end{cases}$$

Combining (a.13) and (a.14) it is proved by induction that

$$\text{(a.15)} \qquad \hat{J}_{fg}(i) = \underline{\hat{A}}_f(i).\underline{\hat{A}}_g(i) \quad \text{for all } f,g,i.$$

By (a.15), the estimate (a.11) applies to the difference $|\hat{J}_{fg}(i) - \underline{A}_f(i).\underline{A}_g(i)|$. This second approach requires per step n+3 additions of multi-length numbers with 2D decimal places.

REMARK. If the inner products are computed as in the second approach, there is no need to carry the array $\underline{\hat{A}}$, unless one has a special interest in the coordinates $\hat{z}_{kj}(i)$ (e.g. when computing inhomogeneous approximations as in section 6A, p. 122).

EXAMPLES. In the example on p. 127 we had $n = 6$, $D = 24$. To see which approach is cheaper we must compare the cost of $n+3 = 9$ additions of numbers in $\hat{J}$ with 48 decimal places (second way) with the cost of $n = 6$ additions of numbers in $\underline{\hat{A}}$ with 24 decimal places, plus $n^2 + n = 42$ single precision multiplications (first way).

When computing the fundamental unit of $\mathbb{Q}(\sqrt[3]{167})$ (see p. 111), we had $n = 2$ and $D = 154$. Therefore we have to compare the cost of $n+3 = 5$ additions of 308-digit numbers with the cost of $n^2 + n = 6$ single precision multiplications and $n = 2$ additions of 154-digit numbers.

General remarks.

1. It is important that all the above computations can be done by means of addition and subtraction only. If we use multi-length arithmetic, this means that the word length can be twice that of ordinary multi-length packages (e.g. 10^{14} instead of 10^7 on an IBM-370 machine), since no multiplication routine is needed.

2. If we do not use a subtractive algorithm, but allow $b > 1$, then we must be prepared to multiply multi-length numbers with the integer b (which will probably be single precision). However, in the practice of e.g. the inner product algorithm, it appears that the frequency of the event $b = 1$ increases with the dimension, so that one will often simply use the subtractive algorithm (cf. the example on p. 120-122).

3. Of course the calculations (a.1), (a.3), (a.8) and (a.14) must be done in full precision. However, the computations needed only to determine s and t (e.g. the check of the conditions (5.19)-(5.25)) may well be done in single precision, using say 12 significant digits. It is rather unlikely that we have to compare two quantities the first 12 significant digits of which agree, but if so we can always at that moment use multiple precision (if we can find no theoretical reason why the two quantities should be exactly equal).

4. Notwithstanding the preceding remark, an expansion should be terminated as soon as the error estimates do not allow the necessary verification of some inequality. This is certainly true for the cofactor inequalities. It is also true for all inequalities that must be checked in the 2-dimensional best approximation algorithm; otherwise it could not be claimed that the machine computed expansion is the expansion by this algorithm, and hence not that it contains all best approximations in a certain height range. However, for the inner product algorithm it is not fundamental that $\underline{A}_t \cdot \underline{A}_s$ indeed minimizes $\underline{A}_k \cdot \underline{A}_s$, $k \in \{0,\ldots,n\}$; therefore it is not necessary to verify (6.3) strictly by means of the error estimates.

Example: the inner product algorithm.

Suppose we wish to compute good simultaneous rational approximations to n given positive real numbers $\alpha_1,\ldots,\alpha_n$ by means of the inner product algorithm. In $\mathbb{R}^{n+1}$ we take $\Omega = \mathbb{Z}^{n+1}$, $\ell_0 = (1, \alpha_1, \ldots, \alpha_n)$. The height function is $h(x_0,x_1,\ldots,x_n) = x_0$, the hyperplane ℓ^* is $h = x_0 = 0$. The inner product given by the quadratic form

$$x_0^2 + \sum_{j=1}^{n} (x_0\alpha_j - x_j)^2$$

defines the Euclidean radius function

$$\rho(x_0,x_1,\ldots,x_n) = \left(\sum_{j=1}^{n} (x_0\alpha_j - x_j)^2\right)^{\frac{1}{2}}.$$

Moreover, $x_1,\ldots,x_n$ are orthonormal coordinates in ℓ^*.

Let $\hat{\alpha}_1,\ldots,\hat{\alpha}_n \in 10^{-D}.\mathbb{Z}$ be such that $|\alpha_k - \hat{\alpha}_k| < 10^{-D}$, $k = 1,\ldots,n$. The computer program receives as input the arrays

$$\hat{a} = (1,\ \hat{\alpha}_1,\ \ldots\ldots,\ \hat{\alpha}_n)$$

and

$$\hat{A} = (1,\ 0,\ \ldots\ldots,\ 0),$$

furthermore

$$\underline{\hat{A}} = \begin{bmatrix} -\hat{\alpha}_1 & -\hat{\alpha}_2 & \ldots\ldots & -\hat{\alpha}_n \\ 1 & 0 & \ldots\ldots & 0 \\ 0 & 1 & \ldots\ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots\ldots & 1 \end{bmatrix} \quad \text{or} \quad \hat{J} = \begin{bmatrix} * & -\hat{\alpha}_1 & -\hat{\alpha}_2 & \ldots\ldots & -\hat{\alpha}_n \\ -\hat{\alpha}_1 & 1 & 0 & \ldots\ldots & 0 \\ -\hat{\alpha}_2 & 0 & 1 & \ldots\ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -\hat{\alpha}_n & 0 & 0 & \ldots\ldots & 1 \end{bmatrix}$$

with $* = 1 + \hat{\alpha}_1^2 + \ldots\ldots + \hat{\alpha}_n^2$. Whether we work with $\underline{\hat{A}}$ or with $\hat{J}$ depends on n and D, as explained before.

Let us now see how large D must be if we wish to obtain an approximation B with $\rho(B) < \varepsilon$ for some $\varepsilon > 0$. Put $m = -{}^{10}\log\ \varepsilon$. To have $\underline{B}$ correct to some significant digits, we need a little more than m correct decimal places at the end of the expansion. Moreover, by theorem 4.3 the height of B may be of the order $\varepsilon^{-n} = 10^{mn}$. Applying (a.9) we see that D has to be somewhat larger than m(n+1). Since it is always possible that cofactors are nearly equal, one is never certain in advance that such a value of D is sufficient for the required precision in the cofactor computations. We may, though, expect it to be, on the following grounds. From $\ell_0 = a_0 A_0 + a_1 A_1 + \ldots + a_n A_n$ it follows that

$$\sum_{k=0}^{n} a_k h(A_k) = h(\ell_0) = 1.$$

Therefore, the cofactors will in the end be of the order $\frac{1}{n}.10^{-mn}$, since the heights are of order 10^{mn}. Furthermore, there are reasonable arguments to support that the $y_{kj}(i)$ in (a.4) will be of the order 10^{m}. From (a.4)

we see that, again, D has to be a little larger than $m(n+1)$.

EXAMPLE. In the example on p. 127 we had $\varepsilon = 0.001$, so that $m = 3$. According to the discussion above we would need a little more than $m(n+1) = 21$ decimal places; in fact we used 24. The heights went up to $10^{19} = 10.10^{mn}$, the cofactors down to 10^{-21}. The highest value of $|y_{kj}(i)|$ was 1807, a little larger than $10^m = 1000$,

EXAMPLE. Our discussion of the required number of decimal places applies to the 2-dimensional best approximation algorithm also. Suppose we wish to compute the fundamental unit $\varepsilon_0 \approx 6.10^{95}$ of $\mathbb{Q}(\sqrt[3]{167})$ (see p. 111). Since the projection of ε_0 has length $|\varepsilon_0'| = \varepsilon_0^{-\frac{1}{2}}$ we must take $m = 48$, so that D has to be slightly larger than $m(n+1) = 48\times 3 = 144$. In fact $D = 154$ was used.

Algebraic methods.

In the last example we could guess at the required value of D because a unit of order 6.10^{95} was known, so that ε_0 would be no larger than that. Normally, however, one has no idea of the size of ε_0 before it is found, and thus it is impossible to tell how large D must be. The following implementation of the 2-dimensional best approximation algorithm may solve this problem.

Let $1, \theta_0, \omega_0$ be a base of the first or second kind for the integer lattice $\Omega_0 = \Omega(\theta)$ in our cubic field $\mathbb{Q}(\theta)$ of negative discriminant Δ. Fix the maximal number D of digits we wish to work with; a good choice of D depends on Δ, among other things. We break up the expansion into parts: Define lattices Ω_i with base $\{1, \theta_i, \omega_i\}$ $(i = 1,2,\ldots.)$ as follows:

1. Apply the best approximation algorithm to $\{1, \theta_{i-1}, \omega_{i-1}\}$ for some time (depending on D). Suppose as a result we have the base $\{\alpha_i, \beta_i, \gamma_i\}$, where we may assume that $|\alpha_i'| \le |\beta_i'| \le |\gamma_i'|$.

2. Perform the *algebraic*, hence exact, divisions

$$\begin{cases} \theta_i = \beta_i/\alpha_i = u_{i0} + u_{i1}\theta_0 + u_{i2}\omega_0, & u_{ij} \in \mathbb{Q} \\ \omega_i = \gamma_i/\alpha_i = v_{i0} + v_{i1}\theta_0 + v_{i2}\omega_0, & v_{ij} \in \mathbb{Q}. \end{cases}$$

3. The coefficients u_{ij} and v_{ij} are used to compute the projections of θ_i, ω_i from those of θ_0 and ω_0. Then take $i := i+1$ and return to 1.

Note that the θ_i and ω_i are not necessarily algebraic integers: The

coefficients u_{ij} and v_{ij} have denominators dividing $\prod_{k=1}^{i} N(\alpha_k)$.

It is easy to prove that a point ϕ is a best approximation of Ω_i if and only if $\phi\alpha_1 \ldots\ldots \alpha_i$ is one of $\Omega_0 = \Omega(\theta)$. To check if we have found a unit we must check if $N(\phi) \cdot \prod_{k=1}^{i} N(\alpha_k) = 1$.

<u>EXAMPLE.</u> Suppose we interrupt the expansion of example 4 on p. 110 after 44 steps. We had started with the base $\{1, \theta_0, \omega_0\} = \{1, \theta, \theta^2\}$, $\theta = \sqrt[3]{29}$. The base $A(44)$ consists of

$$\alpha_1 = 324876 + 1057430 + 344180^2, \quad N(\alpha_1) = 3,$$
$$\beta_1 = 6214280 + 20226690 + 6583530^2, \quad N(\beta_1) = 98,$$
$$\gamma_1 = 2120763 + 6902820 + 2246780^2, \quad N(\gamma_1) = 49.$$

We have $\alpha_1^{-1} = \frac{1}{3}(730 + 1280 - 1190^2)$, and from this we obtain

$$\theta_1 = \beta_1/\alpha_1 = \tfrac{17}{3} + \tfrac{7}{3}\theta + \tfrac{2}{3}\theta^2$$

and

$$\omega_1 = \gamma_1/\alpha_1 = \tfrac{4}{3} + \tfrac{2}{3}\theta + \tfrac{1}{3}\theta^2.$$

Then we start the expansion of $\{1, \theta_1, \omega_1\}$, remembering that every point ϕ in this expansion corresponds to $\phi\alpha_1$ of the original expansion.

<u>REMARK.</u> Details of similar algebraic calculations on a computer may be found in MENDY & JEANS [1] and in WILLIAMS, CORMACK & SEAH [1]. Their work concerns the J.P.A. and Voronoi's method, respectively, but applies to our algorithm as well. However, it is not necessary to divide by each best approximation one finds, as they do; it is sufficient to perform the algebraic divisions merely each time we run out of our decimal precision. In order to keep the norms of the α_k small, it may be preferable to stop a little earlier if we find a best approximation of relatively small norm.

REFERENCES

References are listed in alphabetical order of authors. Joint papers appear under their first-named author. After "cit." one finds the page(s) where the work is cited in this book; "not cit." means, of course, that the paper is not mentioned in the text but has been included for the sake of completeness. An * marks that the paper in question has not been available to the author; when results from such work are used in the text, the source is mentioned.

-v-v-

AGRAWAL, M.K.; COATES, J.H.; HUNT, D.C. & VAN DER POORTEN, A.J.

[1] (cit. 62) *Elliptic curves of conductor 11*, Math. Comput. 35 (1980) p. 991-1002.

ANGELL, I.O.

[1] (cit. 116) *A table of complex cubic fields*, Bull. Lond. Math. Soc. 5 (1973) p. 37-38 and review in Math. Comput. 29 (1975) p. 661-665.

[2] (cit. 152) *A table of totally real cubic fields*, Math. Comput. 30 (1976) p. 184-187.

ARWIN, A.

[1] (not cit.) *Two-dimensional chains*, Am. J. Math. 53 (1931) p. 91-102.

BACHMANN, P.

[1] (cit. 34) *Zur Theorie von Jacobi's Kettenbruch-Algorithmen*, J. Reine Angew. Math. 75 (1873) p. 25-34.

BARBOUR, J.M.

[1] (cit. 43, 44) *Music and ternary continued fractions*, Am. Math. Mon. 55 (1948) p. 545-555.

BATEMAN, H.

*[1] (cit. 33) *The linear difference equation of the third order*, Q. J. Math. Oxf. 41 (1910) p. 302-308.

BEACH, B.D.; WILLIAMS, H.C. & ZARNKE, C.R.

*[1] (cit. 116) *Some computer results on units of quadratic and cubic*

fields, Scientific report 31, Univ. of Manitoba, Winnipeg, 1971, and review in Math. Comput. 29 (1975) p. 330.

BERGMAN, G.M.

[1] (cit. 42, 43, 44) *Notes on Ferguson and Forcade's generalized Euclidean algorithm*, preprint, see Am. Math. Soc. Abstracts 1 (1980) p. 369.

BERGMANN-BULLIG, G.

[1] (cit. 153) *Ein periodisches Verfahren zur Berechnung eines Systems von Grundeinheiten in den totalreellen kubischen Körpern*, Abh. Math. Semin. Univ. Hamb. 12 (1938) p. 369-414.

[2] (cit. 153) *Periodische Ketten linearer Transformationen*, J. Reine Angew. Math. 190 (1952) p. 108-124.

[3] (cit. 153) *Theorie der Netze*, Math. Ann. 149 (1963) p. 361-418.

[4] (cit, 153) *Zur numerischen Bestimmung einer Einheitenbasis*, Math. Ann. 166 (1966) p. 103-105.

[5] (cit. 153) *Ein Beispiel numerischer Einheitenberechnung*, Math. Ann. 167 (1966) p. 143-168.

BERNSTEIN, L.

[1] (cit. 36) *Periodical continued fractions for irrationals of degree n by Jacobi's algorithm*, J. Reine Angew. Math. 213 (1963) p. 31-38.

[2] (cit. 36) *Periodicity of Jacobi's algorithm for a special type of cubic irrationals*, J. Reine Angew. Math. 213 (1963) p. 137-146.

[3] (cit. 36) *Periodische Jacobische Algorithmen für eine unendliche Klasse algebraischer Irrationalzahlen vom Grade n, und einige Klassen kubischer Irrationalzahlen*, J. Reine Angew. Math. 214/215 (1964) p. 76-83.

[4] (cit. 36) *Periodische Jacobi-Perronsche Algorithmen*, Archiv d. Math. 15 (1964) p. 421-429.

[5] (cit. 39, 40) *Rational approximations of algebraic irrationals by means of a modified Jacobi-Perron algorithm*, Duke Math. J. 32 (1965) p. 161-176.

[6] (cit. 36) *Representation of* $\sqrt[n]{D^n - d}$ *as a periodic continued fraction by Jacobi's algorithm*, Math. Nachr. 29 (1965) p. 179-200.

[7] (cit. 36) *New infinite classes of periodic Jacobi-Perron algorithms*, Pac. J. Math. 16 (1965) p. 439-469.

[8] (cit. 36) *A periodic Jacobi-Perron algorithm*, Can. J. Math. 17 (1965) p. 933-945.

[9] (cit. 39) *The modified algorithm of Jacobi-Perron*, Mem. Am. Math. Soc. 67 (1966) p. 1-44.

[10] (cit. 37, 40) *Der B-Algorithmus und seine Anwendung*, J. Reine Angew. Math. 227 (1967) p. 150-177.

[11] (cit. 37) *Ein neuer Algorithmus für absteigende Basispotenzen im kubischen Körper*, Math. Nachr. 33 (1967) p. 257-272.

[12] (not cit.) *The generalized Pellian equation*, Trans. Am. Math. Soc. 127 (1967) p. 76-89.

[13] (cit. 24) *The linear Diophantine equation in n variables and its application to generalized Fibonacci numbers*, Fibonacci Q. 6 (1968) nr.3 p. 1-63.

[14] (cit. 36, 38, 40, 115) *Einheitenberechnung in kubischen Körpern mittels des Jacobi-Perronschen Algorithmus aus der Rechenanlage*, J. Reine Angew. Math. 244 (1970) p. 201-220.

[15] (cit. 33, 36, 39, 115, 152) *The Jacobi-Perron algorithm — its theory and application*, Lect. Notes Math. 207 (1971), Springer, Berlin.

[16] (cit. 115) *Infinite non-solubility classes of the Delaunay-Nagell equation*, J. Lond. Math. Soc. II ser. 3 (1971) p. 118-120.

[17] (cit. 36) *A 3-dimensional periodic Jacobi-Perron algorithm of period length 8*, J. Number Theory 4 (1972) p. 48-69.

[18] (cit. 115) *On units and fundamental units*, J. Reine Angew. Math. 257 (1972) p. 129-145.

[19] (cit. 36) *Units from periodic Jacobi-Perron algorithms in algebraic number fields of degree $n > 2$*, Manuscr. Math. 14 (1974) p. 249-261.

[20] (cit. 115) *Fundamental units from the preperiod of a generalized Jacobi-Perron algorithm*, J. Reine Angew. Math. 268/269 (1974) p. 391-409.

[21] (cit. 36, 115) *Units and periodic Jacobi-Perron algorithms in real algebraic number fields of degree 3*, Trans. Am. Math. Soc. 212 (1975) p. 295-306.

[22] (not cit.) *Der Hasse-Bernsteinsche Einheitensatz für den verallgemeinerten Jacobi-Perronschen Algorithmus*, Abh. Math. Semin. Univ. Hamb. 43 (1975) p. 11-20.

[23] (not cit.) *Gaining units from units*, Can. J. Math. 29 (1977) p. 93-106.

[24] (not cit.) *An algorithm for Halter-Koch units*, Mich. Math. J. 25 (1978) p. 371-377.

BERNSTEIN, L. & HASSE, H.

[1] (cit. 152) *An explicit formula for the units of an algebraic number field of degree $n \geq 2$*, Pac. J. Math. 30 (1969) p. 293-365.

BERWICK, W.E.H.

[1] (cit. 65, 66, 115) *The classification of ideal numbers that depend on a cubic irrationality*, Proc. Lond. Math. Soc. II ser. 12 (1913) p. 393-429.

[2] (cit. 153) *Algebraic number fields with two independent units*, Proc. Lond. Math. Soc. II ser. 34 (1932) p. 360-378.

BILLEVICH, K.K.

*[1] (cit. 153) *"On units of algebraic fields of third and fourth degrees" (russian)*, Mat. Sb. 40 (82) (1956) p. 123-136.

BRUN, V.

[1] (cit. 41) *En generalisation av kjedebrøken I*, Skr. Vidensk. Selsk. Kristiania 1919 nr. 6, mat. nat. kl.

[2] (cit. 41) *En generalisation av kjedebrøken II*, Skr. Vidensk. Selsk. Kristiania 1920 nr. 6, mat. nat. kl.

*[3] (not cit.) *Music and ternary continued fractions*, Kongl. Norske Vidensk. Selsk. Forh. 23 (1950) nr. 10.

[4] (cit. 42, 43) *Algorithmes euclidiens pour trois et quatre nombres*, XIIIième Congr. Math. Scand. (1957) p. 45-64, Helsinki 1957.

[5] (cit. 43) *Mehr-dimensionale Algorithmen welche die Eulersche Kettenbruchentwicklung der Zahl e verallgemeinern*, Sammelband zu Ehren Eulers 250. Geburtstages, p. 87-100, Akad. Verlag, Berlin 1959.

[6] (cit. 44) *Musikk og Euklidske algoritmer*, Nord. Mat. Tidskr. 9 (1961) p. 29-36.

[7] (not cit.) *Euclidean algorithms and musical theory*, Enseign. Math. II sér. 10 (1964) p. 125-137.

BRUNOTTE, H. & HALTER-KOCH, F.

[1] (cit. 154) *Zur Einheitenberechnung in totalreellen kubischen Zahlkörpern nach Godwin*, J. Number Theory 11 (1979) p. 552-559.

CASSELS, J.W.S.

[1] (cit. 115) *The rational solutions of the Diophantine equation* $y^2 = x^3 - D$, Acta Math. 82 (1950) p. 244-273.

[2] (cit. 34) *An introduction to Diophantine approximation*, Cambridge Tracts in Math. and Math. Phys. 45 (1957).

CHABAUTY, C. & PISOT, Ch.

[1] (cit. 39) *Un algorithme pour l'approximation simultanée de deux nombres réels*, C. R. Acad. Sci. Paris sér. A 206 (1938), p. 1069-1071.

COHN, H. & GORN, S.

*[1] (cit. 154) *A computation of cyclic cubic units*, J. Res. Natl. Bur. Stand. 59 (1967) p. 155-168.

COLEMAN, J.B.

[1] (cit. 35) *A test for the type of irrationality represented by a periodic ternary continued fraction*, Am. J. Math. 52 (1930) p. 835-842.

[2] (cit. 35) *The Jacobi algorithm for periodic continued fractions as defining a cubic irrationality*, Am. J. Math. 55 (1933) p. 585-592.

CUSICK, T.W.

[1] (cit. 65) *Diophantine approximation of ternary linear forms I*, Math. Comput. 25 (1971) p. 163-180.

[2] (cit. 65) *Diophantine approximation of ternary linear forms II*, Math. Comput. 26 (1972) p. 977-993.

[3] (cit. 65) *Diophantine approximation of linear forms over an algebraic number field*, Mathematika 20 (1973) p. 16-23.

[4] (cit. 65) *The two-dimensional Diophantine approximation constant*, Monatsh. Math. 78 (1974) p. 297-304.

[5] (cit. 119) *The Szekeres multi-dimensional continued fraction algorithm*, Math. Comput. 31 (1977) p. 280-317.

[6] (cit. 65, 109) *Best Diophantine approximations for ternary linear forms*, J. Reine Angew. Math. 315 (1980) p. 40-52.

DAUS, P.H.

[1] (cit. 37) *Normal ternary continued fraction expansions for the*

cube roots of integers, Am. J. Math. 44 (1922) p. 279-296.

[2] (cit. 37, 115, 153) *Normal ternary continued fraction expansions for cubic irrationalities*, Am. J. Math. 51 (1929) p. 67-98.

DAVID, M.

[1] (cit. 39) *Sur un algorithme voisin de celui de Jacobi*, C. R. Acad. Sci. Paris sér. A 229 (1949) p. 965-967.

[2] (cit. 39) *Sur trois algorithmes associés à l'algorithme de Jacobi*, C. R. Acad. Sci. Paris sér. A 230 (1950) p. 1445-1446.

[3] (not cit.) *Caractérisation algorithmique des irrationnels cubiques*, C. R. Acad. Sci. Paris sér. A 232 (1951) p. 1795-1798.

[4] (cit. 33, 35, 39, 44) *Contribution à l'étude algorithmique des approximations rationnelles simultanées de deux irrationnels* Thèse, Univ. de Paris, 28-4-1955, published in Publ. Sci. Univ. Alger sér. A 3 (1956) p. 1-102.

DEDEKIND, R.

[1] (cit. 109) *Über die Anzahl der Idealklassen in reinen kubischen Zahlkörpern*, J. Reine Angew. Math. 121 (1900) p. 40-123.

DELONE, B.N. (DELAUNAY)

[1] (cit. 116) *Interprétation géométrique de la généralisation de l'algorithme des fractions continues donné par Voronoi*, C. R. Acad. Sci. Paris sér. A 176 (1923) p. 554-556.

DELONE, B.N. & FADDEEV, D.K.

[1] (cit. 107, 115, 116, 131, 148-150, 152) *The theory of irrationalities of the third degree*, Am. Math. Soc. Transl. of Math. Monographs 10 (1964).

DUBOIS, E.

[1] (not cit.) *Approximations simultanées de deux nombres réels*, Sémin. Delange-Pisot-Poitou, 11-12- 1978.

[2] (cit. 65, 116) *Meilleures approximations diophantiennes. Application à la recherche de l'unité fondamentale des corps cubiques non totalement réels*, C. R. Acad. Sci. Paris sér. A 289 (1979) p. 39-41.

[3] (cit. 116) *Best approximations of zero by a cubic linear form. Calculation of the fundamental unit of a not totally real cubic*

field, Proc. Queen's Number Theory Conf., Kingston (1979) p. 205-221.

[4] (cit. 65, 66) *Approximations diophantiennes simultanées de nombres algébriques; calcul des meilleures approximations*, Thèse, Univ. Curie, Paris, 17-3-1980.

[5] (cit. 66, 116) *Calculation of F-best approximations of zero by a ternary form. Computation of units*, submitted to Math. Comput.

[6] (cit. 65, 116) *F-best approximations of zero by a cubic linear form. Properties of approximation and periodicity*, submitted to J. Aust. Math. Soc.

[7] (cit. 118) *Contre-exemple à une conjecture de G. Szekeres*, preprint from the author.

DUBOIS, E. & PAYSANT-LE-ROUX, R.

[1] (cit. 34, 36) *Développement périodique par l'algorithme de Jacobi-Perron et nombre de Pisot-Vijayaraghavan*, C. R. Acad. Sci. Paris sér. A 272 (1971) p. 649-652.

[2] (cit. 36) *De nouvelles classes d'entiers m pour lesquelles on connaît des développements périodiques par l'algorithme de Jacobi-Perron dans* $\mathbb{Q}(\sqrt[n+1]{m})$, C. R. Acad. Sci. Paris sér. A 280 (1975) p. 57-59.

[3] (cit. 36) *Algorithme de Jacobi-Perron dans les extensions cubiques*, C. R. Acad. Sci. Paris sér. A 280 (1975) p. 183-186.

ELSNER, L. & HASSE, H.

[1] (cit. 36) *Numerische Ergebnisse zum Jacobischen Kettenbruch-algorithmus in rein-kubischen Zahlkörpern*, Math. Nachr. 34 (1967) p. 95-97.

EULER, L.

[1] (cit. 24, 32) *De relatione inter ternas pluresve quantitates instituenda*, Commentationes Arithmeticae Collectae, Petersburg 1849, p. 99 sqq.

FERGUSON, H.R.P. & FORCADE, R.W.

[1] (cit. 42, 123, 127) *Generalization of the Euclidean algorithm for real numbers to all dimensions higher than two*, preprint from the authors; see Bull. Am. Math. Soc. new ser. 1 (1979) p. 912-914.

FISCHER, R.

[1] (cit. 33) *Konvergenzgeschwindigkeit beim Jacobi-Algorithmus*, Anz. Österr. Akad. Wiss. Math. Naturwiss. Kl. (1972) p. 156-158.

FISCHER, R. & SCHWEIGER, F.

[1] (not cit.) *The number of steps in a finite Jacobi algorithm*, Manuscr. Math. 17 (1975) p. 291-308.

FOKKER, A.D.

[1] (cit. 42) *Multiple antanâiresis*, Indagationes Math. 25 (1963) p. 1-6.

FROBENIUS, F.G.

[1] (cit. 60) *Über Matrizen aus positiven Elementen*, Sitz. Ber. Kgl. Preuss. Akad. Wiss. Berlin (1908) p. 471-476, and Gesamm. Abh. III p. 404-409.

FÜRSTENAU, E.

*[1] (not cit.) *Über Kettenbrüche höherer Ordnung*, Jahresber. Kgl. Realgymn. Wiesbaden 1874; see Jahrb. Fortschr. Math. 7 (1874).

FURTWÄNGLER, Ph.

[1] (cit. 65, 66) *Über die simultane Approximation von Irrationalzahlen II*, Math. Ann. 99 (1927) p. 71-83.

GODWIN, H.J.

[1] (cit. 154) *The determination of units in totally real cubic fields*, Proc. Camb. Phil. Soc. 56 (1960) p. 318-321.

GODWIN, H.J. & SEMET, P.A.

[1] (not cit.) *A table of real cubic fields*, J. Lond. Math. Soc. 34 (1959) p. 108-110.

GRAS-MONTOUCHET, M.N.

[1] (cit. 154) *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) p. 89-116.

GREITER, G.

[1] (cit. 36, 43) *Mehrdimensionale Kettenbrüche*, Dissertation, Techn. Univ. München, 24-8-1977.

GÜTING, R.

[1] (cit. 38) *Zur Verallgemeinerung des Kettenbruchalgorithmus I*, J. Reine Angew. Math. 278/279 (1975) p. 165-173.

[2] (cit. 39, 115) *Zur Verallgemeinerung des Kettenbruchalgorithmus II*, J. Reine Angew. Math. 281 (1976) p. 184-198.

[3] (cit. 38) *Zur Verallgemeinerung des Kettenbruchalgorithmus III*, J. Reine Angew. Math. 283/284 (1976) p. 384-387.

HASSE, H.

[1] (cit. 154) *Arithmetische Bestimmung von Grundeinheiten und Klassenzahl in zyklischen kubischen und biquadratischen Zahlkörpern*, Abh. Deutsche Akad. Wiss. Berlin 2 (1948) p. 1-95.

HENDY, M.D. & JEANS, N.S,

[1] (cit. 164) *The Jacobi-Perron algorithm in integer form*, Math. Comput. 36 (1981) p. 565-574.

HERMITE, Ch.

[1] (cit. 37) *Correspondance d'Hermite et de Stieltjes II*, lettres 232, 238, 408, Gauthier-Villars, Paris 1905.

HUA KOO LENG & WANG YUAN

[1] (cit. 36) *A note on simultaneous Diophantine approximations to algebraic integers*, Sci. Sin. 20 nr. 5 (1977) p. 563-567.

JACOBI, C.G.J.

[1] (cit. 24, 32) *Über die Auflösing der Gleichung* $\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n = fu$, J. Reine Angew. Math. 69 (1868) p. 1-28.

[2] (cit. 30, 35) *Allgemeine Theorie der kettenbruchähnlichen Algorithmen, in welchen jede Zahl aus drei vorhergehenden gebildet wird*, J. Reine Angew. Math. 69 (1868) p. 29-64.

JEANS, N.S. & HENDY, M.D.

[1] (cit. 111) *Some inequalities related to the determination of the fundamental unit of a pure cubic field*, Math. Comput. 32 (1978) p. 925-935.

JURKAT, W.; KRATZ, W. & PEYERIMHOFF, A.

[1] (cit. 34, 65, 102, 119) *On best two-dimensional Dirichlet approximations and their algorithmic calculation*, Math. Ann. 244 (1979) p. 1-32.

JURKAT, W. & PEYERIMHOFF, A.

[1] (cit. 62) *A constructive approach to Kronecker approximations and*

its application to the Mertens conjecture, J. Reine Angew. Math. 286/287 (1976) p. 322-340.

KHINTCHINE, A. Ya.

[1] (cit. 34) *Über eine Klasse linearer Diophantischer Approximationen*, Rend. Circ. Mat. Palermo 50 (1926) p. 170-195.

[2] (cit. 9, 27) *Continued fractions*, Noordhoff, Groningen 1963, or Univ. of Chicago Press 1964.

KLEIN, F.

[1] (cit. 12) *Ausgewählte Kapitel der Zahlentheorie*, Teubner, Verlag, Leipzig 1907.

KOKSMA, J.F.

[1] (cit. 30, 34) *Diophantische Approximationen*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 4 (1936) p. 407-571.

KOLLROS, L.

*[1] (cit. 50) *Un algorithme pour l'approximation simultanée de deux grandeurs*, Dissertation, Zürich 1905.

[2] (cit. 50) *Sur l'approximation périodique des irrationnelles cubiques*, Bull. Sci. Math. sér. II 29 (1905) p. 21-24; see also Jahrb. Fortschr. Math. 36 (1905) p. 294.

LAGARIAS, J.C.

[1] (cit. 56, 65, 66) *Some new results in simultaneous Diophantine approximation*, Proc. of the Queen's Number Theory Conference 1979, Queen's Papers in Pure and Applied Math. 54 (1980) p. 453-474, Kingston 1980.

[2] (cit. 56) *Best simultaneous Diophantine approximation I. Growth rates of best approximation denominators*, Trans. Am. Math. Soc., to appear.

[3] (cit. 56, 66) *Best simultaneous Diophantine approximation II. Behavior of consecutive best approximations*, Pac. J. Math., to appear.

LEHMER, D.H.

[1] (cit. 24, 38) *A note on the linear Diophantine equation*, Am. Math. Mon. 48 (1941) p. 240-246.

LEHMER, D.N.

[1] (cit. 36) *On Jacobi's extension of the continued fraction algorithm*, Proc. Natl. Acad. Sci. U.S.A. 4 (1918) p. 360-364.

[2] (cit. 24) *The general solution of the indeterminate equation* $Ax + By + Cz + \ldots\ldots = r$, Proc. Natl. Acad. Sci. U.S.A. 5 (1919) p. 111-114.

[3] (cit. 36) *Inverse ternary continued fractions*, Bull. Am. Math. Soc. 37 (1931) p. 565-569.

LEVESQUE, C.

*[1] (cit. 36) *Jacobi-Perron algorithms and units in some algebraic number fields*, Thesis, Chicago 1976.

[2] (cit. 36) *A class of periodic Jacobi-Perron algorithms in pure algebraic number fields of degree* $n \geq 3$, Manuscr. Math. 22 (1977) p. 235-269.

[3] (cit. 36) *A class of fundamental units and some classes of Jacobi-Perron algorithms in pure cubic fields*, Pac. J. Math. 81 (1979) p. 447-466.

MACK, J.M.

[1] (cit. 56) *A note on simultaneous Diophantine approximation*, Bull. Aust. Math. Soc. 3 (1970) p. 81-83.

MARKOV, A.

*[1] (cit. 115) *Sur les nombres entiers dépendants d'une racine cubique d'un nombre entier ordinaire*, Mem. Acad. Imp. Sci. Petersburg (7) 38 (1892) nr. 9 p. 1-37.

MAUNSELL, F.G.

[1] (cit. 33) *Some notes on extended continued fractions*, Proc. Lond. Math. Soc. II ser. 30 (1930) p. 127-132.

MINKOWSKI, H.

[1] (cit. 66, 153) *Zur Theorie der Kettenbrüche*, Gesamm. Abh. I p. 278-292; appeared as *Généralisation de la théorie des fractions continues* in Ann. Sci. Ec. Norm. Supér. sér. III 13 (1896) p. 41-60.

[2] (cit. 61) *Über periodische Approximationen algebraischer Zahlen*, Acta Math. 26 (1902) p. 333-351, or Gesamm. Abh. I p. 357-371.

[3] (cit. 50) *Zur Geometrie der Zahlen*, Verh. III Int. Math. Kongr.

Heidelberg (1904) p. 164-173, or Gesamm. Abh. II p. 43-52.

NAGELL, T.

[1] (cit. 115, 116) *Solution complète de quelques équations cubiques à deux indéterminées*, J. Math. Pures Appl. sér. IX 4 (1925) p. 209-270.

PALEY, R.E.A.C. & URSELL, H.D.

[1] (cit. 34, 44) *Continued fractions in several dimensions*, Proc. Camb. Phil. Soc. 26 (1930) p. 127-144.

PERRON, O.

[1] (cit. 29, 31-36, 38, 60) *Grundlagen für eine Theorie des Jacobischen Kettenalgorithmus*, Math. Ann. 64 (1907) p. 1-76.

[2] (cit. 9) *Die Lehre von den Kettenbrüchen*, Teubner Verlag, Leipzig 1913, Stuttgart 1954.

[3] (cit. 35) *Ein Satz über Jacobi-Ketten zweiter Ordnung*, Ann. R. Sc. Norm. Super. Pisa, fis. mat., (2) 4 (1935) p. 133-138.

PIPPING, N.

[1] (cit. 50) *Ein Kriterium für die reellen algebraischen Zahlen auf eine direkte Verallgemeinerung des Euklidischen Algorithmus gegründet*, Acta Acad. Åbo ser. B 1 (1922) nr. 1 p. 1-16.

[2] (cit. 50) *Über eine Verallgemeinerung des Euklidischen Algorithmus*, Acta Acad. Åbo ser B 1 (1922) nr. 7 p. 1-14.

[3] (cit. 50) *Un critérium pour les nombres algébriques réels, fondé sur une généralisation directe de l'algorithme d'Euclide*, C. R. Acad. Sci. Paris sér. A (1923) p. 1155-1156.

[4] (not cit.) *Arithmetische Kriterien für reelle algebraische Zahlen*, Acta Acad. Åbo ser. B 7 (1933) nr. 4.

[5] (not cit.) *Ein neues Kriterium für die reellen kubischen Irrationalzahlen*, Acta Acad. Åbo ser. B 8 (1935) nr. 6 p. 1-20.

[6] (cit. 50) *Ein neues Kriterium für die algebraischen Zahlen n-ten Grades*, Acta Acad. Åbo ser. B 8 (1935) nr. 7 p. 1-14.

[7] (cit. 39, 50) *Verallgemeinerungen des Euklidischen Algorithmus*, Acta Acad. Åbo ser. B 18 (1952) nr. 5 p. 1-17.

[8] (cit. 39, 44, 50) *Approximation zweier reellen Zahlen durch rationale Zahlen mit gemeinsamen Nenner*, Acta Acad. Åbo ser. B 21 (1957) nr. 1 p. 1-17.

POINCARE, H.

[1] (cit. 40) *Sur une généralisation des fractions continues*, C. R. Acad. Sci. Paris sér. A 99 (1884) p. 1014-1016, or Oeuvres V p. 185-187.

REID, L.W.

*[1] (cit. 115) *Tafel der Klassenanzahlen für kubische Zahlkörper*, Dissertation, Göttingen 1899.

[2] (cit. 115) *Table of class numbers for cubic fields*, Am. J. Math. 23 (1901) p. 68-84.

ROSSER, B.

[1] (cit. 24, 43) *A note on the linear Diophantine equation*, Am. Math. Mon. 48 (1941) p. 662-666.

[2] (cit. 23) *A generalization of the Euclidean algorithm to several dimensions*, Duke Math. J. 9 (1942) p. 59-92.

[3] (cit. 44) *Generalized ternary continued fractions*, Am. Math. Mon. 57 (1950) p. 528-535.

RUDMAN, R.J.

[1] (cit. 115) *On the fundamental unit of a purely cubic field*, Pac. J. Math. 46 (1973) p. 253-256.

RUDMAN, R.J. & STEINER, R.P.

[1] (cit. 131, 154) *A generalization of Berwick's unit algorithm*, J. Number Theory 10 (1978) p. 16-34.

SANG, E.

[1] (cit. 38, 44) *On the extension of Brouncker's method to the comparison of several magnitudes*, Trans. R. Soc. Edinb. 26 (I) (1869) p. 59-67; see also Jahrb. Fortschr. Math. 2 (1869) p. 107.

SCHMIDT, W.M.

[1] (cit. 34) *Flächenapproximation beim Jacobialgorithmus*, Math. Ann. 136 (1958) p. 365-374.

SCHWEIGER, F.

[1] (cit. 35) *Geometrische und elementare metrische Sätze über den Jacobischen Algorithmus*, Sitzungsber. Abt. II d. österr. Akad. Wiss., Math-Naturwiss. Kl. 173 (1964), 59-92.

[2] (cit. 35) *Metrische Sätze über den Jacobischen Algorithmus*, Monatsh. Math. 69 (1965) p. 243-255.

[3] (cit. 35) *Ergodische Theorie des Jacobischen Algorithmus*, Acta Arithm. 11 (1966) p. 451-460.

[4] (cit. 35) *Existenz eines invarianten Maßes beim Jacobischen Algorithmus*, Acta Arithm. 12 (1967) p. 163-168.

[5] (cit. 35) *Induzierte Maße und Jacobischer Algorithmus*, Acta Arithm. 13 (1968) p. 419-422.

[6] (cit. 35) *Mischungseigenschaften und Entropie beim Jacobischen Algorithmus*, J. Reine Angew. Math. 229 (1968) p. 50-56.

[7] (cit. 35) *Ein Kuzminscher Satz über den Jacobischen Algorithmus*, J. Reine Angew. Math. 232 (1968) p. 35-40.

[8] (cit. 35) *The metrical theory of Jacobi-Perron algorithm*, Lect. Notes Math. 334 (1973), Springer Verlag, Berlin.

[9] (cit. 34) *Volumsapproximation beim Jacobialgorithmus*, Math. Ann. 203 (1973) p. 283-288.

[10] (cit. 34) *Volumsapproximation beim Jacobialgorithmus II*, Acta Arithm. 23 (1973) p. 393-400.

[11] (cit. 34) *Der Satz von Hurwitz beim Jacobialgorithmus*, Monatsh. Math. 80 (1975) p. 215-218.

[12] (cit. 34) *Approximation properties of Jacobi's algorithm*, Monatsh. Math. 81 (1976) p. 59-61.

[13] (cit. 38) *Über einen Algorithmus von R. Güting*, J. Reine Angew. Math. 293/294 (1977) p. 263-270.

SELMER, E.S.

[1] (cit. 115) *Tables for the purely cubic field* $K(\sqrt[3]{m})$, Avh. Norske Vid. Akad. Oslo I (1955) nr. 5.

[2] (cit. 26, 42, 49) *Om flerdimensjonal kjedebrøk*, Nord. Mat. Tidskr. 9 (1961) p. 37-43.

STEINER, R.P.

[1] (cit. 131, 154) *On the units in algebraic number fields*, Proc. 6th Manitoba Conf. Numer. Math. (1976) p. 413-435.

STEINER, R.P. & RUDMAN, R.J.

[1] (cit. 153) *On an algorithm of Billevich for finding units in algebraic number fields*, Math. Comput. 30 (1976) p. 598-609.

STENDER, H.J.

[1] (cit. 115) *Über die Grundeinheit für spezielle unendliche Klassen reiner kubischer Zahlkörper*, Abh. Math. Semin. Univ. Hamb. 33 (1969) p. 203-215.

[2] (cit. 152) *Einheiten für eine allgemeine Klasse totalreeller algebraischer Zahlkörper*, J. Reine Angew. Math. 257 (1972) p. 151-178.

SVED, M.

[1] (cit. 111, 115) *Units in pure cubic number fields*, Ann. Univ. Sci. Budap. Rolando Eötvös sect. math. 13 (1970) p. 141-149.

SZEKERES, G.

[1] (cit. 65, 117, 119) *Multidimensional continued fractions*, Ann. Univ. Sci. Budap. Rolando Eötvös sect. math. 13 (1970) p. 113-140.

TE RIELE, H.J.J.

[1] (cit. 127) *Tables of the first 15000 zeros of the Riemann zeta function to 28 significant digits, and related quantities*, Report NW 67/79, Mathematisch Centrum, Amsterdam 1979.

TÖRNQVIST, L.

[1] (cit. 39, 50) *Kriterien für die reellen algebraischen Zahlen, arithmetische Ketten und diophantische Approximationen*, Acta Acad. Åbo ser. B 10 (1936) nr. 7 p. 1-95.

USPENSKY, J.V.

[1] (cit. 115) *A method for finding units in cubic orders of a negative discriminant*, Trans. Am. Math. Soc. 33 (1931) p. 1-22.

VAN DE LUNE, J. & TE RIELE, H.J.J.

[1] (cit. 34, 62, 119) *Explicit computation of special zeros of partial sums of Riemann's zeta function*, Report NW 44/77, Mathematisch Centrum, Amsterdam 1977; also in Studieweek Getaltheorie en Computers, ibid. 1980.

VAUGHAN, T.P.

[1] (cit. 33) *A note on the Jacobi-Perron algorithm*, Pac. J. Math. 72 (1977) p. 261-271.

[2] (cit. 43) *A generalization of the simple continued fraction*

algorithm, Math. Comput. 32 (1978) p. 537-558.

VORONOI, G.F.

*[1] (cit. 65, 66, 115, 131, 148) *"On a generalization of the continued fraction algorithm" (russian)*, Dissertation, Warszawa 1896; see also Jahrb. Fortschr. Math. 27 (1896) p. 170-174.

WADA, H.

[1] (cit. 115) *A table of fundamental units of pure cubic fields*, Proc. Jap. Acad. 46 (1970) p. 1135-1140, and review in Math. Comput. 26 (1972) p. 302-303.

WILLIAMS, H.C.

[1] (cit. 115) *Some results on fundamental units in cubic fields*, J. Reine Angew. Math. 286/287 (1976) p. 75-85.

[2] (cit. 111, 116) *Improving the speed of calculating the regulator of certain pure cubic fields*, Math. Comput. 35 (1980) p. 1423-1434.

[3] (not cit.) *Some results concerning Voronoi's continued fraction over* $\mathbb{Q}(\sqrt[3]{D})$, Math. Comput. 36 (1981) p. 631-652.

WILLIAMS, H.C.; CORMACK, G. & SEAH, E.

[1] (cit. 112, 116, 16) *Calculation of the regulator of a pure cubic field*, Math. Comput. 34 (1980) p. 567-611.

WILLIAMS, H.C. & HOLTE, R.

[1] (cit. 116) *Computation of the solution of* $x^3 + Dy^3 = 1$, Math. Comput. 31 (1977) p. 778-785.

WILLIAMS, H.C. & ZARNKE, C.R.

[1] (cit. 116) *Computation of the solutions of the Diophantine equation* $x^3 + Dy^3 = 1$, Proc. 1st Manitoba Conf. Numer. Math. (1971) p. 671-676.

[2] (cit. 116, 131, 152) *Computer calculation of units in cubic fields*, Proc. 2nd Manitoba Conf. Numer. Math. (1972) p. 433-468.

*[3] (cit. 116, 152) *A table of fundamental units for cubic fields*, Scient. Report 63, Univ. of Manitoba, Winnipeg 1973; review in Math. Comput. 29 (1975) p. 339.

WOLFE, C.

*[1] (cit. 115) *On the indeterminate cubic equation* $x^3 + Dy^3 + D^2z^3 - 3Dxyz = 1$, Calif. Univ. Publ. 1 (1923) nr. 16 p. 359-369; see also Jahrb. Fortschr. Math. 49 (1923) p. 97.

REGISTER

Acceleration, 26, 90, 120, 147
Algebraic units, 29
Baker, method of, 62
Berwick, method of, 153
Best approximation, 1-dimensional, 12
, general, 54
Best approximations, computation of, 69 ff.
, periodicity of, 57 ff.
, sequence of, 56
Branching algorithm, 50
Brun, algorithm of, 22, 30, 40 ff.
Chain of relative minima, 131
, two-sided, 149
Characteristic polynomial, 28
Cofactor, 13, 16
Computer implementation, 155 ff.
Continued fraction algorithm, ordinary, 9 ff.
, multi-dimensional, 19 ff.
Continued fraction expansion, ordinary, 9 ff.
, multi-dimensional, 18 ff.
Convergence, strong, 21, 22
, weak, 21, 22
Convergent, 11
Cubic field, cyclic, 154
, of negative discriminant, 105 ff.
, of positive discriminant, 148 ff.
Daus, algorithm of, 37, 115, 153
Dependence, 17
Dependence rank, 17
Dice, 20
Division algorithm, 20, 26
Empty step, 20
Ferguson and Forcade, algorithm of, 123 ff.

Fundamental unit(s) of cubic fields, 105 ff., 148 ff.
, tables of, 115 ff., 152 ff.
Godwin, method of, 154
Good approximations, 62, 117
Greatest angle algorithm, 94 ff.
Güting, algorithm of, 38 ff., 115
Height function, 21, 54
Independence, 17
Inhomogeneous approximations, 93, 122
Inner product algorithm, multi-dimensional, 119 ff.
, 2-dimensional, 99 ff.
Intermediate convergent, 12
Irreducibility of polynomials, 53
Irreducible lattice, 130
Iteration, 126
Jacobi-Perron algorithm, 29, 30 ff., 115, 119, 152
, convergence of, 33 ff.
, ergodic properties of, 35
, modified, 39
, periodicity of, 35 ff.
Jurkat, Kratz and Peyerimhoff, algorithm of, 102 ff.
Kollros, algorithm of, 50
Lattice, 16
Lattice base, 16
, of the first kind, 71
, of the second kind, 72
, positive, 132
, regular, 132
, regulated, 102
Line, 16
Linear Diophantine equation, 11, 22 ff.
Minkowski, lattice point theorem of, 54
Multi-precision arithmetic, 112, 155 ff.
Partial denominator, 9
Pell's equation, 15
Periodic continued fraction, 14
Periodic expansion, 27

Pipping, algorithm of, 50
Poincaré, algorithm of, 40
Predecessor, 56
P.V.-number, 34, 111
Quadratic field, 15
Quadratic irrationality, 14
Quality, 62
Radius function, 54
Rational line, 17
Relative minimum, 130 ff.
Sang, algorithm of, 38
Selmer, algorithm of, 26, 49
Simultaneous Diophantine approximations, 63 ff.
Splitting algorithm, 123
Step, 19
Subtractive algorithm, 20, 26
Successor, 56, 130
Szekeres, algorithm of, 24, 111, 115, 117 ff.
Termination, 73, 94, 123
Units, 29
Vectorial algorithm, 25 ff.
Voronoi, method of, 150
Zeta-function, 62, 127
$\mathbb{Z}$-relation, 52