THE INHOMOGENEOUS MINIMA
OF INDEFINITE BINARY QUADRATIC FORMS

by

JANE PITMAN

A Thesis submitted for the Degree of Doctor of Philosophy of the University of Sydney, August, 1957.

ERRATA.

Page	Line	For:	Read:
iiii	3	Chapter 1	Chapter 2
2	-3	$=$	$\neq$
12	5	$\Delta_2/\sqrt{(221)}$	$5\Delta_2/\sqrt{(221)}$
	6	$\Delta_3/\sqrt{(1517)}$	$13\Delta_3/\sqrt{(1517)}$
	7	$\Delta_4/\sqrt{(2600)}$	$17\Delta_4/\sqrt{(2600)}$
21	-5	(D_0-A_0)	(B_0-C_0)
	-4	(D_0-A_0)	(B_0-C_0)
22	-1	(D_n-A_n)	(B_n-C_n)
44	-5	given in	given as in
46	-2	$M(f_n;x_r,y_r)$	$M(f_r;x_r,y_r)$
54	7	Δ	$\pm\Delta$
	-2	Δ	$\pm\Delta$
57	-3	Lemma 3.2	Lemma 3.3
59	5	one root whose	both roots with
	6	modulus is	moduli less than or equal to
	7	$\frac{3+\sqrt5}{2}$ $(=r_2)$,	$[3,-r_2]=\frac{9-\sqrt5}{2}<\frac{5+\sqrt5}{2}$;
59	-3	$c>0$	$c<0$
61	-5	so that R_2	so that r_2
90	-5	an a-chain	a permissible a-chain
91	12	$3k+\ell+1=n$	$3(k+1)+\ell+1=n$
92	2	$y=[-3_{n-2}, S]$	$y=[-3_n, S]$

Page	Line	For:	Read:
96	3	$\epsilon_{-1}=1$	$\epsilon_{-1}=-1$
103	8(ii)	$2,-2,3,2,-2,-3,\underline{-3},$	$2,-2,\underline{3},2,-2,-3,-3,$
	9	$0,0,1,0,-1,-1,\pm\underline{1},$	$0,0,\underline{1},0,0,-1,\pm 1,$
108	3	$m \geq p \geq 2$	$p \geq 2,\ m \geq 0$
113	2	$3,2,-2,-2,3,2,-2,-3,$	$3,2,-2,-3,3,2,-2,-3,$
126	11(iii)	$\underline{3},3,2,-2,-3,$	$3,\underline{3},2,-2,-3,$
	12	$\underline{1},1,0,0,\pm 1,$	$1,\underline{1},0,0,\pm 1,$
136	-2	x^2-73y^2	$x^2+xy-18y^2$

THE INHOMOGENEOUS MINIMA

OF INDEFINITE BINARY QUADRATIC FORMS

Summary of a Thesis submitted by Jane Pitman for the Degree of Doctor of Philosophy of the University of Sydney, August, 1957.

Let

$$f(x,y) = ax^2 + bxy + cy^2 \quad (D = b^2 - 4ac > 0)$$

be an indefinite binary quadratic form which does not represent zero, and let $m(f)$, $M(f)$, $M_2(f)$ be its homogeneous minimum and its first and second inhomogeneous minima.

Chapter 1. An introduction to problems on the inhomogeneous minima, some of which are investigated in this thesis.

Chapter 2. An account of the 'divided cell method' for evaluating $M(f)$, which was derived by Barnes and Swinnerton-Dyer [1] and extended by Barnes [2]. The later chapters of the thesis all depend on this method.

Chapter 3. The 'divided cell method' depends on the correspondence between chains of divided cells (cells with one vertex in each quadrant) of an inhomogeneous lattice and chains of 'I-reduced' forms. Chapter 3 is a discussion of the question of whether it is possible to obtain all the chains of I-reduced forms equivalent to f by taking all the chains which contain one particular form — in fact it is possible to obtain all the chains by starting from at most

three forms, and if f is rational it is possible to obtain all the I-reduced forms equivalent to f (though not all chains) by taking all the forms in chains from just one form.

Chapter 4. Let g_n be the symmetric Markov form

$$g_n(x,y) = u_{2n+3}x^2 + v_{2n+3}xy - u_{2n+3}y^2 \ (n \geq 1),$$

where u_i, v_i, $i = 0,1, \ldots$ denote the Fibonacci and Lucas numbers, then for $n \geq 11$ we have

(i) if $n \equiv 0 \pmod 3$, then

$$M(g_n) = \tfrac{1}{4}u_{2n+3} = \tfrac{1}{4}m(g_n)$$

(ii) if $n \equiv 0 \pmod 3$, then

$$M(g_n) = \tfrac{1}{4}(8u_{2n+3} - 3v_{2n+3}) > \tfrac{1}{4}m(g_n),$$

$$M_2(g_n) = \tfrac{1}{4}u_{2n+3}.$$

Chapter 5. It is shown that results similar to those of Chapter 4 hold for the early symmetric Markov forms; the inhomogeneous minimum of the form

$$g(x,y) = x^2 + \sqrt{5}xy - y^2$$

is obtained.

Chapter 6. Let K denote Davenport's constant:

$$K = \sup \ k; \ M(f) > k\Delta \ .$$

The results of Chapter 4 and the methods of Chapter 2 are used to show that

$$\frac{1}{39} \leq K \leq \frac{1}{12} .$$

REFERENCES

[1] BARNES, E.S. and SWINNERTON-DYER, H.P.F. The inhomogeneous minima of binary quadratic forms. III. Acta Math. 92 (1954), 199-234.

[2] BARNES, E.S. The inhomogeneous minima of binary quadratic forms. IV. Acta Math. 92 (1954), 235-264.

CONTENTS

Chapter 1 is introductory. Chapter 2 is expository, and, apart from a certain amount of reorganization, the only original work in Chapter 1 is Theorem 2.2, the proof of Theorem 2.1 and 2.10, and Lemma 2.10. Sections 3.2 and 3.3, the whole of Chapter 4, the proof of Theorem 5.3, and sections 5.3, 5.4, 5.5, 6.3, 6.4, 6.5 are original. Whenever known results are stated, the appropriate references are given. Acknowledgements are made at the end of the Introduction (Chapter 1).

CHAPTER 1

INTRODUCTION

1.1. Definitions and Minkowski's Theorem

Let

$$f(x,y) = ax^2 + bxy + cy^2 \tag{1.1}$$

be an indefinite binary quadratic form with real coefficients and discriminant $D = b^2 - 4ac > 0$, and write $\Delta = +\sqrt{D}$. If there exist integers x,y, not both zero, such that $f(x,y) = K$, the form f is said to <u>represent</u> the number K; this thesis is mainly concerned with forms which do not represent zero (i.e. such that $f(x,y) \neq 0$ for all integer pairs $(x,y) \neq (0,0)$).

The <u>homogeneous minimum</u>, $m(f)$, of the form f is defined by

$$m(f) = \inf\left[|f(x,y)|;\ x,y \text{ integral},\ (x,y) \neq (0,0)\right]. \tag{1.2}$$

If $P = (x_0,y_0)$ is any real point, we define

$$M(f;P) = M(f;\ x_0,y_0) = \inf\left[|f(x+x_0,y+y_0)|;x,y \text{ integral}\right]. \tag{1.3}$$

We now define the <u>inhomogeneous minimum</u>, $M(f)$, of the form f by

$$M(f) = \sup_P M(f;P), \tag{1.4}$$

where the supremum is taken over all real points P. Clearly, if $P' = (x_0',y_0')$ and $P' \equiv P \pmod 1$ (i.e. if $x_0' \equiv x_0, y_0' \equiv y_0$

(mod 1)), then $M(f;P') = M(f;P)$. Hence in (1.4) it is sufficient to take the supremum over any complete set of incongruent points (mod 1) (e.g. the set $|x| \leq \frac{1}{2}$, $|y| \leq \frac{1}{2}$).

Since

$$f(x+\tfrac{1}{2},\ y+\tfrac{1}{2}) = \frac{1}{4}f(2x+1,\ 2y+1),$$

it follows from (1.2) and (1.3) that

$$M(f;\ \tfrac{1}{2},\tfrac{1}{2}) \geq \frac{1}{4}m(f).$$

Hence

$$M(f) \geq \frac{1}{4}m(f). \tag{1.5}$$

It follows from (1.3) and (1.4) that corresponding to any point $P = (x_o,y_o)$ and any given $\varepsilon > 0$, there exists an integer pair (x,y) such that

$$|f(x+x_o,\ y+y_o)| < M(f) + \varepsilon.$$

If in fact corresponding to any point $P = (x_o,y_o)$ there exists an integer pair (x,y) such that

$$|f(x+x_o,\ y+y_o)| \leq M(f),$$

then we shall say that $M(f)$ is an attained minimum.

We define

$$M_2(f) = \sup\left[M(f;P);\ P\in\ C\right],$$

where

$$C = \left[P;\ M(f;P) = M(f)\right].$$

Obviously

$$M_2(f) \leq M(f). \tag{1.6}$$

If strict inequality holds in (1.6), we say that $M(f)$ is an <u>isolated minimum</u> and call $M_2(f)$ a <u>second minimum.</u> Similarly we define a sequence $M_3(f)$, $M_4(f)$, ... , of successive minima which is strictly decreasing until a non-isolated minimum is reached.

We shall say that two indefinite binary quadratic forms $f(x,y)$, $F(x,y)$ are <u>equivalent</u> if there exists an integral unimodular transformation

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} t & u \\ v & w \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}, \tag{1.7}$$

where t, u, v, w are integral, and $tw - uv = \pm 1$, such that

$$f(x,y) = F(X,Y).$$

If further

$$\begin{bmatrix} x_o \\ y_o \end{bmatrix} = \begin{bmatrix} t & u \\ v & w \end{bmatrix} \begin{bmatrix} X_o \\ Y_o \end{bmatrix},$$

then

$$f(x+x_o,\ y+y_o) = F(X+X_o,\ Y+Y_o) \tag{1.8}$$

for all integral x,y, X,Y, and we shall sometimes say that $f(x+x_o,\ y+y_o)$ is equivalent to $F(x+X_o,\ y+Y_o)$, or that f at (x_o,y_o) is equivalent to F at (X_o,Y_o). It is clear from (1.8) that in this case

$$M(f;\ x_o,y_o) = M(F;\ X_o,Y_o).$$

Thus, if the forms $f(x,y)$, $F(x,y)$ are equivalent, then

$$M(f) = M(F).$$

We note that, trivially,

$$M(Kf) = |K| \; M(f)$$

for any real K.

If the form f is given by (1.1), then we may write

$$f(x,y) = (\alpha x + \beta y)(\gamma x + \delta y), \tag{1.9}$$

where

$$|\alpha\delta - \beta\gamma| = \Delta. \tag{1.10}$$

If further we write

$$\xi_o = \alpha x_o + \beta y_o, \qquad \eta_o = \gamma x_o + \delta y_o, \tag{1.11}$$

$$\xi = \alpha x + \beta y + \xi_o, \quad \eta = \gamma x + \delta y + \eta_o, \tag{1.12}$$

then

$$f(x+x_o,\ y+y_o) = \xi\eta, \tag{1.13}$$

so that

$$M(f;\ x_o,y_o) = \inf \left[|\xi\eta|;\ x,y \text{ integral} \right]. \tag{1.14}$$

For any real ξ_o, η_o, the pair of linear forms ξ,η given by (1.12) is called a pair of <u>inhomogeneous linear forms</u> of determinant Δ, where Δ is given by (1.10). If when x,y undergo the integral unimodular transformation (1.7), we have

$$\xi(x,y) = \alpha' X + \beta' Y + \xi_o = \xi'(X,Y),$$

$$\eta(x,y) = \gamma' X + \delta' Y + \eta_o = \eta'(X,Y),$$

$$|\alpha'\delta' - \beta'\gamma'| = \Delta,$$

then we say that the pair of forms ξ',η', is <u>equivalent</u> to the pair of forms ξ,η. The classical result on pairs of inhomogeneous linear forms was found by Minkowski [42], and may be expressed in several different ways:

Theorem 1.1 (Minkowski [42], § 4). (i) If ξ,η are any two inhomogeneous linear forms with determinant Δ, then there exists an integer pair x,y such that

$$|\xi\eta| \leq \frac{\Delta}{4};$$

this is true with strict inequality unless the pair ξ,η or η,ξ is equivalent to a pair of forms,

$$\theta(x + \xi_0), \quad \phi(y + \eta_0),$$

such that

$$|\theta\phi| = \Delta, \quad (\xi_0,\eta_0) \equiv (\tfrac{1}{2},\tfrac{1}{2}) \pmod 1.$$

It follows from (1.13) that this is equivalent to:

(ii) If f(x,y) is any indefinite binary quadratic form with discriminant $D > 0$, and if $\Delta = +\sqrt{D}$, then for any point (x_0,y_0) there exists an integer pair x,y such that

$$|f(x+x_0,\ y+y_0)| \leq \frac{\Delta}{4}; \tag{1.15}$$

this is true with strict inequality unless f(x,y) is equivalent to k xy and

$$(x_0,y_0) \equiv (\tfrac{1}{2},\tfrac{1}{2}) \pmod 1.$$

The form (i) of the theorem led to the conjecture, attributed to Minkowski, that the following result holds for all n:

If ξ_1, ξ_2, ... ξ_n are n inhomogeneous linear forms in n variable $x_1, x_2, \ldots, x_n$, and if their determinant is Δ, then there exist integers $x_1, \ldots, x_n$ such that

$$|\xi_1 \xi_2 \ldots \xi_n| \leq \frac{\Delta}{2^n} .$$

This has been proved for n = 3,4, but for no higher value of n, though a weaker result has been proved for all n (see Hardy and Wright [35], § 24.9). A very large number of proofs of Minkowski's theorem for n = 2 have been produced (see Koksma [40], pp. 18-20) in the hope of finding a method which would generalize easily to higher dimensions.

It was soon realized that for certain forms f a much stronger result than (1.15) holds, and this gave rise to the concept of the inhomogeneous minimum M(f). The form (ii) of Minkowski's theorem implies that for every point (x_o, y_o)

$$M(f;\, x_o, y_o) \leq \frac{\Delta}{4} .$$

Thus Minkowski's theorem may be expressed equivalently as:

(iii) If f(x,y) is any indefinite binary quadratic form with discriminant D > 0, and if $\Delta = +\sqrt{D}$, then

$$M(f) \leq \frac{\Delta}{4}; \qquad (1.16)$$

this is true with strict inequality unless f(x,y) is equivalent to k xy, when

$$M(f) = M(f;\, \tfrac{1}{2}, \tfrac{1}{2}) = \frac{\Delta}{4}.$$

A number of authors have "sharpened" Minkowski's theorem by replacing the "Minkowski constant" $\Delta/4$ in (1.16) by upper bounds for $M(f)$ in terms of the coefficients of f. In section 3.4, I shall obtain such a bound as an obvious consequence of Theorem 3.3, and then I shall discuss the other results of the same kind which have been given by different authors.

1.2. Davenport's Theorem

Davenport [27] has proved the following rather surprising result, which is complementary to Minkowski's theorem:

Theorem 1.2 (Davenport [27]). There exists an absolute constant k such that, if

$$f(x,y) = ax^2 + bxy + cy^2$$

is any indefinite binary quadratic form which does not represent zero and if $\Delta = +\sqrt{(b^2 - 4ac)} > 0$, then

$$M(f) > k\Delta.$$

We define Davenport's constant K by

$$K = \sup \left[k;\ M(f) > k\Delta\right], \tag{1.17}$$

where the supremum is taken over all forms f which do not represent zero.

In Chapter 6 I shall obtain bounds for the value of this constant and discuss bounds which have already been given. Theorem 1.2 actually holds also for forms which represent zero (see Cassels [15,16]), but this case will not be considered in Chapter 6.

Davenport [28,29] has proved theorems similar to Theorem 1.2 for certain factorizable ternary cubic and quaternary quartic forms, and Cassels [15] has improved the constants in these theorems.

Clearly Theorem 1.2 is also closely related to the following result, which is a particular case of a theorem of Cassels [19] (see Ths. X,XI).

There exists an absolute constant $\Gamma_{1,1}$ such that for every θ there is an α for which

$$|x| \; |\theta x - y - \alpha| \geq \Gamma_{1,1}$$

for all integral x,y such that $x \neq 0$.

Cassels [19] states that the best possible value of $\Gamma_{1,1}$ is greater than 1/45.2 and less than or equal to 1/12. Perhaps this best possible value may even be the same as that of K.

1.3. The Euclidean Algorithm in Real Quadratic Fields

A real quadratic number field $k(\sqrt{m})$, where m is a square free positive integer, is said to be Euclidean if it has a Euclidean algorithm, that is, if, corresponding to every number δ (integral or not) of the field, there is an integer ρ of the field such that

$$|\text{norm } (\rho+\delta)| < 1. \qquad (1.18)$$

It can be shown that the 'fundamental theorem of arithmetic' on the unique factorization of integers holds for the integers of any real quadratic field which is Euclidean (see Hardy and Wright [35], §§12.8, 12.9, 14.7).

The elements of a real quadratic field $k(\sqrt{m})$, where m is a square free positive integer, are of the form

$$x + \omega y, \tag{1.19}$$

where x,y are rational and

$$\omega = \sqrt{m} \qquad \text{if } m \equiv 2,\ 3 \pmod 4,$$
$$\omega = \tfrac{1}{2}(\sqrt{m} + 1) \qquad \text{if } m \equiv 1 \pmod 4.$$

The integers of $k(\sqrt{m})$ are numbers (1.19) with x,y rational integers. Then norm $(x + \omega y)$ is an indefinite binary quadratic form $f_m(x,y)$ given by

$$\left.\begin{aligned} f_m(x,y) &= x^2 - my^2 && \text{if } m \equiv 2,\ 3 \pmod 4, \\ f_m(x,y) &= x^2 + xy - \tfrac{1}{4}(m-1)y^2 && \text{if } m \equiv 1 \pmod 4 \end{aligned}\right\}. \tag{1.20}$$

If we write $\delta = x_0 + \omega y_0$, $\rho = x + \omega y$, then (1.18) is equivalent to

$$|f_m(x+x_0,\ y+y_0)| < 1$$

for some integral x,y. Thus $k(\sqrt{m})$ is Euclidean if and only if

$$M(f_m;\ P) < 1 \tag{1.21}$$

for all <u>rational</u> points P. Clearly a sufficient condition for this is

$$M(f_m) < 1. \tag{1.22}$$

For this reason a large part of the work that has been done on the inhomogeneous minimum has been aimed at proving (1.22) or at disproving (1.21) for some rational P, in order to determine whether or not the field $k(\sqrt{m})$ is Euclidean for different values of m. This has led to a special interest in the norm forms f_m given by (1.20) and much effort has been devoted to the precise evaluation of the inhomogeneous minima of these forms. Barnes and Swinnerton-Dyer [9], § 12, tabulated all the results which were known for the first and second minima of forms f_m with $m \leq 101$ up to 1952; since then Barnes and Swinnerton-Dyer [10], Barnes [6], and Godwin [34] have published further results on the minima of norm forms.

Davenport [30] proved that

$$K \geq \frac{1}{128},$$

where K is the constant defined by (1.17), and further that if the form f has rational coefficients then there exists a rational point P such that

$$M(f;\, P) > \frac{\Delta}{128} .$$

If Δ^2 is the discriminant of the norm form f_m then for m large enough Δ is greater than 128; thus the number of Euclidean real quadratic fields $k(\sqrt{m})$ is finite.

The set of Euclidean fields $k(\sqrt{m})$ has now been completely determined. Chatland and Davenport [22] and Barnes and Swinnerton-Dyer [9] settled the last few doubtful cases, thus establishing the result that $k(\sqrt{m})$ is Euclidean if and only if m is one of the numbers

2, 3, 5, 6, 7, 11, 13, 17, 19, 21, 29, 33, 37, 41, 57, 73.

1.4. The Markov Forms

Apart from the norm forms, the group of forms whose inhomogeneous minima have been studied most carefully are the Markov forms. This is due to the special importance of the homogeneous minima of these forms. Markov [41] proved the following theorem.

Theorem 1.3 (Markov [41]). There exists a sequence of forms $\{F_i\}$ $(i \geq 0)$ such that:

if f is any indefinite binary quadratic form with homogeneous minimum $m(f)$ and discriminant $D > 0$, and if $\Delta = +\sqrt{D}$, then

$$m(f) > \frac{\Delta}{3}$$

if and only if f is equivalent to a form proportional to F_i for some i.

The sequence of forms

$$F_i(x,y) = Q_i x^2 + P_i xy + R_i y^2$$

which satisfy Theorem 1.3 and the condition that

$$m(F_i) = Q_i$$

are called the Markov forms. The discriminant of the form F_i is

$$\Delta_i^2 = 9Q_i^2 - 4.$$

The sequence

$$\left\{\frac{m(F_i)}{\Delta_i}\right\} = \left\{\frac{Q_i}{\Delta_i}\right\}$$

is strictly decreasing and tends down to $\frac{1}{3}$ as $i \to \infty$. Dickson [33] (Ch.VII) and Cassels [19] (Ch.II) both prove

Theorem 1.3 and give full accounts of the Markov forms.

The first few of the Markov forms are:

$F_0(x,y) = x^2 + xy - y^2;$ $\qquad m(F_0) = 1 = \Delta_0/\sqrt{5};$

$F_1(x,y) = x^2 + 2xy - y^2;$ $\qquad m(F_1) = 1 = \Delta_1/\sqrt{8};$

$F_2(x,y) = 5x^2 + 11xy - 5y^2;$ $\qquad m(F_2) = 5 = \Delta_2/\sqrt{(221)};$

$F_3(x,y) = 13x^2 + 29xy - 13y^2;$ $\qquad m(F_3) = 13 = \Delta_3/\sqrt{(1517)};$

$F_4(x,y) = 17x^2 + 38xy - 17y^2;$ $\qquad m(F_4) = 17 = \Delta_4/\sqrt{(2600)}.$

Davenport [25] developed a method by which he obtained a denumerably infinite sequence of successive inhomogeneous minima of F_0, and Varnavides [49] used this method to obtain similar results for F_1. In particular,

$$M(F_0) = \tfrac{1}{4} = \tfrac{1}{4}m(F_0),$$

$$M(F_1) = \tfrac{1}{2} > \tfrac{1}{4}m(F_1).$$

Davenport [25] also proved that

$$M(F_2) = \tfrac{5}{4} = \tfrac{1}{4}m(F_2).$$

It was conjectured that for $i \geq 2$,

$$M(F_i) = \tfrac{1}{4}m(F_i). \qquad (1.23)$$

In Chapter 4 I shall consider a sequence $\{g_n\}$ of symmetric forms which is a subsequence of $\{F_i\}$. I shall show that, when n is great and $n \not\equiv 0 \pmod 3$,

$$M(g_n) = \tfrac{1}{4}m(g_n),$$

but, when $n \equiv 0 \pmod 3$,

$$M(g_n) > \tfrac{1}{4}m(g_n).$$

Thus the conjecture (1.23) is certainly not true for all

Markov forms F_i for which $i \geq 2$.

In Chapter 5 I shall discuss the inhomogeneous minima of the forms F_o to F_4 and of the form

$$g(x,y) = x^2 + \sqrt{5}xy - y^2,$$

which may be regarded as the 'limiting symmetric' Markov form.

1.5. A Method for Rational Forms and some General Results

In order to show that $M(f) = k$, say, it is sufficient to show that

$$\sup_{P_o} \left[M(f;P_o);\ P_o \in S \right] = k,$$

where S is the unit square $|x| \leq \frac{1}{2}$, $|y| \leq \frac{1}{2}$. The natural thing to do is to exclude from consideration a set S^* of points P_o such that

$$M(f;P_o) < k, \tag{1.24}$$

and then to examine carefully the values of $M(f;P_o)$ for the points P_o in the region $S - S^*$. The inequality (1.24) holds if and only if there exists a pair of integers x', y' such that

$$|f(x'+x_o,\ y'+y_o)| < k;$$

thus for P_o to be a point of S^*, P_o must be 'covered' by a hyperbolic region

$$|f(x'+x,\ y'+y)| < k \tag{1.25}$$

for some integer pair (x',y').

If

$$T = \begin{bmatrix} t & u \\ v & w \end{bmatrix}$$

is an integral unimodular linear transformation such that (1.7) implies

$$f(x,y) = f(X,Y),$$

then T is called an <u>automorph</u> of f. (If further $tw - uv = +1$, T is a <u>proper</u> automorph of f). If $P_0 = (x_0,y_0)$, $Q_0 = (X_0,Y_0)$, and

$$\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} t & u \\ v & w \end{bmatrix} \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix}, \tag{1.26}$$

we shall write

$$P_0 = T(Q_0).$$

Clearly, if T is an automorph of f, then (1.7) and (1.26) imply that

$$f(x+x_0,\ y+y_0) = f(X+X_0,\ Y+Y_0),$$

so that

$$M(f;T(Q_0) = M(f;Q_0). \tag{1.27}$$

If R is a region all of whose points satisfy (1.24) and the form f has an automorph T, it follows from (1.27) that if the point $T^{\pm n}(Q_0)$ lies in R for some integral n, then the point Q_0 also satisfies (1.24). Thus, by starting from the region R and considering the points $T^{\pm n}(Q_0)$ corresponding to points Q_0 in S − R, we can obtain a larger set of 'covered' points S^*.

Barnes and Swinnerton-Dyer [9] (Theorem J) gave a simple result by which it is possible to determine a 'covered' set of points R in S without any tedious examination of the hyperbolic regions (1.25). Then, starting from the fact that, if an indefinite binary quadratic form f has rational coefficients and does not represent zero, then f has an infinity of automorphs (see Dickson [32], pp. 111-115), and from the ideas of the preceding paragraphs, they developed a general method for evaluating the inhomogeneous minimum of a rational form f. They used this method to evaluate the minima of certain norm forms and then in [10] they extended it to obtain sequences of successive minima of some other norm forms. By the methods of [10] they were able to establish the general results given in Theorem 1.4, where we write $f(P) = f(x,y)$ if $P = (x,y)$.

<u>Theorem 1.4 (Barnes and Swinnerton-Dyer [10], Ths. L, M.)</u>. If $f(x,y)$ is an indefinite binary quadratic form which has rational coefficients and does not represent zero, then

(i) To any point P there corresponds a point P_1 and an integer point Q such that

$$M(f;P) = M(f;P_1) \quad \text{and} \quad M(f;P_1) = f(P_1+Q);$$

(ii) The set of values of $M(f;P)$, as P varies, is closed (so that $M(f)$ is an attained minimum).

(iii) Given any $\varepsilon > 0$ there is a rational point P such that $M(f;P) > M(f) - \varepsilon$.

Barnes and Swinnerton-Dyer [10] conjectured that if f satisfies the conditions of Theorem 1.4, then $M(f)$ is rational and isolated, and there is at least one rational point P such that $M(f;P) = M(f)$.

1.6. Inhomogeneous Lattices

We now consider the problem of the inhomogeneous minimum of a form f which satisfies (1.9) and (1.10) in terms of the ξ,η-plane, where ξ and η are determined by (1.11) and (1.12). If the set of points (ξ,η), where x and y take all integral values, has no point on either of the axes, $\xi = 0$, $\eta = 0$, then this set is called an inhomogeneous lattice corresponding to f and (x_0,y_0), and is denoted by $L = L(\xi_0,\eta_0)$. (If there is a point on one of the axes, then $M(f;x_0,y_0) = 0$, and so from the point of view of the inhomogeneous minimum of f we lose nothing by excluding this case.)

If ξ', η' are a pair of non-homogeneous linear forms equivalent to ξ, η (see sect. 1.1), then ξ', η' determine the same inhomogeneous lattice as ξ, η, and $\xi'\eta'$ determines a form equivalent to f. Also, if $\xi' = K\xi$, $\eta' = \frac{1}{K}\eta$, then the points (ξ', η') form an inhomogeneous lattice corresponding to f and (x_0,y_0), and the values of $|\xi'\eta'|$ coincide with those of $|\xi\eta|$; the inhomogeneous lattice determined by the points (ξ',η') is said to be similar to L.

It follows from (1.14) that

$$M(f;\ x_0,y_0) = \left[\inf\ |\xi\eta|;\ \ (\xi,\eta) \in L\right]. \tag{1.28}$$

A parallelogram whose vertices are lattice points of L is called a cell of the lattice if it contains no lattice points other than the vertices. The area of a parallelogram whose vertices are lattice points is not less than Δ, and it is clear that such a parallelogram is a cell if and only if its area is Δ.

If P, Q are lattice points, then the infinite straight line PQ is called a lattice line; and if the segment PQ contains no lattice points except P and Q, then the directed segment PQ is called a lattice step. Any two lattice steps which form neighbouring edges of a cell may be used with a point of the lattice to generate the whole lattice.

A cell is said to be divided if one of its vertices is in each of the four quadrants. Delauney [31] proved

Theorem 1.5 (Delauney [31]). Every inhomogeneous lattice has at least one divided cell.

Delauney's proof was given by Barnes and Swinnerton-Dyer [11], and Bambah [3] has given a different proof. Sawyer [47], assuming the existence of a divided quadrilateral with lattice points as vertices, gave a new and elegant proof of Minkowski's theorem (Theorem 1.1).

Barnes and Swinnerton-Dyer [11] have developed a powerful method for evaluating the inhomogeneous minimum of an indefinite binary quadratic form f by constructing chains of divided cells of inhomogeneous lattices corresponding to f. All the work in this thesis depends on this method.

1.7. Plan of the Thesis.

In Chapter 2 I shall give an account of the theory of the divided cell method of Barnes and Swinnerton-Dyer [11], because this is the basis of the work in the later chapters. The chains of divided cells corresponding to a form f turn out to be closely related to chains of 'I-reduced' forms equivalent to f (see sect. 2.4). These chains are different from the classical chains of 'reduced' forms in that the number of chains corresponding to f is infinite, and that in general a form has pairs of possible right and left neighbouring forms instead of unique right and left neighbouring forms. The question arises, whether we can obtain either all the chains of I-reduced forms equivalent to f, or at least chains containing all the I-reduced forms equivalent to f, by taking all the chains to which a given form belongs. In Chapter 3 I shall consider this question, which must be answered in order to use the divided cell method to find the inhomogeneous minima of forms for which the number of equivalent I-reduced forms depends on a parameter and so is unbounded (e.g. the forms g_n of Chapter 4).

In Chapter 4 I shall apply the methods of Chapter 2 and the results of Chapter 3 to find the inhomogeneous minima of a sequence of forms $\{g_n\}$ which are a susequence of the symmetric Markov forms (see sect. 1.4). In Chapter 5 I shall consider the early symmetric Markov forms and the limiting symmetric Markov form (see sect. 1.4).

In Chapter 6 I shall use the methods of Chapter 2 to obtain a lower bound for Davenport's constant K (see sect. 1.2).

Numbers in square brackets will refer to the bibliography at the end, and whenever results are not original, the appropriate references will be given.

I wish to thank my supervisor, Dr. E. S. Barnes, very much for all his help; in particular, I am grateful to him for suggesting the problems I have studied and for suggesting the application of the divided cell method of [11] to these problems. The work for this thesis was done while I held a Teaching Fellowship in the University of Sydney, and the computations for Chapters 4, 5, and 6 were done on a Brunsviga provided by the University.

CHAPTER 2

THE DIVIDED CELL METHOD

The results of the later chapters of this thesis all depend on the divided cell method for evaluating the inhomogeneous minimum of an indefinite binary quadratic form. Therefore, for the sake of completeness, I give in this chapter a full account of the theory of the method, which was devised by Barnes and Swinnerton-Dyer [11], and then extended and applied by Barnes [6]. I use the notation introduced by Barnes and Swinnerton-Dyer [11], and, apart from some rearrangements and extensions, I follow the general lines of their discussion. After the headings of theorems and lemmas I list theorems or lemmas of [11] or [6] which contain any of the same results; where proofs are given completely in [11] or [6] I include only those points which are essential to the understanding of the divided cell method.

2.1. The Chain of Divided Cells of an Inhomogeneous Lattice

Throughout this chapter f denotes the indefinite binary quadratic form

$$f(x,y) = (\alpha x + \beta y)(\gamma x + \delta y), \tag{2.1}$$

where α/β, δ/γ are irrational, $|\alpha\delta - \beta\gamma| = \Delta = \sqrt{D}$, and $D > 0$ is the discriminant of the form; we consider an inhomogeneous lattice L (see sect. 1.6) corresponding to f and to the point (x', y'). By Delauney's Theorem (Theorem 1.5), L has at least one divided cell; and since α/β, δ/γ are irrational none of the lattice lines of L can be parallel to either of the axes. The method of Barnes and Swinnerton-Dyer depends on an algorithm, based on these two facts, for constructing a doubly infinite chain of divided cells, $\{S_n\}(-\infty < n < \infty)$.

Suppose A_o, B_o, C_o, D_o are the vertices of the divided cell S_o, and are <u>either</u> in the first, fourth, third, and second quadrants respectively, <u>or</u> in the third, second, first, and fourth quadrants respectively, so that A_oD_o, B_oC_o intersect the η-axis. Then S_1, with vertices A_1, B_1, C_1, D_1, is the cell defined by taking A_1B_1 as the unique lattice step in the line $A_o D_o$ which cuts the ξ-axis, and C_1D_1 as the unique lattice step in the line B_oC_o which cuts the ξ-axis. Thus

$$A_1 = A_o + (h_o + 1)(D_o - A_o),$$

$$B_1 = A_o + h_o(D_o - A_o)$$

where h_o is the unique integer such that $\eta(A_1), \eta(B_1)$ are opposite in sign; and

$$C_1 = C_o + (k_o + 1)(D_o - A_o),$$

$$D_1 = C_o + k_o(D_o - A_o),$$

where k_o is the unique integer such that $\eta(C_1), \eta(D_1)$ are opposite in sign. It is clear that S_1 is again a divided cell, where A_1, C_1 lie one in each of the first and third quadrants, and B_1, D_1 lie one in each of the second and fourth quadrants. The unique divided cell which can thus be constructed from a given divided cell S_o will be called the <u>successor</u> of S_o.

Similarly, we can construct a unique divided cell S_{-1}, the <u>predecessor</u> of S_o, by taking the unique lattice steps in the lines A_oB_o, C_oD_o which cut the η-axis; it follows from the construction that S_o must then be the successor of S_{-1}.

Since there are no lattice lines parallel to the axes, and S_{-1}, S_1 are themselves divided cells, we can use the same constructions again to obtain the divided cells S_{-2}, S_2, and so on indefinitely. In this way, starting from S_o, we get a doubly infinite chain of divided cells $\{S_n\}$ $(-\infty < n < \infty)$ of the lattice L; if now we apply this process to any cell S_n of the chain, we shall obtain exactly the same chain because the constructions for the successor and for the predecessor of a divided cell are mutually inverse. For each n, the vertices A_n, B_n, C_n, D_n of S_n lie in the first, fourth, third, and second quadrants respectively, or in the third, second, first, and fourth quadrants respectively, and satisfy the relations

$$\left.\begin{aligned} A_{n+1} &= A_n + (h_n + 1)(D_n - A_n) \\ B_{n+1} &= A_n + h_n\,(D_n - A_n) \\ C_{n+1} &= C_n + (k_n + 1)(B_n - C_n) \\ D_{n+1} &= C_n + k_n\,(D_n - A_n) \end{aligned}\right\}, \qquad (2.2)$$

where h_n is the unique integer such that $\eta(A_{n+1})$, $\eta(B_{n+1})$ are opposite in sign, and k_n is the unique integer such that $\eta(C_{n+1})$, $\eta(D_{n+1})$ are opposite in sign. If we write

$$\underset{\sim}{V}_n = A_n - D_n = B_n - C_n = B_{n+1} - A_{n+1} = C_{n+1} - D_{n+1}, \tag{2.3}$$

we have

$$\underset{\sim}{V}_{n+1} = -(h_n + k_n)\,\underset{\sim}{V}_n - \underset{\sim}{V}_{n-1}\,. \tag{2.4}$$

The chain of integer pairs h_n, k_n and the vertices of the divided cells S_n satisfy the following lemmas.

Lemma 2.1. For all n, $h_n \neq 0$, $k_n \neq 0$ and h_n, k_n have the same sign.

Proof. $\eta(A_n)$, $\eta(D_n)$ have the same sign, so that $h_n \neq 0$; similarly $k_n \neq 0$. Also $\eta(D_n - A_n)$, $\eta(C_n - B_n)$ have the same sign, so that h_n, k_n have the same sign.

Lemma 2.2 ([11], Lemma 1). Each of the following statements is impossible:

(i) $h_n = -1$ for all $n \geq n_o$ or for all $n \leq -n_o$;

(ii) $k_n = -1$ for all $n \geq n_o$ or for all $n \leq -n_o$;

(iii) $h_{n_o+2r} = k_{n_o+2r+1} = 1$ for all $r \geq 0$ or for all $r \leq 0$.

Proof. The proof is given fully in [11].

Lemma 2.3 ([11], Lemma 2). Each of

$$|\eta(\underset{\sim}{V}_n)|,\ |\eta(A_n)|,\ |\eta(B_n)|,\ |\eta(C_n)|,\ |\eta(D_n)|,$$

$$|\xi(\underset{\sim}{V}_{-n})|,\ |\xi(A_{-n})|,\ |\xi(B_{-n})|,\ |\xi(C_{-n})|,\ |\xi(D_{-n})|$$

tends to zero as $n \longrightarrow +\infty$ and tends to infinity as $n \longrightarrow -\infty$.

Proof. In [11], it is shown in the proof of Lemma 1 that $|\underset{\sim}{V}_n|$ is strictly decreasing , and in the proof of Lemma 2 that the inequality

$$|\eta(\underset{\sim}{V}_n)| < \frac{3}{5}\,|\eta(\underset{\sim}{V}_{n-1})| \tag{2.5}$$

holds for arbitrarily large positive n (though not necessarily for all large positive n). Hence $|\eta(\underset{\sim}{V}_n)|$ tends to zero as $n \to +\infty$. Since

$$|\eta(\underset{\sim}{V}_{n-1})| = |\eta(B_n)| + |\eta(A_n)| = |\eta(C_n)| + |\eta(D_n)|,$$

it now follows that each of $|\eta(A_n)|$, $|\eta(B_n)|$, $|\eta(C_n)|$, $|\eta(D_n)|$ tends to zero as $n \to \infty$. The corresponding results for $|\xi(\underset{\sim}{V}_{-n})|$, $|\xi(A_{-n})|$, $|\xi(B_{-n})|$, $|\xi(C_{-n})|$, $|\xi(D_{-n})|$ may be proved similarly.

By a similar argument to that used in [11], it can be shown that (2.5) holds for arbitrarily large negative n, so that $|\eta(\underset{\sim}{V}_n)| \to \infty$ as $n \to \infty$.

Suppose $|\eta(A_n)|$ does not tend to infinity as $n \to -\infty$, so that there exists $K > 0$ such that $|\eta(A_n)| < K$ for arbitrarily large negative n. Since $|\xi(A_n)| \to 0$ as $n \to -\infty$, this means that, for arbitrarily large negative n, A_n lies in the square $|\xi| < K$, $|\eta| < K$. This square contains only a finite number of lattice points, for each of which $|\xi| > 0$. Thus $|\xi(A_n)|$ cannot tend to zero as $n \to -\infty$, and we have a contradiction. Hence $|\eta(A_n)| \to \infty$ as $n \to -\infty$.

The other results may be proved similarly.

<u>Theorem 2.1</u> ([11], Theorem 5). If L is the inhomogeneous lattice corresponding to f and to the point (x', y'), then there is a doubly infinite chain of divided cells S_n $(-\infty < n < \infty)$ of L whose vertices A_n, B_n, C_n, D_n satisfy (2.2), and

$$M(f;x',y') = \inf\left[|\xi\eta|;\ (\xi,\eta)\in\{A_n,B_n,C_n,D_n\},\ -\infty < n < \infty\right]. \tag{2.6}$$

<u>Theorem 2.2</u>. If a chain $\{S_n\}$ of divided cells satisfies the condition of Theorem 2.1, then it includes all the divided cells of L.

The existence of a chain $\{S_n\}$ was shown above. In [11], the proof of Theorem 5 is sketched; (2.6) is an immediate consequence of this theorem. Theorem 2.2 means that every inhomogeneous lattice corresponding to f has a unique chain of divided cells; this result is not stated explicitly or proved in [11] or [6]. It is convenient to combine the proofs of (2.6) and Theorem 2.2.

<u>Proof of Theorems 2.1 and 2.2</u>. Let $\{S_n\}$ be a chain of divided cells of L, and for each n let P_n, Q_n, R_n, T_n be the vertices of S_n which lie in the first, second, third, and fourth quadrants respectively. If P_n, P_{n+1} are distinct first quadrant vertices of successive cells, then one of the following statements holds:

(i) T_n, P_n, P_{n+1}, Q_{n+1} lie on one lattice line, while T_{n+1}, R_{n+1}, R_n, Q_n lie on the neighbouring lattice line parallel to it and on the other side of the origin O (see fig. 1);

(ii) P_n, T_n, T_{n+1}, R_{n+1} lie on one lattice line, while P_{n+1}, Q_{n+1}, Q_n, R_n lie on the neighbouring lattice line parallel to it and on the other side of O (see fig. 2).

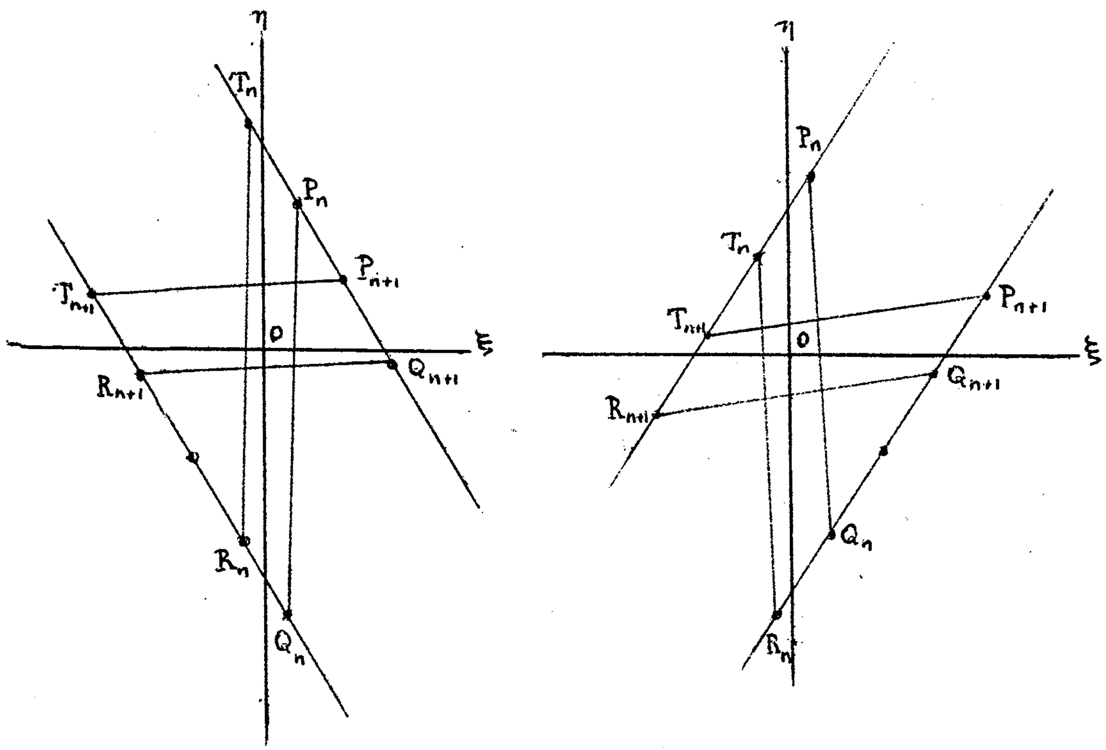

Fig. 1. Case (i) Fig. 2. Case (ii)

By Lemma 1.3 $\xi(P_n)$ tends to zero as $n \longrightarrow -\infty$ and to infinity as $n \longrightarrow +\infty$, while $\eta(P_n)$ tends to infinity as $n \longrightarrow -\infty$ and to zero as $n \longrightarrow +\infty$; and it follows from the construction that, for all n, $\xi(P_{n+1}) \geq \xi(P_n)$,

$\eta(P_{n+1}) \leq \eta(P_n)$. Hence, if P is any first quadrant lattice point, there exists an n such that P lies either on the line OP_n or in the interior of the acute angle determined by OP_n, OP_{n+1}, where $P_{n+1} \neq P_n$.

It is clear from the figures that, whether (i) or (ii) holds, there can be no lattice point in the interior of the triangle determined by the line $P_n P_{n+1}$ and the positive axes.* It now follows from the strict convexity of the region $|\xi\eta| \geq K$ that, if P is on OP_n, then P lies in the region if and only if P_n does, while, if P is in the interior of the acute angle determined by OP_n, OP_{n+1}, then P lies in the region $|\xi\eta| \geq K$ if and only if P_n, P_{n+1} do. Thus $|\xi\eta| \geq K$ for all first quadrant lattice points (ξ,η) if and only if $|\xi\eta| \geq K$ for all points P_n. By applying the same argument to the other quadrants, we have

$$\inf\left[|\xi\eta|;(\xi,\eta)\in L\right]=\left[\inf\ |\xi\eta|;(\xi,\eta)\in\{P_n,Q_n,R_n,T_n\},\ -\infty<n<\infty\right];$$

since A_n, B_n, C_n, D_n are a permutation of P_n, Q_n, R_n, T_n, this proves (2.6).

Now suppose that P, Q, R, T are the first, fourth, third, and second quadrant vertices of a divided cell such that $P \neq P_n$

---

*Note, however, that, if (ii) holds, there may be lattice points on the line P_nP_{n+1} between P_n and the η-axis, or between P_{n+1} and the ξ-axis; it is easily shown that the set of all such points must coincide with P_{n-1}, P_{n-2}, ..., P_{n-r}, for some r, or with P_{n+2}, P_{n+3}, ..., P_{n+s} for some s.

for all n. As PQRT contains no lattice points, P must lie in the interior of the acute angle determined by OP_n, OP_{n+1}, for some n for which $P_{n+1} \neq P_n$.

Suppose (i) holds. Then if either Q or T is on or above the line P_n, P_{n+1}, PQRT contains one of P_n, P_{n+1}. Hence Q, T must be on or below the line $T_{n+1}Q_n$, so that the triangle PQT contains R_n, R_{n+1}. Thus, if (i) holds, PQRT cannot be a cell.

Suppose (ii) holds, so that P cannot lie between the lines P_nT_n, Q_nR_n. If P is on or above the line P_nT_n, then the triangle OPT contains P_n, while, if P is on or below the line Q_nR_n, the triangle OPQ contains P_{n+1}. Thus, if (ii) holds, PQRT cannot be a divided cell.

Thus we get a contradiction unless $P = P_n$ for some n. By applying the same argument to the other quadrants, we see that the set of points

$$\left[\{P_n, Q_n, R_n, T_n\},\ -\infty < n < \infty\right] = \left[\{A_n, B_n, C_n, D_n\},\ -\infty < n < \infty\right]$$

includes all the vertices of divided cells of L.

Let $P = P_n \neq P_{n+1}$, $Q = Q_m \neq Q_{m+1}$ be the first and fourth quadrant vertices of a divided cell, and P', Q' be the first and fourth quadrant vertices of its successor. It follows from the construction that, for any n, if $P_{n+1} \neq P_n$, then $\xi(P_{n+1}) > \xi(P_n)$, $\eta(P_{n+1}) < \eta(P_n)$; similarly, if $P' \neq P$, $\xi(P') > \xi(P)$, $\eta(P') < \eta(P)$. Suppose $P' \neq P$. Then, since P' must coincide with P_r for some r, P' must be one of P_{n+1}, P_{n+2}, ... There can be no lattice point inside the triangle determined by PP' and the positive axes, but this triangle contains P_{n+1} if $P' \neq P_{n+1}$; hence $P' = P_{n+1}$. It is clear from Figs. 1 and 2 that there can be only one pair of successive divided cells with P_n, P_{n+1} as first quadrant vertices, so that $P = P_n$, $P' = P_{n+1}$ implies $Q = Q_n$. If $P' = P$, then $Q' \neq Q$, and by a similar argument $P = P_m$.

Hence, if $P_n \neq P_{n+1}$, $Q_m \neq Q_{m+1}$ are the first and fourth quadrant vertices of a divided cell, the cell must be one of S_n, S_m. By applying the same argument to the other quadrants, we see that $\{S_n\}$ includes all the divided cells of L.

Theorems 2.1 and 2.2 show the importance of chains of divided cells for the evaluation of the inhomogeneous minimum. In order to study the chains of divided cells further, we need to set up a special type of continued fraction, for which the necessary results will now be given.

2.2. A Special Type of Continued Fraction

If $\{a_n\}(n \geq 1)$ is a sequence of integers such that $|a_n| \geq 2$ for $n \geq 1$, we write

$$\left.\begin{aligned} p_{-1} = 0,\ q_{-1} = 1;\ p_0 = 1,\ q_0 = 0;\ p_1 = a_1,\ q_1 = 1; \\ p_{n+1} = a_{n+1}\, p_n - p_{n-1}\ (n \geq 1), \\ q_{n+1} = a_{n+1}\, q_n - q_{n-1}\ (n \geq 1). \end{aligned}\right\} \quad (2.7)$$

The following lemma is easily proved by induction.

<u>Lemma 2.4</u> ([11], Lemma 3). If $\{p_n\}$, $\{q_n\}$ are defined by (2.7), then

(i) for all $n \geq 1$, $p_{n-1}q_n - p_nq_{n-1} = 1$;

(ii) for all $n \geq 1$,

$$|p_n| \geq n+1, \quad |q_n| \geq n, \quad \frac{p_n}{q_n} \geq 1 + \frac{1}{n};$$

(iii) if $a_i > 0$ for $i = 1,2,\ldots,n$, then $p_n > 0$, $q_n > 0$.

It follows from Lemma 2.4 that, for $n \geq 2$,

$$\left|\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}}\right| = \frac{1}{|q_nq_{n-1}|} \leq \frac{1}{n(n-1)},$$

so that the series

$$\sum\left(\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}}\right)$$

is convergent. Hence the sequence $\{p_n/q_n\}$ converges to a limit α, which, by (ii), satisfies $|\alpha| \geq 1$. This justifies the following definition.

<u>Definition 2.1</u>. For any sequence of integers $\{a_n\}$ such that $|a_n| \geq 2$ for $n \geq 1$, $\alpha = [a_1,a_2,a_3,\ldots]$ is defined by

$$\alpha = [a_1,a_2,a_3,\ldots] = \lim_{n\to\infty} p_n/q_n,$$

where p_n, q_n are defined by (2.7) and $|\alpha| \geq 1$.

Clearly,

$$\alpha = a_1 - \frac{1}{a_2-}\,\frac{1}{a_3-}\cdots,$$

so that $[a_1,a_2,a_3,\ldots]$ is easily transformed into a classical semi-regular continued fraction

$$\alpha = a_1 + \frac{\mu_2}{|a_2|} + \frac{\mu_3}{|a_3|} + \dots \qquad (\mu_i = \pm 1), \qquad (2.8)$$

whose convergents have the same value (though the signs of p_n and q_n may be different). We shall never use classical semi-regular continued fraction expansions of the type (2.8); so without confusion we may call $\alpha = [a_1, a_2, a_3, \dots]$ a <u>semi-regular</u> continued fraction expansion of α, to distinguish it from the simple continued fraction expansion of α.

<u>Lemma 2.5</u> ([11], Lemma 4). If $\alpha = [a_1, a_2, a_3, \dots]$, then

(i) $|\alpha| > 1$ if a_n is not constantly equal to 2 or to -2 for large n;

(ii) the convergents p_n/q_n form a strictly decreasing sequence if further $a_n > 0$ for all n.

<u>Proof</u>. A complete proof of (i) is given in [11]; (ii) is an immediate consequence of (i) and (iii) of Lemma 2.4.

In fact we shall consider only semi-regular continued fractions which satisfy condition (i) of Lemma 2.5. Any irrational α has a unique expansion by the nearest integer above; $\alpha = [a_1, a_2, a_3, \dots]$, with $a_n \geq 2$ for $n \geq 2$ and $a_n \geq 3$ for some arbitrarily large n; but, on the other hand, if the a_n are restricted only by the conditions that $|a_n| \geq 2$ and that a_n be not constantly equal to 2 or to -2 for all large n, then any irrational α has infinitely many expansions.

2.3. The Correspondence between Chains of Divided Cells and Pairs of Chains of Integers.

Let $\{S_n\}$ $(-\infty < n < \infty)$ be the doubly infinite chain of divided cells of the inhomogeneous lattice L corresponding to the form f and the point (x',y'), so that, for all n, the vertices of S_n; S_{n+1} satisfy (2.2), (2.3), and (2.4). Let the ξ,η-coordinates of the vertices of S_n be given by

$$A_n = (\xi'_n + \alpha_n + \beta_n, \eta'_n + \gamma_n + \delta_n),\ B_n = (\xi'_n + \alpha_n, \eta'_n + \gamma_n),$$
$$C_n = (\xi'_n, \eta'_n),\ D_n = (\xi'_n + \beta_n, \eta'_n + \delta_n), \tag{2.9}$$

and define the point (x_n, y_n) by

$$\xi'_n = \alpha_n x_n + \beta_n y_n,\ \eta'_n = \gamma_n x_n + \delta_n y_n. \tag{2.10}$$

<u>Theorem 2.3</u> ([11], Theorem 2). For all n, let

$$a_{n+1} = h_n + k_n, \qquad \varepsilon_n = h_n - k_n,$$
$$\theta_n = [a_n, a_{n-1}, a_{n-2}, \dots],\ \phi_n = [a_{n+1}, a_{n+2}, a_{n+3}, \dots]. \tag{2.11}$$

Then

$$\alpha_n/\beta_n = \theta_n, \qquad \delta_n/\gamma_n = \phi_n, \tag{2.12}$$

$$2\xi'_n + \alpha_n + \beta_n = \beta_n\left(\varepsilon_{n-1} + \sum_{r=1}^{\infty} (-1)^r \frac{\varepsilon_{n-r-1}}{\theta_{n-1}\theta_{n-2}\cdots\theta_{n-r}}\right), \tag{2.13}$$

$$2\eta'_n + \gamma_n + \delta_n = \gamma_n\left(\varepsilon_n + \sum_{r=1}^{\infty} (-1)^r \frac{\varepsilon_{n+r}}{\phi_{n+1}\phi_{n+2}\cdots\phi_{n+r}}\right), \tag{2.14}$$

and $f(x+x', y+y')$ is equivalent to

$$f_n(x+x_n, y+y_n) = (\alpha_n x + \beta_n y + \xi'_n)(\gamma_n x + \delta_n y + \eta'_n). \tag{2.15}$$

<u>Proof</u>. Using (2.3), we have

$$\underset{\sim}{V}_r = (\alpha_r, \gamma_r) = (-\beta_{r+1}, -\gamma_{r+1}),$$

so that, by (2.4),

$$\underset{\sim}{V}_{r+1} = - a_{r+1}\underset{\sim}{V}_r - \underset{\sim}{V}_{r-1},$$

$$\alpha_{r+1} = - a_{r+1}\alpha_r - \alpha_{r-1},$$

$$\gamma_{r+1} = - a_{r+1}\gamma_r - \gamma_{r-1}.$$

Thus

$$\underset{\sim}{V}_r = (\alpha_r, \gamma_r) = (-1)^r(\alpha_o p_r - \beta_o q_r, \gamma_o p_r - \delta_o q_r),$$

where

$$p_{-1} = 0,\ q_{-1} = 1;\ p_o = 1,\ q_o = 0;\ p_1 = a_1,\ q_1 = 1;$$

$$p_{r+1} = a_{r+1}p_r - p_{r-1} \qquad (r \geq 1),$$

$$q_{r+1} = a_{r+1}q_r - q_{r-1} \qquad (r \geq 1).$$

It is easily shown by induction that $|q_r| \to \infty$ as $r \to \infty$; and, by Lemma 2.3, $\eta(\underset{\sim}{V}_r) \to 0$. Hence, by Definition 2.1,

$$\delta_o/\gamma_o = \lim_{r \to \infty} p_r/q_r = [a_1, a_2, a_3, \ldots] = \phi_o.$$

Similarly, $\alpha_o/\beta_o = \theta_o$. Thus (2.12) holds for n = 0, and similarly for all n.

It follows from (2.2), (2.3), and (2.4) that, for $n \geq 1$,

$$A_n = A_o - (h_o + 1)\ \underset{\sim}{V}_o - (h_1 + 1)\ \underset{\sim}{V}_1 - \ldots - (h_{n-1} + 1)\ \underset{\sim}{V}_{n-1},$$

$$C_n = C_o + (k_o + 1)\ \underset{\sim}{V}_o + (k_1 + 1)\ \underset{\sim}{V}_1 + \ldots + (k_{n-1} + 1)\ \underset{\sim}{V}_{n-1}.$$

Hence

$$A_o + C_o = A_n + C_n + \sum_{r=0}^{n-1} (h_r - k_r)\ \underset{\sim}{V}_r,$$

so that

$$2\eta_o' + \gamma_o + \delta_o = \eta(A_n) + \eta(C_n) + \sum_{r=0}^{n-1}(-1)^r\varepsilon_r(\gamma_o p_r - \delta_o q_r). \quad (2.16)$$

By (2.11) and Definition 2.1, we have, for all $r \geq 1$,

$$\phi_o = \frac{p_r\phi_r - p_{r-1}}{q_r\phi_r - q_{r-1}},$$

that is,

$$p_{r-1} - \phi_o q_{r-1} = \phi_r(p_r - \phi_o q_r).$$

Thus, since $p_o - \phi_o q_o = 1$,

$$\gamma_o p_r - \delta_o q_r = \gamma_o(p_r - \phi_o q_r) = \frac{\gamma_o}{\phi_1\phi_2 \cdots \phi_r}. \quad (2.17)$$

Since, by Lemma 2.3, $\eta(A_n)$, $\eta(C_n) \to 0$ as $n \to \infty$, (2.14) now follows from (2.16) and (2.17) for $n = 0$, and similarly for all n; (2.13) may be proved in a corresponding way.

The last statement of the theorem follows from (2.9) and (2.10) and the fact that A_n, B_n, C_n, D_n are the vertices of a cell of the inhomogeneous lattice corresponding to f and the point (x', y').

Lemma 2.6 ([6], Lemmas 2.3, 2.4). Let $\{a_n\}$, $(-\infty < n < \infty)$ be a doubly infinite chain of integers such that

(i) $|a_n| \geq 2$ and a_n is not constantly equal to 2 or to -2 for large n of either sign.

Let $\{\varepsilon_n\}$ $(-\infty < n < \infty)$ be any corresponding chain of integers which satisfy the following conditions:

(ii) $|\varepsilon_n| \leq |a_{n+1}| - 2$ and ε_n has the same parity as a_{n+1};

(iii) neither $a_{n+1} + \varepsilon_n$ nor $a_{n+1} - \varepsilon_n$ is constantly equal to -2 for large n of either sign;

(iv) for any n, the relation

$$a_{n+2r+1} + \varepsilon_{n+2r} = a_{n+2r+2} - \varepsilon_{n+2r+1} = 2$$

does not hold either for all $r \geq 0$ or for all $r \leq 0$.

For all n, let

$$\theta_n = [a_n, a_{n-1}, a_{n-2}, \ldots] \;, \phi_n = [a_{n+1}, a_{n+2}, a_{n+3}, \ldots]. \qquad (2.18)$$

Then the series

$$\varepsilon_{n-1} + \sum_{r=1}^{\infty} (-1)^r \frac{\varepsilon_{n-r-1}}{\theta_{n-1}\theta_{n-2}\ldots\theta_{n-r}},$$

$$\varepsilon_n + \sum_{r=1}^{\infty} (-1)^r \frac{\varepsilon_{n+r}}{\phi_{n+1}\phi_{n+2}\ldots\phi_{n+3}}$$

are absolutely convergent with sums whose moduli are less than $|\theta_n| - 1$, $|\phi_n| - 1$, respectively.

Proof. The proof is given in [6] and requires all the conditions (i) to (iv).

Theorem 2.4 ([6] ,§2). Let $\{a_n\}$, $\{\varepsilon_n\}$, θ_n, ϕ_n be defined as in Lemma 2.6, and for each n let

$$2\xi_n + \theta_n + 1 = \varepsilon_{n-1} + \sum_{r=1}^{\infty} (-1)^r \frac{\varepsilon_{n-r-1}}{\theta_{n-1}\theta_{n-2}\ldots\theta_{n-r}}, \qquad (2.19)$$

$$2\eta_n + 1 + \phi_n = \varepsilon_n + \sum_{r=1}^{\infty} (-1)^r \frac{\varepsilon_{n+r}}{\phi_{n+1}\phi_{n+2}\ldots\phi_{n+r}}. \qquad (2.20)$$

Then the four points A_n, B_n, C_n, D_n given by

$$\left.\begin{array}{ll} A_n = (\xi_n + \theta_n + 1, \eta_n + 1 + \phi_n), & B_n = (\xi_n + \theta_n, \eta_n + 1), \\ C_n = (\xi_n, \eta_n), & D_n = (\xi_n + 1, \eta_n + \phi_n) \end{array}\right\} \qquad (2.21)$$

are vertices of a divided cell of an inhomogeneous lattice; the coordinates of the vertices A_{n+1}, B_{n+1}, C_{n+1}, D_{n+1} of the successive divided cell are of the same form:

$$\left.\begin{aligned} A_{n+1} &= (\beta(\xi_{n+1} + \theta_{n+1} + 1),\ \gamma(\eta_{n+1} + 1 + \phi_{n+1})), \\ B_{n+1} &= (\beta(\xi_{n+1} + \theta_{n+1}),\ \gamma(\eta_{n+1} + 1)), \\ C_{n+1} &= (\beta\xi_{n+1},\ \gamma\eta_{n+1}), \\ D_{n+1} &= (\beta(\xi_{n+1} + 1),\ \gamma(\eta_{n+1} + \phi_{n+1})), \end{aligned}\right\} \quad (2.22)$$

where $\beta = -\theta_n$, $\gamma = -1/\phi_{n+1}$.

Proof. By Lemma 2.6, we have

$$|2\xi_n + \theta_n + 1| < |\theta_n| - 1,$$

$$|2\eta_n + 1 + \phi_n| < |\phi_n| - 1,$$

from which we can immediately deduce that

$$\left.\begin{aligned} \operatorname{sgn} \xi_n &= \operatorname{sgn}(\xi_n + 1) = -\operatorname{sgn} \theta_n, \\ \operatorname{sgn}(\xi_n + \theta_n) &= \operatorname{sgn}(\xi_n + \theta_n + 1) = \operatorname{sgn} \theta_n; \end{aligned}\right\}$$

$$\left.\begin{aligned} \operatorname{sgn} \eta_n &= \operatorname{sgn}(\eta_n + 1) = -\operatorname{sgn} \phi_n, \\ \operatorname{sgn}(\eta_n + \phi_n) &= \operatorname{sgn}(\eta_n + \phi_n + 1) = \operatorname{sgn} \phi_n. \end{aligned}\right\}$$

Thus the points A_n, B_n, C_n, D_n given by (2.21) lie one in each of the four quadrants and are the vertices of a divided cell of the lattice

$$\left.\begin{aligned} \xi &= \xi_n + \theta_n x + y \\ \eta &= \eta_n + x + \phi_n y \end{aligned}\right\} . \tag{2.23}$$

By a similar argument, the points A_{n+1}, B_{n+1}, C_{n+1}, D_{n+1} given by (2.22) are the vertices of a divided cell of a lattice.

From (2.18), (2.19), and (2.20) we can deduce that

$$\left.\begin{aligned} -\theta_n\xi_{n+1} &= \xi_n + (k_n + 1)\theta_n, \\ -\eta_{n+1}/\phi_{n+1} &= \eta_n + k_n + 1, \end{aligned}\right\} \tag{2.24}$$

where $2k_n = a_{n+1} - \varepsilon_n$. If now we write $2h_n = a_{n+1} + \varepsilon_n$, then a simple calculation shows that the points (2.22) satisfy

$$A_{n+1} = A_n - (h_n + 1)\,\underset{\sim}{V}_n, \quad B_{n+1} = A_n - h_n\,\underset{\sim}{V}_n,$$

$$C_{n+1} = C_n + (k_n + 1)\,\underset{\sim}{V}_n, \quad D_{n+1} = C_n + k_n\,\underset{\sim}{V}_n,$$

where $\underset{\sim}{V}_n = A_n - D_n$. Since the points are the vertices of a divided cell, it follows from (2.2) and (2.3) that this cell is the successor of the cell determined by A_n, B_n, C_n, D_n.

Thus a pair of chains $\{a_n\}$, $\{\varepsilon_n\}$ which satisfy the conditions of Lemma 2.6 determines the chain of divided cells of a lattice which is unique apart from similarity. We write

$$h_n + k_n = a_{n+1}, \quad h_n - k_n = \varepsilon_n,$$

$$\xi'_n = \beta_n \xi_n, \quad \eta'_n = \gamma_n \eta_n,$$

where

$$\beta_n \gamma_n = \frac{\Delta}{|\theta_n \phi_n - 1|}.$$

Then, conversely, it is clear from Lemmas 2.1, 2.2 and Theorem 2.3 that a chain of divided cells whose vertices are given by (2.9) determines a unique pair of chains of integers $\{a_n\}$, $\{\varepsilon_n\}$ such that the conditions of Lemma 2.6 and the relations (2.19), (2.20) are satisfied.

2.4. I-reduced Forms

We shall say that the form f is inhomogeneously reduced, or I-reduced, if it can be factorized in the form

$$f(x,y) = \pm \frac{\Delta}{|\theta\phi - 1|}(\theta x + y)(x + \phi y), \tag{2.25}$$

where

$$|\theta| > 1, \quad |\phi| > 1.$$

It is well known (see Dickson [32], Th.76) that, corresponding to any indefinite quadratic form F(x,y) which does not represent zero, there is an equivalent Gauss-reduced form, that is a form (2.25) which satisfies the more stringent conditions

$$\theta < -1, \quad \phi > 1.$$

Hence we have

Lemma 2.7([6], Lemma 2.2). If $F(x,y)$ is an indefinite quadratic form which does not represent zero, there exists an I-reduced form equivalent to it.

Since any form which does not represent zero can be written in the form (2.1), Lemma 2.7 also follows from Theorem 2.3; for, by Lemmas 2.1, 2.2, and 2.5, $|\theta_n| > 1$, $|\phi_n| > 1$ for all n, so that, by (2.11), the forms f_n of (2.15) must be I-reduced for all n.

2.5 The Evaluation of the Inhomogeneous Minimum

We now combine our results and apply them to the problem of evaluating the inhomogeneous minimum of an indefinite binary quadratic form f given by (2.1). The notation of this section will be used throughout the thesis.

Let

$$f_o(x,y) = \pm \frac{\Delta}{|\theta_o\phi_o - 1|} (\theta_o x + y)(x + \phi_o y)$$

be any I-reduced form equivalent to f. Then any chain of integers $\{a_n\}$ $(-\infty < n < \infty)$ for which condition (i) of Lemma 2.6 holds and for which

$$\theta_o = [a_o, a_{-1}, a_{-2}, \ldots], \quad \phi_o = [a_1, a_2, a_3, \ldots]$$

is called an <u>a-chain</u> of f. A chain of equivalent I-reduced forms $\{f_n\}$ is defined by

$$f_n(x,y) = \pm \frac{\Delta}{|\theta_n\phi_n - 1|} (\theta_n x + y)(x + \phi_n y), \qquad (2.26)$$

where

$$\theta_n = [a_n, a_{n-1}, a_{n-2}, \ldots], \quad \phi_n = [a_{n+1}, a_{n+2}, a_{n+3}, \ldots]; \tag{2.27}$$

there is clearly a one to one correspondence between the chains $\{a_n\}$ and $\{f_n\}$.

Any chain of integers $\{\varepsilon_n\}$ $(-\infty < n < \infty)$ which satisfies conditions (ii), (iii), and (iv) of Lemma 2.6 is called an <u>ε-chain</u> corresponding to the a-chain $\{a_n\}$ (or, equivalently, to the chain of forms $\{f_n\}$).

We write

$$\sigma_n = \varepsilon_{n-1} + \sum_{r=1}^{\infty} (-1)^r \frac{\varepsilon_{n-r-1}}{\theta_{n-1}\theta_{n-2} \cdots \theta_{n-r}}, \tag{2.28}$$

$$\tau_n = \varepsilon_n + \sum_{r=1}^{\infty} (-1)^r \frac{\varepsilon_{n+r}}{\phi_{n+1}\phi_{n+2} \cdots \phi_{n+r}}; \tag{2.29}$$

and define π_n as

$$\frac{\Delta}{|\theta_n\phi_n - 1|} \min \begin{bmatrix} |(1+\theta_n+\sigma_n)(1+\phi_n+\tau_n)|, & |(-1+\theta_n+\sigma_n)(1-\phi_n+\tau_n)|, \\ |(-1-\theta_n+\sigma_n)(-1-\phi_n+\tau_n)|, & |(1-\theta_n+\sigma_n)(-1+\phi_n+\tau_n)| \end{bmatrix}. \tag{2.30}$$

We define (x_n, y_n) $(-\infty < n < \infty)$ by

$$\left.\begin{aligned} \theta_n x_n + y_n &= \tfrac{1}{2}(-1 - \theta_n + \sigma_n) \\ x_n + \phi_n y_n &= \tfrac{1}{2}(-1 - \phi_n + \tau_n) \end{aligned}\right\}. \tag{2.31}$$

<u>Theorem 2.5</u>. If $\{a_n\}$ is an a-chain of f, and $\{\varepsilon_n\}$ a corresponding ε-chain, and if we put

$$M(\{a_n\}, \{\varepsilon_n\}) = \inf_n \pi_n/4, \tag{2.32}$$

then there exists a point P such that (for all n).

$$M(P) = M(f;P) = M(f_n;\, x_n, y_n) = M(\{a_n\},\ \{\varepsilon_n\}); \qquad (2.33)$$

and if $M(f)$ is the inhomogeneous minimum, then

$$M(f) = \sup M(\{a_n\},\ \{\varepsilon_n\}), \qquad (2.34)$$

where the supremum is taken over all the a-chains of f and all possible corresponding ε-chains.

Proof. In Theorem 2.4,

$$\beta\gamma = \theta_n/\phi_{n+1} = \frac{\Delta/(\theta_n\phi_n - 1)}{\Delta/(\theta_{n+1}\phi_{n+1} - 1)},$$

and, by (2.27) and (2.28),

$$\left.\begin{array}{l} 2\xi_n + \theta_n + 1 = \sigma_n \\ 2\eta_n + 1 + \phi_n = \tau_n \end{array}\right\} \qquad (2.35)$$

Hence, by Theorem 2.4, the set of values of $\pi_n/4$, $(-\infty < n < \infty)$ given by (2.30) coincides with the set of values

$$\min\left[|\xi\eta|;\ (\xi,\eta) \in \{A_n, B_n, C_n, D_n\}\right] \quad (-\infty < n < \infty),$$

where A_n, B_n, C_n, D_n are the vertices of the divided cell S_n, and $\{S_n\}$ is the chain of divided cells of the inhomogeneous lattice which, by (2.23), (2.31), and (2.35) corresponds to the form f_n and the point (x_n, y_n) for any n. Since all the f_n are equivalent to f, (2.33) now follows from Theorem 2.1.

By Lemmas 2.1 and 2.2 and Theorem 2.3, the chains of divided cells of all possible inhomogeneous lattices corresponding to f are included by taking all a-chains of f with all possible corresponding ε-chains. This proves (2.34).

Lemma 2.8 ([6], §3). The points (x_n, y_n), (x_{n+1}, y_{n+1}) determined by (2.31) satisfy the equations

$$\left.\begin{aligned} x_{n+1} &= y_n \\ y_{n+1} &= -(x_n + a_{n+1}y_n + 1 + k_n) \end{aligned}\right\}, \tag{2.36}$$

where

$$2k_n = a_{n+1} - \varepsilon_n.$$

Also, if

$$\left.\begin{aligned} x' &= y \\ y' &= -x - a_{n+1}y + k_n + 1 \end{aligned}\right\}, \tag{2.37}$$

then

$$f_{n+1}(x' + x_{n+1},\ y' + y_{n+1}) = f_n(x + x_n,\ y + y_n).$$

Proof. By (2.31) and (2.35)

$$\left.\begin{aligned} \theta_n x_n + y_n &= \xi_n \\ x_n + \phi_n y_n &= \eta_n \end{aligned}\right\};$$

the lemma now follows from (2.24). A detailed proof is given in [6].

Corollary. If the chains $\{a_n\}$, $\{\varepsilon_n\}$ are both periodic, then the points (x_n, y_n) are rational.

This follows immediately from (2.36).

The success of this approach to the problem of the inhomogeneous minimum depends on the rapid convergence of the series (2.28) and (2.29). Estimates of the error made in replacing these series by partial sums are given in Lemmas 2.9 and 2.10 below.

We here introduce the permanent notation $\|x\|$ for a quantity whose modulus does not exceed $|x|$.

<u>Lemma 2.9</u> ([6], §3). If $\{a_n\}$ is an a-chain of f, and $\{\varepsilon_n\}$ $(-\infty < n < \infty)$ is a chain of integers which satisfies condition (ii) of Lemma 2.6, then

$$\sigma_n = \varepsilon_{n-1} - \frac{\varepsilon_{n-2}}{\theta_{n-1}} + \dots + (-1)^r \frac{\varepsilon_{n-r-1}}{\theta_{n-1} \dots \theta_{n-r}} + \left\| \frac{1}{\theta_{n-1} \dots \theta_{n-r}} \left(1 - \frac{1}{|\theta_{n-r-1}|} \right) \right\|, \quad (2.38)$$

$$\tau_n = \varepsilon_n - \frac{\varepsilon_{n+1}}{\phi_{n+1}} + \dots + (-1)^r \frac{\varepsilon_{n+r}}{\phi_{n+1} \dots \phi_{n+r}} + \left\| \frac{1}{\phi_{n+1} \dots \phi_{n+r}} \left(1 - \frac{1}{|\phi_{n+r+1}|} \right) \right\|. \quad (2.39)$$

<u>Proof</u>. First we note that, by (2.28),

$$\tau_n = \varepsilon_n - \frac{\varepsilon_{n+1}}{\phi_{n+1}} + \dots + \frac{(-1)^r \varepsilon_{n+r}}{\phi_{n+1} \dots \phi_{n+r}} + \frac{(-1)^{r+1} \tau_{n+r+1}}{\phi_{n+1} \dots \phi_{n+r+1}}. \quad (2.40)$$

Since $|\varepsilon_n| \le |a_{n+1}| - 2$ for all n, we have, from (2.28),

$$|\tau_n| \le |a_{n+1}| - 2 + \sum_{r=1}^{\infty} \frac{|a_{n+r+1}| - 2}{|\phi_{n+1} \dots \phi_{n+r+1}|}. \quad (2.41)$$

By (2.27),

$$\phi_n = a_{n+1} - \frac{1}{\phi_{n+1}}, \tag{2.42}$$

so that

$$|\phi_n| - 1 \geq |a_{n+1}| - \frac{1}{|\phi_{n+1}|} - 1 = |a_{n+1}| - 2 + \frac{|\phi_{n+1}| - 1}{|\phi_{n+1}|}.$$

Repeated application of this inequality gives

$$|\phi_n| - 1 \geq |a_{n+1}| - 2 + \frac{|a_{n+2}| - 2}{|\phi_{n+1}|} + \frac{|a_{n+3}| - 2}{|\phi_{n+1}\phi_{n+2}|} + \dots ,$$

so that, for all n, by (2.40)

$$|\tau_n| \leq |\phi_n| - 1. \tag{2.43}$$

From (2.40) and (2.43) we now get (2.39); (2.38) may be proved similarly.

<u>Lemma 2.10</u> (not given in [11] or [6]). If $\{a_n\}$, $\{\varepsilon_n\}$ are given in Lemma 2.9, then

(i) if further θ_{n-r-1}, θ_{n-r-2} differ in sign (i.e. if a_{n-r-1}, a_{n-r-2} differ in sign), then

$$\sigma_n = \varepsilon_{n-1} - \frac{\varepsilon_{n-2}}{\theta_{n-1}} + \dots + (-1)^r \frac{\varepsilon_{n-r-1}}{\theta_{n-1} \dots \theta_{n-r}} + \left\| \frac{1}{\theta_{n-1} \dots \theta_{n-r}} \left(1 - \frac{1}{|\theta_{n-r-1}|} - \frac{2}{|\theta_{n-r-1}\theta_{n-r-2}|} \right) \right\| ; \tag{2.44}$$

(ii) if further ϕ_{n+r+1}, ϕ_{n+r+2} differ in sign (i.e. if a_{n+r+2}, a_{n+r+3} differ in sign), then

$$\tau_n = \varepsilon_n - \frac{\varepsilon_{n+1}}{\phi_{n+1}} + \dots + (-1)^r \frac{\varepsilon_{n+r}}{\phi_{n+1} \dots \phi_{n+r}}$$

$$+ \left\| \frac{1}{\phi_{n+1} \dots \phi_{n+r}} \left(1 - \frac{1}{|\phi_{n+r+1}|} - \frac{2}{|\phi_{n+r+1}\phi_{n+r+2}|} \right) \right\| . \quad (2.45)$$

Proof. If ϕ_n, ϕ_{n+1} differ in sign, (2.42) gives

$$|\phi_n| = |a_{n+1}| + \frac{1}{|\phi_{n+1}|} ,$$

so that

$$|\phi_n| - 1 - \frac{2}{|\phi_{n+1}|} = |a_{n+1}| - 2 + \frac{|\phi_{n+1}| - 1}{|\phi_{n+1}|} .$$

By (2.40) and (2.43)

$$|\tau_n| \leq |\varepsilon_n| + \frac{|\tau_{n+1}|}{|\phi_{n+1}|} \leq |a_{n+1}| - 2 + \frac{|\phi_{n+1}| - 1}{|\phi_{n+1}|} .$$

Hence

$$|\tau_n| \leq |\phi_n| - 1 - \frac{2}{|\phi_{n+1}|} .$$

Applying this result to τ_{n+r+1} and using (2.40), we get (2.45); (2.44) may be proved similarly.

When we wish to evaluate $M(f)$, we try first to find and reject those a-chains (with the corresponding chains of forms $\{f_n\}$) for which $M(\{a_n\}, \{\varepsilon_n\})$ is low for any corresponding ε-chain, and then to examine the other chains more closely. At this preliminary stage, strict inequalities are not needed, and it is unnecessary to decide whether $\{\varepsilon_n\}$ satisfies all the conditions of Lemma 2.6; therefore, it is convenient in Lemmas 2.9 and 2.10 to assume only that $\{\varepsilon_n\}$ satisfies condition (ii) of Lemma 2.6. It is possible to eliminate a large number of chains $\{f_n\}$ with the corresponding a-chains by using Lemmas 2.11, 2.12 below.

We first introduce a symbol which will be used repeatedly in later chapters.

Definition 2.2. For any indefinite binary quadratic form f, given by

$$f(x,y) = ax^2 + bxy + cy^2,$$

we define

$$\lambda = \lambda(f) = \min |a \pm b + c| = \min |f(1,\pm 1)|.$$

In particular, for forms f_n given by (2.26),

$$\lambda(f_n) = \frac{\Delta}{|\theta_n \phi_n - 1|} \min \left[|(\theta_n - 1)(\phi_n - 1)|,\ |(\theta_n + 1)(\phi_n + 1)| \right]. \quad (2.46)$$

Lemma 2.11 ([6], Lemmas 3.2, 3.3). If $\{f_n\}$ is a chain of I-reduced forms equivalent to f, and $\{a_n\}$ is the corresponding a-chain, then for every corresponding ε-chain and for every r we have

$$M(f;\, P) = M(\{a_n\},\{\varepsilon_n\}) = M(f_n;\, x_r, y_r) \le \pi_r/4 \le \lambda(f_r)/4.$$

Proof. The proof is given in [6].

Lemma 2.12. In Lemma 2.11, we can have

$$M(f;\ P) = \lambda(f_r)/4$$

if and only if conditions (i) and (ii) are both satisfied:

(i) $\lambda(f_r) = \inf_n \lambda(f_n)$;

(ii) $(x_r, y_r) \equiv (\frac{1}{2}, \frac{1}{2}) \pmod 1$.

The condition (ii) implies that

(iii) the chain $\{a_n\}$ is even, i.e. a_n is even for all n;

(iv) $\varepsilon_n = 0$ for all n;

(v) $P \equiv (\frac{1}{2}, 0)$ or $(0, \frac{1}{2})$ or $(\frac{1}{2}, \frac{1}{2}) \pmod 1$.

Proof. It is clear from (2.32) and (2.46) that $M(f;\ P) = \lambda(f_r)/4$ if and only if (i) and (ii) hold; if (ii) holds, then $\sigma_r = \tau_r = 0$, which implies (iii) and (iv); (v) follows from (ii) also.

As a corollary to the two lemmas we have a simple inequality for $M(f)$:

Corollary. If the form $f(x,y)$ is I-reduced, then

$$M(f) \leq \lambda(f)/4;$$

equality can occur only if $M(f) = M(f;\ P)$, where $P \equiv (\frac{1}{2}, \frac{1}{2}) \pmod 1$.

It is clear that this result is closely related to the estimate for $M(f)$ given in Barnes [5]. The connection between these two results and the relation of these results to others of the same type will be discussed in section 3.4.

Finally, to avoid unnecessary enumeration of cases, we need another lemma of [6] .

Lemma 2.13 ([6], Lemma 3.1). If $\{\varepsilon_n\}$ is any ε-chain corresponding to the a-chain $\{a_n\}$, the value of $M(P) = M(\{a_n\}, \{\varepsilon_n\})$ is unaltered by any of the following operations:

(i) reversing the chains $\{a_{n+1}\}$, $\{\varepsilon_n\}$ about the same point;

(ii) changing the signs of all ε_n;

(iii) changing the signs of all a_n and of alternate ε_n.

Proof. The proof is given in [6].

CHAPTER 3

EQUIVALENT I-REDUCED FORMS

In order to use the method of section 2.5 for evaluating the inhomogeneous minimum of an indefinite binary quadratic form g which does not represent zero, we must be able to determine all the a-chains of g. In this chapter I consider the problem of determining all possible chains $\{f_n\}$ of I-reduced forms equivalent to g (and hence all possible a-chains of g).

In section 3.1 I give, as Theorem 3.1, a condition for a form to be I-reduced which corresponds to the condition for a form to be Gauss-reduced given by Inkeri [39] (see sect. 2.4). Barnes [6] showed that there is only a finite number of I-reduced forms equivalent to a form g with integral coefficients; for the sake of completeness I include this result, which is easily derived from Theorem 3.1, as Theorem 3.2.

In section 3.2 I give some results on equivalence which are needed in the following section.

It follows from Theorem 3.2 that, if g is proportional to a form with integral coefficients, we can obtain all the

I-reduced forms equivalent to g by a finite number of trials. However, this method becomes laborious if the discriminant of g is large and of course breaks down altogether if the number of I-reduced forms equivalent to g is infinite. Thus another method is needed.

The natural thing to do is to start from a particular I-reduced form equivalent to g, say f, where

$$f(x,y) = \pm \frac{\Delta}{|\theta\phi - 1|}(\theta x + y)(x + \phi y) \qquad (|\theta| > 1,\ |\phi| > 1),$$

and, by expanding θ, ϕ in all possible ways as semi-regular continued fractions, to obtain all the chains $\{f_n\}$ to which f belongs. The questions then arise, whether we can obtain all the I-reduced forms equivalent to g in this way, by starting from just one form, and whether we can get all the different chains of forms in some such way. In section 3.3 I consider these questions.

As an immediate consequence of Lemma 2.11 and 2.12 and of Theorem 3.3 of section 3.3, I obtain a bound for the inhomogeneous minimum of a Gauss-reduced form f (see sect. 2.4) in terms of the coefficients of f which is the same as the bound for the inhomogeneous minimum of any indefinite binary quadratic form given by Barnes [5]. In section 3.4 I discuss these results and others of the same type obtained by different authors.

3.1. The Condition for a Form to be I-reduced and a Theorem on Forms with Integral Coefficients

Theorem 3.1 (cf. Inkeri [39], §5). An indefinite form

$$f(x,y) = ax^2 + bxy + cy^2$$

is I-reduced if and only if

$$|b| > |a + c| \tag{3.1}$$

(i.e. if and only if $a + b + c$, $a - b + c$ differ in sign).

Proof. We have

$$\left.\begin{aligned} f(x,y) &= \frac{2ac}{b+\Delta}\left(\frac{b+\Delta}{2c}x + y\right)\left(x + \frac{b+\Delta}{2a}y\right) \\ &= \frac{2ac}{b-\Delta}\left(x + \frac{b-\Delta}{2a}y\right)\left(\frac{b-\Delta}{2c}x + y\right) \end{aligned}\right\} \tag{3.2}$$

where $\Delta = +\sqrt{(b^2 - 4ac)}$. It now follows from the definition of an I-reduced form (see sect. 2.4) that $ax^2 + bxy + cy^2$ is I-reduced if and only if $ax^2 - bxy + cy^2$ is. Hence, without loss of generality, we may take $b \geq 0$. In this case it is clear that $f(x,y)$ is I-reduced if and only if

$$b + \Delta > 2|a|, \quad b + \Delta > 2|c|. \tag{3.3}$$

If, now, (3.1) is satisfied, then

$$b > |a + c|. \tag{3.4}$$

Hence

$$\Delta^2 = b^2 - 4ac > (a + c)^2 - 4ac = (a - c)^2,$$

so that

$$\Delta > |a - c|. \tag{3.5}$$

By (3.4) and (3.5)

$$b + \Delta > |a + c| + |a - c| = 2 \max \left[|a|, |c| \right].$$

Thus (3.3) holds, and $f(x,y)$ is I-reduced.

If, on the other hand, (3.1) does not hold, then

$$b \leq |a + c|,$$

and by a similar argument we get

$$b + \Delta \leq 2 \max \left[|a|, |c| \right],$$

so that (3.3) does not hold and $f(x,y)$ is not I-reduced.

Thus $f(x,y)$ is I-reduced if and only if (3.1) holds.

<u>Theorem 3.2</u> (Barnes [6], Lemma 2.1). If an indefinite quadratic form $f(x,y)$ has integral coefficients, there is only a finite number of I-reduced forms equivalent to it.

<u>Proof</u>. We suppose that $f(x,y)$ has discriminant Δ^2 and show that there is only a finite number of I-reduced forms with integral coefficients and discriminant Δ^2. If $ax^2 + bxy + cy^2$ has discriminant Δ^2 and is I-reduced, then, by Theorem 3.1,

$$0 < b^2 - (a + c)^2 = (b - a - c)(b + a + c) = \Delta^2 - (a - c)^2.$$

Thus $|a - c|$ is less than Δ, and for each of the finite number of possible integral values of $|a - c|$ there is only a finite number of possible integral values of $b - a - c$, $b + a + c$, and so of a, b, c.

3.2. Some Results on Equivalences for I-reduced Forms

If the form F is equivalent to the form f under an integral unimodular linear transformation

$$T = \begin{bmatrix} t & u \\ v & w \end{bmatrix},$$

where t, u, v, w are integral, and $tw - uv = \pm 1$ (that is, if

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} t & u \\ v & w \end{bmatrix}\begin{bmatrix} X \\ Y \end{bmatrix}$$

gives

$$f(x,y) = F(X,Y)),$$

then we shall write

$$F = fT = f\begin{bmatrix} t & u \\ v & w \end{bmatrix}.$$

With this notation $(fT_1)T_2 = f(T_1T_2)$. If $tw - uv = +1$, the forms will be called <u>properly equivalent</u>; and if neither of the statements $t = w = 0$, $u = v = 0$ holds, the transformation T will be called <u>non-trivial</u>.

Throughout this chapter we denote by $f = (a,b,c)$ an I-reduced form which does not represent zero:

$$f(x,y) = ax^2 + bxy + cy^2,$$

where $D = b^2 - 4ac$ is the discriminant of the form, and $+\sqrt{D} = \Delta$. It follows from (3.2) that $f(x,y)$ is I-reduced if and only if $f(x,-y)$, $f(y,x)$, and $f(y,-x)$ are I-reduced; also, any chain containing one of these forms can be converted into a chain containing $f(x,y)$ by reversing the chain $\{a_n\}$ ($\{a_{-n}\}$ is the <u>reverse</u> of $\{a_n\}$) or by replacing $\{a_n\}$ by

$\{-a_n\}$ (its <u>negative</u>) or both (see Lemma 2.13). It would therefore be sufficient to consider only those I-reduced forms (a,b,c) with $b > 0$, $|a| \leq |c|$; here we adopt the convention of considering only I-reduced forms for which $b > 0$. With this convention, the I-reduced form $f = (a,b,c)$ can be factorized as

$$f(x,y) = ax^2 + bxy + cy^2 = \frac{\Delta}{|r_1r_2 - 1|}(r_1x + y)(x + r_2y), \quad (3.6)$$

where

$$r_1 = \frac{b + \Delta}{2c}, \quad r_2 = \frac{b + \Delta}{2a},$$

$$b > 0, \quad |r_1| > 1, \ |r_2| > 1,$$

and r_1, r_2 are irrational. We shall call r_1 and r_2 the <u>first and second roots</u> of f respectively.

If $f = (a,b,c)$, we shall call the form

$$(c,b,a) = f\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

the <u>reverse</u> of f.

We denote by T a non-trivial linear transformation

$$T = \begin{bmatrix} t & u \\ v & w \end{bmatrix}, \quad (3.7)$$

where t, u, v, w are integral, $w \geq 0$, and $tw - uv = 1$; and by $F = (A,B,C)$ the I-reduced form

$$\left.\begin{aligned} F(x,y) &= Ax^2 + Bxy + Cy^2 \quad (B > 0,\ B^2 - 4AC = \Delta^2) \\ &= \frac{\Delta}{|R_1R_2 - 1|}(R_1x + y)(x + R_2y) \end{aligned}\right\}. \quad (3.8)$$

whose first and second roots are

$$R_1 = \frac{B + \Delta}{2C}, \qquad R_2 = \frac{B + \Delta}{2A},$$

so that

$$|R_1| > 1, \quad |R_2| > 1.$$

We note that if $F = fT$, where T is given by (3.7) then

$$\left.\begin{aligned} A &= at^2 + btv + cv^2 \\ B &= 2atu + b(tw + uv) + 2cvw \\ C &= au^2 + buw + cw^2 \end{aligned}\right\}. \tag{3.9}$$

Thus $fT = f(-T)$ and there is no loss of generality in assuming in (3.7) that $w \geq 0$.

We now prove five lemmas which are needed for the next section.

<u>Lemma 3.1</u>. If $F = fT$, where T is given by (3.7), then

$$R_1 = \frac{tr_1 + v}{ur_1 + w}, \quad R_2 = \frac{wr_2 + u}{vr_2 + t}. \tag{3.10}$$

<u>Proof</u>. The relations (3.10) follow from (3.6), (3.8), and (3.9).

<u>Lemma 3.2</u>. If $F = fT$, where T is given by (3.7), and if

$$R_2 = \left[a_1, \frac{w_1r_2 + u_1}{v_1r_2 + t_1}\right],$$

where

$$\pm \begin{bmatrix} t_1 & u_1 \\ v_1 & w_1 \end{bmatrix} = \begin{bmatrix} a_1t-u & t \\ a_1v-w & v \end{bmatrix},$$

then

$$\frac{t_1r_1 + v_1}{u_1r_1 + w_1} = [a_1, R_1].$$

Proof. The lemma follows immediately from the expression for R_1 given in (3.10).

Lemma 3.3. If $F = fT$, where T is given by (3.7), then

(i) $|u| > |t|$ implies $w \geq |u|$; $|v| \geq |t|$;

(ii) $|v| > w$ implies $|t| \geq |v|$, $|u| \geq w$.

Proof. From (3.10) we have

$$r_1 = \frac{-wR_1 + v}{uR_1 - t} = -\frac{w}{u} - \frac{1}{u(uR_1 - t)} = -\frac{v}{t} - \frac{R_1}{t(uR_1 - t)} \quad (3.11)$$

If $|t| = 0$, then $|u| = |v| = 1$, and since T is non-trivial, $w \geq 1$, so that (i) holds. If $|u| > |t| \geq 1$, then, since $|R_1| > 1$,

$$|uR_1 - t| \geq (|t| + 1)|R_1| - |t| > |R_1|;$$

thus $|r_1| < 1$ by (3.11) unless $w \geq |u|$, $|v| \geq |t|$. This proves (i) for all cases, and (ii) is proved similarly.

Lemma 3.4. If t, u, v, w are integers such that $tw - uv = 1$, and it is not true that $t = w = 0$ or that $u = v = 0$, then exactly one of the following sets of relations holds:

$$\begin{cases} w = |v| = 1, & |u| > |t|; \quad (3.12) \\ w > |v|, & |u| \geq |t|; \quad (3.13) \end{cases}$$

$$\begin{cases} w = |v| = 1, & |u| < |t|; \quad (3.14) \\ w < |v|, & |u| \leq |t|; \quad (3.15) \end{cases}$$

Proof. Clearly we cannot have $w = |v|$, $|u| = |t|$, as this contradicts $tw - uv = 1$.

If $w > |v|$, $|u| > |t|$, then

$$|tw - uv| \geq (|u| + 1)(|v| + 1) - |uv| = |u| + |v| + 1;$$

since we cannot have $u = v = 0$, this contradicts $tw - uv = 1$. Similarly, $w < |v|$, $|t| < |u|$ leads to a contradiction.

Thus (3.12) to (3.15) cover all the possibilities; as they are mutually exclusive, this proves the lemma.

Lemma 3.5. If $F = fT$, where T is given by (3.7), and if (3.12) or (3.13) holds, then T must be one of the following matrices (where k is a positive integer):

$$\begin{bmatrix} 0 & \mp 1 \\ \pm 1 & 1 \end{bmatrix}, \tag{3.16}$$

$$\begin{bmatrix} 1 & \pm 1 \\ \pm(k-1) & k \end{bmatrix} \quad (k \geq 3), \tag{3.17}$$

$$\left.\begin{matrix} \begin{bmatrix} t & u \\ v & w \end{bmatrix} \\ (|u| > |t| > 0,\ w > |v|,\ w > |u|,\ |v| > |t|) \end{matrix}\right\}, \tag{3.18}$$

$$\begin{bmatrix} 0 & \mp 1 \\ \pm 1 & k \end{bmatrix} \quad (k \geq 2), \tag{3.19}$$

$$\begin{bmatrix} 1 & \pm(k-1) \\ \pm 1 & k \end{bmatrix} \quad (k \geq 2), \tag{3.20}$$

$$\begin{bmatrix} 1 & \pm 1 \\ 0 & 1 \end{bmatrix}. \tag{3.21}$$

Proof. We use (i) of Lemma 3.2, and the fact that $tw - uv = 1$.

If (3.12) holds, T must be given by (3.16).

Now suppose that (3.13) holds, so that $w > |v|$, $|u| \geq |t|$. If $|u| = |t|$ $(=1)$, then for $|v| > 1$, T is given by (3.17), for $|v| = 1$, T is given by (3.20) with $k = 2$, and for $v = 0$, T is given by (3.21). If $|u| > |t|$, then $w \geq |u|$, $|v| \geq |t|$; for $|v| = |t|$ $(=1)$, T is then given by (3.20) with $k > 3$, and for $|v| > |t|$, T is given by (3.18) if $t \neq 0$, and by (3.19) if $t = 0$.

This covers all ppssibilities.

3.3. Chains of I-reduced Forms Equivalent to a given I-reduced Form

An a-chain $\{a_n\}$, $(-\infty < n < \infty)$ of F such that

$$R_1 = [a_0, a_{-1}, a_{-2}, \ldots], \quad R_2 = [a_1, a_2, a_3, \ldots]$$

will be called an <u>a-chain from F</u>, and the corresponding chain $\{f_n\}$ of I-reduced forms will be described as <u>from</u> F.

We now turn to the problem of determining all the I-reduced forms equivalent to a given I-reduced form f. We note first that there exist forms f for which it is not possible to obtain all the I-reduced forms equivalent to f by taking all the forms in all the chains from f. For example, in section 5.5 we shall consider the form

$$g = (1, \sqrt{5}, -1)$$

with roots

$$r_1 = -r_2,$$
$$r_2 = \frac{3+\sqrt{5}}{2} = [3, r_2] = [2, -2, -r_2].$$

The equivalent form

$$G = g \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = (1,\ 2 + \sqrt{5},\ \sqrt{5})$$

has roots

$$R_1 = \frac{1 + \sqrt{5}}{2}, \quad R_2 = \frac{5 + \sqrt{5}}{2}.$$

Since every form in any chain from g has one root whose modulus is

$$\frac{3 + \sqrt{5}}{2} \ (= r_2),$$

it is clear that G cannot belong to any chain of I-reduced forms which contains g.

In this section we prove the following theorem.

Theorem 3.3. Let $f = (a,b,c)$ $(b > 0)$ be a Gauss-reduced form given by (3.6) (so that $r_1 < -1$, $r_2 > 1$ — see sect. 2.4), and let $F = (A,B,C)$ $(B > 0)$ be an I-reduced form which is properly equivalent to f under the non-trivial linear transformation T given by (3.7). Then any chain of I-reduced forms which contains F must contain at least one of the three forms

$$f = (a,b,c),$$

$$f \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = (a,\ 2a + b,\ a + b + c), \tag{3.22}$$

$$f \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = (a - b + c,\ b - 2c,\ c). \tag{3.23}$$

Since f is given by (3.6) and is Gauss-reduced, we have

$$a > 0,\ b > 0,\ c > 0,$$

and so, by Theorem 3.1,

$$a + b + c > 0, \quad a - b + c < 0.$$

Thus

$$2a + b + c + (2a + b) > a + b + c > 0,$$

while

$$2a + b + c - (2a + b) = c < 0,$$

so that, by Theorem 3.1, the form (3.22) is always I-reduced when f is Gauss-reduced. Similarly, the form (3.23) is always I-reduced when f is Gauss-reduced.

If F is equivalent to f under a non-trivial transformation

$$\begin{bmatrix} t & u \\ v & w \end{bmatrix}$$

for which $tw - uv = -1$, then F is properly equivalent to the reverse of f under the non-trivial transformation

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} t & u \\ v & w \end{bmatrix} .$$

Hence we can include the case of improper equivalence by replacing 'f' by 'the reverse of f' in Theorem 3.3.

Thus Theorem 3.3 means that if $f = (a,b,c)(b > 0)$ is a Gauss-reduced form, then we can obtain all the chains of I-reduced forms equivalent to f by taking all the chains from f and from the two forms (3.22) and (3.23). Since there is at least one Gauss-reduced form equivalent to any indefinite binary quadratic form g which does not represent zero, it follows that we can obtain all the chains of I-reduced forms equivalent to g by taking all the a-chains from at most three forms equivalent to g. This enables us to apply the method of section 2.5 to sets of forms whose coefficients depend on a parameter for which the number of equivalent I-reduced forms is unbounded (e.g. the forms g_n discussed in Ch.4), as well as to forms which have an infinite number of equivalent I-reduced forms (e.g. the form g discussed in sect. 5.5).

We shall say that the a-chain $\{a_n\}$ (or, equivalently, the corresponding chain of I-reduced forms) from F <u>leads forwards</u> to f if, for some n,

$$R_2 = [a_1, a_2, \dots, a_n, r_2] ,$$
$$r_1 = [a_n, a_{n-1}, \dots, a_1, R_1] .$$

We shall say that an a-chain from F <u>leads backwards</u> to f if it leads forwards from f to F. We shall say that all a-chains from F <u>lead forwards without choice</u>* to f if, for some k > 0, either

$$r_1 = [2_k, R_1], \quad R_2 = [2_k, r_2], \quad r_2 > 0,$$

or

$$r_1 = [-2_k, R_1], \quad R_2 = [-2_k, r_2], \quad r_2 < 0,$$

so that R_2 must be the 'tail' of any semi-regular continued

---

*We say 'without choice' because it follows from Defn. 2.1 that, if $1 < \alpha < 2$, then $a_1 = 2$ for <u>every</u> semi-regular continued fraction expansion $\alpha = [a_1, a_2, a_3, \dots]$ of α.

fraction expansion of R_2. Similarly, we shall say that all a-chains from F <u>lead backwards without choice</u> to f if, for some $k > 0$,

$$R_1 = \left[(\pm 2)_k, r_1\right], \quad r_2 = \left[(\pm 2)_k, R_2\right], \ \pm r_1 > 0.$$

We now give a sequence of four lemmas, from which we deduce the proof of Theorem 3.3. In these lemmas we take f as given by (3.6) and do not assume that $r_1 < 0$, $r_2 > 0$. We suppose that $F = fT$, where F is given by (3.8) and T by (3.7), and that the elements of T satisfy (3.12) or (3.13), so that, by Lemma 3.5, T must be one of the matrices (3.16) to (3.21). We show that if T is one of the special matrices (3.16) or (3.17) or if T is of the general type (3.18) then any a-chain from $F = fT$ must lead forwards to a form fU where U is either the matrix

$$\begin{bmatrix} 1 & 0 \\ \pm 1 & 1 \end{bmatrix} \qquad (3.24)$$

or one of the matrices (3.19), (3.20), and (3.21); and we show that if $F = fT$, where T is given by (3.19) or (3.20), then any a-chain from F must lead forwards to a form fU where U is one of the matrices (3.21), (3.24). In the lemmas we prove rather more than this because more precise results will be needed to prove Theorem 3.4.

Lemma 3.6. If $F = fT$, then

(i) if T is the matrix (3.16), then all a-chains from F lead forwards without choice to fU, and all a-chains from fU lead backwards without choice to F, where U is given by (3.24);

(ii) if T is the matrix (3.17), then all a-chains from F lead forwards without choice to fU, where U is the matrix

$$\begin{bmatrix} 1 & \pm 1 \\ \pm 1 & 2 \end{bmatrix} \tag{3.25}$$

(i.e. where U is given by (3.20) with $k = 2$).

Proof. If T is given by (3.16), then by (3.10)

$$R_1 = \frac{\pm 1}{\mp r_1 + 1}, \quad R_2 = \frac{r_2 \mp 1}{\pm r_2},$$

where $\pm r_1 > 0$, $\pm r_2 < 0$ (so that $|R_1| > 1$, $|R_2| > 1$). Thus

$$R_2 = \pm 1 - \frac{1}{r_2} = \left[\pm 2, \frac{r_2}{\pm r_2 + 1}\right]$$

and

$$r_1 \pm 1 = \pm 2 - (-r_1 \pm 1) = [\pm 2, R_1].$$

By (3.10), this proves (i).

We note that $y = |2_k, x|$ if and only if

$$y = \frac{(k+1)x - k}{kx - (k-1)} \tag{3.26}$$

where k is any positive integer.

If T is the matrix (3.17), then by (3.10) and (3.26)

$$R_2 = \frac{kr_2 \pm 1}{\pm(k-1)r_2 + 1} = \left[\pm 2_{k-2}, \frac{2r_2 \pm 1}{\pm r_2 + 1}\right].$$

By Lemmas 3.1 and 3.2, this gives (ii).

Lemma 3.7. If $F = fT$, where T is the matrix (3.18), then every a-chain from F leads forwards to a form fU, where U is one of the matrices (3.19), (3.20).

Proof. By (3.10), we have

$$R_2 = \frac{wr_2 + u}{vr_2 + t} = \frac{w}{v} - \frac{1}{v(vr_2+t)} = \frac{u}{t} + \frac{r_2}{t(vr_2+t)},$$

so that

$$R_2 = \frac{w}{v} + \frac{h}{v} = \frac{u}{t} + \frac{h'}{t} \tag{3.27}$$

where, by (3.18), $|h| < 1$, $|h'| < 1$, and w/v is not integral.

By Definition 2.1, for any semi-regular continued fraction expansion $R_2 = |a_1, a_2, a_3, \ldots|$ of R_2, a_1 is an integer such that $|a_1| \geq 2$ and $|R_2 - a_1| < 1$. For any such a_1, we have

$$R_2 = \left[a_1, \frac{w_1 r_2 + u_1}{v_1 r_2 + t_1}\right],$$

where

$$\pm\begin{bmatrix} t_1 & u_1 \\ v_1 & w_1 \end{bmatrix} = \begin{bmatrix} a_1 t - u & t \\ a_1 v - w & v \end{bmatrix}.$$

It follows from (3.27) that

$$|w/v - a_1| < 1, \quad |u/t - a_1| \leq 1;$$

and without loss of generality $w_1 > 0$. Hence, by Lemma 3.2, every a-chain from F leads forwards to a form fT_1, where

$$T_1 = \begin{bmatrix} t_1 & u_1 \\ v_1 & w_1 \end{bmatrix} \quad (t_1 w_1 - u_1 v_1 = 1),$$

$$|u_1| \geq |t_1|,\ w_1 > |v_1|,\ w_1 > |u_1|,\ u_1 \neq 0,\ v_1 \neq 0,$$

and

$$|t_1| \leq |t|,\ |u_1| < |u|,\ |v_1| < |v|,\ w_1 < w.$$

We use the same type of argument as in the proof of Lemma 3.5. Suppose first that $|u_1| = |t_1|$. If $|v_1| = 1$, T_1 is the matrix (3.20) with $k = 2$; while if $|v_1| > 1$, T is the matrix (3.17), so that, by (ii) of Lemma 3.6, all a-chains from fT_1 lead forwards without choice to fU, where U is the matrix (3.20) with $k = 2$. Suppose now that $|u_1| > |t_1|$. If $|v_1| = |t_1|$, then T_1 is given by (3.20); while if $|v_1| > |t_1|$ and $t_1 = 0$, then T_1 is given by (3.19).

The only other possibility is that

$$|u_1| > |t_1| > 0,\ w_1 > |v_1|,\ w_1 > |u_1|,\ |v_1| > |t_1|.$$

If this is so, T_1 satisfies the same conditions as T, and we can apply the same type of argument again.

Thus we obtain a sequence of matrices

$$T_r = \begin{bmatrix} t_r & u_r \\ v_r & w_r \end{bmatrix}$$

such that every a-chain from the form fT_{r-1} leads forwards either to a form fU (where U is given by (3.19) or (3.20) or to a form fT_r, where

$$|u_r| \geq |t_r|,\ w_r > |v_r|,\ w_r > |u_r|,\ u_r \neq 0,\ v_r \neq 0,$$

and

$$|t_r| \leq |t_{r-1}|, \ |u_r| < |u_{r-1}|, \ |v_r| < |v_{r-1}|, \ w_r < w_{r-1}.$$

It follows that we must eventually reach a T_r for which $|u_r| = |t_r|$ or $|v_r| = |t_r|$ or $t_r = 0$, that is a T_r such that fT_r either is, or leads forwards without choice to, a form of U, where U is given by (3.19) or (3.20).
This proves the lemma.

<u>Lemma 3.8</u>. If $F = fT$, where T is the matrix (3.19) then there is an a-chain from F which leads forwards to f, and every a-chain from F leads forwards to f or to fU, where U is one of the matrices (3.24).

<u>Proof</u>. By (3.10) we have

$$R_2 = \frac{\pm k r_2 - 1}{r_2} = \pm k - \frac{1}{r_2} = [\pm k, r_2].$$

If $\pm k \neq -2$ and $r_2 < 0$, then

$$R_2 = \left[\pm k + 1, \frac{r_2}{r_2 + 1}\right],$$

while if $\pm k \neq 2$ and $r_2 > 0$, then

$$R_2 = \left[\pm k - 1, \frac{r_2}{-r_2 + 1}\right].$$

The lemma now follows from Lemma 3.2.

<u>Lemma 3.9</u>. If $F = fT$, where T is given by (3.20) and if $\pm r_2 > 0$, then there is an a-chain from F which leads forwards to the form fU, where

$$U = \begin{bmatrix} 1 & \pm 1 \\ 0 & 1 \end{bmatrix};$$

and every a-chain from F leads forwards either to fU or to fV, where

$$V = \begin{bmatrix} 1 & 0 \\ \mp 1 & 1 \end{bmatrix}.$$

Proof. By (3.10)

$$R_2 = \frac{kr_2 \pm (k-1)}{\pm r_2 + 1} = \pm k - \frac{1}{r_2 \pm 1} = \left[\pm k, r_2 \pm 1\right] \qquad (\pm r_2 > 0),$$

or

$$R_2 = \left[\pm k \mp 1, \frac{r_2 \pm 1}{\mp r_2}\right] \qquad (\pm k \neq 2, \pm r_2 > 0).$$

The lemma now follows from Lemma 3.1, 3.2, and 3.6(i).

Proof of Theorem 3.3. Let the I-reduced forms f, F satisfy the conditions of Theorem 3.3, so that $r_1 < 0$, $r_2 > 0$. We first note that, since $r_1 < 0$, $r_2 > 0$, it follows from Lemma 3.1 that none of the following forms is I-reduced:

$$f\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad f\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \quad f\begin{bmatrix} 1 & -1 \\ -(k-1) & k \end{bmatrix}, \quad f\begin{bmatrix} 1 & -(k-1) \\ -1 & k \end{bmatrix}$$

(where k is an integer greater than one).

We first suppose that either (3.12) or (3.13) holds. If we exclude possibilities which would give non-I-reduced forms, and use Lemmas 3.5 to 3.9, then we see that either any a-chain from F must lead forwards to f or to one of the forms (3.22), (3.23) or F itself is one of the forms (3.22), (3.23).

By Lemma 3.4, if (3.12) and (3.13) do not hold, then

(3.14) or (3.15) holds. In this case the reverse of F is

$$F\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} = f\begin{bmatrix}t & u\\ v & w\end{bmatrix}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} = f\begin{bmatrix}u & t\\ w & v\end{bmatrix} = f\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}w & v\\ u & t\end{bmatrix},$$

where

$$f\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$$

is the reverse of f. Without loss of generality we may take $t \geq 0$ instead of $w \geq 0$, so that either $t = |u| = 1$, $|v| > |w|$, or $t > |u|$, $|v| \geq |w|$. Then, by an argument exactly similar to that given above, it follows that <u>either</u> every a-chain from the reverse of F must lead forwards to the reverse of f or to one of the forms

$$\begin{aligned} f\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}1 & -1\\ 0 & 1\end{bmatrix} &= f\begin{bmatrix}1 & 0\\ -1 & 1\end{bmatrix}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix},\\ f\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 1 & 1\end{bmatrix} &= f\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}, \end{aligned} \tag{3.28}$$

<u>or</u> the reverse of F is itself one of the forms (3.28), which are the reverses of the forms (3.22), (3.23). This is equivalent to saying that, if (3.12) and (3.13) do not hold, then either every a-chain from F leads backwards to f or to one of the forms (3.22), (3.23), or F is itself one of the forms (3.22), (3.23).

Thus the theorem holds in all cases.

If f and F are given by (3.6) and (3.8) respectively, then there exist integers a_i, $i = 1, \ldots, n$ ($|a_i| \geq 2$), such that

$$R_2 = [a_1, \dots , a_n, r_2],$$
$$r_1 = [a_n, \dots , a_1, R_1].$$

Hence, by Lemma 3.1, F = fT, where T is the matrix

$$\begin{bmatrix} -q_{n-1} & -p_{n-1} \\ q_n & p_n \end{bmatrix} \quad (n \geq 1), \qquad (3.29)$$

and p_n, q_n are given by (2.7).

Thus, if there is an a-chain from f which leads forwards to one of the forms (3.22), (3.23), then f is equivalent to f under one of the transformations

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} T, \quad \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} T, \qquad (3.30)$$

where T is given by (3.29). Similarly, if there is an a-chain from the form (3.22) to the form (3.23), then f is equivalent to f under the transformation

$$\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} T \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \qquad (3.31)$$

where T is given by (3.29). It follows from (2.7) and Lemma 2.4 that, for $n \geq 1$,

$$|p_{n-1} \pm q_{n-1}| \geq 1, \quad |p_n \pm p_{n-1}| \geq 1, \quad |q_n \pm q_{n-1}| \geq 1, \quad |p_{n-1}| \geq 1.$$

Hence the transformations (3.30) and (3.31) are non-trivial.

If we denote the negative of f by $\bar{f}$, then for any transformation

$$\begin{bmatrix} t & u \\ v & w \end{bmatrix} \qquad (tw - uv = 1),$$

the negative of the form

$$f\begin{bmatrix} t & u \\ v & w \end{bmatrix} \tag{3.32}$$

is the form

$$\bar{f}\begin{bmatrix} -t & u \\ v & -w \end{bmatrix} ;$$

and the reverse of the form (3.32) is the form

$$f\begin{bmatrix} t & u \\ v & w \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = f\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} w & v \\ u & t \end{bmatrix}.$$

By using these results and arguing as in the previous paragraph, we see that if there is a chain of I-reduced forms which contains any two of the forms f, (3.22), (3.23), or their negatives, or their reverses, or the reverses of their negatives, then f or f reversed or $\bar{f}$ or $\bar{f}$ reversed is equivalent to f under a non-trivial integral unimodular linear transformation, that is, f has a non-trivial automorph, U, say; if U is of infinite order, then f must be proportional to a form with integral coefficients.

We now turn to forms with integral coefficients, for which we can prove a stronger result than Theorem 3.3. Theorem 3.4, which will be given below, really means that we can obtain all the I-reduced forms equivalent to a given integral form g by taking all the forms which belong to chains of I-reduced forms from any Gauss-reduced form f equivalent to g. However, although we can obtain all forms equivalent to g by starting from such a form f, we cannot always obtain all chains of forms $\{f_n\}$ (or equivalently all a-chains) in this way. For example, in Chapter 4 we shall consider Gauss-reduced forms g_n whose first and second roots are

$$R_1 = -R_2,$$
$$R_2 = [3_n, 2, -2, R_2] = \left[3_n, \frac{5R_2+2}{2R_2+1}\right] = [3_{n+1}, 2, R_2+1].$$

The form

$$g_n\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix} \tag{3.33}$$

has first and second roots

$$\frac{R_1}{R_1+1} = \frac{R_2}{R_2-1}, \quad R_2+1.$$

Since g_n has roots of opposite signs, it cannot belong to the chain of forms from the form (3.33) determined by the periodic semi-regular continued fraction expansion of $R_2 + 1$:

$$R_2 + 1 = [4, 3_n, 2, R_2+1].$$

First we give a lemma which is needed to prove Theorem 3.4, and then we give Theorem 3.4.

Lemma 3.10. Let $f = (a,b,c)$ $(b > 0)$ be an integral Gauss-reduced form given by (3.6); then at least one a-chain from f leads forwards to f.

Proof. Since f is an integral form, it has a non-trivial proper automorph T (see sect. 1.5) such that $f = fT$, where

$$T = \begin{bmatrix}t & u\\ v & w\end{bmatrix},$$

t, u, v, w are integral, $w \geq 0$, and $tw - uv = 1$.

Also, $f = fT^{-1}$, where

$$T^{-1} = \begin{bmatrix} w & -u \\ -v & t \end{bmatrix},$$

so that by Lemma 3.4 we may assume without loss of generality that the elements of T satisfy (3.12) or (3.13).

If T is the matrix (3.16), then, by (3.10),

$$r_1 = \frac{\pm 1}{\mp r_1 + 1}, \quad r_2 = \frac{r_2 \mp 1}{\pm r_2},$$

so that

$$\mp r_1^2 + r_1 \mp 1 = 0, \ \pm r_2^2 - r_2 \pm 1 = 0.$$

This is impossible for real r_1, r_2, so T cannot be given by (3.16).

It now follows from Lemmas 3.5 to 3.9 that there is at least one a-chain from f which leads forwards either to f or to the form (3.22).

Similarly we can use the fact that

$$f\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = fT^{-1}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

where the elements of T^{-1} satisfy (3.14) and (3.15) to show that there is at least one a-chain from f which leads backwards to f or to the form (3.23).

If there is an a-chain from f which leads backwards or forwards to f, the lemma is proved. If not, then there is an a-chain from the form (3.23) which leads forwards to f, and an a-chain from f which leads forwards to the form (3.22). Then one of the following statements must hold for a set of integers a_i, $i = 1, \ldots , n$ $(|a_i| \geq 2)$:

(i) $$r_2 = [a_1, \ldots , a_n, r_2 + 1] \qquad (a_n \neq 2),$$

$$\frac{r_1}{r_1 + 1} = [a_n, \ldots , a_1, r_1];$$

(ii) $$r_2 = [a_1, \ldots , a_k, 2_\ell, r_2 + 1] \quad (k \geq 1, \ell \geq 1),$$
$(a_k \neq 2)$
$$\frac{r_1}{r_1 + 1} = [2_\ell, a_k, \ldots , a_1, r_1];$$

(iii) $$r_2 = [2_k, r_2 + 1] = [2_k, 3, 2_{k-1}, r_2 + 1] \quad (k \geq 1),$$

$$\frac{r_1}{r_1 + 1} = [2_k, r_1] = [2_{k-1}, 3, 2_k, r_1].$$

If (i) holds, then

$$r_2 = \left[a_1, \ldots , a_n - 1, \frac{r_2 + 1}{-r_2}\right] = \left[a_1, \ldots , a_n - 1, -2, \frac{r_2}{-r_2 + 1}\right],$$
$$r_1 - 1 = \left[-2, \frac{-1}{r_1 + 1}\right] = \left[-2, \frac{r_1}{r_1 + 1} \mp 1\right],$$

so that there is an a-chain from f which leads forwards to the form (3.23); as there is an a-chain from the form (3.23) which leads forward to f, this means that there is an a-chain from f which leads forwards to f.

If (ii) holds, then by using (3.26) we can show that

$$r_2 = [a_1, \ldots , a_k - 1, -(\ell + 1), r_2],$$
$$r_1 = \frac{1}{(r_1 + 1)/r_1 - 1} = [-(\ell + 1), a_k - 1, \ldots , a_1, r_1],$$

so that there is a chain from f which leads forwards to f.

If (iii) holds with $k = 1$, then (i) holds, and if (iii) holds with $k > 1$, then (ii) holds. Thus in all cases there is an a-chain from f which leads forwards to f.

Theorem 3.4. Let $f = (a,b,c)$ $(b > 0)$ be an integral Gauss-reduced form given by (3.6) and let $F = (A,B,C)$ $(B > 0)$ be an I-reduced form which is properly equivalent to f under a non-trivial transformation

$$T = \begin{bmatrix} t & u \\ v & w \end{bmatrix},$$

where t, u, v, w are integral, $w \geq 0$, and $tw - uv = 1$. Then there is an a-chain from f which leads either forwards or backwards to F.

Proof. By Lemma 3.10 there exist integers a_i, $i = 1, \ldots, n$ $(|a_i| \geq 2)$, such that

$$r_2 = [a_1, \ldots, a_n, r_2],$$
$$r_1 = [a_n, \ldots, a_1, r_1].$$

As f is Gauss-reduced, we have $r_1 < 0$ and therefore $a_n < 0$, so that also

$$r_2 = \left[a_1, \ldots, a_n - 1, \frac{r_2}{-r_2 + 1}\right],$$
$$r_1 - 1 = [a_n - 1, \ldots, a_1, r_1];$$

thus there is an a-chain from the form (3.23) which leads backwards to f. Similarly, since $r_2 > 0$ and therefore $a_1 > 0$, there is an a-chain from the form (3.22) which leads forwards to f.

If we exclude possibilities which give non-I-reduced forms and use Lemmas 3.5 to 3.9, we see (by arguing in the same way as in the proof of Theorem 3.3) that one of the following statements holds:

(i) all a-chains from F lead forwards without choice to the form (3.23) and all a-chains from this form lead backwards without choice to f (see Lemma 3.6);

(ii) all a-chains from F lead backwards without choice to the form (3.22) and all a-chains from this form lead forwards without choice to F;

(iii) there is an a-chain from F which leads forwards to f or to the form (3.22);

(iv) there is an a-chain from F which leads backwards to f or to the form (3.23).

Since there is an a-chain from the form (3.22) which leads forwards to f, (ii) and (iii) imply the existence of an a-chain from F which leads forwards to f. Similarly, (i) and (iv) imply the existence of an a-chain from F which leads backwards to f.

As with Theorem 3.3, the case of improper equivalnces can easily be included by using the reverse of f.

3.4. A Bound for M(f) in Terms of the Coefficients of f

For any indefinite binary quadratic form $f = (a,b,c)$ we define

$$\begin{aligned}\mu(f) &= \max\left[|a|,\ |c|,\ \min|a \pm b + c|\right] \\ &= \max\left[|a|,\ |c|,\ \lambda(f)\right] \quad \text{(see Defn. 2.2)} \\ &= \max\left[|f(1,0)|,\ |f(0,1)|,\ \min|f(1,\pm 1)|\right]. \quad (3.34)\end{aligned}$$

As an immediate consequence of Theorem 3.3, we have

Theorem 3.5. If $f = (a,b,c)$ is a Gauss-reduced form which does not represent zero, then

$$M(f) \leq \mu(f)/4;$$

equality can occur only when $M(f) = M(f;\ P)$ and $2P \equiv 0 \pmod 1$.

Proof. If we write

$$F_0 = f\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad F_1 = f\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix},$$

then, by (3.22),

$$\lambda(F_0) = \min\left[|c|,\ |4a + 2b + c|\right].$$

Since f is Gauss-reduced, we may suppose $a > 0$, $b > 0$, $c < 0$, so that, by Theorem 3.1,

$$|b| = b > |a + c|.$$

Thus

$$4a + 2b + 2c > 2(a + b + c) > 0,$$

and

$$4a + 2b + c > -c = |c|,$$

so that $\lambda(F_0) = |c|$. Similarly, $\lambda(F_1) = |a|$.

The result now follows from Theorem 3.3 and Lemmas 2.11 and 2.12.

It is convenient to note here the following results, which will be used in the next chapter:

$$M(f;\ \tfrac{1}{2},0) = \frac{1}{4}\inf_n \lambda(f_n) \leq \frac{1}{4}|a|, \tag{3.35}$$

where $\{f_n\}$ is the chain of I-reduced forms corresponding to the even chain from the form F_1 (given by (3.23));

$$M(f;\ 0,\tfrac{1}{2}) = \frac{1}{4}\inf_n \lambda(f_n) \leq \frac{1}{4}|c|, \tag{3.36}$$

where $\{f_n\}$ is the chain of I-reduced forms corresponding to the even chain from the form F_0 (given by (3.22));

$$M(f;\ \tfrac{1}{2},\tfrac{1}{2}) = \frac{1}{4}\inf_n \lambda(f_n) \leq \frac{1}{4}\lambda(f), \tag{3.37}$$

where $\{f_n\}$ is the chain of I-reduced forms corresponding to the even a-chain from f.

Theorem 3.5 is clearly a special case of the following theorem given by Barnes [5]:

(i) If $f = (a,b,c)$ is any indefinite binary quadratic form, then

$$M(f) \leq \mu(f)/4\ ;$$

equality can occur only when $M(f) = M(f;\ P)$ and

$$2P \equiv 0 \pmod 1.$$

Similar bounds for $M(f)$ in terms of the coefficients of f had been obtained by Heinhold [36], Davenport [24], and Inkeri [38].

Inkeri [39] proved the following result:

(ii) If $f = (a,b,c)$ is any indefinite binary quadratic form, then there exists a 'reduced' form $q = (A,B,C)$ which is equivalent to f and for which

$$\mu(q) \leq \mu(f).$$

Inkeri [39] then proved the result (i) for 'reduced' forms, and, by using (ii), showed that the corresponding results of Heinhold [36], Davenport [24], Inkeri [38], and Barnes [5] could be derived from this.

For forms which do not represent zero, Inkeri's 'reduced' forms are Gauss-reduced forms. Hence it follows from (ii) that if f does not represent zero then the best bound for $M(f)$ that can be obtained from (i) by considering forms equivalent to f can also be obtained from Theorem 3.5 by considering only Gauss-reduced forms equivalent to f.

Inkeri [39] showed that the best bound for $M(f)$ obtainable in this way is greater than or equal to $\Delta/4\sqrt{5}$ (where $\Delta = +\sqrt{(b^2 - 4ac)}$), because always

$$\mu(f) \geq \Delta/\sqrt{5}.$$

For the first Markov form (see sect. 1.4), $f = (1,1,-1)$, it is well known that

$$M(f) = \frac{1}{4} = \frac{1}{4}\,\frac{\Delta}{\sqrt{5}} = \frac{1}{4}\,\mu(f).$$

Theorem 3.5 and (i) are best possible in the sense that there exist many forms for which $M(f) = \frac{1}{4}\,\mu(f)$. For example, Davenport [24] showed that if $f = (1,2k,-1)$, where k is a positive integer, then $M(f) = \frac{1}{2}k = \frac{1}{4}\,\mu(f)$. However, in Chapter 4 we shall consider a set of forms $f = (a,b,c)$ for many of which $M(f) = \frac{1}{4}|a|$, where $|a|$ is much smaller than $\mu(f)$, so that for these forms $\mu(f)$ does not give a good bound for $M(f)$.

More recently, Rogers [46] has given a geometrical proof of a more general result from which (i) can be deduced:

(iii) If $f(x,y)$ is a continuous function such that the region $|f(x,y)| \leq K$ has two asymptotes and satisfies certain other conditions, then, for any real (x_0,y_0), there exist $(x,y) \equiv (x_0,y_0) \pmod 1$ such that

$$|f(x,y)| \leq \max\left[|f(\tfrac{1}{2},0)|,\ |f(0,\tfrac{1}{2})|,\ \min|f(\tfrac{1}{2}, \pm\tfrac{1}{2})|\right];$$

equality can occur only when

$$(2x_0,2y_0) \equiv (0,0) \pmod 1.$$

Bambah [2], Chalk [21], Mordell [44], and Bambah and Rogers [4] have proved similar results for differently shaped regions $|f(x,y)| \leq K$.

The results discussed by Inkeri [39] were originally put forward as attempts to sharpen Minkowski's Theorem (Theorem 1.1), and in fact this theorem can be deduced from any of them. In particular, if f is Gauss-reduced, then, by Theorem 3.1,

$$\Delta^2 = b^2 + 4|a||c| > b^2 > (a + c)^2,$$

and

$$\Delta^2 > \Delta^2 - (a - c)^2 = b^2 - (a + c)^2 > 0,$$

so that

$$\Delta > |a + c|, \quad \Delta > |a - c|, \quad \Delta > \min |a \pm b + c|.$$

Hence

$$\Delta > \mu(f),$$

and Theorem 3.5 (with (ii)) implies Minkowski's Theorem for forms which do not represent zero.

CHAPTER 4

A SEQUENCE OF SYMMETRICAL MARKOV FORMS

4.1. Introduction — Definition of the Forms g_n and Statement of Theorem 4.1

Let $\{g_n\}$ $(n \geq 1)$ be the subsequence of the symmetric Markov forms (see sect. 1.4) defined by

$$g_n(x,y) = u_{2n+3}x^2 + v_{2n+3}xy - u_{2n+3}y^2 {}^{*} \quad (n \geq 1), \qquad (4.1)$$

where u_r, $r = 0,1, \ldots$, denote the Fibonacci numbers ($u_0 = 0$, $u_1 = 1$, $u_{r+1} = u_r + u_{r-1}$ for $r \geq 1$), and v_r, $r = 0,1, \ldots$, denote the Lucas numbers ($v_0 = 2$, $v_1 = 1$, $v_{r+1} = v_r + v_{r-1}$ for $r \geq 1$); and let $M(g_n)$, $M_2(g_n)$ denote the first and second inhomogeneous minima of g_n, and $m(g_n)$ the homogeneous minimum of g_n.

In this chapter I prove

---

*The simple and semi-regular continued fraction expansions of a root of this form (see (4.6) and (4.15)) are simply related to the number n; hence we call the form g_n even though this makes the definition of g_n in terms of its coefficients seem rather clumsy.

<u>Theorem 4.1</u>. For $n \geq 11$ the following statements hold:

(i) If $n \not\equiv 0 \pmod 3$, then

$$M(g_n) = \tfrac{1}{4}u_{2n+3} = \tfrac{1}{4}m(g_n)$$

(ii) if $n \equiv 0 \pmod 3$, then

$$M(g_n) = \tfrac{1}{4}(8u_{2n+3} - 3v_{2n+3}) > \tfrac{1}{4}m(g_n),$$

$$M_2(g_n) = \tfrac{1}{4}u_{2n+3}\,.$$

In the next chapter I shall discuss the behaviour of the first few of the forms g_n.

Since $v^2_{2n+3} = 5u^2_{2n+3} - 4$ (see Hardy and Wright [35], § 10.14, for properties of the Fibonacci and Lucas numbers), the discriminant of the form g_n is $9u^2_{2n+3} - 4$, as we should expect for a Markov form. If we write

$$\Delta = +\sqrt{(9u^2_{2n+3} - 4)},$$

$$S = \frac{\Delta + v_{2n+3}}{2u_{2n+3}},$$

then the first and second roots (see (3.6) and the definition which follows it) of g_n are

$$R_1 = -S, \quad R_2 = S.$$

In section 4.2 I obtain the simple continued fraction expansion of S and deduce some results which are needed for the proof of Theorem 4.1; in the course of the discussion I verify that g_n is in fact a Markov form with $m(g_n) = u_{2n+3}$.

In section 4.2 I discuss the semi-regular continued fraction expansion (see Defn. 2.1) of S and obtain some further results needed for the proof of Theorem 4.1. Throughout this chapter

$$\alpha = (a_1, a_2, a_3, \dots)$$

denotes the simple continued fraction expansion of α, while

$$\alpha = [a_1, a_2, a_3, \dots]$$

denotes a semi-regular continued fraction expansion of α.

The proof of Theorem 4.1 depends on the fact that, by Theorem 3.1, since g_n is Gauss-reduced, every chain of I-reduced forms equivalent to g_n must contain at least one of the forms

$$g_n = (u_{2n+3}, v_{2n+3}, -u_{2n+3}),$$

$$g_n \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = (u_{2n+3}, 2u_{2n+3} + v_{2n+3}, v_{2n+3}), \qquad (4.2)$$

$$g_n \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = (-v_{2n+3}, 2u_{2n+3} + v_{2n+3}, -u_{2n+3}). \qquad (4.3)$$

If f is either of the forms (4.2), (4.3), then $\lambda(f) = u_{2n+3}$; it now follows from Lemmas 2.11 and 2.12 that, if $\{a_r\}$ is an a-chain from one of these forms and is <u>not even</u> (i.e. not all a_r are even), then for every corresponding ε-chain

$$M(P) = M(g_n; P) = M(\{a_r\}, \{\varepsilon_r\}) < \tfrac{1}{4}u_{2n+3}.$$

In section 4.4 I shall show that if $\{a_r\}$ is an a-chain from g_n for which not all a_r are even, then for every corresponding ε-chain

$$M(P) < \tfrac{1}{4}u_{2n+3}.$$

This, with the results of sections 4.2 and 4.3, will complete the proof of Theorem 4.1.

4.2. The Simple Continued Fraction Expansion of a Root of g_n

Since $t = 3u_{2n+3}$, $u = 1$ satisfy $t^2 - \Delta^2 u^2 = 4$, the transformation

$$T = \begin{bmatrix} \frac{1}{2}(3u_{2n+3} - v_{2n+3}) & u_{2n+3} \\ u_{2n+3} & \frac{1}{2}(3u_{2n+3} + v_{2n+3}) \end{bmatrix}$$

is a proper automorph of g_n (see Dickson[32],§69). From the relationships between the Fibonacci and the Lucas numbers we obtain

$$v_{2n+3} = u_{2n+4} + u_{2n+2} = u_{2n+3} + 2u_{2n+2}\,. \qquad (4.4)$$

Hence

$$T = \begin{bmatrix} u_{2n+1} & u_{2n+3} \\ u_{2n+3} & u_{2n+5} \end{bmatrix}.$$

Since $R_2 = S$, we have, by Lemma 3.1,

$$S = \frac{u_{2n+5}S + u_{2n+3}}{u_{2n+3}S + u_{2n+1}}\,. \qquad (4.5)$$

Also we have

$$\frac{u_{2n+5}}{u_{2n+3}} = (2, 1_{2n+2}) = (2, 1_{2n}, 2), \qquad \frac{u_{2n+3}}{u_{2n+1}} = (2, 1_{2n}).$$

It follows that

$$S = (2, 1_{2n}, 2, S), \qquad (4.6)$$

so that g_n is in fact a Markov form (see Dickson [33], Ch. VII).

Lemma 4.1. For integral $(x,y) \neq (0,0)$ we have

$$|g_n(x,y)| \geq u_{2n+3};$$

and if in addition $|g_n(x,y)| \neq u_{2n+3}$, then

$$|g_n(x,y)| \geq 8u_{2n+3} - 3v_{2n+3}.$$

Proof. It follows from Lagrange's Theorem (Dickson [32], Th. 85) that all the values of $|f_n(x,y)|$ less than $\Delta/2$ for coprime integers (x,y) are given by the set of values of Δ/z, where

$$z \in [(1_r,2,S) + (0,1_{2n-r},2,S);\ r = 0,1,\ \ldots\ ,2n]. \quad (4.7)$$

If $a < b$, then $(1,1,a) < (1,1,b)$, and if $a > 2$, then $(1,1,a) < a$. Thus, if $a > 2$, we have

$$(1_{2r+2},a) < (1_{2r},a) < \ldots < (1,1,a) < a,$$

and

$$(0,1_{2r+2},a) > (0,1_{2r},a) > \ldots > (0,1,1,a) > (0,a).$$

Hence, for $n-1 \geq r \geq 0$,

$$(1_{2r+1},2,S) + (0,1_{2n-2r-1},2,S) = (0,1_{2r},2,S) + (1_{2n-2r},2,S)$$
$$\leq (0,1_{2n-2},2,S) + (1,1,2,S); \quad (4.8)$$

and, for $n \geq r \geq 1$,

$$(1_{2r},2,S) + (0,1_{2n-2r},2,S) \leq (1,1,2,S) + (0,1_{2n-2},2,S); \quad (4.9)$$

and

$$(1,1,2,S) + (0,1_{2n-2},S) < (2,S) + (0,1_{2n},2,S). \quad (4.10)$$

Also

$$(2,S) + (0,1_{2n},2,S) = S + \frac{1}{S} = \Delta/u_{2n+3} \quad (4.11)$$

and

$$(1,1,2,S) + (0,1_{2n-2},2,S) = (0,1,2,S) + (1_{2n-1},2,S)$$
$$= \frac{2S+1}{3S+1} + \frac{2-S}{S-3}$$
$$= \Delta/(8u_{2n+3} - 3v_{2n+3}) \quad (4.12)$$

It follows from (4.8) to (4.11) that if z satisfies (4.7) then always

$$z \leq \Delta/u_{2n+3},$$

and from (4.12) that if $z \neq \Delta/u_{2n+3}$, then

$$z \leq \Delta/(8u_{2n+3} - 3v_{2n+3}).$$

This proves the lemma.

Since $g_n(1,0) = u_{2n+3}$, we have the following

Corollary. $m(g_n) = u_{2n+3}$.

Lemma 4.2. If $n \equiv 0 \pmod 3$ and x,y are both odd integers, then

$$|g_n(x,y)| \geq 8u_{2n+3} - 3v_{2n+3}.$$

Proof. We note first that $g_n(3,1) = 8u_{2n+3} - 3v_{2n+3}$, so that equality is possible. By Lemma 4.1, it is now sufficient to show that if $n \equiv 0 \pmod 3$, and x,y are both odd, then $|g_n(x,y)| \neq u_{2n+3}$. We write $n = 3\ell$, $x = 2X+1$, $y = 2Y+1$, and use the facts that $u_r | u_{rs}$ for every s, and that u_r, u_s are coprime if r,s are coprime.

By (4.4),

$$v_{2n+3} = 4u_{2n+1} + 3u_{2n} = 4u_{6\ell+1} + 3u_{6\ell};$$

also, $8 = u_6$, so that $8 | u_{6\ell}$. Therefore

$$4 | v_{2n+3}. \qquad (4.13)$$

From (4.4) we have

$$u_{2n+3} = v_{2n+3} - 2u_{2n+2} = v_{2n+3} - 2u_{6\ell+2}.$$

Since $2 = u_3$, $2 \nmid u_{6\ell+2}$; it follows that

$$2 | u_{2n+3}, \quad 4 \nmid u_{2n+3}. \qquad (4.14)$$

Now

$$g_n(2X+1, 2Y+1) = u_{2n+3}\{(2X+1)^2 - (2Y+1)^2\} + v_{2n+3}(2X+1)(2Y+1),$$

and $2 | \{(2X+1)^2 - (2Y+1)^2\}$. Hence, by (4.13) and (4.14) $4 | g_n(2X+1, 2Y+1)$, so that

$$g_n(2X+1, 2Y+1) \neq \pm u_{2n+3}.$$

This proves the lemma.

The following lemma is an immediate deduction from Lemmas 4.1 and 4.2 and the relations (3.35) and (3.36).

<u>Lemma 4.3</u>. For all n

$$M(g_n;\ \tfrac{1}{2}, \tfrac{1}{2}) \geq M(g_n;\ \tfrac{1}{2}, 0) = M(g_n;\ 0, \tfrac{1}{2}) = \tfrac{1}{4}u_{2n+3};$$

and if $n \equiv 0 \pmod 3$ then

$$M(g_n;\ \tfrac{1}{2}, \tfrac{1}{2}) = \tfrac{1}{4}(8u_{2n+3} - 3v_{2n+3}).$$

4.3. The Semi-regular Continued Fraction Expansions of a Root of g_n

We have

$$\frac{u_{2n+5}}{u_{2n+3}} = [3_{n+1},2] = [3_n,2,-2], \quad \frac{u_{2n+3}}{u_{2n+1}} = [3_n,2],$$

so that, by (4.5)

$$S = [3_n,2,-2,S]. \tag{4.15}$$

It follows from the last paragraph of section 4.1 that

$$M(P) = M(\{a_r\},\{\varepsilon_r\}) \leq \tfrac{1}{4}u_{2n+3}$$

for all a-chains (and all corresponding ε-chains) except possibly for those a-chains from g_n which do not lead backwards or forwards to one of the forms (4.2),(4.3), and which we shall call permissible. By Lemma 3.1, the roots of the forms (4.2) and (4.3) are, respectively,

$$-S/(-S+1), \quad S+1, \tag{4.16}$$

and

$$-(S+1), \quad S/(-S+1); \tag{4.17}$$

since g_n has integral coefficients, any form which has one of the numbers (4.16), (4.17) as a root must be one of the forms (4.2), (4.3). Hence permissible a-chains are determined by semi-regular continued fraction expansions of $R_1 = -S$, $R_2 = S$ which are not, for any n, of the form

$$[a_1,a_2, \ldots ,a_n,z],$$

where z is any of the numbers (4.16), (4.17); we shall call such expansions of $\pm S$ <u>permissible expansions</u>.

We note the following results

$$[-2,S] = [-3,S/(-S+1)], \qquad [2,-2,S] = [3,2,S+1], \quad (4.18)$$

$$[3,2,-2,S] = [2,-2,-3,S] = [2,-2,-4,S/(-S+1)], \quad (4.19)$$

$$[3,3,2,-2,S] = [2,-2,-3,-3,S] = [2,-2,-2,2,2,S+1]. \quad (4.20)$$

Also, for any z such that $|z| > 1$, we have

$$[3,3,z] = \left[2,-2,\frac{2z-1}{-z+1}\right];$$

and from this we deduce that, for $k \geq 0$,

$$[3_{k+3},2,-2,S] = [2,-2,-3_{k+3},S], \quad (4.21)$$

$$[-3_{k+3},S] = [-2,2,3_k,2,-2,S], \quad (4.22)$$

$$[3_{k+3},2,-2,S] = [2,-2,-2,2,3_k,2,-2,S]. \quad (4.23)$$

We can now prove

<u>Lemma 4.4</u>. If $n \not\equiv 0 \pmod 3$, then

$$M(g_n;\ \tfrac{1}{2},\tfrac{1}{2}) = \tfrac{1}{4}u_{2n+3}.$$

<u>Proof</u>. By (4.19) and (4.23), if $n \equiv 1 \pmod 3$, the even a-chain from g_n is determined by the expansion

$$S = \left[[2,-2,-2,2]_k,2,-2,-4,S/(-S+1)\right],$$

and so is not permissible. Similarly, by (4.20) and (4.23), if $n \equiv 2 \pmod 3$, the even a-chain from g_n is not permissible. The lemma now follows from Lemma 4.3. (We note that, by (4.23), if $n \equiv 0 \pmod 3$, the even a-chain from g_n is permissible, which explains why $M(g_n; \frac{1}{2},\frac{1}{2})$ is larger in this case.)

As there are infinitely many semi-regular continued fraction expansions of any given number, we need a notation which indicates which particular expansion we are using; therefore we write

$$\alpha \equiv [a_0, a_1, \ \ldots \ , a_r, z],$$

when we mean that $\alpha = [a_0, a_1, \ \ldots \ , a_r, z]$ <u>and</u> that we are choosing expansions of α whose first $r + 1$ partial quotients are $a_0, a_1, \ \ldots \ , a_r$. We note that, if $\alpha = [\pm 2, z]$ and $|\alpha| < 2$, then $\alpha \equiv [\pm 2, z]$ and no expansion of α can begin in any other way.

<u>Lemma 4.5</u>. Let $\{a_r\}$ be an a-chain from g_n which is not even. Then $\{a_r\}$ (or its negative or its reverse or its negative reversed) contains a subchain determined by pairs of expansions of the roots $R_1 = -S$, $R_2 = S$ of g_n which begin in one of the following ways:

(i) $-S$ arbitrary, $S \equiv [3_k, 2, -2, y]$,

where $y = [-3_{n-k}, S]$, $3 \le k \le n$;

(ii) $-S$ arbitrary, $S \equiv [3, 3, 2, -2, y]$,

where $y = [-3_{n-2}, S]$;

(iii) $-S \equiv [-3, -2, 2, -y]$, $S \equiv [3, 2, -2, y]$,

where $y = [-3_{n-1}, S]$;

(iv) $-S \equiv [-2, 2, x]$, $S \equiv [3, 2, -2, y]$,

where $x = [3_n, -S]$, $y = [-3_{n-1}, S]$;

(v) $-S \equiv [-2, 2, x]$, $S \equiv [[2, -2, -2, 2]_k, 2, -2, -3, y]$,

where $x = [3_n, -S]$, $y = [-3_\ell, S]$, $3k+\ell+1 = n$, $k \ge 0$;

(vi) $-S \equiv [-2, 2, x]$, $S \equiv [2, -2, [-2, 2, 2, -2]_k, -2, 2, 3, y]$,

where $x = [3_n, -S]$, $y = [3_\ell, 2, -2, S]$, $3k+\ell+1 = n$, $k \ge 0$.

<u>Proof</u>. By (4.18), $[-2, S]$ and $[2, -2, S]$ have no permissible alternative expansions; it now follows from (4.15) and equations (4.19) to (4.23) that any permissible expansion of S must begin in one of the following ways.

$S \equiv [3_k, 2, -2, y]$,

where $y = [-3_{n-k}, S]$, $0 < k \le n$;

$S \equiv [[2, -2, -2, 2]_k, 3, y]$,

where $y = [3_\ell, 2, -2, S]$, $3k+\ell+1 = n$, $k > 0$;

$S \equiv [[2, -2, -2, 2]_k, 2, -2, -3, y]$,

where $y = [-3_\ell, S]$, $3k+\ell+1 = n$, $k \ge 0$;

$$S \equiv [2,-2,y],$$

Where $$y = [-3_{n-2}, \bar{S}].$$

(We note that the last expansion includes the two previous ones as special cases, and that of course many expansions which are not permissible may begin in these ways also). Lemma 4.5 now follows from the symmetry of g_n and the fact that the a-chains are assumed not to be even.

4.4 A Sequence of Lemmas leading to the Proof of Theorem 4.1

Theorem 4.1 now follows immediately from Lemmas 4.3 and 4.4 and the following lemma.

Lemma 4.6. If $n \geq 11$ and $\{a_r\}$ is a permissible a-chain from g_n which is not even, then for every corresponding ε-chain

$$M(P) = M(\{a_r\},\{\varepsilon_r\}) < \tfrac{1}{4}u_{2n+3}.$$

In Lemmas 4.7 to 4.15 we prove that, if $n \geq 11$, then for a-chains which contain certain subchains and for certain corresponding ε-chains, we have, for some r,

$$\pi_r < \Delta/3 \qquad \text{(see (2.30), (2.32), (2.33)), so that}$$

so that

$$M(P) \leq \tfrac{1}{4}\pi_r < \tfrac{1}{4}\tfrac{\Delta}{3} = \tfrac{1}{4}\sqrt{(u^2_{2n+3} - \tfrac{4}{9})} < \tfrac{1}{4}u_{2n+3}.$$

We then prove Lemma 4.6 by using Lemma 4.5 to show that the a-chains and ε-chains considered in Lemmas 4.7 to 4.15 include all permissible a-chains from g_n which are not even, and all corresponding ε-chains.

In the proofs of Lemmas 4.7 to 4.15 we use the notation and results of section 2.5. We introduce the following notation: by a <u>chain pair</u>

$$\begin{array}{l} \ldots\ ,p,q,\underline{r},s,t,u,\ \ldots \\ \ldots\ ,a,b,\underline{c},d,e,f,\ \ldots \end{array}$$

we mean an a-chain $\{a_r\}$ such that $a_1 = r$, $a_2 = s$, $a_0 = q$, $a_{-1} = p$, ... , with a corresponding ε-chain $\{\varepsilon_r\}$ such that $\varepsilon_0 = c$, $\varepsilon_1 = d$, $\varepsilon_{-1} = b$, $\varepsilon_{-2} = a$, ... If in addition the values of one θ and one ϕ are given, then θ_0, ϕ_0 are determined and $\{a_r\}$ is an a-chain from f_0 which contains the subchain determined by the pair of expansions

$$\theta_0 \equiv \left[q,p,\theta_{-2}\right], \quad \phi_0 \equiv \left[r,s,t,u,\phi_4\right],$$

and hence also $\theta_{-1}, \theta_{-2}, \phi_1, \phi_2$, ... are determined. Sometimes, for the sake of clarity, the values of two θ's or two ϕ's are given, though only one of each is needed to determine the subchain.

By (4.5),

$$\begin{aligned} S &= \frac{u_{2n+5}}{u_{2n+3}} - \frac{1}{u_{2n+3}(u_{2n+3}S + u_{2n+1})} \\ &> \frac{u_{2n+5}}{u_{2n+3}} - \frac{1}{u_{2n+3}(2u_{2n+3} + u_{2n+1})} . \end{aligned}$$

Hence, for $n \geq 4$, we have

$$2.61803 < S < 2.61804.$$

We give in Table 1 some numerical information which is valid for $n \geq 4$ and which is needed for the proofs of Lemmas 4.7 to 4.15. As we shall want to use this information for obtaining inequalities, we adopt the following convention.

If $\alpha = \alpha(x)$ (where x is given in the first column) is one of the numbers tabulated, then the value corresponding to α given in the table will be

$$a_o . a_1 a_2 a_3 a_4,$$

where

$$a_o . a_1 a_2 a_3 a_4 < \alpha < a_o . a_1 a_2 a_3 a_4 + 10^{-4} \text{ if } \alpha > 0,$$

and

$$a_o . a_1 a_2 a_3 a_4 > \alpha > a_o . a_1 a_2 a_3 a_4 - 10^{-4} \text{ if } \alpha < 0.$$

TABLE 1

x	x	1/x	$[-2,x]$	$1/[-2,x]$	$[2,-2,x]$	$1/[2,-2,x]$
S	2.6180	0.3819	-2.3819	-0.4198	2.4198	0.4132
$[-3,S]$	-3.3819	-0.2956	-1.7043	-0.5867	2.5867	0.3865
$[-3_2,S]$	-2.7043	-0.3697	-1.6302	-0.6134	2.6134	0.3826
$[-3_3,S]$	-2.6302	-0.3801	-1.6198	-0.6173	2.6173	0.3820
$[-3_4,S]$	-2.6198	-0.3817	-1.6183	-0.6179	2.6179	0.3819
$[-3_5,S]$	-2.6183	-0.3819	-1.6180	-0.6180	2.6180	0.3819
$[-3_k,S]$ $(k \geq 6)$	-2.6180	-0.3819				

If $\alpha = [a_o, a_1, a_2, \ldots]$, then $-\alpha = [-a_o, -a_1, -a_2, \ldots]$; and by (4.19), (4.20), and (4.21) we have

$$[3_k, 2, -2, -3_m, S] = [2, -2, -3_{k+m}, S] \qquad (k \geq 0,\ m \geq 0,\ k+m \leq n).$$

Hence it is sufficient to tabulate $[2,-2,-3_k,S]$.

All inequalities given in the course of the proofs of Lemmas 4.7 to 4.15 are strict inequalities.

Lemma 4.7. If $n \geq 11$, then $\pi_o < \Delta/3$ for the chain pairs

$$\begin{matrix} \ldots, 3, 3, \underline{3}, 2, -2, \ldots \\ \ldots, 1, 1, \underline{1}, 0, 0, \ldots \end{matrix} \quad \text{and} \quad \begin{matrix} \ldots, 3, 3, \underline{3}, 2, -2, \ldots \\ \ldots, \pm 1, -1, \underline{1}, 0, 0, \ldots \end{matrix},$$

where

$$\theta_{-2} = [3_m, -S] \quad (m \geq 0),$$
$$\phi_3 = [-3_k, S] \quad (k \geq 0).$$

Proof.

$$\tau_o = 1 + \left\| \frac{1}{\phi_1\phi_2}\left(1 - \frac{1}{|\phi_3|}\right) \right\|.$$

For $k = 0$,

$$\tau_o = 1 + \|.414 \times .420 \times .619\|. \qquad (4.24)$$

For $k \geq 2$,

$$\tau_o = 1 + \|.3827 \times .6181 \times .6303\|.$$

For $k = 1$, we must have $\phi_3 \equiv [-3, S]$ as no other expansion of ϕ_3 is permissible, so that ϕ_3, ϕ_4 are opposite in sign, and, by Lemma 2.10,

$$\begin{aligned} \tau_o &= 1 + \left\| \frac{1}{\phi_1\phi_2}\left(1 - \frac{1}{|\phi_3|} - \frac{2}{|\phi_3\phi_4|}\right) \right\| \\ &= 1 + \|.387 \times .587 \times (1 - .295 - .295 \times .762)\| \\ &= 1 + \|.387 \times .587 \times .481\|. \end{aligned} \qquad (4.25)$$

Thus for all k

$$\tau_o = 1 + \|.1491\| > .8509. \qquad (4.26)$$

Also

$$2.586 < \phi_o < 2.6181. \qquad (4.27)$$

Hence

$$|1 - \phi_o + \tau_o| < .7672. \tag{4.28}$$

If $\varepsilon_{-1} = 1$, $\sigma_o < 0$, and if $\varepsilon_{-1} = \varepsilon_{-2} = 1$, then

$$\sigma_o = 1 - \frac{1}{\theta_{-1}} + \left\| \frac{1}{\theta_{-1}} \left(1 - \frac{1}{|\theta_{-2}|}\right) \right\|.$$

Hence in either case,

$$\sigma_o < 1 - \frac{1}{\theta_{-1}} + \left| \frac{1}{\theta_{-1}} \left(1 - \frac{1}{|\theta_{-2}|}\right) \right|.$$

For $m = 0$,

$$\sigma_o < 1 - .295 + .296 \times .6181,$$

$$2.704 < \theta_o < 2.705.$$

Therefore

$$\left.\begin{aligned} &|-1 + \theta_o + \sigma_o| < 2.593, \\ &3/|\theta_o\phi_o - 1| < 3/(2.704 \times 2.586 - 1) < .501 \end{aligned}\right\} \tag{4.29}$$

For $m \geq 2$,

$$\sigma_o < 1 - .380 + .3820 \times .631$$

$$2.618 < \theta_o < 2.620.$$

Therefore

$$\left.\begin{aligned} &|- 1 + \theta_o + \sigma_o| < 2.482, \\ &3/|\theta_o\phi_o - 1| < 3/(2.618 \times 2.586 - 1) < .520 \end{aligned}\right\} \tag{4.30}$$

For $m = 1$, by the same type of argument as that used to get (4.25),

$$\begin{aligned} \sigma_o &< 1 - \frac{1}{\theta_{-1}} + \left| \frac{1}{\theta_{-1}} \left(1 - \frac{1}{|\theta_{-2}|} - \frac{2}{|\theta_{-2}\theta_{-3}|}\right) \right| \\ &< 1 - .369 + .370 \times .481, \end{aligned}$$

$$2.630 < \theta_o < 2.631.$$

Therefore

$$\left.\begin{array}{l} |-1+\theta_o+\sigma_o| < 2.440 \\ 3/|\theta_o\phi_o - 1| < .520 \end{array}\right\} \qquad (4.31)$$

From (4.28), (4.29), (4.30), and (4.31), it now follows that, for all k,m,

$$3\pi_o/\Delta < .7672 \times 1.2991 < .997 < 1.$$

Lemma 4.8. If $n \geq 11$, then $\pi_{-1} < \Delta/3$ for the chain pair

$$\begin{array}{l} \ldots, 3, 3, \underline{3}, 2, -2, \ldots \\ \ldots, -1, 1, \underline{1}, 0, 0, \ldots \end{array}$$

where

$$\theta_{-2} = [3_m, -S] \quad (m \geq 0),$$
$$\phi_3 = [-3_k, S] \quad (k \geq 0).$$

Proof.

$$\tau_{-1} = 1 - \frac{1}{\phi_o}\tau_o.$$

Hence, using (4.26) and (4.27),

$$\tau_{-1} > .555,$$

$$2.613 < \phi_{-1} < 2.6181.$$

Therefore

$$|1 - \phi_{-1} + \tau_{-1}| < 1.0631. \qquad (4.32)$$

$$\sigma_{-1} = -1 + \left\|1 - \frac{1}{|\theta_{-2}|}\right\|.$$

For $m \geq 1$,

$$\sigma_{-1} = -1 + \|.705\|,$$

$$2.618 < \theta_{-1} < 2.705.$$

Therefore

$$\left.\begin{aligned} &|-1 + \theta_{-1} + \sigma_{-1}| < 1.410, \\ &3/|\theta_{-1}\phi_{-1} - 1| < .520. \end{aligned}\right\} \qquad (4.33)$$

For m = 0,

$$\sigma_{-1} = -1 + \|.6181\|,$$

$$3.3819 < \theta_{-1} < 3.3820.$$

Therefore

$$\left.\begin{aligned} &|-1 + \theta_{-1} + \sigma_{-1}| < 2.0001, \\ &3/|\theta_{-1}\phi_{-1} - 1| < 3/(3.3819 \times 2.613 - 1) < .383 \end{aligned}\right\} \qquad (4.34)$$

It now follows from (4.32), (4.33), and (4.34) that

$$3\pi_{-1}/\Delta < .815 < 1.$$

<u>Lemma 4.9</u>. If $n \geq 11$, then $\pi_0 < \Delta/3$ for the chain pair

(i) $$\begin{matrix} \ldots, & 3, & \underline{3}, & 2, & -2, & -2, & 2, & \ldots \\ \ldots, & \pm1, & \underline{1}, & 0, & 0, & 0, & 0, & \ldots \end{matrix}$$

where

$$\theta_{-1} = -S, \ \phi_{-1} = S, \quad \phi_5 = [3_k, 2, -2, S] \qquad (k \geq 6),$$

and for the chain pair

(ii) $$\begin{matrix} \ldots, & 3, & \underline{3}, & 2, & -2, & -3, & -3, & \ldots \\ \ldots, & \pm1, & \underline{1}, & 0, & 0, & -1, & \pm1, & \ldots \end{matrix}$$

where

$$\theta_{-1} = -S, \ \phi_{-1} = S, \quad \phi_5 = [-3_m, S] \qquad (m \geq 7).$$

<u>Proof</u>. For the chain pair (i),

$$\tau_o = 1 + \left\| \frac{1}{\phi_1\phi_2\phi_3\phi_4}\left(1 - \frac{1}{|\phi_5|}\right) \right\|$$

$$= 1 + \left\|(.618)^3 \times (.3821)^2\right\| = 1 + \|.035\|. \qquad (4.35)$$

For the chain pair (ii),

$$\tau_o = 1 + \frac{1}{\phi_1\phi_2\phi_3} + \left\|\frac{1}{\phi_1\phi_2\phi_3}\left(1 - \frac{1}{|\phi_4|}\right)\right\| > 1$$

Hence in both cases

$$\tau_o > .965$$

$$2.6180 < \phi_o < 2.6181.$$

Therefore

$$|1 - \phi_o + \tau_o| < .6531. \qquad (4.36)$$

Also, in both cases

$$\sigma_o = \pm 1 + \left\|1 - \frac{1}{|\theta_{-1}|}\right\| < 1.6181,$$

$$3.3819 < \theta_o < 3.3820.$$

Thus

$$\left.\begin{aligned} &|-1 + \theta_o + \sigma_o| < 4.0001, \\ &3/|\theta_o\phi_o - 1| < 3/(3.3819 \times 2.6180 - 1) < .382. \end{aligned}\right\} \quad (4.37)$$

It now follows from (4.36) and (4.37) that, in both cases,

$$3\pi_o/\Delta < .998 < 1.$$

<u>Lemma 4.10</u>. If $n \geq 11$, then $\pi_0 < \Delta/3$ for the chain pair

$$\begin{matrix} \ldots, & 3, & 2, & -2, & \underline{-3}, & -2, & 2, & \ldots \\ \ldots, & \pm 1, & 0, & 0, & \underline{1}, & 0, & 0, & \ldots \end{matrix}$$

where

$$\theta_{-3} = -S \text{ or } [3,-S], \quad \phi_{-3} = S \text{ or } [3_{n-1},2,-2,S],$$

$$\phi_3 = [3_k,2,-2,S], (k \geq 5) \quad (\text{see } (4.22)).$$

<u>Proof</u>. As in (4.28), we get

$$|1 + \phi_0 + \tau_0| < .7672. \tag{4.38}$$

Also

$$\sigma_0 = \left\| \frac{1}{\theta_{-1}} \left(1 - \frac{1}{|\theta_{-2}|}\right) \right\| < .614 \times .705,$$

$$2.586 < |\theta_0| < 2.614.$$

Therefore

$$\left.\begin{aligned} &|1 + \theta_0 + \sigma_0| < 2.047 \\ &3/|\theta_0\phi_0 - 1| < .520 \text{ (as in (4.30))} \end{aligned}\right\} \tag{4.39}$$

It follows from (4.38) and (4.39) that

$$3\pi_0/\Delta < .817 < 1.$$

<u>Lemma 4.11</u>. If $n \geq 11$, then $\pi_0 < \Delta/3$ for the chain pair

$$\begin{matrix} \ldots, & 3, & 2, & -2, & -3, & \underline{-3}, & \ldots \\ \ldots, & 1, & 0, & 0, & 1, & \underline{\pm 1}, & \ldots \end{matrix}$$

where

$$\theta_{-4} = -S \text{ or } [3,-S], \quad \phi_{-4} = S \text{ or } [3_{n-1},2,-2,S],$$

$$\phi_1 = [-3_k, S] \quad (k \geq 8).$$

Proof.

Since

$$\tau_0 = \pm 1 + \left\| 1 - \frac{1}{|\phi_1|} \right\|,$$

$$2.6180 < |\phi_0| < 2.6181,$$

we have

$$|1 + \phi_0 + \tau_0| < 3.2362 \tag{4.40}$$

Also

$$\sigma_0 = 1 - \frac{1}{\theta_{-1}\theta_{-2}\theta_{-3}} + \left\| \frac{1}{\theta_{-1}\theta_{-2}\theta_{-3}} \left(1 - \frac{1}{|\theta_{-4}|}\right) \right\|.$$

Thus if $\theta_{-4} = -S$,

$$\sigma_0 > 1 + .386 \times .586 \times .295 - .387 \times .587 \times .296 \times .619,$$

while if $\theta_{-4} = [3, -S]$,

$$\sigma_0 > 1 + .382 \times .613 \times .369 - .383 \times .614 \times .370 \times .705.$$

Hence in both cases

$$\sigma_0 > 1.022,$$

$$2.613 < |\theta_0| < 2.618.$$

Therefore

$$\left.\begin{aligned} &|1 + \theta_0 + \sigma_0| < .596 \\ &3/|\theta_0\phi_0 - 1| < 3/(2.613 \times 2.618 - 1) < .514 \end{aligned}\right\} \tag{4.41}$$

It now follows from (4.40) and (4.41) that

$$3\pi_0/\Delta < .992 < 1.$$

Lemma 4.12. If $n \geq 11$, then $\pi_0 < \Delta/3$ for the chain pair

$$\begin{array}{l} \ldots, 2, -2, -3, \underline{3}, 2, -2, \ldots \\ \ldots, 0, \ 0, \pm 1, \underline{1}, 0, \ 0, \ldots \end{array}$$

where

$$-\theta_0 = \phi_0 = S, \quad -\theta_{-3} = \phi_3 = [-3_k, S] \quad (k \geq 10).$$

Proof. As in (4.26),

$$\tau_0 = 1 + \| .1491 \|.$$

Similarly,

$$\sigma_0 = \pm 1 + \| .1491 \|.$$

Also

$$2.6180 < |\theta_0| < 2.6181, \quad 2.6180 < \phi_0 < 2.6181. \quad (4.42)$$

Thus, if $\varepsilon_{-1} = 1$, so that $\sigma_0 > 0$,

$$|(-1 + \theta_0 + \sigma_0)(1 - \phi_0 + \tau_0)| < 2.7672 \times .7672,$$

while, if $\varepsilon_{-1} = -1$, so that $\sigma_0 < 0$, (4.43)

$$|(-1 - \theta_0 + \sigma_0)(-1 - \phi_0 + \tau_0)| < .7672 \times 2.7672.$$

By (4.42),

$$3/|\theta_0\phi_0 - 1| < 3/((2.618)^2 + 1) < .382. \quad (4.44)$$

It now follows from (4.43) and (4.44) that

$$3\pi_0/\Delta < .811 < 1.$$

Lemma 4.13. If $n \geq 11$, then $\pi_o < \Delta/3$ for the chain pair

(i)
$$\begin{array}{l} \ldots, 2, -2, \underline{3}, 2, -2, -2, 2, \ldots \\ \ldots, 0, \ 0, \underline{1}, 0, \ 0, \ 0, 0, \ldots \end{array}$$

where

$$-\theta_o = \phi_o = S,$$

$$\theta_{-2} = [3_m, -S] \ (m \geq 11), \qquad \phi_5 = [3_k, 2, -2, S] \ (k \geq 7).$$

and for the chain pair

(ii)
$$\begin{array}{l} \ldots, 2, -2, 3, 2, -2, -3, \underline{-3}, \ldots \\ \ldots, 0, \ 0, 1, 0, -1, -1, \underline{\pm 1}, \ldots \end{array}$$

where

$$-\theta_o = \phi_o = S,$$

$$\theta_{-2} = [3_m, -S] \ (m \geq 11), \qquad \phi_5 = [-3_k, S] \ (k \geq 8).$$

Proof. As in (4.36), in both cases we have

$$|1 - \phi_o + \tau_o| < .6531 \tag{4.45}$$

Also, in both cases,

$$\sigma_o = \left\| \frac{1}{\theta_{-1}} \left(1 - \frac{1}{|\theta_{-2}|}\right) \right\| < .3821,$$

$$2.6180 < \theta_o < 2.6181.$$

Thus

$$\left. \begin{array}{l} |-1 + \theta_o + \sigma_o| < 4.0002, \\ 3/|\theta_o \phi_o - 1| < .382 \ \ \text{(as in (4.44).} \end{array} \right\} \tag{4.46}$$

By (4.45) and (4.46)

$$3\pi_o/\Delta < .9983 < 1.$$

Lemma 4.14. If $n \geq 11$, then $\pi_o < \Delta/3$ for the chain pair

$$\text{(i)}\quad \begin{array}{l} \ldots, 2,-2, [2,-2,-2,2]_k, 2,-2,-3,\underline{-2},2, \ldots \\ \ldots, 0, 0, [0, 0, 0,0]_k, 0, 0, 1, \underline{0},0, \ldots \end{array} \quad (k \geq 0),$$

where

$$-\theta_{-4k-3} = \phi_{-4k-3} = S,$$

$$\theta_{-4k-5} = [3_n, -S], \quad \phi_2 = [3_m, 2, -2, S] \quad (m \geq 0),$$

and for the chain pair

$$\text{(ii)}\quad \begin{array}{l} \ldots, 2,-2, [2,-2,-2,2]_k, 2,-2,-3,\underline{-3},-2,2, \ldots \\ \ldots, 0, 0, [0, 0, 0,0]_k, 0, 0, 1,\underline{\pm 1}, 0,0, \ldots \end{array} \quad (k \geq 0),$$

where

$$-\theta_{-4k-3} = \phi_{-4k-3} = S,$$

$$\theta_{-4k-5} = [3_n, -S], \quad \phi_3 = [3_m, 2, -2, S] \quad (m \geq 0).$$

Proof. For the chain pair (i)

$$|\tau_o| < 1,$$

$$2.6180 < |\phi_o| < 2.6303;$$

and for the chain pair (ii)

$$\tau_o = \pm 1 + \left\| \frac{1}{\phi_1 \phi_2}\left(1 - \frac{1}{\phi_3}\right) \right\| = \pm 1 + \| .382 \times .6303 \times .6181 \|, \quad (4.47)$$

$$2.618 < |\phi_o| < 2.620.$$

Hence in both cases,

$$|1 + \phi_o + \tau_o| < 2.769. \quad (4.48)$$

Also in both cases,

$$\sigma_o = 1 + \left\| \frac{1}{\theta_{-1}\theta_{-2}\theta_{-4k-3}\theta_{-4k-4}} \left(1 - \frac{1}{|\theta_{-4k-5}}\right) \right\|$$

$$= 1 + \| .414 \times .382 \times (.6181)^3 \| = 1 + \| .038 \|, \quad (4.49)$$

$$2.586 < |\theta_o| < 2.6181.$$

Thus

$$\left.\begin{aligned} |1 + \theta_o + \sigma_o| &< 6561, \\ 3/|\theta_o\phi_o - 1| &< .520 \quad (\text{as in } (4.30)). \end{aligned}\right\} \qquad (4.50)$$

It now follows from (4.48) and (4.50) that

$$3\pi_o/\Delta < .945 < 1.$$

<u>Lemma 4.15</u>. If $n \geq 11$, then $\pi_o < \Delta/3$ for the chain pair

$$\text{(i)} \quad \begin{array}{l} \ldots, 2, -2, 2, -2, \left[-2, 2, 2, -2\right]_k, -2, 2, 3, \underline{2}, -2, \ldots \\ \ldots, 0, \ 0, 0, \ 0, \left[0, \ 0, 0, \ 0\right]_k, \ 0, 0, 1, \underline{0}, \ 0, \ldots \end{array} \quad (k \geq 0),$$

where

$$-\theta_{-4k-5} = \phi_{-4k-5} = S,$$

$$\theta_{-4k-7} = [3_n, -S], \qquad \phi_2 = [-3_m, S] \quad (m \geq 3).$$

and for the chain pair

$$\text{(ii)} \quad \begin{array}{l} \ldots, 2, -2, 2, -2, \left[-2, 2, 2, -2\right]_k, -2, 2, 3, \underline{3}, 2, -2, \ldots \\ \ldots, 0, \ 0, 0, \ 0, \left[0, 0, 0, \ 0\right]_k, \ 0, 0, 1, \underline{\pm 1}, 0, \ 0, \ldots \end{array} \quad (k \geq 0),$$

where

$$-\theta_{-4k-5} = \phi_{-4k-5} = S,$$

$$\theta_{-4k-7} = [3_n, -S], \qquad \phi_3 = [-3_m, S] \quad (m \geq 3).$$

<u>Proof</u>. For the chain pair (i)

$$|\tau_o| < 1,$$

and for the chain pair (ii), in a way similar to (4.47), we get

$$\tau_o = \pm 1 + \|.147\|.$$

In both cases,

$$2.6173 < \phi_o < 2.6181,$$

so that

$$|-1 + \phi_o + \tau_o| < 2.766. \qquad (4.51)$$

Also, for both chain pairs, (4.49) holds, and

$$2.6180 < \theta_o < 2.6199.$$

Hence

$$\left.\begin{aligned} &|1 - \theta_o + \sigma_o| < .658, \\ &3/|\theta_o\phi_o - 1| < 3/(2.6180 \times 2.6173 - 1) < 513. \end{aligned}\right\} \qquad (4.52)$$

It now follows from (4.51) and (4.52) that

$$3\pi_o/\Delta < .933 < 1.$$

<u>Proof of Lemma 4.6</u>. By Lemma 2.13 and the discussion following the statement of Lemma 4.6, it is sufficient to show that if $\{a_r\}$ is a permissible a-chain which contains one of the subchains (i) to (vi) of Lemma 4.5, and $\{\varepsilon_r\}$ is a corresponding ε-chain, then $\{a_r\}$ or its negative or its reverse or its negative reversed is one of the a-chains considered in Lemmas 4.7 to 4.15 and $\{\varepsilon_r\}$ or its negative is one of the corresponding ε-chains considered in these lemmas.

The chain pairs of Lemmas 4.7 and 4.8 include all the a-chains containing (i) and all corresponding ε-chains or their negatives.

If $y = [-3_k, S]$, then by (4.22), any semi-regular continued fraction expansion of y must begin in one of the following ways:

$$y \equiv [-2,2,z], \qquad z = [3_{k-3},\ 2,-2,S] \qquad (k \geq 3);(4.53)$$

$$y \equiv [-3,-2,2,z], \qquad z = [3_{k-4},\ 2,-2,S] \qquad (k \geq 4);(4.54)$$

$$y \equiv [-3,-3,z], \qquad z = [-3_{k-2},S] \qquad (k \geq 2).(4.55)$$

Hence the chain pairs of Lemmas 4.9, 4.10, and 4.11 include all a-chains containing (ii) and all corresponding ε-chains or their negatives.

The chain pairs of Lemma 4.12 consist of all a-chains containing (iii) with all corresponding ε-chains or their negatives.

For the subchain (iv), y must be given by one of (4.53), (4.54), and (4.55). Hence the chain pairs of Lemmas 4.10, 4.11, and 4.13 include all a-chains containing (iv) and all corresponding ε-chains or their negatives.

For the subchain (v), y must be given by one of (4.53), (4.54), or (4.55). If y satisfies (4.53) or (4.54), then all a-chains containing (v) and all corresponding ε-chains or their negatives are included in the chain pairs of Lemma 4.14. If y satisfies (4.55), then the chain pairs of Lemmas 4.7 and 4.8 include the reverses of the negatives of all a-chains containing (v) and all corresponding ε-chains or their negatives.

For the subchain (vi), the semi-regular continued fraction expansion of y must begin in one of the following ways (see (4.21)):

$$y \equiv [2,-2,z], \qquad z = [-3_m,S] \qquad (m \geq 0), \qquad (4.56)$$

$$y \equiv [3,2,-2,z], \qquad z = [-3_m,S] \qquad (m \geq 0), \qquad (4.57)$$

$$y \equiv [3_p,2,-2,z], \qquad z = [-3_m,S] \quad (m \geq p \geq 2). \qquad (4.58)$$

If y is given by (4.56) or (4.57) with $m \geq 3$, then all a-chains containing (vi) and all corresponding ε-chains or their negatives are included in the chain pairs of Lemma 4.15. If y satisfies (4.56) or (4.57) with m = 0,1,2, then it follows from (4.18), (4.19), and (4.20) that every permissible a-chain containing (vi) must be the reverse of the negative of an a-chain containing one of the subchains (i) to (v). If y satisfies (4.58), then all a-chains containing (vi) and all corresponding ε-chains are included in the chain pairs of Lemmas 4.7 and 4.8.

This completes the proof of Lemma 4.6 and so of Theorem 4.1.

CHAPTER 5

SOME SPECIAL FORMS

In Chapter 4 I considered the subsequence $\{g_n\}$ $(n \geq 1)$ of the symmetric Markov forms defined by (4.1) and obtained the inhomogeneous minimum of g_n for $n \geq 11$. In sections 5.1 to 5.4 of this chapter I shall discuss the first few of the forms g_n, and then in section 5.5 I shall discuss the form $g = (1, \sqrt{5}, -1)$, which may be regarded as the limiting symmetric Markov form. In this chapter I shall use the notation and results of section 2.5 and of Chapter 4.

5.1. The Early Symmetric Markov Forms

The first two Markov forms F_o, F_1 (see sect. 1.4) do not belong to the sequence $\{g_n\}$ for $n \geq 1$ and were not included in the general discussion of chapter 4 because the continued fraction expansions, both simple and semi-regular, of their roots are rather special; however the coefficients of F_o, F_1 are of the same form as those of the g_n, and so we may write

$$g_{-1}(x,y) = x^2 + xy - y^2 = F_o(x,y)$$

$$g_o(x,y) = 2x^2 + 4xy - 2y^2 = 2F_1(x,y).$$

The inhomogeneous minima of F_o and F_1 have been studied exhaustively by Davenport [24,25] and Varnavides [49]. For the sake of completeness I include the following two theorems.

Theorem 5.1. For the form $g_{-1} = (1,1,-1) = F_o$, we have

$$M(g_{-1}) = \tfrac{1}{4} = \tfrac{1}{4}m(g_{-1}).$$

Proof. (Barnes [5]). Clearly $m(g_{-1}) = 1$, and so

$$M(g_{-1}) \geq \tfrac{1}{4}m(g_{-1}) = \tfrac{1}{4};$$

and by Theorem 3.5,

$$M(g_{-1}) \leq \tfrac{1}{4}\mu(g_{-1}) = \tfrac{1}{4}.$$

Theorem 5.2. For the form $g_o = (2,4,-2) = 2F_1$, we have

$$M(g_o) = 1,$$

$$M_2(g_o) = \tfrac{1}{2} = \tfrac{1}{4}m(g_o).$$

Proof. Clearly $m(g_o) = 2$, and, for integral x,y,

$$|g_o(2x+1,2y+1)| = |2(2x+2y+2)(2x-2y) + 4(2x+1)(2y+1)| \geq 4.$$

Hence

$$M(g_o;\ \tfrac{1}{2},0) = M(g_o;\ 0,\tfrac{1}{2}) = \tfrac{1}{2},$$

$$M(g_o;\ \tfrac{1}{2},\tfrac{1}{2}) = 1.$$

The only I-reduced form equivalent to g_o is $f = (2,8,4)$. Since $\lambda(f) = 2$, it follows from Lemma 2.11 that $M(P) \leq \frac{1}{2}$ for every chain of I-reduced forms containing f and every corresponding ε-chain. The only chain of I-reduced forms, equivalent to g_0 which does not contain f is the one corresponding to the even a-chain from g_0 for which, by (3.37),

$$M(P) = M(g_0;\ \tfrac{1}{2}.\tfrac{1}{2}).$$

The result now follows.

Since $4 = 8\times 2 - 3\times 4$, Theorems 5.1 and 5.2 mean that Theorem 4.1 holds for $n = -1, 0$. In sections 5.2, 5.3, and 5.4 I shall show that Theorem 4.1 holds for $n = 1, 2$, and 3. This strongly suggests that the theorem holds also for $4 \leq n \leq 10$ and so for all $n \geq -1$, but the details of the proof for $4 \leq n \leq 10$ would be very tedious.

The inhomogeneous minimum of the form g_1 was obtained by Davenport [25] by a different method.

All the results of sections 4.1, 4.2, and 4.3 except Lemma 4.5 hold for $n = 1,2,3$; hence, in order to prove Theorem 4.1 for $n = 1,2,3$, it is sufficient to show that if $n = 1,2,3$ and $\{a_r\}$ is a permissible a-chain from g_n which is not even, then for every corresponding ε-chain

$$M(P) = M(\{a_r\},\ \{\varepsilon_r\}) < \tfrac{1}{4}u_{2n+3}.$$

5.2. The Form $g_1 = (5,11,-5) = F_2$

Theorem 5.3. For the form $g_1 = (5,11,-5) = F_2$ we have

$$M(g_1) = \frac{5}{4} = \frac{1}{4}m(g_1).$$

<u>Proof</u>. By (4.15)

$$S = [3,2,-2,S].$$

Hence, by (4.19), the only permissible expansions of S are

$$S \equiv [3,2,-2,S],$$

$$S \equiv [2,-2,-3,S].$$

It now follows from the symmetry of the form g_1, whose roots are $R_1 = -S$, $R_2 = S$, that any permissible a-chain from g_1 must be an arrangement of either or both of the blocks of numbers

$$A = 3,2,-2,$$

$$B = 2,-2,-3, .$$

In Lemmas 5.1 and 5.2 we shall show that, if an a-chain contains BA or AAAA, then, for every corresponding ε-chain, we have, for some r,

$$M(P) \leq \tfrac{1}{4}\pi_r < \tfrac{5}{4}.$$

Since B is the negative of A reversed, this result holds also for a-chains containing BBBB. An arrangement of either or both of A,B which does not contain AAAA or BBBB must contain BA. Thus the a-chains of Lemmas 5.1 and 5.2 include all permissible a-chains from g_1 or their negatives reversed. Hence the proof of Theorem 5.3 will follow from Lemmas 5.1 and 5.2.

For the proofs of Lemmas 5.1 and 5.2 we need the numerical information given in Table 2, in which the same conventions are used as in Table 1.

TABLE 2

x	x	$1/x$	$[-2,x]$	$1/[-2,x]$	$[2,-2,x]$	$1/[2,-2,x]$.
S	2.5866	0.3866	−2.3866	−0.4190	2.4190	0.4133
$[-3,S]$	−3.3866	−0.2952	−1.7047	−0.5866	2.5866	0.3866

$$14.86606 < \Delta < 14.86607$$

The proofs of Lemmas 5.1 and 5.2 and of the other lemmas of this chapter follow exactly the same lines as those of Lemmas 4.7 to 4.15. Therefore less detail will be given in the computations than was given in Chapter 4.

By Lemma 2.13, we can fix the sign of one ε_r in a given ε-chain without loss of generality; we do so without comment in the lemmas of this chapter.

<u>Lemma 5.1</u>. $\pi_o < 5$ for the chain pair

$$\begin{array}{c} \ldots,2,-2,-3,\underline{3},2,-2,\ldots \\ \ldots,0,\ 0,\pm 1,\underline{1},0,\ 0,\ldots \end{array}$$

where

$$\theta_{-3} = -S, \quad \phi_3 = S.$$

<u>Proof</u>. We have

$$\tau_o = 1 + \left\| \frac{1}{\phi_1\phi_2}\left(1 - \frac{1}{|\phi_3|}\right) \right\| = 1 + \|.107\|.$$

Similarly,

$$\sigma_o = \pm 1 + \|.107\|.$$

Also

$$2.586 < \phi_o < 2.587, \quad 2.586 < |\theta_o| < 2.587. \qquad (5.1)$$

Thus, if $\varepsilon_{-1} = 1$ so that $\sigma_0 > 0$,

$$\left.\begin{aligned} &|(-1+\theta_0+\sigma_0)(1-\phi_0+\tau_0)| < 2.694 \times .694; \\ &\text{while, if } \varepsilon_{-1} = -1 \text{ so that } \sigma_0 < 0, \\ &|(-1-\theta_0+\sigma_0)(-1-\phi_0+\tau_0)| < .694 \times 2.694. \end{aligned}\right\} \quad (5.2)$$

$$\Delta/|\theta_0\phi_0 - 1| < 14.867/((2.586)^2 + 1) < 1.935. \quad (5.3)$$

It follows from (5.2) and (5.3) that

$$\pi_0 < 3.7 < 5.$$

<u>Lemma 5.2</u>. $\pi_0 < 5$ for the chain pair

$$\begin{matrix} \ldots\ ,3,2,-2,\ 3,2,-2,\underline{3},2,-2,\ 3,2,-2,\ \ldots \\ \ldots\ ,\pm1,0,\ 0,\pm1,0,\ 0,\underline{1},0,\ 0,\pm1,0,\ 0,\ \ldots \end{matrix}$$

where

$$\theta_{-6} = -S, \qquad \phi_6 = S.$$

<u>Proof</u>. We have

$$\tau_0 = 1 \pm \frac{1}{\phi_1\phi_2\phi_3} + \left\| \frac{1}{\phi_1\phi_2\phi_3\phi_4\phi_5}\left(1 - \frac{1}{|\phi_6|}\right) \right\|$$

$$> 1 - .068 - .008 = .924$$

Since θ_0, ϕ_0 satisfy (5.1), it follows that (5.3) holds and

$$|1 - \phi_0 + \tau_0| < .663 \quad (5.4)$$

Also

$$\sigma_0 = \pm\frac{1}{\theta_{-1}\theta_{-2}} \pm \frac{1}{\theta_{-1}\theta_{-2}\theta_{-3}\theta_{-4}\theta_{-5}} + \left\| \frac{1}{\theta_{-1}\theta_{-2}\theta_{-3}\theta_{-4}\theta_{-5}}\left(1 - \frac{1}{|\theta_{-6}|}\right) \right\|$$

$$< .174 + (.174 \times .068) + (.174 \times .068 \times .614) < .194$$

Hence

$$|-1 + \theta_o + \sigma_o| < 3.781. \tag{5.5}$$

It follows from (5.3),(5.4), and (5.5) that

$$\pi_o < 4.86 < 5.$$

By the argument following the statement of Theorem 5.3, this completes the proof of Theorem 5.3.

5.3. The Form $g_2 = (13,29,-13) = F_3$

Theorem 5.4. For the form $g_2 = (13,29,-13) = F_3$, we have

$$M(g_2) = \frac{13}{4} = \frac{1}{4}m(g_2).$$

Proof. By (4.15),

$$S = [3,3,2,-2,S].$$

Hence, by (4.19) and (4.20), the only permissible expansions of S are

$$S \equiv [3,3,2,-2,S],$$
$$S \equiv [3,2,-2,-3,\bar{S}],$$
$$S \equiv [2,-2,-3,-3,\bar{S}].$$

It now follows from the symmetry of the form g_2, whose roots are $R_1 = S$, $R_2 = S$, that any permissible a-chain from g_2 must be an arrangement of some or all of the three blocks of numbers

$$A = 3,3,2,-2,$$
$$B = 3,2,-2,-3,$$
$$C = 2,-2,-3,-3, \text{.}$$

In Lemmas 5.3 to 5.7 we shall show that, if an a-chain contains AA, BB, AC, BAB, or BCAB, then for every corresponding ϵ-chain,

$$M(P) < \frac{13}{4}.$$

Since C is the negative of A reversed and B is its own negative reversed, this result holds also for a-chains containing CC and BCB. Any arrangement of some or all of A,B, and C which does not contain BB must contain A or C; any arrangement which contains A but none of AA, BB, AC, CC, must contain BAB or BCAB, and similarly any arrangement which contains C but none of AA, BB, AC, CC must contain BCB or BCAB. Thus the chains of Lemmas 5.3 to 5.7 include all permissible a-chains from g_2 or their negatives reversed. Hence the proof of Theorem 5.4 will follow from Lemmas 5.3 to 5.7.

For the proofs of these lemmas we need the numerical information given in Table 3, in which the same conventions are used as in Table 1.

TABLE 3

x	x	$1/x$	$[-2,x]$	$1/[-2,x]$	$[2,-2,x]$	$1/[2,-2,x]$.
S	2.6134	0.3826	-2.3826	-0.4197	2.4197	0.4132
$[-3,S]$	-3.3826	-0.2956	-1.7043	-0.5867	2.5867	0.3865
$[-3_2,S]$	-2.7043	-0.3697	-1.6302	-0.6134	2.6134	0.3826

$$38.94868 < \Delta < 38.94869.$$

<u>Lemma 5.3</u>. $\pi_o < 13$ for the chain pair

$$\begin{array}{l} \ldots,\ 3,\ 3, 2, -2,\ 3, \underline{3}, 2, -2,\ \ldots \\ \ldots, \pm 1, \pm 1, 0,\ 0, \pm 1, \underline{1}, 0,\ 0,\ \ldots \end{array}$$

where

$$-\theta_{-5} = \phi_3 = S.$$

<u>Proof</u>. We have

$$\tau_o = 1 + \left\| \frac{1}{\phi_1} \left(1 - \frac{1}{|\phi_2|} - \frac{2}{|\phi_2 \phi_3|}\right) \right\|$$

$$= 1 + \|.109\| > .891,$$

$$2.586 < \phi_o < 2.587. \tag{5.6}$$

Therefore

$$|1 - \phi_o + \tau_o| < .696. \tag{5.7}$$

Also

$$\sigma_o = \pm 1 + \left\| 1 - \frac{1}{|\theta_{-1}|} - \frac{2}{|\theta_{-1}\theta_{-2}|} \right\|$$

$$= \pm 1 + \|.150\| < 1.150,$$

$$3.382 < \theta_o < 3.383. \tag{5.8}$$

Thus

$$|-1 + \theta_o + \sigma_o| < 3.533, \tag{5.9}$$

$$\Delta / |\theta_o \phi_o - 1| < 38.949/(2.586 \times 3.382 - 1) < 5.03 \tag{5.10}$$

It follows from (5.7), (5.9), and (5.10) that

$$\pi_o < 12.37 < 13.$$

<u>Lemma 5.4</u>. $\pi_o < 13$ for the chain pair

$$\begin{array}{l} \ldots\ ,\ 3,2,-2,-2,\underline{3},2,-2,-3,\ \ldots \\ \ldots\ ,\pm 1,0,\ 0,\pm 1,\underline{1},0,\ 0,\pm 1,\ \ldots \end{array}$$

where

$$-\theta_{-4} = \phi_4 = S.$$

<u>Proof</u>. We have

$$\tau_o = 1 + \left\| 1 - \frac{1}{|\phi_1|} - \frac{2}{|\phi_1\phi_2|} \right\| = 1 + \|.162\|.$$

Similarly,

$$\sigma_o = \pm 1 + \|.162\|.$$

Also

$$2.613 < |\theta_o| < 2.614, \quad 2.613 < \phi_o < 2.614.$$

Thus, if $\varepsilon_{-1} = 1$ so that $\sigma_o > 0$,

$$|(-1 + \theta_o + \sigma_o)(1 - \phi_o + \tau_o)| < 2.776 \times .776;$$

while, if $\varepsilon_{-1} = -1$ so that $\sigma_o < 0$,

$$|(-1 - \theta_o + \sigma_o)(-1 - \phi_o + \tau_o)| < .776 \times 2.776. \tag{5.11}$$

Also

$$\Delta/|\theta_o\phi_o - 1| < 4.977. \tag{5.12}$$

It follows from (5.11) and (5.12) that

$$\pi_o < 10.8 < 13.$$

<u>Lemma 5.5</u>. $\pi_0 < 13$ for the chain pair

$$\begin{matrix} \dots, & 3, \underline{3}, 2, -2, 2, -2, -3, -3, & \dots \\ \dots, & \pm 1, \underline{1}, 0, 0, 0, 0, \pm 1, \pm 1, & \dots \end{matrix}$$

where

$$-\theta_{-1} = \phi_7 = S, \quad \phi_5 = [-3, -3, S].$$

<u>Proof</u>. We have

$$\tau_0 = 1 + \left\| \frac{1}{\phi_1 \phi_2} \left(1 - \frac{1}{|\phi_3|} - \frac{2}{|\phi_3 \phi_4|}\right) \right\|$$

$$= 1 + \| .027 \|.$$

θ_0, ϕ_0 satisfy (5.6) and (5.8); hence (5.10) holds, and

$$|1 - \phi_0 + \tau_0| < .614 \tag{5.13}$$

Also

$$\sigma_0 = \pm 1 + \left\| 1 - \frac{1}{|\theta_{-1}|} \right\| = \pm 1 + \| .618 \|,$$

so that

$$|-1 + \theta_0 + \sigma_0| < 4.001 \tag{5.14}$$

It follows from (5.10), (5.13), and (5.14) that

$$\pi_0 < 12.4 < 13.$$

Lemma 5.6. $\pi_0 < 13$ for the chain pair

(i) $\quad \ldots, 3, 2, -2, -3, 3, \underline{3}, 2, -2, 3, 2, -2, -3, \ldots$
$\quad \ldots, \pm 1, 0, 0, 1, \pm 1, \underline{\pm 1}, 0, 0, \pm 1, 0, 0, \pm 1, \ldots$

where

$$-\theta_{-1} = \phi_{-1} = S, \quad \phi_0 = [3, 2, -2, S],$$

and for the chain pair

(ii) $\ldots, 3, 2, -2, -3, 2, -2, -3, -3, 3, \underline{3}, 2, -2, 3, 2, -2, -3, \ldots$
$\quad \ldots, \pm 1, 0, 0, \pm 1, 0, 0, -1, 1, \pm 1, \underline{\pm 1}, 0, 0, \pm 1, 0, 0, \pm 1, \ldots$

where

$$-\theta_{-1} = \phi_{-1} = S, \quad \phi_0 = [3, 2, -2, S].$$

Proof. If $\varepsilon_0 = 1$, then

$$\tau_0 = \pm \frac{1}{\phi_1 \phi_2 \phi_3} + \left\| \frac{1}{\phi_1 \phi_2 \phi_3} \left(-\frac{1}{|\phi_4|} - \frac{2}{|\phi_4 \phi_5|} \right) \right\|$$

$$> 1 - .067 + \| .067 \times .162 \| > .921; \tag{5.15}$$

also θ_0, ϕ_0 satisfy (5.6) and (5.8). Hence (5.10) holds and

$$|1 - \phi_0 + \tau_0| < .666. \tag{5.16}$$

For the chain pair (i),

$$\sigma_0 = \pm 1 - \frac{1}{\theta_{-1}} + \left\| \frac{1}{\theta_{-1} \theta_{-2} \theta_{-3}} \left(-\frac{1}{|\theta_{-4}|} \right) \right\|,$$

$$|\sigma_0| < 1 + .383 + .062 < 1.445;$$

and for the chain pair (ii),

$$\sigma_0 = \pm 1 - \frac{1}{\theta_{-1}} - \frac{1}{\theta_{-1} \theta_{-2}} + \left\| \frac{1}{\theta_{-1} \theta_{-2}} \left(-\frac{1}{|\theta_{-3}|} \right) \right\|,$$

$$|\sigma_0| < 1 - \frac{1}{\theta_{-1}}.$$

Thus in both cases

$$|-1 + \theta_0 + \sigma_0| < 3.828. \tag{5.17}$$

Hence, if $\varepsilon_0 = 1$, then by (5.16), (5.17), and (5.10),

$$\pi_0 < 12.9 < 13;$$

if $\varepsilon_0 = -1$, we get the same result by considering the product

$$|(-1 + \phi_0 + \tau_0)(1 - \theta_0 + \sigma_0)|.$$

<u>Lemma 5.7</u>. $\pi_{-2} < 13$ for the chain pair

$$\begin{array}{l} \ldots, 3,2,-2,-3,2,-2,-3,-3,\ 3,\ \underline{3},2,-2,\ 3,2,-2,-3,\ \ldots \\ \ldots, \underline{+}1,0,\ 0,\underline{+}1,0,\ 0,\ 1,\ 1,\underline{+}1,\underline{\underline{+}1},0,\ 0,\underline{+}1,0,\ 0,\underline{+}1,\ \ldots \end{array}$$

where

$$-\theta_{-1} = \phi_{-1} = S, \qquad -\theta_{-2} = \phi_0 = [3,2,-2,S].$$

<u>Proof</u>. The relation (5.10) holds, since

$$2.586 < |\theta_{-2}| < 2.587, \qquad 3.382 < |\phi_{-2}| < 3.383.$$

By the same argument as was used to obtain (5.15) we have

$$\sigma_{-2} > .921,$$

so that

$$|1 + \theta_{-2} + \sigma_{-2}| < .666. \tag{5.18}$$

Also

$$\tau_{-2} = 1 + \left\| 1 - \frac{1}{|\phi_{-1}|} \right\| > .295,$$

so that

$$|1 + \phi_{-2} + \tau_{-2}| < 2.088. \tag{5.19}$$

It follows from (5.18), (5.19), and (5.10) that

$$\pi_{-2} < 7 < 13.$$

By the remarks following the statement of Theorem 5.4, this completes the proof of Theorem 5.4

5.4. The Form $g_3 = (34,76,-34) = 2F_4$.

Theorem 5.5. For the form $g_3 = (34,76,-34) = 2F_4$ we have

$$M(g_3) = 11,$$
$$M_2(g_3) = \frac{17}{2} = \frac{1}{4}m(g_3).$$

Proof. By (4.15)

$$S = [3,3,3,2,-2,S].$$

Using the relations (4.18) to (4.23) we deduce that the only permissible expansions of S are

$$S \equiv [3,3,3,2,-2,S],$$
$$S \equiv [2,-2,-3,-3,-3,S],$$
$$S \equiv [3,3,2,-2,-3,S],$$
$$S \equiv [3,2,-2,-3,-3,S],$$
$$S \equiv [2,-2,-2,2,2,-2,S].$$

It now follows from the symmetry of the form g_3, whose roots are $R_1 = -S$, $R_2 = S$, that any permissible a-chain from g_3 must be an arrangement of some or all of the five blocks of numbers:

$$A = 3,3,3,2,-2,$$

$$B = 2,-2,-3,-3,-3,$$

$$C = 3,3,2,-2,-3,$$

$$D = 3,2,-2,-3,-3,$$

$$E = 2,-2,-2,2,2,-2, \; .$$

In Lemmas 5.8 and 5.9 we shall show that if an a-chain from g_3 contains A, then, for every corresponding ε-chain

$$M(P) < \frac{17}{2};$$

since B is the negative of A reversed, this is true also for a-chains from g_3 containing B. In Lemma 5.10 we shall show that the same result holds for a-chains from g_3 containing EC or DC; and we shall deduce from Lemmas 5.10 to 5.12 that it holds also for a-chains from g_3 containing CC. Any permissible a-chain from g_3 which is not even and does not contain A or B must contain C or D (which is the negative of C reversed) and must therefore contain one of EC, DC, CC or their negatives reversed. Thus, by the remarks at the end of section 5.1, the proof of Theorem 5.5 will follow from Lemmas 5.8 to 5.12.

The result actually proved in each lemma is that, for the chain-pair considered and for some r,

$$\pi_r < \frac{\Delta}{3} \, .$$

Since $\Delta/3 < 34$, this implies that $M(P) < 17/2$.

For the proofs of these lemmas we need the numerical information given in Table 4, in which the same conventions are used as in Table 1.

TABLE 4

x	x	$1/x$	$[-2,x]$	$1/[-2,x]$	$[2,-2,x]$	$1/[2,-2,x]$.
S	2.6173	0.3820	−2.3820	−0.4198	2.4198	0.4132
$[-3,S]$	−3.3820	−0.2956	−1.7043	−0.5867	2.5867	0.3865
$[-3_2,S]$	−2.7043	−0.3697	−1.6302	−0.6134	2.6134	0.3826
$[-3_3,S]$	−2.6302	−0.3801	−1.6198	−0.6173	2.6173	0.3820

Lemmas 5.8 and 5.9 are similar to Lemmas 4.7 and 4.8.

<u>Lemma 5.8</u>. $3\pi_o/\Delta < 1$ for the chain pairs

$$\begin{matrix} \ldots, 3, 3, \underline{3}, 2, -2, \ldots \\ \ldots, 1, 1, \underline{1}, 0, 0, \ldots \end{matrix} \quad \text{and} \quad \begin{matrix} \ldots, 3, 3, \underline{3}, 2, -2, \ldots \\ \ldots, \pm 1, -1, \underline{1}, 0, 0, \ldots \end{matrix}$$

where

$$-\theta_{-2} = \phi_3 = S.$$

<u>Proof</u>. We have

$$\tau_o = 1 + \left\| \frac{1}{\phi_1\phi_2} \left(1 - \frac{1}{|\phi_3|}\right) \right\| > .892 , \tag{5.20}$$

$$2.586 < \phi_o < 2.587. \tag{5.21}$$

Thus

$$|1 - \phi_o + \tau_o| < .695. \tag{5.22}$$

If $\varepsilon_{-1} = -1$, $\sigma_o < 0$, and if $\varepsilon_{-1} = \varepsilon_{-2} = 1$, then

$$\sigma_o = 1 - \frac{1}{\theta_{-1}} + \left\| \frac{1}{\theta_{-1}} \left(1 - \frac{1}{|\theta_{-2}|}\right) \right\|.$$

Hence in both cases

$$\sigma_o < .888,$$

$$2.704 < \theta_o < 2.705.$$

Therefore

$$\left.\begin{aligned} &|-1 + \theta_o + \sigma_o| < 2.593, \\ &3/|\theta_o\phi_o - 1| < .501 \end{aligned}\right\} \tag{5.23}$$

It follows from (5.22) and (5.23), that

$$3\pi_o/\Delta < .903 < 1.$$

<u>Lemma 5.9</u>. $3\pi_{-1}/\Delta < 1$ for the chain pair

$$\begin{matrix} \ldots, & 3, 3, \underline{3}, 2, -2, & \ldots \\ \ldots, & -1, 1, \underline{1}, 0, \ 0, & \ldots \end{matrix}$$

where

$$-\theta_{-2} = \phi_3 = S.$$

<u>Proof</u>.

$$\tau_{-1} = 1 - \frac{1}{\phi_o}\tau_o;$$

hence, by (5.20) and (5.21),

$$\tau_{-1} > .571.$$

Also

$$2.613 < \phi_{-1} < 2.614,$$

so that

$$|1 - \phi_{-1} + \tau_{-1}| < 1.043. \tag{5.24}$$

We have

$$\sigma_{-1} = -1 + \left\| 1 - \frac{1}{|\theta_{-2}|} \right\| < -.382,$$

$$3.382 < \theta_{-1} < 3.383.$$

Hence

$$\left.\begin{aligned} |-1 + \theta_{-1} + \sigma_{-1}| &< 2.001, \\ 3/|\theta_{-1}\phi_{-1} - 1| &< .383. \end{aligned}\right\} \quad (5.25)$$

It follows from (5.24) and (5.25) that

$$3\pi_0/\Delta < .8 < 1.$$

<u>Lemma 5.10</u>. $3M(P)/\Delta < 1$ for the following chain pairs

(i) $\begin{array}{l} \ldots, 3, \underline{3}, 2, -2, -3, \ldots \\ \ldots, -1, \underline{1}, 0, 0, \pm 1, \ldots \end{array}$

(ii) $\begin{array}{l} \ldots, 2, -2, -2, 2, 2, -2, 3, \underline{3}, 2, -2, -3, \ldots \\ \ldots, 0, 0, 0, 0, 0, 0, 1, \underline{1}, 0, 0, \pm 1, \ldots \end{array}$

(iii) $\begin{array}{l} \ldots, 3, 2, -2, -3, -3, \underline{3}, 3, 2, -2, -3, \ldots \\ \ldots, \pm 1, 0, 0, 1, -1, \underline{1}, 1, 0, 0, \pm 1, \ldots \end{array}$

(iv) $\begin{array}{l} \ldots, 3, 2, -2, -3, -3, 3, \underline{3}, 2, -2, -3, \ldots \\ \ldots, \pm 1, 0, 0, -1, 1, 1, \underline{1}, 0, 0, \pm 1, \ldots \end{array}$

where, in each case,

$$-\theta_{-1} = \phi_{-1} = S, \qquad \phi_0 = [3, 2, -2, -3, \bar{S}].$$

<u>Proof</u>. For the chain pairs (i), (ii), and (iii) we have

$$\tau_0 = 1 + \left\| \frac{1}{\phi_1\phi_2}\left(1 - \frac{1}{|\phi_3|} - \frac{2}{|\phi_3\phi_4|}\right) \right\| > .890,$$

$$2.613 < \phi_0 < 2.614. \quad (5.26)$$

Therefore

$$|1 - \phi_0 + \tau_0| < .724. \quad (5.27)$$

For the chain pair (i), $\sigma_0 < 0$; for the chain pair (iii),

$$\sigma_0 = 1 + \frac{1}{\theta_{-1}} + \left\| \frac{1}{\theta_{-1}}\left(1 - \frac{1}{|\theta_{-2}|}\right) \right\| < 1;$$

and for the chain pair (ii)

$$\sigma_o = 1 + \left\| \frac{1}{\theta_{-1}\theta_{-2}\ldots\theta_{-6}}\left(1 - \frac{1}{|\theta_{-7}|}\right)\right\|.$$

Thus in all three cases

$$\sigma_o < 1.010,$$

$$3.382 < \theta_o < 3.383. \tag{5.28}$$

Hence

$$|-1 + \theta_o + \sigma_o| < 3.393, \tag{5.29}$$

$$3/|\theta_o\phi_o - 1| < .383. \tag{5.30}$$

It follows from (5.27), (5.29) and (5.30) that, for the chain pairs (i), (ii), and (iii),

$$3\pi_o/\Delta < .95 < 1.$$

Similarly, by considering the product

$$|(-1 - \theta_{-2} + \sigma_{-2})(-1 - \phi_{-2} + \tau_{-2})|,$$

we can show that, for the chain pair (iv),

$$3\pi_{-2}/\Delta < 1.$$

Since $3M(P)/\Delta < 1$ for the chain pair (i), it follows that $3M(P)/\Delta < 1$ also for the negative of its reverse:

(v) $$\begin{array}{l} \ldots, \underline{3}, 2, -2, -3, -3, \ldots \\ \ldots, \underline{\pm 1}, 0, 0, 1, 1, \ldots \end{array}$$

where

$$-\theta_o = \phi_o = S.$$

Thus in Lemma 5.10 we have considered every possible ε-chain (or its negative) corresponding to the a-chains of the pairs (ii), (iii), and (iv).

<u>Lemma 5.11</u>. $3\pi_0/\Delta < 1$ for the chain pair

$$\begin{array}{l} \ldots ,3,\underline{3},2,-2,-3, \ldots \\ \ldots ,1,\underline{1},0,\ 0,-1, \ldots \end{array}$$

where

$$-\theta_{-1} = \phi_{-1} = S, \quad \phi_0 = [3,2,-2,-3,S].$$

<u>Proof</u>. The relation (5.26) holds, and

$$\tau_0 = 1 + \frac{1}{\phi_1\phi_2\phi_3} + \left\| \frac{1}{\phi_1\phi_2\phi_3}\left(1 - \frac{1}{\lceil\phi_4\rceil}\right) \right\| > 1;$$

hence

$$|1 - \phi_0 + \tau_0| < .614. \tag{5.31}$$

Also (5.28) holds and

$$\sigma_0 = 1 + \left\| 1 - \frac{1}{\lceil\theta_{-1}\rceil} \right\| < 1.618;$$

thus (5.30) holds and

$$|-1 + \theta_0 + \sigma_0| < 4.001. \tag{5.32}$$

It follows from (5.30),(5.31), and (5.32) that

$$3\pi_0/\Delta < .95 < 1.$$

<u>Lemma 5.12</u>. $3\pi_0/\Delta < 1$ for the chain pair

$$\begin{array}{l} \ldots ,3,3,2,-2,-3,3,\underline{3},2,-2,-3, \ldots \\ \ldots ,1,1,0,\ 0,\ 1,1,\underline{1},0,\ 0,\ 1, \ldots \end{array}$$

where

$$-\theta_{-1} = \phi_{-1} = S, \quad \phi_0 = [3,2,-2,-3,S].$$

Proof. The inequalities (5.26) and (5.28) hold, so that (5.30) holds.

We have

$$\tau_o = 1 - \frac{1}{\phi_1\phi_2\phi_3} + \left\| \frac{1}{\phi_1\phi_2\phi_3}\left(1 - \frac{1}{|\phi_4|}\right) \right\| < .976.$$

Thus

$$|-1 + \phi_o + \tau_o| < 2.590. \qquad (5.33)$$

Also

$$\sigma_o = 1 - \frac{1}{\theta_{-1}} + \frac{1}{\theta_{-1}\theta_{-2}} + \left\| \frac{1}{\theta_{-1}\theta_{-2}}\left(1 - \frac{1}{|\theta_{-3}|}\right) \right\|$$

$$> 1.382;$$

thus

$$|1 - \theta_o + \sigma_o| < 1.001. \qquad (5.34)$$

It follows from (5.30), (5.33), and (5.34) that

$$3\pi_o/\Delta < .993 < 1.$$

The ε-chains of the chain pairs (i) and (v) of Lemma 5.10 and of the chain pairs of Lemma 5.11 and 5.12 include every possible ε-chain (or its negative) corresponding to an a-chain from g_3 which contains CC, where C is the block of numbers

$$C = 3,3,2,-2,-3, .$$

By the argument following the statement of Theorem 5.5, this completes the proof of the theorem.

5.5. Note on the Limiting Symmetric Markov Form

I now consider the form

$$g(x,y) = x^2 + \sqrt{5}xy - y^2.$$

If $F(x,y)$ is any symmetric Markov form, then

$$F(x,y) = Qx^2 + Pxy - Qy^2,$$

where $P^2 = 5Q^2 - 4$ (see sect. 1.4), so that $F(x,y)$ is proportional to the form

$$f(x,y) = x^2 + \sqrt{(5 - 4/Q^2)}xy - y^2.$$

As $Q \to \infty$, $f(x,y)$ tends to the form $g(x,y)$. Thus we may regard $g(x,y)$ as the limit of forms proportional to the symmetric Markov forms, and, in particular, as the limit as $n \to \infty$ of forms proportional to the forms $g_n(x,y)$ discussed in Chapter 4. The form g behaves in some ways like one of the forms g_n for which $n \equiv 0 \pmod 3$. This is illustrated by the following lemma.

Lemma 5.13. For the form $g = (1,\sqrt{5},-1)$, we have

$$m(g) = 1,$$
$$M(g;\ \tfrac{1}{2},\tfrac{1}{2}) = \frac{8 - 3\sqrt{5}}{4}.$$

Proof. The form g has determinant 3 and first and second roots $R_1 = -R$, $R_2 = R$, where

$$R = \frac{3 + \sqrt{5}}{2} = (2, 1\infty). \tag{5.36}$$

As in the proof of Lemma 4.1, it can now be deduced from (5.36) and Lagrange's Theorem that for integral $(x,y \neq (0,0)$ we have

$$|g(x,y)| \geq 1,$$

and that if $|g(x,y)| \neq 1$, then

$$|g(x,y)| \geq 8 - 3\sqrt{5}.$$

For integral X,Y

$$g(2X+1,\ 2Y+1) = (2X+1)^2 - (2Y+1)^2 + (2X+1)(2Y+1)\sqrt{5} \neq \pm 1.$$

Since $g(1,0) = 1$ and $g(3,-1) = 8 - 3\sqrt{5}$, this completes the proof of the lemma.

The following theorem can be proved geometrically; however I shall prove it by using the methods of section 2.5 and the results of Chapters 3 and 4, to illustrate the fact that these methods can be applied to forms whose coefficients are not rational.

Theorem 5.6. For the form $g = (1,\sqrt{5},-1)$ we have

$$M(g) = \frac{8 - 3\sqrt{5}}{4}.$$

Proof. By Theorem 3.1 any chain of I-reduced forms equivalent to g must contain one of the forms

$$g = (1,\sqrt{5},-1),$$

$$\left.\begin{aligned} g\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix} &= (1, 2+\sqrt{5}, \sqrt{5}),\\ g\begin{bmatrix}1 & 0\\ -1 & 1\end{bmatrix} &= (-\sqrt{5}, 2+\sqrt{5}, -1).\end{aligned}\right\} \tag{5.37}$$

If f is either of the forms (5.37), then $\lambda(f) = 1$; therefore it follows from Lemmas 2.11 and 2.12 that if

$\{a_r\}$ is an a-chain from either of these forms, then, for every corresponding ε-chain,

$$M(P) \leq \frac{1}{4}.$$

It follows from (5.36) that any expansion of R must begin in one of the following ways:

$$\text{(i)} \quad R \equiv [2,-2,-R],$$
$$\text{(ii)} \quad R \equiv [3_k,2,-2,-R],$$
$$\text{(iii)} \quad R \equiv [3_\infty].$$

If $\{a_r\}$ is an a-chain from g such that (i) holds, then $\{a_r\}$ leads forwards to the form

$$h = (3-2\sqrt{5}, 4-\sqrt{5}, 1),$$

whose roots are

$$[2,-R], \quad [-2,-R].$$

Since $\lambda(h) = 8 - 3\sqrt{5}$, this means that, for any corresponding ε-chain,

$$M(P) \leq \frac{8 - 3\sqrt{5}}{4}.$$

If g_n is one of the forms (4.1) with first and second roots $-S$ and S, and if n and m are sufficiently great, then

$$S, \quad [3_m,-S], \quad [3_m,2,-2,S]$$

are arbitrarily close to R. Hence it follows from Lemmas 4.7 to 4.13 that if $\{a_r\}$ is an a-chain from g such that (ii) holds, then, for any corresponding ε-chain,

$$M(P) < \frac{1}{4}\frac{\Delta}{3} = \frac{1}{4}.$$

In Lemma 5.16 we show that if $\{a_r\}$ is an a-chain from g such that (iii) holds, then, for any corresponding ε-chain

$$M(P) < \frac{8 - 3\sqrt{5}}{4}.$$

Lemma 5.14. $\pi_o < 1$ for the chain pairs

(i) $$\begin{matrix} \ldots, & 3,3,3,\underline{3},3,3, & \ldots \\ \ldots, & 1,1,1,\underline{1},1,1, & \ldots \end{matrix},$$

(ii) $$\begin{matrix} \ldots, & 3,\ 3,\underline{3},\ 3, & \ldots \\ \ldots, & \underline{+}1,-1,\underline{1},\underline{+}1, & \ldots \end{matrix},$$

where, in each case,

$$\theta_o = [3_k, -R] \quad (k \geq 7), \qquad \phi_o = R.$$

Proof. In both cases

$$2.6180 < \theta_o < 2.6181, \qquad 2.6180 < \phi_o < 2.6181,$$

$$3/|\theta_o\phi_o - 1| < .513 \tag{5.38}$$

If (i) holds, then

$$\tau_o = 1 - \frac{1}{\phi_1} + \frac{1}{\phi_1\phi_2} + \left\| \frac{1}{\phi_1\phi_2}\left(1 - \frac{1}{|\phi_3|}\right) \right\|,$$

so that

$$.672 < \tau_o < .856.$$

Similarly $$.672 < \sigma_o < .856$$

Hence

$$(-1 + \theta_o + \sigma_o)(1 - \phi_o + \tau_o) < 2.475 \times .947 < 2.344 \tag{5.39}$$

If (ii) holds, then

$$\tau_o = 1 + \left\| 1 - \frac{1}{|\phi_1|} \right\| = 1 + \|.619\|.$$

Similarly,

$$\sigma_o = -1 + \|.619\|.$$

Hence

$$(-1 + \theta_o + \sigma_o)(1 - \phi_o + \tau_o) < 1.238 \times 1.238 < 1.533, \tag{5.40}$$

It follows from (5.38), (5.39) and (5.40) that in both cases

$$\pi_0 < 1.21 < 8 - 3\sqrt{5}.$$

Theorem 5.6 now follows from Lemmas 5.13 and 5.14.

CHAPTER 6

DAVENPORT'S CONSTANT

6.1. Introduction

Davenport's constant K is defined by

$$K = \sup[k;\ M(f) > k\Delta],$$

where the supremum is taken over all indefinite binary quadratic forms

$$f(x,y) = ax^2 + bxy + cy^2$$

which do not represent zero, and $\Delta = +\sqrt{(b^2 - 4ac)}$. The existence of K follows from Theorem 1.2 , and in section 1.2 I mentioned a number of results related to this theorem; in this chapter I discuss the value of K.

By Theorem 4.1, if the forms g_n are defined by (4.1), then, for arbitrarily large $n \not\equiv 0 \pmod 3$, we have

$$M(g_n) = \tfrac{1}{4}m(g_n).$$

Since $m(g_n)$ tends down to $\Delta/3$ as $n \to \infty$, this means that there are forms g_n with $M(g_n)$ arbitrarily close to $\Delta/12$. Thus we have an upper bound for the value of K:

$$K \leq \frac{1}{12}.$$

Davenport [30] showed that $K \geq 1/128$, and Cassels [15,16] has improved this result to about $K \geq 1/45.2$. In this chapter I use the method of

Chapter 2 to obtain

$$K > .0256 > \frac{1}{39}.$$

By Lemma 2.11, if a chain of I-reduced forms contains a form f for which $\lambda(f)$ is small, then $M(P)$ is small for all corresponding ε-chains; and the examination of chain pairs in [6] and in Chapters 4 and 5 shows that, if $\{a_r\}$ contains certain combinations of 2's, 3's, and 4's, then $M(P)$ is small for all corresponding ε-chains. For any form f it is always possible to find a chain $\{f_r\}$ of I-reduced forms such that, for all r, $\lambda(f_r)$ is not too small (e.g. $\lambda(f_r) > \Delta/3$ for the Hurwitz chain defined in sect. 6.2); but alternative expansions to blocks of 2's, 3's, and 4's usually contain 2's, 3's, and 4's again or lead to forms with small λ's. Since the a-chains of a Markov form consist only of 2's, 3's, and 4's, this suggests that the Markov forms are among those with the smallest inhomogeneous minima. For all Markov forms F,

$$M(F) \geq \frac{1}{4}m(F) \geq \frac{\Delta}{12};$$

thus the existence of symmetric Markov forms g_n with $M(g_n)$ arbitrarily close to $\Delta/12$ supports the conjecture that K may in fact be 1/12. Apart from the symmetric Markov forms, the form with the smallest known inhomogeneous minimum is the norm form

$$f(x,y) = x^2 - 73y^2,$$

for which Godwin [34] showed that

$$M(f) = \frac{1541}{2136} = .721 \ldots,$$

while

$$\frac{\Delta}{12} = \frac{\sqrt{73}}{6} = .712 \ldots.$$

(It can be shown that all the a-chains of f contain 'bad' combinations of 2's, 3's, and 4's, so we should expect M(f) to be small). Thus all the known evidence supports the conjecture that the value of K is 1/12.

An advantage of the method of Chapter 2 is that it is quite general, and so could perhaps be used eventually for the precise evaluation of K. To obtain a bound for K, we would need to choose an a-chain of any given form which is not too 'bad' (where a 'bad' a-chain is one such that M(P) is small for all corresponding ε-chains) and then to choose a corresponding ε-chain such that $\inf_r \pi_r$ is as large as possible. Before we could hope to get a good bound for K, we would therefore need to know which a-chains are particularly bad, and which ε-chains corresponding to a given a-chain will make M(P) small or large. It might be possible to obtain some information of this kind by using a high speed computer. The main difficulty is that this information would be local, and an ε-chain which is 'good' at one place may be very 'bad' at another — it might be impossible to obtain an ε-chain which is everywhere 'good' (and so makes M(P) large) by joining subchains which are locally 'good'.

The work of this chapter is only a preliminary attack on the problem for the purpose of showing that the method of Chapter 2 will yield a bound for K. I have not attempted to find the ideal a-chain but have used the Hurwitz chain, of which I give an account in section 6.2; this chain was also used by Davenport [30]. The Hurwitz chain has the advantage that, for every form f_r in the chain, $\lambda(f_r) > \Delta/3$, and a further reason for using it is that it has simple properties which make the calculations easier. The corresponding a-chain cannot contain ...,2,k,..., $(k > 0)$; this excludes long sequences of 2's, which are bad, but may exclude sequences which are good (e.g. for the forms g_n of Ch.4 the Hurwitz chain gives $M(P) < \Delta/12$, while, for $n \equiv 0 \pmod 3$, $M(f)$ is much greater than $\Delta/12$). The work done here on the Hurwitz chain shows which subchains are particularly bad and so provides information which would be useful for choosing a better chain.

In section 6.3 I describe a set of rules for choosing an ε-chain corresponding to the Hurwitz chain. These rules were suggested by experience with the symmetric Markov forms $\{g_n\}$ and are designed so that π_r will not obviously be very small for any r. Their main advantage is that they are simple — they are certainly not best possible.

Then in sections 6.4 and 6.5 I show that if $\{a_r\}$ is any Hurwitz chain and $\{\varepsilon_r\}$ is the corresponding ε-chain chosen according to the rules of section 6.3, then for all r

$$\pi_r/\Delta > .1025,$$

so that

$$M(P) > \frac{.1025}{4}\,\Delta > .0256\,\Delta > \frac{\Delta}{39}$$

This will prove

<u>Theorem 6.1</u>. If

$$f(x,y) = ax^2 + bxy + cy^2$$

is an indefinite binary quadratic form which does not represent zero, and $\Delta = +\sqrt{(b^2 - 4ac)}$, then

$$M(f) > .0256\,\Delta > \Delta/39.$$

If the form $f(x,y)$ has rational coefficients, then its Hurwitz chain is periodic and the ε-chain chosen according to the rules of section 6.3 is periodic also (though its period may be twice that of the Hurwitz chain). Therefore, by the Corollary to Lemma 2.8, we have the following

<u>Corollary</u>. If the form f of Theorem 6.1 has rational coefficients, then there exists a rational point (x',y') such that

$$M(f;\, x',y') > .0256\Delta > \Delta/39.$$

In this chapter I exclude from the discussion forms which represent zero because I use the methods of Chapter 2, which were given explicitly for forms whose roots are irrational. However, as Barnes [8] has shown, the methods of Chapter 2 can be modified to include singly infinite or finite chains of divided cells, which correspond to forms with one or two rational roots. It should be possible, by these modified methods, to obtain results for forms which represent zero corresponding to those of this chapter, and in particular to obtain bounds for the constant $\Gamma_{1,1}$ defined by Cassels [19] (see sect. 1.2), which may in fact have the same value as Davenport's constant K.

6.2. The Hurwitz Chain

The Hurwitz chain is a special type of chain of equivalent forms which was defined and used by Hurwitz [37]; as the argument in the rest of this chapter depends on the properties of the Hurwitz chain, I now give an account of them.

It is convenient to represent the numbers $\frac{3+\sqrt{5}}{2}$, $\frac{1+\sqrt{5}}{2}$, by the first few digits of their decimal expansions:

$$\frac{3+\sqrt{5}}{2} = 2.6180\ldots ,$$

$$\frac{1+\sqrt{5}}{2} = 1.6180\ldots ;$$

we note that the reciprocals of these numbers are

$$\frac{2}{3+\sqrt{5}} = \frac{3-\sqrt{5}}{2} = .3819\ldots ,$$

$$\frac{2}{1+\sqrt{5}} = \frac{\sqrt{5}-1}{2} = .6180\ldots .$$

A form f(x,y) which does not represent zero is said to be Hurwitz-reduced or <u>H-reduced</u> if it can be factorized in the form

$$f(x,y) = \pm \frac{\Delta}{|\theta\phi - 1|} (\theta x + y)(x + \phi y),$$

where θ, ϕ are irrational and

$$\left.\begin{array}{c} |\phi| > 2, \\ |\theta| > 1.618\ldots \text{ if } \theta,\phi \text{ differ in sign,} \\ |\theta| \geq 2.618\ldots \text{ if } \theta,\phi \text{ have the same sign.} \end{array}\right\} \quad (6.1)$$

Equivalently, f is H-reduced if θ, ϕ are irrational and

$$\left.\begin{array}{ll} -\frac{1}{2} < \frac{1}{\phi} < \frac{1}{2}, & \\ -.6180\ldots < \frac{1}{\theta} \leq .3819\ldots & \text{if } \phi > 0, \\ -.3819\ldots \leq \frac{1}{\theta} < .6180\ldots & \text{if } \phi < 0. \end{array}\right\} \quad (6.2)$$

Clearly a form which is H-reduced is I-reduced.

For any chain of I-reduced forms $\{f_n\}$ and for any corresponding a-chain $\{a_n\}$ $(|a_n| \geq 2)$ we have, for $-\infty < n < \infty$,

$$f_n(x,y) = \frac{\pm\Delta}{|\theta_n\phi_n - 1|}(\theta_n x + y)(x + \phi_n y),$$

$$\left.\begin{array}{l} \phi_n = a_{n+1} - \frac{1}{\phi_{n+1}} \\ \theta_{n+1} = a_{n+1} - \frac{1}{\theta_n} \end{array}\right\} \quad (6.3)$$

$$(|\theta_n| > 1,\ |\phi_n| > 1,\quad |\theta_{n+1}| > 1,\ |\phi_{n+1}| > 1).$$

If the form f_o is H-reduced, the Hurwitz chain of forms from f_o is the chain of I-reduced forms $\{f_n\}$ which corresponds to the a-chain $\{a_n\}$ determined by the expansions of ϕ_o, θ_o such that

$$-\frac{1}{2} < \frac{1}{\phi_n} < \frac{1}{2} \quad (n \geq 1), \tag{6.4}$$

$$\left.\begin{aligned} -.6180\ldots < \frac{1}{\theta_n} \leq .3819\ldots \quad \text{if } \theta_{n+1} > 0, \\ -.3819\ldots \leq \frac{1}{\theta_n} < .6180\ldots \quad \text{if } \theta_{n+1} < 0. \end{aligned}\quad (n \leq -1)\right\} \tag{6.5}$$

The chain $\{a_n\}$ is uniquely determined and is called the H-chain from f_o. It is easily deduced from (6.2) and (6.3) that, if f_r is H-reduced and (6.4) holds for $n = r+1$, then (6.5) holds for $n = r+1$; similarly if f_r is H-reduced and (6.5) holds for $n = r-1$, then (6.4) holds for $n = r-1$. Thus if $\{f_n\}$ is the Hurwitz chain from f_o, then (6.4) and (6.5) hold for all n, so that all the forms of the chain are H-reduced, and the Hurwitz chain from any of them is $\{f_n\}$. If f is a Gauss-reduced form equivalent to a given form then either f or at least one of the equivalent forms (3.22), (3.23) is H-reduced. Thus there is at least one H-reduced form equivalent to any given form g, and g has at least one H-chain. (It can be shown that the H-chain of a given form g is unique, apart from taking its negative, but this result is not used here.)

We note that the definition of the Hurwitz chain given here is slightly different from the classical one because we use the semi-regular continued fractions given by Definition 2.1 instead of classical semi-regular continued fractions.

For the remainder of this chapter $\{a_n\}$ is an H-chain of any given form f. We shall frequently use without comment the relations (6.1) or (6.2) and the fact that, since $|\phi_n| > 2$ for all n, the chain cannot contain any of the subchains:

$$\ldots,2,k,\ldots\ ,\quad \ldots,-2,-k,\ldots\ ,\quad \ldots,2_m,\ldots\ ,\quad \ldots,-2_m,\ldots\ ,$$

where $k > 0$, $m \geq 2$.

6.3. Rules for Choosing the ε-chain corresponding to a given H-chain

In this section and in the remainder of this chapter we use the notation and results of section 2.5.

Given an H-chain $\{a_n\}$, we choose a corresponding ε-chain $\{\varepsilon_n\}$ according to the following rules:

$$\text{(i)}\ |\varepsilon_n| \leq 1 \quad \text{for all } n;$$
$$\text{(ii)}\ \operatorname{sgn} \varepsilon_n = \operatorname{sgn} \frac{\tau_{n+1}}{\phi_{n+1}} \quad \text{if } \varepsilon_n \neq 0.$$

By the definition of an ε-chain (Lemma 2.6,(ii), (iii), and (iv)), it follows from (i) that:

$$\varepsilon_n = 0 \quad \text{if } a_{n+1} \text{ is even,}$$

$$\varepsilon_n = \pm 1 \quad \text{if } a_{n+1} \text{ is odd.}$$

For all n,

$$\begin{aligned}\tau_{n+1} &= \varepsilon_{n+1} + \left\| 1 - \frac{1}{|\phi_{n+2}|} \right\| \\ &= \varepsilon_{n+1} - \frac{\varepsilon_{n+2}}{\phi_{n+2}} + \ldots + \frac{(-1)^{r-1}\varepsilon_{n+r}}{\phi_{n+2}\ldots\phi_{n+r}} \\ &\qquad + \left\| \frac{1}{\phi_{n+2}\ldots\phi_{n+r}} \left(1 - \frac{1}{|\phi_{n+r+1}|}\right) \right\|.\end{aligned}$$

Hence, if $\varepsilon_{n+1} \neq 0$, then sgn τ_{n+1} = sgn ε_{n+1}, and if $\varepsilon_{n+1} = \varepsilon_{n+2} = \dots = \varepsilon_{n+r-1} = 0$, $\varepsilon_{n+r} \neq 0$, then

$$\operatorname{sgn} \tau_{n+1} = \operatorname{sgn} \{(-1)^{r-1}\phi_{n+2} \dots \phi_{n+r}\} \operatorname{sgn} \varepsilon_{n+r}.$$

Thus if we fix the sign of one $\varepsilon_n = 0$, then the rule (ii) uniquely determines the sign of every other non-zero ε_r, and so (i) and (ii) uniquely determine the ε-chain. By Lemma 2.13 we may fix the sign of one ε_n without loss of generality.

From (ii) we deduce in particular that, if $\varepsilon_n \neq 0$, $\varepsilon_{n+1} \neq 0$, then

$$\left.\begin{aligned} \text{(iii)}\quad \operatorname{sgn} \varepsilon_n &= \operatorname{sgn} \varepsilon_{n+1} \quad \text{when } a_{n+1} > 0,\ a_{n+2} > 0, \\ \operatorname{sgn} \varepsilon_n &= -\operatorname{sgn} \varepsilon_{n+1} \quad \text{when } a_{n+1} < 0,\ a_{n+2} < 0. \end{aligned}\right\}$$

The rule (i) is satisfied in any case by subchains of $\{\varepsilon_n\}$ corresponding to most bad subchains of $\{a_n\}$, and was therefore chosen for the sake of simplicity. If (iii) does not hold, π_n is always small (e.g. see the chain pair (i) of Lemma 5.10), so that an ε-chain which does not satisfy (iii) is certainly bad. It turns out that, if $\{a_n\}$ is an H-chain, then π_n is likely to be small if $|\tau_n|$ is large, so that a good ε-chain must ensure that $|\tau_n|$ is small. Since

$$\tau_n = \varepsilon_n - \frac{\tau_{n+1}}{\phi_{n+1}} = \varepsilon_n + \left\| 1 - \frac{1}{|\phi_{n+1}|} \right\| \qquad (6.6)$$

the rules (i) and (ii) mean that $|\tau_n|$ is always small. This and the fact that (ii) implies (iii) were the main reasons for the choosing the rule (ii).

From (i) and (ii) and the properties (6.1), (6.2) of the Hurwitz chain we deduce the following lemmas.

<u>Lemma 6.1</u>. If $\{a_n\}$ is an H-chain and $\{\varepsilon_n\}$ is the corresponding ε-chain which satisfies (i), then for all n,

$$\frac{|\sigma_n|}{|\theta_n|} \leq 0.6180\ldots .$$

<u>Proof</u>. We have

$$\sigma_n = \varepsilon_{n-1} + \sum_{r=1}^{\infty} (-1)^r \frac{\varepsilon_{n-r-1}}{\theta_{n-1}\cdots\theta_{n-r}}$$

$$= \varepsilon_{n-1} - \frac{\varepsilon_{n-2}}{\theta_{n-1}} + \ldots \pm \frac{\varepsilon_{n-r-1}}{\theta_{n-1}\,\cdots\,\theta_{n-r}}$$

$$+ \left\| \frac{1}{\theta_{n-1}\,\cdots\,\theta_{n-r}} \left(1 - \frac{1}{|\theta_{n-r-1}|}\right) \right\| .$$

Hence, if $|\theta_n| \geq 2.6180\ldots$, and $|\theta_{n-r}| \geq 2.6180\ldots$ for all $r \geq 1$, then

$$\frac{|\sigma_n|}{|\theta_n|} \leq \frac{1}{2.6180\ldots} \left(1 + \frac{1}{2.6180\ldots} + \frac{1}{(2.6180\ldots)^2} + \ldots\right)$$

$$= 0.6180\ldots ;$$

while if $|\theta_n| > 2.6180\ldots$, $|\theta_{n-1}| \geq 2.6180\ldots$, ... ,

$|\theta_{n-r}| \geq 2.6180\ldots$, and $1.6180\ldots < |\theta_{n-r-1}| < 2.6180\ldots$, then

$$\frac{|\sigma_n|}{|\theta_n|} \leq \frac{1}{2.6180\ldots}\left(1 + \frac{1}{2.6180\ldots} + \cdots + \frac{1}{(2.6180\ldots)^r} + \left\| \frac{1}{(2.6180\ldots)^r}\left(1 - \frac{1}{2.6180\ldots}\right)\right\|\right)$$

$$\leq 0.6180\ldots .$$

If $1.6180\ldots < |\theta_n| < 2.6180\ldots$, then $|a_n| = 2$, so that $\varepsilon_{n-1} = 0$ and $|\sigma_n| < 1$; thus

$$\left|\frac{\sigma_n}{\theta_n}\right| < 0.6180\ldots .$$

in this case also. This completes the proof of the lemma.

Lemma 6.2. If $\{a_n\}$ is an H-chain and $\{\varepsilon_n\}$ is the corresponding ε-chain which satisfies (i) and (ii), then for all n,

$$|\tau_n| \leq 1 \quad \text{and} \quad \left|\frac{\tau_n}{\phi_n}\right| < 0.4.$$

Proof. It follows from (i), (ii) and (6.6) that $|\tau_n| \leq 1$ for all n. Since $|\phi_n| > 2$ this implies that

$$\frac{|\tau_n|}{|\phi_n|} < 0.5.$$

If $|\phi_n| > 2.5$, then

$$\frac{|\tau_n|}{|\phi_n|} < 0.4;$$

and if $2 < |\phi_n| < 2.5$, then $\varepsilon_n = 0$ and

$$\frac{|\tau_n|}{|\phi_n|} = \frac{|\tau_{n+1}|}{|\phi_n \phi_{n+1}|} < \frac{0.5}{2} = 0.25.$$

Thus the lemma holds in all cases.

We shall use the following deductions from Lemmas 6.1 and 6.2.

Lemma 6.3. If $\{a_n\}$ is an H-chain and $\{\varepsilon_n\}$ is the corresponding ε-chain which satisfies (i) and (ii), then for all n

$$\sigma_n = \varepsilon_{n-1} + \| 0.6180\ldots \|,$$

$$\tau_n = \varepsilon_n + \| 0.4 \|,$$

$$|\tau_n| < 1.$$

Proof. We have

$$\sigma_n = \varepsilon_{n-1} - \frac{\sigma_{n-1}}{\theta_{n-1}}. \tag{6.7}$$

The lemma follows from Lemmas 6.1 and 6.2 by (6.6) and (6.7).

The bounds for σ_n and τ_n used in sections 6.4 and 6.5 are all obtained by using (6.6), (6.7), the rules (i) and (ii), and the results of Lemma 6.3. To avoid too much computation, we shall mostly use the cruder bounds for σ_n given by

$$\sigma_n = \varepsilon_{n-1} + \| 0.619 \|.$$

6.4. Proof of Theorem 6.1 — the Case when $\theta\phi > 0$

In this section and section 6.5 we shall prove Theorem 6.1 by showing that if $\{a_n\}$ is any H-chain and $\{\varepsilon_n\}$ is the corresponding ε-chain which satisfies (i) and (ii) of section 6.3, then, for all n,

$$\pi_n/\Delta > 0.1025;$$

clearly it is sufficient to show that always

$$\pi_0/\Delta > 0.1025.$$

In this section we show that this is true when θ_0, ϕ_0 have the same sign ($\theta_0\phi_0 > 0$), and in section 6.5 we show that it is true also when θ_0, ϕ_0 differ in sign ($\theta_0\phi_0 < 0$). As in both cases we shall always consider π_0, we shall write

$$\pi = \pi_0, \ \theta = \theta_0, \ \phi = \phi_0, \ \sigma = \sigma_0, \ \tau = \tau_0,$$

where no confusion arises.

Without loss of generality we may assume in this section that $\theta > 0$, $\phi > 0$, so that, by (6.1),

$$\theta \geq 2.6180\ldots\ , \quad \phi > 2; \tag{6.8}$$

and we may also fix the sign of one ε_n. If $\varepsilon_n \neq 0$, we take $\varepsilon_0 > 0$, and if $\varepsilon_0 = 0$, $\varepsilon_{-1} \neq 0$, we take $\varepsilon_{-1} > 0$; it follows from (i) and (ii) that one of the following possibilities must occur:

$$\left.\begin{array}{l} \varepsilon_{-1} = 0,\ \varepsilon_0 = 1; \\ \varepsilon_{-1} = \varepsilon_0 = 1; \\ \varepsilon_{-1} = 1,\ \varepsilon_0 = 0,\ \tau > 0; \\ \varepsilon_{-1} = \varepsilon_0 = 0. \end{array}\right\} \qquad (6.9)$$

By (2.30), π/Δ is given by

$$\frac{1}{|\theta\phi - 1|} \min \begin{bmatrix} |(1+\theta+\sigma)(1+\phi+\tau)|, & |(-1+\theta+\sigma)(1-\phi+\tau)|, \\ |(-1-\theta+\sigma)(-1-\phi+\tau)|, & |(1-\theta+\sigma)(-1+\phi+\tau)| \end{bmatrix}.$$

By Lemma 6.3 $|\tau| \leq 1$, and by (2.43)

$$|\sigma| \leq |\theta| - 1, \qquad |\tau| \leq |\phi| - 1.$$

Hence

$$\frac{\pi}{\Delta} = \frac{1}{\theta\phi - 1} \min\left[(\theta-1+\sigma)(\phi-1-\tau),\ (\theta-1-\sigma)(\phi-1+\tau)\right] \qquad (6.10)$$

$$> \min\left[\frac{(\theta-1+\sigma)}{\theta}\,\frac{(\phi-1-\tau)}{\phi},\ \frac{(\theta-1-\sigma)}{\theta}\frac{(\phi-1+\tau)}{\phi}\right]. \qquad (6.11)$$

In most cases we can show that $\pi/\Delta > .1025$ by considering the products (6.11).

If

$$0 < \underline{\theta} \leq \theta,\ \underline{\sigma} \leq \sigma \leq \bar{\sigma}, \text{ where } \underline{\sigma} \leq 1,\ \bar{\sigma} \geq -1,$$

then

$$\frac{\theta - 1 + \sigma}{\theta} \geq \frac{\underline{\theta} - 1 + \underline{\sigma}}{\underline{\theta}},$$

$$\frac{\theta - 1 - \sigma}{\theta} \geq \frac{\underline{\theta} - 1 - \bar{\sigma}}{\underline{\theta}}.$$

These results were used to obtain Table 5, where, for each range of values of θ and corresponding range of values σ, values less than $(\theta - 1 \pm \sigma)/\theta$ are given.

TABLE 5

CASE WHEN $\theta\phi > 0$

θ type	θ	ε_{-1}	σ	$(\theta-1+\sigma)/\theta$ $>$	$(\theta-1-\sigma)/\theta$ $>$
1	$2.6180\ldots \leq \theta < 3.6180\ldots$	1			
1.1	$\theta \equiv [3, k, \ldots]$ (k odd)	1	$0.381 < \sigma \leq 1$	0.76	0.235
1.2	$\theta \equiv [3, k, \ldots]$ (k even, $\neq 2$)	1	$= 1 + \left\Vert \frac{0.619}{3.618} \right\Vert$ $0.828 < \sigma < 1.172$	0.93	0.169
1.3	$3 < \theta \equiv [3, -3, \ldots]$	1	$0.381 < \sigma < 1.619$	0.79	0.127
1.4	$3 < \theta \equiv [3, -k, \ldots]$ (k odd, $\neq -3$)	1	$\sigma = 1 + \left\Vert \frac{1.619}{4.618} \right\Vert$ $0.648 < \sigma < 1.352$	0.88	0.216
1.5	$3 < \theta \equiv [3, -k, \ldots]$ (k even)	1	$\sigma = 1 + \left\Vert \frac{0.619}{1.618} \right\Vert$ $0.616 < \sigma < 1.384$	0.87	0.205
2	$3.6180\ldots \leq \theta < 4.6180\ldots$	0	$-0.619 < \sigma < 0.619$	0.552	0.552
3	$4.6180\ldots \leq \theta$	1,0	$-0.619 < \sigma < 1.619$	0.64	0.431

k denotes a positive integer, $k \geq 2$.

The ranges of values of σ were obtained by using (6.1), Lemma 6.3, and the relations (6.6) to (6.9). Table 6, which gives values less than $(\phi - 1 \mp \tau)/\phi$ is similar.

TABLE 6

CASE WHEN $\theta\phi > 0$

ϕ type	ϕ	ε_{-1}	ε_0	τ	$(\phi-1-\tau)/\phi$	$(\phi-1+\tau)/\phi$
1.1	$2.0 < \phi < 2.5$	1	0	$0 < \tau < 0.4$	0.3	0.5
1.2		0	0	$-0.4 < \tau < 0.4$	0.3	0.3
2	$2.5 < \phi < 3.5$	1,0	1	$0.6 < \tau \leq 1$	0.2	0.84
3.1	$3.5 < \phi < 4.5$	1	0	$0 < \tau < 0.4$	0.6	0.714
3.2		0	0	$-0.4 < \tau < 0.4$	0.6	0.6
4	$4.5 < \phi < 5.5$	1,0	1	$0.6 < \tau \leq 1$	0.55	0.9
5.1	$5.5 < \phi$	1	1,0	$0 < \tau \leq 1$	0.63	0.818
5.2		0	1,0	$-0.4 < \tau \leq 1$	0.63	0.745

It follows from (6.9) that if θ is of any of the types 1.1 to 1.5 (Table 5), then ϕ is of one of the types 1.1, 2, 3.1, 4, 5.1 (Table 6); while if θ is of type 2, then ϕ is of one of the types 1.2, 2, 3.2, 4, 5.2, and if θ is of type 3, then ϕ may be of any type. Calculation of the products shows that the products (6.11) are greater than 0.1025 for all these cases except when θ is of one of the types 1.2, 1.3 and ϕ is of the type 1.1, or when θ is of the type 1.3 and ϕ is of the type 3.1. In each of these cases

$$\frac{\theta - 1 + \sigma)}{\theta} \frac{(\phi - 1 - \tau)}{\phi} > 0.1025.$$

so we only need to show, by using more careful methods, that

$$p = \frac{(\theta - 1 - \sigma)(\phi - 1 + \tau)}{\theta\phi - 1} > 0.1025.$$

If θ is of type 1.2 and ϕ is of type 1.1, then, by Tables 5 and 6 ,

$$p > \frac{(\theta - 1 - 1.172)(\phi - 1)}{\theta\phi - 1},$$

which is an increasing function of θ and ϕ. Hence

$$p > \frac{(2.618 - 2.172)}{4.236} > 0.103.$$

If θ is of type 1.3, then

$$\theta = [3, -3, \ldots] > 3.276.$$

If now ϕ is of type 1.1, then, arguing as in the previous paragraph, we get

$$p > \frac{(3.276 - 2.619)}{5.552} > 0.118.$$

If θ is of type 1.3 and ϕ is of type 3.1, then by a similar argument we get

$$p > \frac{(3.276 - 2.619)2.5}{10.466} > 0.156.$$

Hence in all cases, if θ, ϕ have the same sign, we have

$$\pi/\Delta > 0.1025.$$

(It is clear that a very much stronger result could be obtained by more precise analysis; however this would not be worth while because, for our choice of $\{a_n\}$ and $\{\varepsilon_n\}$, it is not possible to obtain easily a much stronger result for the case when θ,ϕ differ in sign.)

6.5. Proof of Theorem 6.1 — the Case when $\theta\phi < 0$

In this section we show that

$$\pi/\Delta > 0.1025 \tag{6.12}$$

when θ,ϕ differ in sign ($\theta\phi < 0$). In this case, by (6.1),

$$|\theta| > 1.6180\ldots, \quad |\phi| > 2. \tag{6.13}$$

By (2.43)

$$|\sigma| \leq |\theta| - 1, \quad |\tau| \leq |\phi| - 1.$$

Hence it follows from (2.30) that π/Δ is greater than or equal to

$$\frac{1}{|\theta\phi|+1} \min\left[(|\theta|+1-|\sigma|)(|\phi|-1-|\tau|),\ (|\theta|-1-|\sigma|)(|\phi|+1-|\tau|)\right]$$

$$= \frac{|\theta\phi|}{|\theta\phi|+1} \min\left[\frac{(|\theta|+1-|\sigma|)}{|\theta|}\,\frac{(|\phi|-1-|\tau|)}{|\phi|},\ \frac{(|\theta|-1-|\sigma|)}{|\theta|}\,\frac{(|\phi|+1-|\tau|)}{|\phi|}\right]$$

If $|\theta| > \underline{\theta}$, $|\sigma| < \bar{\sigma}$, then

$$\frac{|\theta| + 1 - |\sigma|}{|\theta|} > \min\left[1,\ \frac{\underline{\theta} + 1 - \bar{\sigma}}{\underline{\theta}}\right],$$

$$\frac{|\theta| - 1 - |\sigma|}{|\theta|} > \frac{\underline{\theta} - 1 - \bar{\sigma}}{\underline{\theta}}.$$

These results were used to obtain Table 7, where, for each range of values of $|\theta|$ and corresponding upper bound of $|\sigma|$, values less than $(|\theta| \pm 1 - |\sigma|)/|\theta|$ are given.

TABLE 7

CASE WHEN $\theta\phi < 0$

θ type	$\lvert\theta\rvert$	$\lvert\sigma\rvert$ $<$	$\frac{\lvert\theta\rvert + 1 - \lvert\sigma\rvert}{\lvert\theta\rvert}$ $>$	$\frac{\lvert\theta\rvert - 1 - \lvert\sigma\rvert}{\lvert\theta\rvert}$ $>$
1	$1.6180\ldots < \lvert\theta\rvert < 2$			
1.1	$\pm\theta \equiv [2, k, \ldots]$ (k even, $\neq 2$)	$\frac{0.619}{3.618} < 0.172$	1	0.275
1.2	$\pm\theta \equiv [2, k, \ldots]$ (k even, $\neq 2, 4$) (k odd, $\neq 3$)	$\frac{1.619}{4.618} < 0.352$	1	0.164
1.3	$\pm\theta \equiv [2, 3, k, \ldots]$ $1.642 < \lvert\theta\rvert$	$\frac{1 + \frac{0.619}{5.618}}{2.8} < 0.397$	1	0.149
1.4	$\pm\theta \equiv [2, 3, 4, \ldots]$ $1.632 < \lvert\theta\rvert$	$\frac{1 + \frac{0.619}{3.618}}{2.723} < 0.431$	1	0.123
1.5	$\pm\theta \equiv [2, 3, k, \ldots]$ (k odd)	$\frac{1}{2.618} < 0.382$	1	0.145
1.6	$\pm\theta \equiv [2, 3, -2, \ldots]$ $1.704 < \lvert\theta\rvert$	$\frac{1 + \frac{0.619}{1.618}}{3.381} < 0.410$	1	0.172
1.7	$\pm\theta \equiv [2, 3, -k, \ldots]$ $1.666 < \lvert\theta\rvert$	$\frac{1 + \frac{1.619}{4.618}}{3} < 0.451$	1	0.129
1.8	$\pm\theta \equiv [2, 3, -4, \ldots]$ $1.666 < \lvert\theta\rvert$	$\frac{1 + \frac{0.619}{3.618}}{3} < 0.391$	1	0.165

k denotes a positive integer, $k \geq 2$.

TABLE 7 — CONTINUED

θ type	$\lvert\theta\rvert$	$\lvert\sigma\rvert$ $<$	$\frac{\lvert\theta\rvert+1-\lvert\sigma\rvert}{\lvert\theta\rvert}$ $>$	$\frac{\lvert\theta\rvert-1-\lvert\sigma\rvert}{\lvert\theta\rvert}$ $>$
1.9	$\pm\theta \equiv [2,3,-3,\ldots]$ $1.694 < \lvert\theta\rvert$	$\dfrac{1 + \frac{1.619}{2.618}}{3.276} < 0.495$	1	0.117
2	$2 < \lvert\theta\rvert < 2.6180\ldots$	0.619	1	0.190
3	$2.6180\ldots \leq \lvert\theta\rvert < 3$			
3.1	$\pm\theta \equiv [3,k,\ldots]$ (k odd)	1	1	0.235
3.2	$\pm\theta \equiv [3,k,\ldots]$ (k even, $\neq 2$)	$1 + \frac{0.619}{3.618} < 1.172$	0.931	0.169
4	$3 < \lvert\theta\rvert$	1.619	0.793	0.127

k denotes a positive integer, $k \geq 2$.

The upper bounds for $\lvert\sigma\rvert$ in Table 7 , were obtained by using (i) and (ii) of section 6.3, Lemma 6.3, and the relations (6.1), (6.6), (6.7), and (6.13). The expansions of $\pm\theta$ given in this table cover all possibilities since those which are omitted cannot occur in an H-chain (e.g. $\pm\theta \equiv [3,2,\ldots]$).

Table 8 , which gives values less than $(\lvert\phi\rvert \mp 1 - \lvert\tau\rvert)/\lvert\phi\rvert$, is similar to Table 7 .

TABLE 8

CASE WHEN $\theta\phi < 0$

ϕ type	$\lvert\phi\rvert$	$\lvert\tau\rvert$ <	$(\lvert\phi\rvert-1-\lvert\tau\rvert)/\lvert\phi\rvert$ >	$(\lvert\phi\rvert+1-\lvert\tau\rvert)/\lvert\phi\rvert$ >
1	$2.0 < \lvert\phi\rvert < 2.5$	0.4	0.3	1.24
2	$2.5 < \lvert\phi\rvert < 3.5$	1.0	0.2	1.0
3	$3.5 < \lvert\phi\rvert < 4.5$	0.4	0.6	1.133
4	$4.5 < \lvert\phi\rvert$	1.0	0.555	1.0

If (6.13) holds, then

$$\frac{|\theta\phi|}{|\theta\phi| + 1} > 0.764. \tag{6.14}$$

It follows from Tables 7 and 8 and from (6.14) that in all cases

$$\frac{|\theta\phi|}{|\theta\phi| + 1} \frac{(|\theta| + 1 - |\sigma|)}{|\theta|} \frac{(|\phi| - 1 - |\tau|)}{|\phi|} > 0.121$$

Thus we now need only to show that

$$p = \frac{|\theta\phi|}{|\theta\phi| + 1} \frac{(|\theta| - 1 - |\sigma|)}{|\theta|} \frac{(|\phi| + 1 - |\tau|)}{|\phi|} > 0.1025.$$

The results given in Table 9 were derived from Tables 7 and 8 .

By calculating p from Table 9 , we see that

$$p > 0.1028$$

except when θ is of one of the types 1.4, 1.9 and ϕ is of type 2. These cases must be considered separately.

TABLE 9

CASE WHEN $\theta\phi < 0$

θ type	ϕ type	$\frac{\lvert\theta\phi\rvert}{\lvert\theta\phi\rvert + 1}$ >	$\frac{\lvert\theta\rvert - 1 - \lvert\sigma\rvert}{\lvert\theta\rvert}$ >	$\frac{\lvert\phi\rvert + 1 - \lvert\tau\rvert}{\lvert\phi\rvert}$ >
any	1	0.764	0.117	1.24
1.1, 1.2, 1.3, 1.5, 1.6, 1.8	2	0.764	0.145	1
1.4	2	0.803	0.123	1
1.7	2	0.806	0.129	1
1.9	2	0.808	0.117	1
2, 3.1, 3.2, 4	2	0.833	0.127	1
any	3	0.849	0.117	1.333
any	4	0.879	0.117	1

If θ is of type 1.4, then one of the following statements holds (where k is a positive integer, $k \geq 2$):

(i) $\pm\theta \equiv [2,3,4,-k, \ldots]$,

(ii) $\pm\theta \equiv [2,3,4,k, \ldots]$ (k even, $k \neq 2$),

(iii) $\pm\theta \equiv [2,3,4,k, \ldots]$ (k odd).

If (i) holds, then

$$|\sigma| < \frac{1 + \frac{0.619}{4}}{2.75} < 0.420;$$

if (ii) holds, then

$$|\sigma| < \frac{1 + \frac{0.619}{(3.618)^2}}{2.723} < 0.385;$$

and if (iii) holds, then it follows from rule (ii) of section 6.3 that

$$\operatorname{sgn} \varepsilon_{-4} = -\operatorname{sgn} \varepsilon_{-2},$$

so that $|\sigma_{-1}| < 1$ and

$$|\sigma| = |\sigma_0| < \frac{1}{2.723} < 0.368.$$

Thus if θ is of type 1.4, then

$$|\sigma| < 0.420;$$

if now ϕ is of type 2, it follows from Tables 7 and 8 that

$$p > \frac{(|\theta| - 1 - 0.420)(|\phi| + 1 - 1)}{|\theta\phi| + 1},$$

which is an increasing function of $|\theta|$ and $|\phi|$. Hence

$$p > \frac{(1.632 - 1.420)2.5}{5.08} > 0.1043 > 0.1025.$$

If θ is of type 1.9, then one of the following statements holds (where k is a positive integer, $k \geq 2$):

(iv) $\pm\theta \equiv [2,3,-3,k, \ldots]$,

(v) $\pm\theta \equiv [2,3,-3,-k, \ldots]$ (k even, $k \neq 2$),

(vi) $\pm\theta \equiv [2,3,-3,-k, \ldots]$ (k odd).

If (iv) holds, then

$$|\sigma| < \frac{1 + \frac{1.619}{3}}{3.276} < 0.470;$$

if (v) holds, then

$$|\sigma_{-1}| < 1 + \frac{1}{2.618} + \frac{0.619}{2.618 \times 3.618} < 1.448,$$

$$|\sigma| < \frac{1.448}{3.276} < 0.443;$$

and if (vi) holds, then $|\sigma_{-2}| < 1$, so that

$$|\sigma| < \frac{1 + \frac{1}{2.618}}{3.276} < 0.422.$$

Thus if θ is of type 1.9, then

$$|\sigma| < 0.470.$$

If now ϕ is of type 2, then, by an argument similar to that of the previous paragraph, we get

$$p > \frac{(1.694 - 1.470)2.5}{5.235} > 0.106 > 0.1025.$$

This completes the proof of Theorem 6.1.

BIBLIOGRAPHY

[1] BAMBAH, R.P. Non-homogeneous binary quadratic forms. I, II. Acta Math. 86 (1951), 1-29, 32-56.

[2] ———, Non-homogeneous binary cubic forms. Proc. Cambridge Philos.Soc. 47 (1951), 457-460.

[3] ———, Divided cells. Res.Bull.Panjab Univ. 81 (1955), 173-174.

[4] BAMBAH, R.P. and ROGERS, K. An inhomogeneous minimum for nonconvex star regions with hexagonal symmetry. Canadian J.Math. 7 (1955), 337-346.

[5] BARNES, E.S. Non-homogeneous binary quadratic forms. Quart.J.Math., Oxford Ser. (2),1 (1950),199-210.

[6] ———, The inhomogeneous minima of binary quadratic forms (IV). Acta Math. 92 (1954), 235-264.

[7] ———, The inhomogeneous minimum of a ternary quadratic form. Acta Math. 92 (1954), 13-33.

[8] ———, On linear inhomogeneous Diophantine approximation. J. London Math.Soc. 31 (1956), 73-79.

[9] BARNES, E. S. and SWINNERTON-DYER, H.P.F. The inhomogeneous minima of binary quadratic forms.I. Acta Math. 87 (1952), 259-323.

[10] ------, The inhomogeneous minima of binary quadratic forms.II. Acta Math. 88 (1952), 279-316.

[11] ------, The inhomogeneous minima of binary quadratic forms.III. Acta Math. 92 (1954), 199-234.

[12] BIRCH, B.J. A grid with no split parallelopiped. Proc. Cambridge Philos.Soc. 53 (1957), 536.

[13] CASSELS, J.W.S. The lattice properties of asymmetric hyperbolic regions.II, III. Proc. Cambridge Philos.Soc. 44 (1948), 145-154, 457-462.

[14] ------, The Markoff chain. Ann. of Math. 50 (3) (1949), 676-685.

[15] ------, The inhomogeneous minimum of binary quadratic, ternary cubic and quaternary quartic forms. Proc. Cambridge Philos.Soc. 48 (1952), 72-86.

[16] ------, Addendum to the paper, The inhomogeneous minimum of binary quadratic, ternary cubic and quaternary quartic forms. Proc.Cambridge Philos.Soc. 48 (1952), 519-520.

[17] CASSELS, J.W.S. Yet another proof of Minkowski's theorem on the product of two inhomogeneous linear forms. Proc.Cambridge Philos.Soc. 49 (1953), 365-366.

[18] ------, On the product of two inhomogeneous linear forms. J.Reine Angew.Math. 193 (1954), 65-83.

[19] ------, An introduction to Diophantine Approximation. Cambridge University Press (1957).

[20] CHABAUTY, C. and LUTZ, E. Sur les approximations linéaires réelles.I,II. C.R.Acad.Sci. Paris 231 (1950), 887-888, 938-939.

[21] CHALK, J.H.H. The minimum of a non-homogeneous binary cubic form. Proc.Cambridge Philos.Soc. 48 (1952), 392-401.

[22] CHATLAND, H. and DAVENPORT, H. Euclid's algorithm in real quadratic fields. Canad.J.Math. 2 (1950), 289-296.

[23] DAVENPORT, H. A simple proof of Remak's theorem on the product of three linear forms. J.London Math.Soc. 14 (1939), 47-51.

[24] ------, Non-homogeneous binary quadratic forms,I. Nederl.Akad.Wetensch.Proc.Ser.A, 49 (1946), 815-821.

[25] DAVENPORT, H. Non-homogeneous binary quadratic forms, II-V. Nederl.Akad.Wetensch.Proc. Ser.A,50 (1947), 378-389, 484-491, 741-749, 909-917.

[26] ------, Non-homogeneous ternary quadratic forms. Acta Math. 80 (1948), 65-95.

[27] ------, Indefinite binary quadratic forms. Quart. J. Math., Oxford Ser. (2),1 (1950), 54-62.

[28] ------, Euclid's algorithm in cubic fields of negative discriminant. Acta Math. 84 (1950), 159-179.

[29] ------, Euclid's algorithm in certain quartic fields. Trans.Amer.Math.Soc. 68 (1950), 508-532.

[30] ------, Indefinite binary quadratic forms and Euclid's algorithm in real quadratic fields. Proc.London Math.Soc. (2) 53 (1951), 65-82.

[31] DELAUNEY, B.N. An algorithm for divided cells. Izvestiya Akad.Nauk.SSSR 11 (1947), 505-538 (in Russian).

[32] DICKSON, L.E. Introduction to the Theory of Numbers. University of Chicago Press (1929),pp.99-116.

[33] ------, Studies in the Theory of Numbers. Univ. of Chicago Press (1930),pp.79-107.

[34] GODWIN, H.J. On the inhomogeneous minima of certain norm-forms. J.London Math.Soc. 30 (1955), 114-119.

[35] HARDY, G.H. and WRIGHT, E.M. Introduction to the Theory of Numbers. Oxford University Press, 3rd edition, (1954).

[36] HEINHOLD, J. Verallgemeinerung und Verschaerfung eines Minkowskischen Satzes. I and II. Math.Z.44 (1939), 659-688; 45 (1939), 176-184.

[37] HURWITZ, A. Ueber eine besondere Art der Kettenbruchentwicklung reeller Groessen. Acta Math. 12 (1889), 367-405. - Math.Werke II, 84-115.

[38] INKERI, K. On the Minkowski constant in the theory of binary quadratic forms. Ann.Acad.Fennicae, Ser.A I 66 (1950), 1-34.

[39] ------, Ueber einige Verschaerfungen eines Minkowskischen Satzes. Ann.Acad.Sci.Fennicae Ser.A I 136 (1952), 1-16.

[40] KOKSMA, J.F. Diophantische Approximationen. Springer, Berlin (1936) (repr. Chelsea Publishing Company, New York).

[41] MARKOV, A. Sur les formes quadratiques binaires indéfinies. Math. Ann. 15 (1879), 381-406; 17 (1880), 379-399.

[42] MINKOWSKI, H. Ueber die Annaeherung an eine reelle Groesse durch rationale Zahlen. Math. Ann. 54 (1900), 91-124.

[43] MORDELL, L. J. The product of two non-homogeneous linear forms. IV. J. London Math. Soc. 26 (1951), 93-95.

[44] ------, The minima of some inhomogeneous functions of two variables. Duke Math. J. 19 (1952) 519-527.

[45] ------, Note on Sawyer's paper, "The product of two non-homogeneous linear forms". J. London Math. Soc. 28 (1953), 510-512.

[46] ROGERS, K. The minima of some inhomogeneous functions of two variables. J. London Math. Soc. 28 (1953), 394-402.

[47] SAWYER, D. B. The product of two non-homogeneous linear forms. J. London Math. Soc. 23 (1948), 250-251.

[48] VARNAVIDES, P. Note on non-homogeneous quadratic forms. Qart. J. Math., Oxford Ser. (2), 19 (1948), 54-58.

[49] VARNAVIDES, P. Non-homogeneous quadratic forms, I,II. Nederl. Akad.Wetensch.Proc. Ser.A, 51 (1948), 396-404; 470-481.

[50] ------, On the quadratic form $x^2 - 7y^2$. Proc. Roy. Soc. London Ser.A, 197 (1949), 256-268.

[51] ------, The Euclidean real quadratic fields. Nederl. Akad. Wetensch. Proc. Ser.A 55 = Indagationes Math. 14 (1952), 111-122.

[52] ------, The Minkowski constant of the form $x^2 - 11y^2$. Bull. Soc.Math. Grèce 26 (1952), 14-23.