SOME LATTICE POINT PROBLEMS

by

Lewis Low M.Sc. (Sydney)

A Thesis submitted for the Degree of

Doctor of Philosophy

in The University of Adelaide,

Department of Pure Mathematics,

September, 1978.

Awarded November 1979

CONTENTS

STATEMENT

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university and, to the best of my knowledge and belief, contains no material previously published or written by another person, except when due reference is made in the thesis.

_____

ACKNOWLEDGEMENTS

I would like to thank Dr. Jane Pitman, who suggested the various topics of research, and showed a detailed interest in my work. I am indebted to her for her helpful criticisms and suggestions.

I would also like to thank A. Hesterman for information about his calculations in connection with a conjecture on quadratic Farey sequences.

Finally, I would like to thank Katherine Halsey for typing this thesis.

GENERAL INTRODUCTION.

1. The scope of the thesis.

The material in this thesis falls into three parts which are linked by their use of ideas from the geometry of numbers and their emphasis on 3-dimensional lattices.

The first part of the thesis is in the first five chapters and concerns problems on subdivision of lattice cones related to a lattice point problem of Schinzel. (A lattice cone is a cone with vertex at the origin and edges determined by lattice points.) After establishing the necessary preliminary geometrical and arithmetical results in Chapters 1 and 2, I will introduce the subdivision problems at the beginning of Chapter 3. These problems will be considered for lattice cones in n dimensions, 2 dimensions and 3 dimensions in Chapters 3, 4 and 5, respectively. Applications of the 2-dimensional results to another problem of Schinzel and of the 3-dimensional results to lattice octahedra will also be given.

The second part of this thesis concerns work arising from a problem of Mahler on the quadratic analogue of the classical Farey sequences, and my results are presented in Chapter 6. To prove one of the results there (Theorem 6C.3) another application is made of one of the main results on 3-dimensional lattices in Chapter 5.

The third part of this thesis concerns an index problem of Cassels. Our investigations leads to bounds on the index of the sublattice of a 3-dimensional lattice generated by points with minimal $|x_1| + |x_2| + |x_3|$ in each octant. The results comprise Chapter 7.

2. Preliminaries and notation. In this thesis, references are given by a number in square brackets, and refer to the bibliography at the end of the thesis. The numbering on displayed formulae, equations, and so on begins afresh with each section of a chapter. In each section of a chapter the theorems, corollaries, lemmas and propositions, and notes where the numbering includes the section number, will be numbered in one sequence. For example, in section §7D of Chapter 7, Lemma 7D.2 is followed by Theorem 7D.3, and in section §3A of Chapter 3, Note 3A.1 is followed by Lemma 3A.2. In each section of a chapter, definitions, where numbered, will be numbered in a separate sequence to that of the theorems, and so on.

Throughout this thesis, Z will denote the set of all integers. Some notation of a geometrical or topological nature such as lin A, relint A is given in §1A. The notation $\ell(\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n)$ for a lattice with basis $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n$, is as mentioned in §3B. Most of the definitions and notation needed for Chapters 3 to 5 appear in §3A or §3B which introduce the problem of subdivision of cones.

CHAPTER 1

GEOMETRICAL PRELIMINARIES

This chapter is on geometrical preliminaries. In §1A we discuss background material on convex sets and pointed polyhedral cones. In §1B we introduce a major tool for our later study of subdivision of lattice cones, namely, barycentric subdivision. In §1C we discuss an Euler-type combinatorial result which has applications to Chapter 5.

§1A. Convex sets and cones.

In this section we will discuss the background material on convex sets and pointed polyhedral cones which is assumed in the rest of this chapter, and in Chapters 3 to 5. We will give some lemmas for later use.

As far as possible, we will, in this thesis, use for this material, the vocabulary and definitions of Grünbaum [5], Chapter 2 and Sections 3.1, 3.2. Thus we will assume the definition and basic properties given there for the following:

Convex set, polyhedral set, polytope, face, facet, vertex, edge, extreme point, pointed cone with apex $\underset{\sim}{a}$ (and hence with vertex $\underset{\sim}{a}$).

We will say that a hyperplane H in R^d cuts a set A in R^d if each of the open half-spaces determined by H contains a point of A. (Grünbaum [5], §2.2.)

We will also use the following notation of Grünbaum [5], where A is a subset of d-dimensional real space R^d:

lin A : the linear subspace spanned by A,

aff A : the affine subspace spanned by A,

dim A : the dimension of aff A, that is, one less than the maximum number of affinely independent points of A,

cone A : the cone with apex $\underset{\sim}{b}$ generated by A,
$\underset{\sim}{b}$

conv A : the convex hull of A,

relint A : the relative interior of A (relative to aff A),

relbd A : the relative boundary of A (relative to aff A).

We will say that two sets A, B in R^d are <u>non-overlapping</u> if their relative interiors are disjoint, that is if

$$\text{relint } A \cap \text{relint } B = \emptyset .$$

We will also say that a collection S of sets in R^d is non-overlapping, or that S has non-overlapping members, if each pair of distinct sets in S are non-overlapping.

<u>Lemma 1A.1</u>. Let A, B be convex sets in R^d such that A and B are non-overlapping, and aff $A \subseteq$ aff B. Then

$$A \cap \text{relint } B = \emptyset .$$

<u>Proof</u>. Suppose to the contrary that $A \cap \text{relint } B \neq \emptyset$, and let $\underset{\sim}{x} \in A \cap \text{relint } B$. Since $\underset{\sim}{x} \in \text{relint } B$, there is an open set N of R^d such that

$$\underset{\sim}{x} \in N \cap \text{aff } B \subseteq \text{relint } B.$$

Since A is convex, and $\underset{\sim}{x} \in N \cap A$, there is a point $\underset{\sim}{y}$ in $N \cap \text{relint } A$ (and in fact A, $N \cap A$ and $N \cap \text{relint } A$ all have the same dimension). Hence $\underset{\sim}{y} \in N \cap \text{aff } A \subseteq N \cap \text{aff } B \subseteq \text{relint } B$. Thus $\underset{\sim}{y} \in \text{relint } A \cap \text{relint } B$, so A and B overlap, a contradiction.

Lemma 1A.2. Let A, B be distinct convex sets in R^d such that $A \cap B$ is a face of each of A and B. Then A, B are non-overlapping.

Proof. Suppose to the contrary that there is a point $\underset{\sim}{x}$ in $\text{relint } A \cap \text{relint } B$. Let $F = A \cap B$. Then $\underset{\sim}{x} \in F \cap \text{relint } B$. Since F is a face of B, it follows that $F = B$. Similarly, $F = A$, so $A = B$, a contradiction.

The pointed cones that we will be concerned with are pointed cones with apex $\underset{\sim}{0}$, and by a pointed polyhedral cone we will mean a polyhedral set which is a pointed cone with apex $\underset{\sim}{0}$. Let C be a pointed polyhedral cone in R^d. Then for some $m \times d$ matrix A of rank d,

$$C = \{\underset{\sim}{x} \mid A\underset{\sim}{x} \leqslant \underset{\sim}{0}\} .$$

Also, if C has r edges and $\underset{\sim}{a}_1, \ldots, \underset{\sim}{a}_r$ are non-zero points on distinct edges, then

$$C = \{\lambda_1 \underset{\sim}{a}_1 + \ldots + \lambda_r \underset{\sim}{a}_r \mid \lambda_i \geqslant 0 \quad \forall\ i\}$$

$$= \text{cone}_{\underset{\sim}{0}} (\underset{\sim}{a}_1, \ldots, \underset{\sim}{a}_r) ,$$

and $\dim C \leqslant r$.

Definition 1A.1. A pointed polyhedral cone C for which $\dim C$ is equal to the number of edges of C will be called a simplicial cone.

We comment that if P is an n-dimensional polytope in R^d and $\underset{\sim}{b} \notin \text{aff } P$, then $C = \text{cone}_{\underset{\sim}{b}} (P)$ is an $(n+1)$-

dimensional pointed polyhedral cone, and every pointed polyhedral cone in R^d is of this form. Apart from the empty face of C, there is a one-one correspondence between the faces F of P and the faces $\text{cone}_{\underset{\sim}{b}}(F)$ of C. The cone C is a simplicial cone if and only if P is a simplex.

We complete this section by stating two lemmas which will be used in the next section. If $P \subseteq R^d$, we will write $\text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a},P)$ to mean $\text{cone}_{\underset{\sim}{o}}(\{\underset{\sim}{a}\} \cup P)$.

Lemma 1A.3. If P is an n-dimensional pointed polyhedral cone with apex $\underset{\sim}{b}$ in R^d, and $\underset{\sim}{a}$ is a point in R^d such that $\underset{\sim}{a} \notin \text{aff } P$, then $\text{cone}_{\underset{\sim}{b}}(\underset{\sim}{a},P)$ is an (n+1)-dimensional pointed polyhedral cone with apex $\underset{\sim}{b}$, and the faces of $\text{cone}_{\underset{\sim}{b}}(\underset{\sim}{a},P)$ are the $\text{cone}_{\underset{\sim}{b}}(\underset{\sim}{a},F)$, where F is a face of P, together with the faces of P.

The lemma is easily proved.

Lemma 1A.4. (Grünbaum [5], Theorem 2.6.1). If P is a polyhedral set, then a face of a face of P is again a face of P.

§1B. Barycentric subdivision of cones.

In this section we will introduce the main geometric tool to be used in Chapter 3 for subdivision of lattice cones, namely barycentric subdivision. Since it turns out that this method leads to complexes of cones, we consider subdivision of complexes of cones into complexes of subcones.

Definition. A finite collection C of polyhedral sets in R^d is a complex (in R^d) if any two members of C meet in a face of each of them.

Sometimes "complex" is defined with the additional requirement that any face of a member is again a member. (See, for example, Grünbaum [5], §3.2.) The next lemma, (which is also needed later in this section) shows that the difference is not essential.

Definition. If C is a complex, a face of C is a face of a member of C.

Lemma 1B.1. Let C be a complex in R^d. Then any two faces of C meet in a face of each of them.

Proof. Let $C_1, C_2 \in C$, and let F_1, F_2 be faces of C_1, C_2 respectively. By symmetry, it is sufficient to prove that $F_1 \cap F_2$ is a face of F_1. By Lemma 1A.4 it is sufficient to prove that

(i) $F_1 \cap F_2$ is a face of $F_1 \cap C_2$, and

(ii) $F_1 \cap C_2$ is a face of F_1.

To prove (i), we may suppose that $F_2 = C_2 \cap H$, where H is a hyperplane which does not cut C_2, and hence does not cut $F_1 \cap C_2$. Thus $F_1 \cap F_2 = (F_1 \cap C_2) \cap H$ is a face of $F_1 \cap C_2$. To prove (ii), $F_1 \cap C_2 = (F_1 \cap C_1) \cap C_2 = F_1 \cap (C_1 \cap C_2)$ $= F_1 \cap (C_1 \cap K)$, where K is R^d or is a hyperplane which does not cut C_1 and hence does not cut F_1. Thus $F_1 \cap C_2 = F_1 \cap K$ is a face of F_1 as required.

Definition 1B.1. Let C be a complex of polyhedral sets in R^d. Then a collection $\mathcal{D}$ of polyhedral sets is a subdivision of C if

(a) $\cup \mathcal{C} = \cup \mathcal{D}$,

(b) each member of $\mathcal{C}$ is a union of members of $\mathcal{D}$,

(c) $\mathcal{D}$ is a complex.

The next theorem introduces barycentric subdivision, which is an important way of constructing subdivisions. We will formulate the idea for pointed polyhedral cones with apex $\underset{\sim}{0}$, for this is what we will need in our later investigation of lattice cones. (An analogous theory on "triangulation of polytopes" may be inferred from the correspondence between polytopes and cones that was mentioned in §1A.)

Notation. If S is a collection of sets in R^d, and $\underset{\sim}{a} \in R^d$, then $\text{cone}_{\underset{\sim}{0}}(\underset{\sim}{a}, S)$ denotes the set of all $\text{cone}_{\underset{\sim}{0}}(\underset{\sim}{a}, A)$ with A in S.

Theorem 1B.2. (On barycentric subdivision.) Let S be a complex of pointed polyhedral cones in R^d with apex $\underset{\sim}{0}$, and let $\underset{\sim}{a} \in \cup S$, $\underset{\sim}{a} \neq \underset{\sim}{0}$. Let $S(\underset{\sim}{a})$ be the collection of cones obtained from S by replacing each cone C in S containing $\underset{\sim}{a}$ by all the cones $\text{cone}_{\underset{\sim}{0}}(\underset{\sim}{a}, F)$, where F is a facet of C not containing $\underset{\sim}{a}$. (Notice that if $\underset{\sim}{a} \in C$, then $\underset{\sim}{a} \notin F \Leftrightarrow \underset{\sim}{a} \notin \text{aff } F$.) Then $S(\underset{\sim}{a})$ is a subdivision of S into pointed polyhedral cones with apex $\underset{\sim}{0}$; and if S consists of n-dimensional cones, so does $S(\underset{\sim}{a})$.

Proof. We need only prove that $S(\underset{\sim}{a})$ is a subdivision of S, as the rest follows from Lemma 1A.3. For C in S such that $\underset{\sim}{a} \in C$, let us write

$$F_C = \{F \mid F \text{ is a facet of } C \text{ such that } \underset{\sim}{a} \notin F\},$$

and write

$$S_1 = \bigcup_C F_C, \quad S_2 = \{C \in S \mid \underset{\sim}{a} \notin C\}.$$

Then

$$S(\underset{\sim}{a}) = S_2 \cup \text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, S_1).$$

We are required to prove (a), (b), (c) in Definition 1B.1 with S for C, and $S(a)$ for $\mathcal{D}$. It is obvious that $\cup S(a) \subseteq S$. Hence (a) will follow from (b).

Proof of (b) : Let $C \in S$. If $\underset{\sim}{a} \notin C$, then $C \in S(\underset{\sim}{a})$, so suppose that $\underset{\sim}{a} \in C$. It suffices to prove that

(1) $$C = \cup \text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, F_C).$$

We need only prove that the left hand side is included in the right hand side, as the reverse inclusion is trivial.

Firstly, we prove that the ray $\underset{\sim}{o}\,\underset{\sim}{a}$ in C belongs to the right hand side of (1). Since $\underset{\sim}{o}\,\underset{\sim}{a}$ is contained in each cone of the union, we need only prove that $F_C \neq \emptyset$. If $C = \underset{\sim\sim}{oa}$, then $\underset{\sim}{o} \in F_C$, so $F_C \neq \emptyset$. On the other hand, if $C \neq \underset{\sim\sim}{oa}$, then the facets of C do not all pass through $\underset{\sim\sim}{oa}$, as $\underset{\sim}{o}$ is an extreme point of C. Hence some facet does not contain $\underset{\sim}{a}$, and again $F_C \neq \emptyset$.

Secondly, suppose that $\underset{\sim}{y} \in C$, $\underset{\sim}{y} \notin \underset{\sim\sim}{oa}$. Then $C \neq \underset{\sim\sim}{oa}$ and, as we have just seen, the facets of C do not all pass through $\underset{\sim\sim}{oa}$, and hence do not all contain the triangle $\underset{\sim\sim\sim}{oay}$. Rotate a ray from $\underset{\sim\sim}{oa}$ through $\underset{\sim\sim}{oy}$ till it meets a facet F of C not containing $\underset{\sim\sim\sim}{oay}$. Then $F \in F_C$, and $\underset{\sim}{y} \in \text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, F)$. This completes the proof of (1), and of (b).

Proof of (c) : Let $D_1, D_2 \in S(\underset{\sim}{a})$. We must prove that

(2) $D_1 \cap D_2$ is a face of each of D_1 and D_2.

If D_1, D_2 do not contain a, then $D_1, D_2 \in S$, so (2) follows from the fact that S is a complex. Thus we may

suppose that for some $i = 1$ or 2, $\underset{\sim}{a} \in D_i$, and

$$(3)\begin{cases} D_i = \text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, F_i) \subseteq C_i \in S, \\ F_i \in F_{C_i}, \quad \underset{\sim}{a} \notin F_i, \quad \underset{\sim}{a} \in C_i. \end{cases}$$

Without loss of generality suppose that (3) holds with $i = 1$. There are two cases to consider:

(i) $\underset{\sim}{a} \notin D_2$, in which case $D_2 = C_2 \in S$, or

(ii) $\underset{\sim}{a} \in D_2$, in which case (3) also holds for $i = 2$.

To prove (2) in case (i), it is sufficient, by Lemma 1A.3, to prove that

$$(4) \qquad \text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, F_1) \cap C_2 = F_1 \cap C_2,$$

for by Lemma 1B.1, $F_1 \cap C_2$ is a face of each of F_1 and C_2. To prove (2) in case (ii), it is sufficient by Lemma 1A.3, to prove that

$$(5) \qquad \text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, F_1) \cap \text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, F_2) = \text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, F_1 \cap F_2),$$

for by Lemma 1B.1, $F_1 \cap F_2$ is a face of each of F_1 and F_2.

Now to prove (4) in case (i), it is clearly sufficient to prove that if $\underset{\sim}{x}_2 \in \dot{C}_2$, then

$$(6) \qquad \underset{\sim}{x}_2 \notin \text{relint cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, \underset{\sim}{x}_1)$$

for any $\underset{\sim}{x}_1 \in \dot{F}_1$. So suppose $\underset{\sim}{x}_2 \in \dot{C}$, $\underset{\sim}{x}_1 \in \dot{F}_1$. We will prove (6) by contradiction. Suppose (6) is false. Then $\underset{\sim}{x}_2 \in C_1 \cap C_2$, and any hyperplane which contains $\underset{\sim}{x}_2$ but does not cut C_1 must contain $\text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, x_1)$ and hence $\underset{\sim}{a}$. But $C_1 \cap C_2$ is a face of C_1 since S is a complex. Hence $C_1 \cap C_2$ contains $\text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, x_1)$ and hence $\underset{\sim}{a}$, contradicting the fact that $\underset{\sim}{a} \notin C_2$. This completes the proof of (4).

Now to prove (5) in case (ii), we note that

$$\text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, F_i) = \bigcup_{\underset{\sim}{x}_i \in \dot{F}_i} \text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, \underset{\sim}{x}_i) \qquad (i = 1,2).$$

Hence the left hand side of (5) is equal to the union of intersections

$$\text{cone}_{\underset{\sim}{o}}\ (\underset{\sim}{a},\underset{\sim}{x}_1) \cap \text{cone}_{\underset{\sim}{o}}\ (\underset{\sim}{a},\underset{\sim}{x}_2)\ ,$$

where $\underset{\sim}{x}_1 \in \dot{F}_1$, $\underset{\sim}{x}_2 \in \dot{F}_2$. Thus it is sufficient to prove that (6) holds for $\underset{\sim}{x}_1 \in \dot{F}_1$, $\underset{\sim}{x}_2 \in \dot{F}_2$, and we do this by contradiction. Suppose (6) is false. Then, as in case (i), $C_1 \cap C_2$ contains $\text{cone}_{\underset{\sim}{o}}\ (\underset{\sim}{a},\underset{\sim}{x}_1)$. Hence $\text{cone}_{\underset{\sim}{o}}\ (\underset{\sim}{a},\underset{\sim}{x}_1) \subseteq C_2$, contradicting the assumption that the point $\underset{\sim}{x}_2$ in relint $\text{cone}_{\underset{\sim}{o}}\ (\underset{\sim}{a},\underset{\sim}{x}_1)$ lies on a facet F_2 of C_2 such that $\underset{\sim}{a} \notin \text{aff}\ F_2$. This completes the proof of (5), and the proof of the theorem.

Definition 1B.2. The complex $S(\underset{\sim}{a})$ in Theorem 1B.2 is said to be obtained from S by barycentric subdivision with division point $\underset{\sim}{a}$.

Barycentric subdivision of a simplicial cone. We will describe this case in terms of generators. Let

$$C = \text{cone}_{\underset{\sim}{o}}\ (\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n)$$

be an n-dimensional simplicial cone (Definition 1A.1) with non-zero points $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n$ on distinct edges, and let

$$\underset{\sim}{a} = \lambda_1\underset{\sim}{a}_1 + \ldots + \lambda_n\underset{\sim}{a}_n \in C, \quad \underset{\sim}{a} \neq \underset{\sim}{o}.$$

Let

$$I = \{i \mid \lambda_i > 0\},$$

and define

$$C_i = \text{cone}_{\underset{\sim}{o}}\ (\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_{i-1},\underset{\sim}{a},\underset{\sim}{a}_{i+1},\ldots,\underset{\sim}{a}_n)$$

for $i \in I$. It is easily seen that barycentric subdivision of C with division point $\underset{\sim}{a}$ gives

$$C(\underset{\sim}{a}) = \{C_i \mid i \in I\},$$

and that if

$$\underset{\sim}{b} = \mu_1 \underset{\sim}{a}_1 + \ldots + \mu_n \underset{\sim}{a}_n \in C, \qquad \underset{\sim}{b} \neq \underset{\sim}{o},$$

then

$$\underset{\sim}{b} \in C_i \Leftrightarrow \frac{\mu_i}{\lambda_i} = \min_{j \in I} \frac{\mu_j}{\lambda_j} .$$

Definition. A complex of pointed polyhedral cones is simplicial if all its members are simplicial.

Theorem 1B.3. (Simplicial subdivision.) Let S be a complex of n-dimensional pointed polyhedral cones in R^d with apex $\underset{\sim}{o}$. Then

(i) S is simplicial if and only if $S(\underset{\sim}{a}) = S$ whenever $\underset{\sim}{a}$ is a division point on an edge of S.

(ii) Some finite sequence of successive barycentric subdivisions yields a subdivision T of S into n-dimensional simplicial cones such that the edges of T are just the edges of S.

Proof. (i) It is clearly sufficient to prove the result in the case when S consists of a single cone, say $S = \{C\}$. The necessity of the condition is obvious. We will prove sufficiency by mathematical induction. The result is trivially true when $n = 1$ or 2. Suppose that the result is true when $n = k \geqslant 2$, and let $n = k + 1$. We assume that $C(\underset{\sim}{a}) = C$ whenever $\underset{\sim}{a}$ is a non-zero point on an edge of C, and we are required to prove that C is simplicial. Suppose not. Then C has more than $k+1$ edges. Choosing any $\underset{\sim}{a}$ as above, we deduce from $C(\underset{\sim}{a}) = C$ that there is a facet D of C such that $\underset{\sim}{a} \notin \text{aff } D$ and $C = \text{cone}_{\underset{\sim}{o}} (\underset{\sim}{a}, D)$. Then D is k dimensional with more than k edges, and so is not simplicial. By induction hypothesis, there is a non-zero point $\underset{\sim}{b}$ on an edge of D

such that $D(\underset{\sim}{b})$ has at least two members D_1, D_2. Thus $C(\underset{\sim}{b})$ has at least two members $\text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, D_1)$, $\text{cone}_{\underset{\sim}{o}}(\underset{\sim}{a}, D_2)$, contradicting the fact that $C(\underset{\sim}{b}) = C$.

(ii) Beginning with the complex S, make repeated use of barycentric subdivision, where the division point is on an edge of the complex. The set of edges remains unchanged. By (i), the number of members of the complex increases at each step unless a simplicial complex is reached. This eventually happens because the number of simplexes whose edges are edges of S is finite.

The theory of subdivision cannot be simplified by choosing only division points in the relative interiors of cones. For example, if P is a 3-dimensional square pyramid in R^4 such that $\underset{\sim}{o} \notin \text{aff } P$, then $\text{cone}_{\underset{\sim}{o}}(P)$ cannot be subdivided into simplicial cones by repeated barycentric subdivision using only division points in the relative interiors of cones.

§1C. A combinatorial result.

We present here a combinatorial result which essentially belongs to topology. We will need it in Chapter 5.

Lemma 1C.1. Let S be a complex of 3-dimensional simplicial cones in R^3 such that US is simply connected. Then

$$2E_i + E_b = N + 2,$$

where E_i is the number of edges of S in the relative interior of US, E_b is the number of edges on the relative boundary of US, and N is the number of members of S.

Proof. The result is easily proved by induction on

N. It is of course equivalent to the well-known result that the number of primitive triangles in a primitive triangulation of a simple polygon is

$$2V_i + V_b - 2,$$

where V_i, V_b are the number of vertices in the interior and boundary respectively of the polygon. (See, for example, Gaskell, Klamkin and Watson [4].)

CHAPTER 2

AUXILIARY RESULTS ON RESIDUES

In this chapter we will obtain results on residue sets which are needed later (for the proofs of Proposition 3E.1 and Lemma 5B.1). The main result is Proposition 2.6 below.

Let m be a fixed positive integer greater than 1, and for any integer x let $\bar{x}$ denote the least non-negative residue of x modulo m, so that

$$0 \leqslant \bar{x} < m, \qquad x \equiv \bar{x} \pmod{m}.$$

Definition 2.1. For x in Z such that $1 \leqslant x \leqslant m-1$, let ρ_x be the function with domain Z such that for r in Z,

$$\rho_x(r) = \begin{cases} 1 & \text{if } \overline{rx} = \overline{(r-1)x} + x - m \\ 0 & \text{if } \overline{rx} = \overline{(r-1)x} + x \,. \end{cases}$$

We note that for r in Z

$$\overline{rx} - \overline{(r-1)x} = x - m\rho_x(r), \tag{1}$$

and the following properties of ρ_x are thus immediate.

Lemma 2.1. For x and r in Z and $1 \leqslant x \leqslant m-1$ we have:

(i) $\rho_x(r) = \rho_x(\bar{r})$;

(ii) $\rho_x(r) = 1 \Leftrightarrow \overline{rx} < x$, and
$= 0 \Leftrightarrow \overline{rx} \geqslant x$;

(iii) $\sum_{r=1}^{m} \rho_x(r) = x$;

(iv) $\rho_1(r) = 0$ if $\bar{r} \neq 0$;

(v) $\rho_x(0) = 1$;

(vi) $\rho_x(1) = 0$;

(vii) $\rho_x(2) = 1 \Leftrightarrow x \geqslant \frac{m}{2}$.

In view of (i), (v), (vi), we now consider the restriction $\rho_x|S$ of ρ_x to the subset S of Z given by

(2) $$S = \{2,3,\ldots,m-1\} .$$

Lemma 2.2. Let S be as in (2), and let ρ_x be as in Definition 2.1.

(i) $\rho_1|S = 0$ (the zero function on S).

(ii) $\rho_{m-1}|S = 1$ (the constant function 1 on S).

(iii) If $(a,m) = 1$, $1 \leq a \leq m-1$, then

$$(\rho_a|S) + (\rho_{m-a}|S) = 1 \ (=\rho_{m-1}|S).$$

(iv) For $1 \leq x \leq m-1$,

$$\sum_{r \in S} \rho_x(r) = x - 1.$$

(v) Suppose $a_1,a_2,\ldots,a_k,b$ are integers such that $1 \leq a_i \leq m-1$ $(i = 1,\ldots,k)$, $1 \leq b \leq m-1$, and

$$\sum_{i=1}^{k} (\rho_{a_i}|S) = \rho_b|S.$$

Then

$$\sum_{i=1}^{k} a_i = b + k - 1.$$

(vi) Suppose $a_1,a_2,\ldots,a_k$ are integers such that $1 \leq a_i \leq m-1$ $(i = 1,\ldots,k)$, and

$$\sum_{i=1}^{k} a_i > (k-1)m - (k-2).$$

Then for some r in S,

$$\rho_{a_1}(r) = \ldots = \rho_{a_k}(r) = 1,$$

that is,

$$\overline{ra_1} < a_1, \ \ldots \ , \ \overline{ra_k} < a_k.$$

(Part (vi) is needed later in §3E.)

Proof. Parts (i), (ii) are immediate from the definitions, (iv) is immediate from Lemma 2.1(iii), (v), (vi). Part (v) is immediate from (iv).

(iii) Since $\overline{ra} + \overline{r(m-a)} \equiv 0 \pmod{m}$, $\overline{ra} + \overline{r(m-a)}$ equals 0 if both summands are 0, and equals m otherwise. If $(a,m) = 1$, the latter is the case for all r in S, and the result follows from Lemma 2.1(ii).

(vi) $$\sum_{r\in S}\left(\rho_{a_1}(r)+\ldots+\rho_{a_k}(r)\right) = \sum_{i=1}^{k} a_i - k \quad \text{by (iv)}$$
$$> (k-1)(m-2) \quad \text{by hypothesis,}$$
$$= (k-1)\ \#\ S,$$

so for some r in S,

$$\rho_{a_1}(r) + \ldots + \rho_{a_k}(r) = k.$$

Thus $\rho_{a_1}(r) = \ldots = \rho_{a_k}(r) = 1$, and by Lemma 2.1(ii), $\overline{ra_1} < a_1, \ldots, \overline{ra_k} < a_k$. This completes the proof of Lemma 2.2.

Let $[y]^+$ denote the least integer greater than or equal to y, that is,

$$[y]^+ - 1 < y \leqslant [y]^+.$$

Lemma 2.3 (i) For all x (such that $1 \leqslant x \leqslant m-1$), the t^{th} smallest positive value of r for which $\rho_x(r) = 1$ is

$$\left[\frac{tm}{x}\right]^+ .$$

(ii) $$\left[\frac{(t+1)m}{x}\right]^+ = \left[\frac{tm}{x}\right]^+ + \left[\frac{m}{x}\right]^+ - \varepsilon$$

where $\varepsilon = 0$ or 1.

Proof. (i) This is immediate from the definition of ρ_x.

(ii) This is just a special case of the general result that for any real numbers y, z,

$$[y+z]^+ = [y]^+ + [z]^+ - \varepsilon ,$$

where $\varepsilon = 0$ or 1.

Clearly the behaviour of $\rho_x|S$, where S is as in (2), can be formulated in terms of the set

$$S_x = \{r \mid r \in S, \quad \rho_x(r) = 1\}$$
$$= \{r \mid r \in S, \quad \overline{rx} < x\} \qquad (1 \leqslant x \leqslant m-1),$$

of which $\rho_x | S$ is the indicator. The following lemma is most easily formulated in this way, and is the essence of the proof of Proposition 2.6(ii).

Lemma 2.4. Let b,c be integers such that $1 \leqslant b \leqslant m-1$, $1 \leqslant c \leqslant m-1$. Suppose that $S_b \subseteq S_c$, $S_b \neq S_c$, and $\left[\frac{m}{b}\right]^+ = \left[\frac{m}{c}\right]^+$. Then $\left[\frac{m}{c}\right]^+ = 2$ (that is, $\rho_c(2) = 1$, or $c \geqslant \frac{m}{2}$).

Proof. The condition $S_b \subseteq S_c$ means that $\rho_b(r) \leqslant \rho_c(r)$ for all positive r. Since $S_b \neq S_c$, there is some $r = r_0$ in S such that $\rho_b(r_0) = 0$, $\rho_c(r_0) = 1$, and $\rho_b(r) = \rho_c(r)$ for r in S with $r < r_0$. If r_0 is the (t_0+1)-th smallest integer in S for which $\rho_c(r_0) = 1$, then by Lemma 2.3(i),

$$\left[\frac{tm}{b}\right]^+ = \left[\frac{tm}{c}\right]^+ \qquad \text{for} \quad t = 1,\ldots,t_0,$$

and

$$\left[\frac{(t_0+1)m}{b}\right]^+ = \left[\frac{(t_0+1)m}{c}\right]^+ + \eta = r_0 + \eta = r_1 \quad \text{say} \quad (\leqslant m),$$

where $\eta > 0$, and by assumption, $t_0 \geqslant 1$. Applying Lemma 2.3(ii) with $t = t_0$, and $x = b,c$, we get

$$\left[\frac{(t_0+1)m}{b}\right]^+ = \left[\frac{t_0 m}{b}\right]^+ + \left[\frac{m}{b}\right]^+ - \varepsilon_1 ,$$
$$\left[\frac{(t_0+1)m}{c}\right]^+ = \left[\frac{t_0 m}{c}\right]^+ + \left[\frac{m}{c}\right]^+ - \varepsilon_2 ,$$

where $\varepsilon_1, \varepsilon_2 = 0$ or 1. Hence $\eta = \varepsilon_2 - \varepsilon_1 = 1$ (and $\varepsilon_2 = 1$, $\varepsilon_1 = 0$) and $r_1 = r_0 + 1$. Since $\rho_b(r_1) = 1$ (by Lemma 2.3(i)), $\rho_c(r_1) = 1$, and so $\left[\frac{(t_0+2)m}{c}\right]^+ = r_1 = r_0 + 1$. But by Lemma 2.3(ii) with $t = t_0 + 1$, $x = c$,

$$r_1 = r_0 + \left[\frac{m}{c}\right]^+ - \varepsilon_3,$$

where $\varepsilon_3 = 0$ or 1. Hence $\left[\frac{m}{c}\right]^+ = 1 + \varepsilon_3$, which must be equal to 2 since $m > c$.

We shall need one further lemma for our main proposition.

Lemma 2.5. Let S be as in (2), and let $a_1,\ldots,a_k$ be integers such that $1 \leqslant a_i \leqslant m-1$ for $i = 1,\ldots,k$. Then the statement

$$(3) \qquad \sum_{i=1}^{k} (\rho_{a_i}|S) = 1 \qquad (= \rho_{m-1}|S)$$

(that is, $\sum_{i=1}^{k} \rho_{a_i}(r) = 1 \quad (= \rho_{m-1}(r))$ for $r = 2,\ldots,m-1$) is equivalent to

$$(4) \qquad \sum_{i=1}^{k} \overline{ra_i} = m + r(k-2) \qquad \text{for } r = 1,2,\ldots,m-1.$$

Proof. From equation (1) (following Definition 2.1) we obtain the identity

$$(5) \qquad \sum_{i=1}^{k} \overline{ra_i} - \sum_{i=1}^{k} \overline{(r-1)a_i} = \sum_{i=1}^{k} a_i - m \sum_{i=1}^{k} \rho_{a_i}(r).$$

Suppose that (3) holds. Then Lemma 2.2(v) gives

$$(6) \qquad \sum_{i=1}^{k} a_i = m + k - 2,$$

which is (4) with $r = 1$. Condition (4) follows for all r by induction using (5).

Conversely, suppose that (4) holds. Then (3) follows from (5) on applying (4) with r replaced by r, $r-1$ and 1 $(2 \leqslant r \leqslant m-1)$.

The following is the main result of this chapter.

Proposition 2.6. Let S be as in (2), and let $a_1,\dots,a_k$ be integers such that

$$1 \leqslant a_1 \leqslant \dots \leqslant a_k \leqslant m-1,$$

and

$$(4) \qquad \sum_{i=1}^{k} \overline{ra_i} = m + r(k-2) \qquad \text{for } r = 1,2,\dots,m-1.$$

(i) Then

$$(a_i,m) = 1 \qquad \text{for } i = 1,\dots,k.$$

(ii) Also,

$$1 = a_1 = \dots = a_{k-2} \leqslant a_{k-1} \leqslant a_k,$$

$$a_{k-1} + a_k = m,$$

and $a_{k-1} = a_k$ only if $m = 2$.

Proof. (i) Suppose that for $i = k$,

$$(7) \qquad (a_k,m) = \delta > 1.$$

Let $\delta' = m/\delta$. For integral x, let $\langle x\rangle$ be given by

$$\langle x\rangle \equiv x \pmod{\delta}, \qquad 0 \leqslant \langle x\rangle \leqslant \delta - 1,$$

while, as before, $\bar{x}$ is given by

$$\bar{x} \equiv x \pmod{m}, \qquad 0 \leqslant \bar{x} \leqslant m-1.$$

Let s be an integer such that $1 \leqslant s \leqslant \delta - 1$. From (7),

$$(8) \qquad \overline{s\delta' a_k} = 0.$$

Also, as is easily checked,

$$(9) \qquad \langle sx\rangle = \overline{s\delta' x/\delta'}\ .$$

From (4) with $r = s\delta'$, we obtain on using (8), dividing by δ', and using (9), that

$$(10) \qquad \sum_{i=1}^{k-1} \langle sa_i\rangle = \delta + s(k-2) \qquad \text{for } 1 \leqslant s \leqslant \delta - 1.$$

Let us see if this condition is of the same form as (4). We may of course replace each a_i in (10) by $\langle a_i\rangle$, and

$$0 \leqslant \langle a_i\rangle \leqslant \delta - 1 \qquad \text{for } i = 1,\dots,k-1.$$

Also, $\langle a_k \rangle = 0$ by (7). Now if $1 \leqslant i \leqslant k$, and $\langle a_i \rangle = 0$, then $\delta | a_i$, $m | \delta' a_i$, so $\rho_{a_i}(\delta') = 1$. Thus, if there were more than one value of i with $1 \leqslant i \leqslant k$ for which $\langle a_i \rangle = 0$, then

$$\sum_{i=1}^{k} \rho_{a_i}(\delta') > 1 .$$

By Lemma 2.5 (the (4) $\Rightarrow$ (3) part), this contradicts (4). Thus

$$1 \leqslant \langle a_i \rangle \leqslant \delta - 1 \qquad \text{for } i = 1,\ldots,k-1.$$

Thus, (10) is a similar condition to (4), but with δ in place of m, $k-1$ in place of k, and $k-2$ instead of the corresponding $k-3$ on the right hand side. By the corresponding argument to that of the proof in Lemma 2.5 that (4) $\Rightarrow$ (3) $\Rightarrow$ (6), we obtain

$$\langle a_1 \rangle + \ldots + \langle a_{k-1} \rangle = \delta + (k-3),$$

which now contradicts (10) when $s = 1$. This contradiction shows that (i) is true for $i = k$.

(ii) By Lemma 2.5, we have

$$\sum_{i=1}^{k} (\rho_{a_i} | S) = 1 = \rho_{m-1} | S. \tag{3}$$

If $a_{k-1} = 1$, then for $i \leqslant k-1$, $a_i = 1$, so $\rho_{a_i} | S = 0$, so $\rho_{a_k} | S = \rho_{m-1} | S$, and $a_k = m - 1$ (by Lemma 2.1(iii), for example), and the result follows. Now suppose that $a_{k-1} > 1$. Then $\rho_{a_{k-1}} | S \neq 0$, and $\rho_{a_k} | S \neq \rho_{m-1} | S$. From (3) and the assumption that $a_1 \leqslant \ldots \leqslant a_k$, we see that

$$\left[\frac{m}{a_k}\right]^{+} = \left[\frac{m}{m-1}\right]^{+} = 2, \qquad a_k \geqslant \frac{m}{2} . \tag{11}$$

Since $(a_k, m) = 1$ by (i) for $i = k$, Lemma 2.2(iii) gives

$$\rho_{a_k} | S + \rho_{m-a_k} | S = 1.$$

Hence (3) gives

$$\sum_{i=1}^{k-1} (\rho_{a_i}|S) = \rho_{m-a_k}|S.$$

Now suppose, if possible, that $a_1,\ldots,a_{k-2}$ are not all equal to 1. Then $\rho_{a_{k-1}}|S \neq \rho_{m-a_k}|S$, and from Lemma 2.4 with $b = a_{k-1}$ and $c = m - a_k$, it follows that a_{k-1}, $m - a_k \geqslant m/2$, so $a_k \leqslant m/2$. From (11), $a_k = m/2$. As $(a_k,m) = 1$, it follows that $m = 2$. This is a contradiction, as (ii) (and (i)) are trivially true for $m = 2$. Thus $a_1,\ldots,a_{k-2}$ are all equal to 1, so (3) becomes

$$\rho_{a_{k-1}}|S + \rho_{a_k}|S = 1.$$

Hence, by Lemma 2.2(v),

$$a_{k-1} + a_k = m$$

and since $(a_k,m) = 1$, $a_{k-1} \neq a_k$ unless $m = 2$. This completes the proof of (ii).

Finally, (i) for $1 \leq i \leq k-1$ follows immediately from (i) for $i = k$ and (ii).

Note 1. In Chapter 5, we will be using Proposition 2.6 in the case when $k = 3$. If we were to prove the result in this case only, the only part of the proof which simplifies is the proof of (i) : If $k = 3$, and $(a_3,m) = \delta > 1$, then $\delta' = \frac{m}{\delta}$,

$$\overline{\delta' a_1} + \overline{\delta' a_2} = m + \delta' > m$$

$$\overline{(m-\delta')a_1} + \overline{(m-\delta')a_2} = m + (m-\delta') > m,$$

contradicting the fact that the sum of the left hand sides is less than or equal to $2m$.

Note 2. Proposition 2.6 says that the only decompositions of $\rho_{m-1}|S$ into the sum of two or more $\rho_x|S$ are of the form

$$\rho_{m-1}|S = \rho_a|S + \rho_{m-a}|S + 0 + 0 + \ldots$$

where $(a,m) = 1$. Since condition (3) is equivalent to

$$\sum_{i=1}^{k-1} \rho_{a_i}|S = \rho_{m-a_k}|S$$

provided $(a_k,m) = 1$, it follows that for a_k such that $(a_k,m) = 1$ and $a_k \neq m-1$, $\rho_{a_k}|S$ has only the trivial decompositions

$$\rho_{a_k}|S = \rho_{a_k}|S + 0 + 0 + \ldots$$

It is an obvious question whether there is a non-trivial decomposition

(12) $$\rho_{a_1}|S + \ldots + \rho_{a_{k-1}}|S = \rho_{a_k}|S$$

when $(a_k,m) = \delta > 1$. For this it is clearly necessary that at most one of the (a_i,m), $i \leq k-1$, exceeds 1. In fact we can prove by methods similar to those used in the proof of Proposition 2.6(i), that for any such decomposition (12), $a_{i_0} = h_{i_0}\delta$ for some $i_0 \neq k$, $a_i = h_i\delta + 1$ for $i \neq i_0,k$, and $a_k = (h_1+\ldots+h_{k-1})\delta$.

CHAPTER 3

SUBDIVISION OF n-DIMENSIONAL LATTICE CONES

In §3A I will introduce the subdivision problem for lattice cones, and outline the material covered in the rest of Chapter 3, and in Chapters 4, and 5. I will begin with some necessary definitions, show how a main result of Chapter 1 may be applied to lattice cones, and then discuss a problem of Schinzel, which can be solved by subdivision of lattice cones.

§3A Introduction to subdivision.

§3A.1. Preliminary definitions and results.

We will assume the geometrical preliminaries of Chapter 1 concerning complexes of polyhedral sets in R^d, particularly the results concerning subdivision of complexes of pointed polyhedral cones. We will also use basic results concerning indices of sublattices of point lattices in R^d, which will be stated in §3B.

An s-dimensional (point) lattice in R^d is defined to be the set of all linear combinations with integer coefficients, of s linearly independent points in R^d. A point $\underset{\sim}{a}$ is said to be a primitive point of the lattice L if $r\underset{\sim}{a} \in L$ if and only if $r \in Z$.

Let L be a fixed lattice in R^d.

Definition. An (n-dimensional) L-cone in R^d is an (n-dimensional) pointed polyhedral cone C (with apex $\underset{\sim}{0}$) in R^d, such that each edge of C contains a non-zero point (and hence a unique primitive point) of L.

The simplest type of L-cone is now defined, and then a special case.

Definition. A simplicial L-cone is a simplicial cone which is also an L-cone.

Definition. A basic L-cone is an n-dimensional simplicial L-cone C in R^d such that the set of primitive points on the n edges of C are a basis for the n-dimensional lattice $L' = L \cap \text{lin } C$.

Thus, C is a basic L-cone if and only if C is a basic L'-cone. Also, the faces of a simplicial (or basic) L-cone are clearly simplicial (or basic) L-cones.

We notice that if we perform barycentric subdivision on an L-cone, using points of L as division points, the resulting subcones will be L-cones.

Note 3A.1. Henceforth we will consider only cones which are L-cones. We will be considering complexes of n-dimensional L-cones, and the only subdivisions of these that we will consider are subdivisions into n-dimensional L-cones.

From Theorem 1B.3 on subdivision of a complex of cones, we immediately have the following result.

Lemma 3A.2. Let S be a complex of n-dimensional L-cones in R^d. Then by repeated barycentric subdivision, we may obtain a subdivision S_1 into n-dimensional simplicial L-cones, such that the edges of S_1 are just the edges of S.

The significance of an n-dimensional simplicial L-cone C being basic, is explained in the following lemma, which is an immediate consequence of the definitions.

Lemma 3A.3. If C is basic, then every point of $L \cap C$ is expressible as a linear combination of the n

primitive points on the edges of C, with non-negative integral coefficients. (The converse is also true, as will be seen from Proposition 3B.2(ii) below.)

§3A.2. The subdivision problem, and Schinzel's problem.

Our investigation of subdivision of n-dimensional L-cones in R^d will solve the following problems, and provide various bounds associated with them.

The Subdivision Problem: To find a basic subdivision of a given n-dimensional L-cone, that is, a subdivision into n-dimensional basic L-cones. (See Note 3A.1. above.)

Schinzel's Problem: Given an n-dimensional L-cone C in R^d, to find a finite subset B of $L \cap C$ such that every point $\underset{\sim}{x}$ in $L \cap C$ is expressible as a non-negative integral linear combination of an n-element subset $B_{\underset{\sim}{x}}$ of B.

According to W. Schmidt [12], A. Schinzel considered the special case of the latter problem, where $d = n$, L is an n-dimensional sublattice of Z^n, and C is the non-negative orthant in R^n, that is, the simplicial cone generated by the primitive points of L lying on the positive coordinate axes of R^n. Schinzel conjectured that finite B existed in this case, and Schmidt [12] proved this conjecture by a (non-constructive) compactness argument using induction on n.

It is clear from Lemma 3A.3 that a solution of the subdivision problem will yield a solution of Schinzel's problem, since we may take B to be the set of primitive lattice points on the edges of the subdivision and $B_{\underset{\sim}{x}}$ to be the lattice basis determined by a basic cone of the

subdivision containing $\underset{\sim}{x}$. Moreover, Lemma 3A.2 reduces the subdivision problem (and hence Schinzel's problem) for general L-cones to that for simplicial L-cones. I obtained a simple constructive solution of Schinzel's problem by this method, and a preliminary account of this work is given in Low [9].

Remark 1. If in Schinzel's problem, $B_{\underset{\sim}{x}}$ is an (n+1)-element subset of B, rather than an n-element subset, the problem becomes trivial. For in the case of simplicial C, to which the problem is reduced by Lemma 3A.2, if $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n$ is the set of primitive lattice points on the edges of C, and $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m$ are the coset representatives of the sub-lattice of $L \cap \text{lin } C$ generated by $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n$ chosen in the form

$$\alpha_1\underset{\sim}{a}_1 + \ldots + \alpha_n\underset{\sim}{a}_n$$

where $0 \leqslant \alpha_i < 1$ for $i = 1,\ldots,n$, then any $\underset{\sim}{x}$ in C is expressible in the form

$$\underset{\sim}{x}_i + \lambda_1\underset{\sim}{a}_1 + \ldots + \lambda_n\underset{\sim}{a}_n,$$

where the λ_j are non-negative integers.

Remark 2. If ϕ is a non-singular linear transformation on R^d, then the subdivision problem and Schinzel's problem for the L-cone C are clearly equivalent to the corresponding problems for the $\phi(L)$-cone $\phi(C)$. When C is a simplicial L-cone, this idea may be applied to put C in some sort of standard form, for example, that in which the primitive points of L on the edges of C are the natural unit vectors $\underset{\sim}{e}_1,\ldots,\underset{\sim}{e}_n$. Although such a standard form would be convenient in giving examples of L-cones, it does not help us in solving the subdivision problem, since subcones of a cone in standard form would not themselves be in standard form.

§3A.3. Various definitions.

We shall now introduce some concepts needed to continue our discussion of Schinzel's problem and the subdivision problem.

Let $\mathcal{C}$ and $\mathcal{D}$ be finite collections of n-dimensional L-cones such that $U\mathcal{C} = U\mathcal{D}$. We call $\mathcal{D}$ a decomposition of $\mathcal{C}$ if each cone in $\mathcal{C}$ is a union of members of $\mathcal{D}$. We call $\mathcal{D}$ a dissection of $\mathcal{C}$ if it is a decomposition of $\mathcal{C}$ into non-overlapping L-cones. (The term non-overlapping was defined in §1A.) A subdivision of a complex $\mathcal{C}$ (see Definition 1B.1) is thus a decomposition of $\mathcal{C}$ which is also a complex. By Lemma 1A.2, any subdivision of $\mathcal{C}$ is also a dissection of $\mathcal{C}$.

Now in order to solve Schinzel's problem it is not necessary to restrict consideration to subdivisions into basic cones. Decompositions into basic cones would suffice, with the sets B, $B_{\underset{\sim}{x}}$ being derived in the same way as before. In fact, Schmidt used decompositions, and in Low [9], for the sake of simplicity, I stated and proved the results in terms of dissections (although the word dissection was not used there), even though the method used, namely, barycentric subdivision, led naturally to subdivisions. Dissections are easier to handle than subdivisions because a dissection of $\mathcal{C}$ is just the union of dissections of the members of $\mathcal{C}$, and so to obtain dissections, it suffices to consider barycentric subdivisions of L-cones, rather than of complexes of L-cones, even if some division point used is not in the relative interior of the relevant cone.

Definition. Let C be an L-cone and let $\dot{L} = L \sim \{\underset{\sim}{0}\}$. A point $\underset{\sim}{x}$ in $\dot{L} \cap C$ is decomposable in $\dot{L} \cap C$ (or simply in C) if there exist $\underset{\sim}{x}_1$, $\underset{\sim}{x}_2$ in $\dot{L} \cap C$ such that $\underset{\sim}{x} = \underset{\sim}{x}_1 + \underset{\sim}{x}_2$. Otherwise, $\underset{\sim}{x}$ is indecomposable in $\dot{L} \cap C$ (or in C).

The notation $I(C)$ will be used for the set of indecomposable points of C.

Definition. For $\underset{\sim}{a} = (a_1, a_2, \ldots, a_d)$ in R^d the height $H(\underset{\sim}{a})$ of $\underset{\sim}{a}$ is defined to be

$$H(\underset{\sim}{a}) = \max_i |a_i| .$$

§3A.4. Plan of the rest of Chapter 3 and of Chapters 4 and 5.

In the remainder of Chapter 3 I will consider n-dimensional L-cones for general n. In §3B I will give the necessary preliminary results on the (outer) index of an L-cone, and on special L-cones. In §3C I will solve the subdivision problem and obtain crude bounds which in the case of Schinzel's problem are bounds for

(i) the number of points in B

(ii) the number of distinct $B_{\underset{\sim}{x}}$

(iii) the height of points in B.

In §3D we will study a refinement of the method of §3C, which would appear to lead to more economical bounds (and certainly does in the case of 2- and 3-dimensional L-cones, as we shall see later). In §3E we collect together results on indecomposable points.

Let S be a decomposition of an n-dimensional L-cone C into basic L-cones, and let $B(S)$ be the set of primitive points on the edges (of the members) of S. Then it is obvious that $B(S) \supseteq I(C)$. In Chapters 4 and 5 I will discuss 2-, and 3-dimensional L-cones, and will find in

each case a subdivision S of C into basic L-cones for which $B(S) = I(C)$. These results will yield sharp bounds for the numbers in (i), (ii) and (iii) for Schinzel's problem (in fact the best possible bounds for the case of subdivisions). I will also discuss there the equivalence of the concepts of decomposition, dissection and subdivision under certain conditions.

Note (added August, 1979).

Schinzel's problem is also considered in the (unpublished) thesis
R.A. Lee, A quantitative solution of a problem of Schinzel, Ph.D. thesis, University of Colorado (1970) (available through University Microfilms, London), which concentrates mainly on a different aspect of the problem.

§3B. Outer index of an L-cone, and special L-cones.

In this section, we discuss concepts needed for tackling the subdivision problem. Again, L will be a fixed lattice in R^d of dimension less than or equal to d. We will review the basic results on the index of an n-dimensional sublattice of an n-dimensional lattice. We will define the base points of an L-cone, the parallelotope of a simplicial L-cone, and the index of simplicial and the (outer) index of general L-cones, and give some results in terms of these ideas. The section ends with a discussion of special cones, which are relevant to §3D. All these concepts are needed for our investigation of subdivisions of L-cones in Chapters 3 to 5.

Notation for lattices. If $\underset{\sim}{b}_1,\ldots,\underset{\sim}{b}_n$ are linearly independent points in R^d, then the n-dimensional lattice $N = \{\lambda_1\underset{\sim}{b}_1+\ldots+\lambda_n\underset{\sim}{b}_n \mid \lambda_i \in Z \ \forall i\}$ with basis $\underset{\sim}{b}_1,\ldots,\underset{\sim}{b}_n$ is denoted by $N = \ell(\underset{\sim}{b}_1,\ldots,\underset{\sim}{b}_n)$.

Index of sublattice.

Definition 3B.1. If $M = \ell(\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n)$ is a sublattice of $N = \ell(\underset{\sim}{b}_1,\ldots,\underset{\sim}{b}_n)$ in R^d, then the index (N:M) is the subgroup index, that is, the number of cosets of M in N.

Proposition 3B.1. Let M,N be as in Definition 3B.1, and let the integral matrix $\Lambda = (\lambda_{ij})$ be given by

$$\underset{\sim}{a}_j = \sum_{i=1}^{n} \lambda_{ij}\underset{\sim}{b}_i , \qquad j = 1,\ldots,n.$$

Then

$$(N:M) = |\det \Lambda| ,$$

and in the particular case when $d = n$,

$$(N:M) = \frac{|\det(\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n)|}{|\det(\underset{\sim}{b}_1,\ldots,\underset{\sim}{b}_n)|}$$

The result is easily seen from Cassels [3] I.2.2, even though Cassels considers only those lattices in R^d which are d-dimensional (that is, d = n), and in fact takes the last equation as a definition, and proves the coset result as a lemma.

In order to state the lemma connecting the index of a sublattice, and linear combinations, and to facilitate our study of subdivision of cones, we now introduce parallelotopes.

Definition 3B.2. If $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n$ are linearly independent points in R^d, then the parallelotope of $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n$ is defined as

$$P = P(\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n) = \{\lambda_1\underset{\sim}{a}_1+\ldots+\lambda_n\underset{\sim}{a}_n \mid 0 \leqslant \lambda_i < 1 \ \forall\, i\}.$$

Note that P is a half-open parallelotope which does not include the vertices $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n$. The following result is easy to see.

Proposition 3B.2. Let N,M be n-dimensional lattices with M a sublattice of N, let $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n$ be a basis for M, and let P be the parallelotope of $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n$. Then

(i) $N \cap P$ is a complete set of representatives for the cosets of M in N.

(ii) $N = M \Leftrightarrow (N:M) = 1 \Leftrightarrow \dot{N} \cap P = \emptyset$.

Base points of, and notation for L-cones. If C is an n-dimensional L-cone in R^d with r edges, then C is determined by the r primitive points $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_r$ of L on the edges of C.

Definition 3B.3. The points $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_r$ are called a base for C, and to indicate this, we will use the notation

$$C = c(\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_r).$$

More generally, the primitive points on the edges of a collection S of L-cones will be called the base points of S, and $B(S)$ will denote the set of all these points.

Index of a simplicial L-cone.

Definition 3B.4. If $C = c(\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n)$ is a simplicial L-cone, the index of C (in L) is the index $(L':M)$, of the sublattice $M = \ell(\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n)$ in $L' = L \cap \operatorname{lin} C$, and is denoted by $m(C)$.

Thus, C is a basic L-cone (as defined in §3A.1) if and only if its index is 1. Notice that $\dim L' = \dim C = n$, and that the index of C in L is equal to its index in L'. We do not suppose that $L' = L$, that is $\dim L = \dim C$, as we wish to be able to consider a complex of L-cones C such that the $\operatorname{lin} C$ do not all coincide.

Definition 3B.5. If C is a simplicial L-cone, then the parallelotope of C is the parallelotope $P(\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n)$ of its base. Denote it by $P(C)$.

From Proposition 3B.2 we now have the following result in terms of $P(C)$ (which is equivalent to Lemma 3A.3 together with its converse).

Lemma 3B.3. If C is a simplicial L-cone, then C is basic if and only if $\dot{L} \cap P(C) = \emptyset$ (that is, if and only if the only points of L in the closure $\bar{P}(C)$ of $P(C)$ are the 2^n vertices).

We remark that if C is a simplicial L-cone, and $\underset{\sim}{x}$ is indecomposable in $L \cap C$, then $\underset{\sim}{x}$ is either a base point of C, or belongs to $P(C)$.

Outer index of a general L-cone. This is a notion which will be used in expressing bounds in Theorem 3C.2. (The notion of inner index will be introduced in §3D.)

Definition 3B.6. If $C = c(\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_r)$ is an n-dimensional L-cone, the (outer) index of C (in L) is

$$m(C) = v(C)/v_0(C),$$

where $v(C)$ is the (n-dimensional) volume of

(1) $$S(C) = \mathrm{conv}(\underset{\sim}{0},\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_r),$$

and $v_0(C)$ is the volume of the simplex generated by $\underset{\sim}{0}$ and any basis of the lattice $L' = L \cap \mathrm{lin}\, C$.

Clearly, if C is simplicial, then its index (as defined in Definition 3B.4) is equal to its outer index. (The index is also equal to the inner index defined later in §3D.)

Lemma 3B.4.

Let S be a complex of n-dimensional L-cones in R^d. Then

(i) By repeated barycentric subdivision, we may obtain a simplicial subdivision S' of S (see Note 3A.1.) such that the base points of S' are precisely those of S.

(ii) Any subdivision S_1 of S (whether simplicial or not) whose base points all lie in the union of the $S(C)$ given by (1) for $C \in S$ satisfies

(2) $$\sum_{D \in S_1} m(D) \leqslant \sum_{C \in S} m(C)$$

Proof. (i) is just Lemma 3A.2, expressed in terms of base points. (ii) is clear from volume considerations for C in S.

It can be shown that by making a subdivision S_1 of S whose cones correspond to the facets of the $S(C)$ not through $\underset{\sim}{0}$, and then using barycentric subdivision, one may obtain equality in (2). Another 'facet construction' is considered later in §3D, where we will minimise, rather than maximise the left hand side of (2) (thus obtaining what will be called a minimal special subdivision of S).

Special cones.

We now introduce the notion of special cones in terms of deleted simplexes. Special cones will result from the construction in §3D, and are needed for Chapters 4 and 5.

Definition 3B.7. If $S = \text{conv}(\underset{\sim}{a}_0,\ldots,\underset{\sim}{a}_n)$ is the simplex with the $n+1$ affinely independent points $\underset{\sim}{a}_0,\ldots,\underset{\sim}{a}_n$ as vertices, then the deleted simplex is

$$S^*(\underset{\sim}{a}_0,\ldots,\underset{\sim}{a}_n) = S \sim \{\underset{\sim}{a}_0,\ldots,\underset{\sim}{a}_n\}$$

Definition 3B.8. If $C = c(\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n)$ is an L-cone, then the deleted simplex $S^*(C)$ is

$$S^*(C) = S^*(\underset{\sim}{0},\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n) = \{\Sigma\lambda_i\underset{\sim}{a}_i \in P(C) \sim \{\underset{\sim}{0}\} \mid \Sigma\lambda_i \leqslant 1\}.$$

Definition 3B.9. If $C = c(\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n)$ is a simplicial L-cone, then C is special if

$$L \cap S^*(C) = \emptyset,$$

that is, if

$$\left.\begin{array}{l} \underset{\sim}{a} = \lambda_1\underset{\sim}{a}_1+\ldots+\lambda_n\underset{\sim}{a}_n \in L \cap C, \\ \lambda_1+\ldots+\lambda_n \leqslant 1 \end{array}\right\} \Rightarrow \begin{array}{l} \underset{\sim}{a} = \underset{\sim}{0}, \text{ or} \\ \underset{\sim}{a} = \underset{\sim}{a}_i \text{ for some } i. \end{array}$$

Lemma 3B.5. A 2-dimensional L-cone C is special if and only if it is basic.

Proof. Let $C = c(\underset{\sim}{a}_1, \underset{\sim}{a}_2)$ be a 2-dimensional simplicial cone. By Lemma 3B.3, C is basic if and only if $\dot{L} \cap P(C) = \emptyset$. Hence if C is basic, it is obviously special. Conversely, suppose C is special. To prove that C is basic, suppose to the contrary, that $\dot{L} \cap P(C) \neq \emptyset$, so that some

$$\underset{\sim}{a} = \lambda_1 \underset{\sim}{a}_1 + \lambda_2 \underset{\sim}{a}_2 \in \dot{L} \cap P(C).$$

Then

$$\underset{\sim}{a}' = (1-\lambda_1)\underset{\sim}{a}_1 + (1-\lambda_2)\underset{\sim}{a}_2 \in \dot{L}.$$

If $\lambda_1 + \lambda_2 \leqslant 1$, then $\underset{\sim}{a} \in L \cap S^*(C)$, and if $\lambda_1 + \lambda_2 \geqslant 1$, then $\underset{\sim}{a}' \in L \cap S^*(C)$. This contradicts the assumption that C is special.

Lemma 3B.5 is Lemma 6, of Chapter III of Cassels [3], and is also (essentially) Theorem 34 of Hardy and Wright [6]. The result does not hold for n-dimensional cones with $n > 2$. In Chapter 5 on 3-dimensional L-cones we will obtain and use a characterisation of special cones.

It is clear that if C is a special L-cone, then any face of C (of dimension at least one) is also special.

§3C. Basic subdivisions of a complex of L-cones.

In this section, L will again be a fixed lattice in R^d of dimension $\leqslant d$. We will solve the subdivision problem, and provide the relevant bounds. Then we will give applications to points of a sublattice of Z^n which lie in the non-negative orthant, and (Corollary 3C.3) solve Schinzel's problem, in its original form.

The main theorem of this section, Theorem 3C.2 on subdivision, is based on the following key lemma on barycentric subdivision of a complex of simplicial L-cones.

Lemma 3C.1. Let S be a complex of n-dimensional simplicial L-cones in R^d. Let C be any cone in S of index $m(C) > 1$. Then the following hold.

(i) There is a point $\underset{\sim}{a}$ in $\dot{L} \cap P(C)$.

(ii) If D is any cone in S containing the point $\underset{\sim}{a}$ (of $\dot{L} \cap P(C)$), then $\underset{\sim}{a} \in \dot{L} \cap P(D)$, and the index $m(D) > 1$.

(iii) In barycentric subdivision of S with division point $\underset{\sim}{a}$, the (at most n) cones D_i replacing D all have index less than $m(D)$.

Proof (i) This is immediate from Lemma 3B.3.

(ii) Since S is a complex, $\underset{\sim}{a}$ belongs to the common face $F = C \cap D$. But F is a simplicial L-cone, and its base is a subset of the base of C, and is also a subset of the base of D. Thus $\underset{\sim}{a} \in P(C) \cap F = P(F) = P(D) \cap F \subseteq P(D)$. This shows that $\underset{\sim}{a} \in \dot{L} \cap P(D)$, and so by Lemma 3B.3 again, $m(D) > 1$.

(iii) The lattice point $\underset{\sim}{a}$ is given by

$$\underset{\sim}{0} \neq \underset{\sim}{a} = \lambda_1 \underset{\sim}{a}_1 + \ldots + \lambda_n \underset{\sim}{a}_n,$$

where $0 \leqslant \lambda_i < 1$ for all i, and $\underset{\sim}{a}_1, \ldots, \underset{\sim}{a}_n$ is the base of D. In barycentric subdivision, each cone D is replaced by the cones $D_i = c(\underset{\sim}{a}_i, \ldots, \underset{\sim}{a}_{i-1}, \underset{\sim}{a}, \underset{\sim}{a}_{i+1}, \ldots, \underset{\sim}{a}_n)$, where $\lambda_i > 0$. The cone D_i has index $\lambda_i m(D)$ by definition of the index of a simplicial L-cone (Definition 3B.4.) and by Proposition 3B.1. Thus D is replaced by at most n cones D_i, all of index less than $m(D)$. This completes the proof of Lemma 3C.1.

If S is a collection of L-cones, let

$$N(S) = \#\{C \mid C \in S\},$$

$$E(S) = \#\{\underset{\sim}{b} \mid \underset{\sim}{b} \text{ is a base point of } S\},$$

$$H(S) = \max\{H(\underset{\sim}{b}) \mid \underset{\sim}{b} \text{ is a base point of } S\},$$

where $H(\underset{\sim}{b})$ is the height of $\underset{\sim}{b}$ (see definition in §3A.3), and let $m(C)$ denote the outer index of an L-cone C (Definition 3B.6). Notice that these numbers correspond respectively to (ii), (i), (iii) in §3A.4.

Theorem 3C.2. (Subdivision Theorem.)

Let S be a complex of n-dimensional L-cones in R^d, where $n \geqslant 2$. Let S_1 be (as in Lemma 3B.4) a simplicial subdivision of S such that the base points of S_1 all lie in the union of the $S(C)$ for C in S. Then we obtain a basic subdivision T of S_1 (and so of S) by the following process:

Perform a sequence of barycentric subdivisions, where at each step the division point may be chosen to be any point in $\dot{L} \cap P(C)$, where C is any (simplicial) L-cone of the subdivision reached with index $m(C)$ greater than 1, and stop when all cones of the subdivision reached are basic.

Then the inequalities below apply, where $f(a,b)$ denotes the generalised Fibonacci sequence defined by

$$\begin{cases} f(a,b) = 1 & \text{for } b = -a+1, -a+2, \ldots, 0, \\ f(a,b) = f(a,b-a) + \ldots + f(a,b-1) & \text{for } b \geqslant 1, \end{cases}$$

and $N(S)$, $E(S)$, $H(S)$ are defined as above.

For the simplest case where S consists of just one simplicial L-cone C of index m, we have

(i) $N(T) \leqslant n^{m-1}$

(ii) $E(T) \leqslant n + \frac{n^{m-1}-1}{n-1} = E(C) + \frac{n^{m-1}-1}{n-1}$

(iii) $H(T) \leqslant H(C)\, f(n,m-1)$.

In the case of arbitrary S, we have, on letting

$$N' = \sum_{C \in S} n^{m(C)-1} ,$$

(i)' $N(T) \leqslant N'$

(ii)' $E(T) \leqslant E(S) + \frac{1}{n-1}\left(N'-N(S)\right)$

(iii)' $H(T) \leqslant \max_{C \in S}\left(H(C) f\left(n,m(C)-1\right)\right)$

Proof. (We note that there is no subdivision problem for $n = 1$, as all 1-dimensional L-cones are basic.) The key Lemma 3C.1 shows that the processes described above all terminate with basic L-cones, and leads to the bounds in (i), (ii), (iii). Parts (i)' to (iii)' follow easily. The proof of (i)' uses Lemma 3B.4(ii) and uses the inequality

$$n^{m_1-1} + \ldots + n^{m_s-1} \leqslant n^{m_1+\ldots+m_s-1} \quad (m_i \geqslant 1 \ \forall i, \text{ and } n \geqslant 2).$$

The proof of (ii)' uses the fact that $N(S_1) \geqslant N(S)$.

Corollary 3C.3. (Solution of the special case of Schinzel's problem of §3A.2.)
Let L be an n-dimensional sublattice of Z^n such that the index $(Z^n : L) = r$. Then the non-negative orthant

$$E^+ = \{\underset{\sim}{x}=(x_i) \in R^n \mid x_i \geqslant 0 \ \forall i\}$$

has a subdivision T into basic L-cones such that

$$N(T) \leqslant n^{r^{n-1}-1},$$

$$E(T) \leqslant n + \frac{n^{r^{n-1}-1}}{n-1},$$

$$H(T) \leqslant r f_0(n, r^{n-1}-1),$$

where $f_0(a,b)$ is the generalised Fibonacci sequence $f(a,b)$ modified (diminished) by taking the initial values to be $0,\ldots,0,1$.

Proof. By Cramer's rule, the points $r\underset{\sim}{e}_i$, where $\underset{\sim}{e}_i$ is the ith natural unit vector, belong to L. Thus, E^+ is an L-cone. Its primitive base points are $r_1\underset{\sim}{e}_1,\ldots,r_n\underset{\sim}{e}_n$, where $r_i \geqslant 1$, and $r_i \mid r$ for $i = 1,\ldots,n$. Also,

$$M = \ell(r_1\underset{\sim}{e}_1,\ldots,r_n\underset{\sim}{e}_n) \subseteq L \subseteq Z^n .$$

The index of E^+ in L is defined to be the index $(L:M)$.

But

$$(Z^n:M) = (Z^n:L)\ (L:M), \quad \text{or}$$

$$r_1 \ldots r_n = r(L:M).$$

Hence $(L:M)$, equal to m say, divides r^{n-1}. The result now follows by applying the theorem, with f_0 replacing f in part (iii) to fit in with the special circumstances here, when all but one of the coordinates of each base point of S are zero. (The modified part (iii) is seen in the same way as the original.)

Corollary 3C.4. (Generalisation of Schinzel's problem to non-full lattices in R^d.) Let L be a sub-lattice of Z^d of dimension possibly less than d. Let E^+ be the non-negative orthant in R^d, let $L^+ = L \cap E^+$, and let $\dim L^+ = n$. Then there is a finite subset B of L^+ such that every point $\underset{\sim}{x}$ in L^+ is expressible as a non-negative integral linear combination of an n-element subset of B.

Proof. Let $C = (\text{lin } L^+) \cap E^+$ (which can easily be shown to equal $(\text{lin } L) \cap E^+$). The result will follow once we have shown that

(i) $L^+ = L \cap C$

(ii) C is an n-dimensional L-cone.

Proof of (i) : $L^+ = L^+ \cap \text{lin } L^+ = (L \cap E^+) \cap \text{lin } L^+$

$= L \cap ((\text{lin } L^+) \cap E^+)$

$= L \cap C.$

Proof of (ii) : The fact that the polyhedral cone C is pointed with apex $\underset{\sim}{0}$ follows from the fact that E^+ is pointed with apex $\underset{\sim}{0}$. Since $L^+ \subseteq C \subseteq \text{lin } L^+$, C is n-dimensional. There remains to prove that each edge of

C contains a point of L. Now C is defined by inequalities with rational coefficients. Hence each edge of C contains a point $\underset{\sim}{a}$ of $Z^d \cap \text{lin } L^+$, and to prove that it contains a point $r\underset{\sim}{a}$ of L ($r \in Z$), it is sufficient to prove that both $Z^d \cap \text{lin } L^+$ and the sublattice $L \cap \text{lin } L^+$ are n-dimensional. But $\text{lin } L^+$ has a basis of n points in L^+, and $L^+ \subseteq L$, so $\dim(L \cap \text{lin } L^+) = n$, and $Z^d \cap \text{lin } L^+ \subseteq \text{lin } L^+$, so $\dim(Z^d \cap \text{lin } L^+) \leqslant n$, and the result follows.

Note (added August, 1979).

The above corollary can be used to obtain a refinement of Theorem 2 of the paper

A. Schinzel, Reducibility of lacunary polynomials. I, Acta Arith. 16 (1969), 123-159.

§3D. Minimal special subdivisions, and inner index of a complex of L-cones.

The bounds obtained in Theorem 3C.2 are quite crude ones. We shall find sharp bounds in Chapters 4, 5 for 2, 3-dimensional cones. The method will use the results in the present section, which concern a choice of the S_1 in Theorem 3C.2, and involve the facets of the convex hull of the non-zero lattice points in an L-cone.

It is plausible that to obtain an "economical" subdivision, that is, one with low bounds, one should start off the subdivision of S by selecting new base points from within the sets $S(C)$ (see §3B(1)) for C in S, while minimising the sum of the outer indices (see Definition 3B.6) of the resulting cones. New base points outside any $S(C)$ are then used only if any non-basic cones result. In fact, this idea does work in 2 and 3 dimensions, as we will see in the applications of this section to Chapters 4, 5. The lemma and theorem below are fairly obvious results in 2, 3 dimensions.

Lemma 3D.1. Let C be an n-dimensional L-cone in R^d $(n \geqslant 2)$, let $I(C)$ denote (as usual) the set of points indecomposable in $\dot{L} \cap C$, and let

$$V(C) = \cup\{F \mid F \text{ is a facet of } \operatorname{conv}(\dot{L} \cap C), \text{ and } \underset{\sim}{0} \notin \operatorname{aff} F\}.$$

Then

(i) $\operatorname{conv}(\dot{L} \cap C)$ is an n-dimensional convex polyhedral set and does not contain $\underset{\sim}{0}$.

(ii) Each facet F of $\operatorname{conv}(\dot{L} \cap C)$ such that $0 \notin \operatorname{aff} F$ is an (n-1)-dimensional polytope whose vertices are points of L.

(iii) The set $V(C)$ consists precisely of the points of $\mathrm{conv}(\dot{L} \cap C)$ which are visible from $\underset{\sim}{0}$ (that is, points $\underset{\sim}{x}$ in $\mathrm{conv}(\dot{L} \cap C)$ such that the line segment $\underset{\sim}{0}\underset{\sim}{x}$ contains no point of $\mathrm{conv}(\dot{L} \cap C)$ other than $\underset{\sim}{x}$). Every ray in C from $\underset{\sim}{0}$ passes through exactly one point of $V(C)$.

(iv) Let F be a facet of $\mathrm{conv}(\dot{L} \cap C)$ such that $\underset{\sim}{0} \notin \mathrm{aff}\ F$. Let $\underset{\sim}{n} = \underset{\sim}{n}(F)$ be the unique vector such that

$$\underset{\sim}{n} \cdot \underset{\sim}{x} = 1 \ \forall \underset{\sim}{x} \in F, \text{ and } \underset{\sim}{n} \cdot \underset{\sim}{x} \geqslant 1 \ \forall \underset{\sim}{x} \in \mathrm{conv}(\dot{L} \cap C).$$

Then

$$\{\underset{\sim}{x} \mid \underset{\sim}{x} \in \dot{L} \cap C,\ \underset{\sim}{n} \cdot \underset{\sim}{x} < 2\} \subseteq I(C).$$

(v) $L \cap V(C) \subseteq I(C)$.

Proof. Parts (i) - (iii) are easily seen.

(iv) If $\underset{\sim}{x} \in \dot{L} \cap C$, but $\underset{\sim}{x} \notin I(C)$, then $\underset{\sim}{x} = \underset{\sim}{x}_1 + \underset{\sim}{x}_2$, where $\underset{\sim}{x}_1, \underset{\sim}{x}_2 \in \dot{L} \cap C$, so $\underset{\sim}{n} \cdot \underset{\sim}{x} = \underset{\sim}{n} \cdot \underset{\sim}{x}_1 + \underset{\sim}{n} \cdot \underset{\sim}{x}_2 \geqslant 2$ by (iv).

(v) This follows from (iv), since if $\underset{\sim}{x} \in L \cap V(C)$, then $\underset{\sim}{x} \in L$, and $\underset{\sim}{x} \in F$ for some facet F of $\mathrm{conv}(\dot{L} \cap C)$ with $\underset{\sim}{0} \notin \mathrm{aff}\ F$.

In the following theorem, $m(D)$ will again denote the outer index of an L-cone D (as defined in Definition 3B.6).

Theorem 3D.2. Let S be a complex of n-dimensional L-cones in R^d $(n \geqslant 2)$. For C in S, let $V(C)$ be as defined in Lemma 3D.1. Let

$$S_0 = \{\mathrm{cone}_{\underset{\sim}{Q}}(F) \mid \underset{\sim}{Q} \notin \mathrm{aff}\ F, \text{ and } F \text{ is a facet of } \mathrm{conv}(\dot{L} \cap C) \text{ for some } C \in S\},$$

and let

$$T = \bigcup_{C \in S} \big(L \cap V(C)\big) = L \cap \bigcup_{C \in S} V(C).$$

Then

(i) S_0 is a subdivision of S.

(ii) There is a special subdivision S_1 of S_0 (that is, a subdivision into n-dimensional special L-cones) such that T is the set of base points of S_1. (Special L-cones were defined in Definition 3B.9.)

(iii) For any S_1 as in (ii), and any subdivision V of S,

$$\sum_{D\in S_1} m(D) = \sum_{D\in S_0} m(D) \leq \sum_{D\in V} m(D) \leq \sum_{D\in S} m(D)$$

Proof. (i) It is easy to see that the $\text{conv}(\dot{L}\cap C)$ form a complex, and hence the F do too. Lemma 3D.1(iii) then implies that any ray of C from Q passes through exactly one point of $\bigcup_{C\in S} V(C)$. It follows that S_0 is a complex. The other details are easily checked. (Use, for example, Lemma 3D.1(ii).)

(ii) A subdivision S_1 of the type required is obtained by simply applying repeated barycentric subdivision to S_0 with points of T as division points.

(iii) This follows easily by considering volumes, or indices, for each C in S, using the visibility result in part (iii) of the lemma for the first inequality.

We note that to economically subdivide a given complex, we could try forming S_1 as in the theorem, and then subdividing the resulting special cones into basic cones. This is what we do in Chapters 4 and 5, although in the 2-dimensional case (that is, Chapter 4), the second step is trivial.

Definition 3D.1. A subdivision of the type S_1 in Theorem 3D.2 of a complex S of n-dimensional L-cones will be called a minimal special subdivision of S. The integer

$$\sum_{D\in S_1} m(D) = \sum_{D\in S_0} m(D)$$

will be called the inner index of S.

Notation. We denote the inner index of a complex S of n-dimensional L-cones by

$$m_i(S) .$$

Thus, for simplicial C,

$$m_i(C) = m_i(\{C\}) = m(C),$$

that is, the inner index coincides with the outer index (and the ordinary index). It is also clear that

$$m_i(S) = \sum_{D\in S} m_i(D) .$$

Remark. We have already remarked that the bounds obtained in Theorem 3C.2 are crude ones. However, to illustrate the concept of the inner index, we note that (as is easily seen) the result (i)' therein may be sharpened to

$$N(T) \leqslant n^{m_i(C)-1}$$

in the case when $S = \{C\}$ consists of a single L-cone.

§3E. Results on Indecomposable points.

In this section we gather together for future reference some results on the set $I = I(C)$ of indecomposable points of an L-cone C.

As already mentioned, the points of this set either belong to $P(C)$ or are base points of C, and they are included amongst the base points of any subdivision (or even decomposition) of C into basic L-cones.

The proof of the next result will use a result from Chapter 2 on residues (Lemma 2.2(vi)).

Proposition 3E.1. Let $C = c(\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n)$, $n \geqslant 2$, be a simplicial L-cone in R^d, and let

(1) $$\underset{\sim}{a} = \frac{1}{m}(\alpha_1\underset{\sim}{a}_1+\ldots+\alpha_n\underset{\sim}{a}_n)$$

be in $L \cap C$, where the α_i are non-negative integers. Suppose that

(2) $$\alpha_1 +\ldots+ \alpha_n > (n-1)m - (n-2).$$

Then $\underset{\sim}{a}$ is decomposable in $L \cap C$.

Proof. The result is trivial for $m = 1$. So suppose $m \geqslant 2$. Since the right hand side of (2) is greater than or equal to m, $\underset{\sim}{a} \neq \underset{\sim}{a}_i$ for any i. Hence, if some $\alpha_i \geqslant m$, then $\underset{\sim}{a} = \underset{\sim}{a}_i + (\underset{\sim}{a}-\underset{\sim}{a}_i)$ is a non-trivial decomposition of $\underset{\sim}{a}$, and so $\underset{\sim}{a}$ is decomposable. Also, if some $\alpha_i = 0$, then since the right hand side of (2) is equal to $(n-1)(m-1) + 1$, some $\alpha_j \geqslant m$, so by the previous case, $\underset{\sim}{a}$ is again decomposable. Thus, there remains only the case when $1 \leqslant \alpha_i \leqslant m-1$ for $i = 1,\ldots,n$. By Lemma 2.2(vi), there is an integer r such that $2 \leqslant r \leqslant m-1$, and $\overline{r\alpha_i} < \alpha_i$ for $i = 1,\ldots,n$, where $\bar{x}$ denotes the least non-negative residue of x mod m. Let

$\underset{\sim}{b} = \frac{1}{m}(\overline{r\alpha_1}\underset{\sim}{a}_1 +\ldots+ \overline{r\alpha_n}\underset{\sim}{a}_n)$, and $\underset{\sim}{c} = \underset{\sim}{a} - \underset{\sim}{b}$. Then

$\underset{\sim}{b}, \underset{\sim}{c} \in L \cap C$, and $\underset{\sim}{c} \neq \underset{\sim}{0}$. Now consider the special case when the greatest common divisor of $\alpha_1,\ldots,\alpha_n,m$ is 1. Then $\underset{\sim}{b} \neq \underset{\sim}{0}$, so $\underset{\sim}{a} = \underset{\sim}{b} + \underset{\sim}{c}$ is a non-trivial decomposition, showing that $\underset{\sim}{a}$ is decomposable. In the general case when the greatest common divisor is $d > 1$, (2) gives

$$\frac{\alpha_1}{d}+\ldots+\frac{\alpha_n}{d} > (n-1)\frac{m}{d} - \frac{n-2}{d} \geqslant (n-1)\frac{m}{d} - (n-2) ,$$

and (1) gives

$$\underset{\sim}{a} = \frac{1}{m/d}\left(\frac{\alpha_1}{d}\,\underset{\sim}{a}_1+\ldots+\frac{\alpha_n}{d}\,\underset{\sim}{a}_n\right) ,$$

and indecomposability of $\underset{\sim}{a}$ follows from the result in the special case applied to $\underset{\sim}{a}$ with m/d in place of m. This completes the proof of the proposition.

We know that every point $\underset{\sim}{a}$ of $L \cap C$ is of the form (1), where m is the index of C. Of particular interest in Chapters 4, 5 are the cases $n = 2,3$, which yield the following corollaries of Proposition 3E.1.

Corollary 3E.2. Let $C = c(\underset{\sim}{a}_1,\underset{\sim}{a}_2)$ be a 2-dimensional L-cone, and let $\underset{\sim}{a} = \lambda_1\underset{\sim}{a}_1 + \lambda_2\underset{\sim}{a}_2 \in \dot{L} \cap C$. If $\lambda_1 + \lambda_2 > 1$, then $\underset{\sim}{a}$ is decomposable.

Corollary 3E.3. Let $C = c(\underset{\sim}{a}_1,\underset{\sim}{a}_2,\underset{\sim}{a}_3)$ be a 3-dimensional L-cone, and let $\underset{\sim}{a} = \lambda_1\underset{\sim}{a}_1 + \lambda_2\underset{\sim}{a}_2 + \lambda_3\underset{\sim}{a}_3 \in \dot{L} \cap C$. If $\lambda_1 + \lambda_2 + \lambda_3 \geqslant 2$, then $\underset{\sim}{a}$ is decomposable.

Finally, we give a result on indecomposable points of complexes, which will be used in §5D.

Lemma 3E.4. (i) Let $\underset{\sim}{a} \in \dot{L} \cap F$, where F is a face of the L-cone C. Then $\underset{\sim}{a}$ is indecomposable in C if and only if $\underset{\sim}{a}$ is indecomposable in F.

(ii) Let S be a complex of L-cones in R^d, let $C,D \in S$, and let $\underset{\sim}{a} \in \dot{L} \cap C \cap D$. Then $\underset{\sim}{a}$ is

indecomposable in C if and only if $\underset{\sim}{a}$ is indecomposable in D.

Proof. (i) Obvious.

(ii) This follows from (i).

Thus, we make the following definition.

Definition 3E.1. If S is a complex of L-cones, then an indecomposable point of S is an indecomposable point of a member of S. The notation $I(S)$ will be used for the set of indecomposable points of S.

Note 3E.5. If S is a basic complex, that is, a complex of basic cones, and $B(S)$ is the set of base points of S, then $B(S) = I(S)$. For general S, $B(S) \subseteq I(S)$. If T is a subdivision of S, then $I(S) \subseteq I(T)$.

CHAPTER 4

SUBDIVISION OF 2-DIMENSIONAL LATTICE CONES

Introduction

In this chapter I will consider fairly completely the subdivision problem for a 2-dimensional L-cone C in R^d. A subdivision of C into basic (2-dimensional) L-cones will be called a basic subdivision of C.

After giving some preliminary lemmas in §4A, I will show, in §4B, that C has a unique basic subdivision T_0 whose base points are indecomposable in $\dot{L} \cap C$, and hence I will derive bounds for Schinzel's problem in this case. In §4C I will consider other basic subdivisions of C, and show that any basic subdivision is obtained from the "minimal" basic subdivision T_0 by constructing a sequence of "mediants". In §4D I will relate the notion of subdivision of 2-dimensional L-cones to Farey sequences, convergents and best approximations; and in §4E I will apply the ideas of this chapter to a lemma of Schinzel.

§4A. Preliminary.

In this section I will give some preliminary lemmas, and I will discuss the relation between decompositions, dissections and subdivisions into arbitrary cones, basic cones and basic cones with indecomposable base points. Hence it will appear that there is no real loss of generality in investigating basic subdivisions of a 2-dimensional L-cone, rather than basic decompositions.

Any 2-dimensional L-cone is clearly simplicial, that is of the form

$$C = c(\underset{\sim}{a}_1, \underset{\sim}{a}_2),$$

where $\underset{\sim}{a}_1$, $\underset{\sim}{a}_2$ are the base points of C (see Definition 3B.3).

If C_1, C_2 are two 2-dimensional L-cones lying in the same plane, then clearly a necessary and sufficient condition that C_1, C_2 have no relative interior point in common is that $C_1 \cap C_2$ is a face of C_1, C_2 (which may be an edge, or $\underset{\sim}{0}$). Hence the concepts of dissection and subdivision of a 2-dimensional L-cone (see §3A.3) are identical.

We next consider the relation between dissections and decompositions, assuming that our subcones are basic. For this we need the following lemma.

Lemma 4A.1. Let $C_1 = c(\underset{\sim}{a}_1, \underset{\sim}{a}_2)$, $C_2 = c(\underset{\sim}{b}_1, \underset{\sim}{b}_2)$ be basic 2-dimensional L-cones included in some 2-dimensional cone C. Then either one of C_1, C_2 is contained in the other, or their relative interiors are disjoint.

Proof. Suppose to the contrary that C_1, C_2 overlap (that is, their relative interiors are not disjoint), and that neither of C_1, C_2 is contained in the other. Then the four edges of C_1, C_2 are distinct, and, without loss of generality, we may suppose that they are ordered by angular displacement in the order $\underset{\sim}{0}\underset{\sim}{a}_1$, $\underset{\sim}{0}\underset{\sim}{b}_1$, $\underset{\sim}{0}\underset{\sim}{a}_2$, $\underset{\sim}{0}\underset{\sim}{b}_2$, where the last three rays are in the same one of the two open half-planes of $\operatorname{lin} C$ bounded by the line $\underset{\sim}{0}\underset{\sim}{a}_1$. Let φ be the linear transformation from $\operatorname{lin} C$ to R^2 for

which

$$\varphi(\underset{\sim}{a}_1) = \underset{\sim}{e}_1, \qquad \varphi(\underset{\sim}{a}_2) = \underset{\sim}{e}_2,$$

where $\underset{\sim}{e}_1$, $\underset{\sim}{e}_2$ are the unit vectors, and suppose that

$$\varphi(\underset{\sim}{b}_1) = \underset{\sim}{c}_1, \qquad \varphi(\underset{\sim}{b}_2) = \underset{\sim}{c}_2,$$

so that

(1) The points $\underset{\sim}{c}_1$, $\underset{\sim}{c}_2$ are in the first and second open quadrants respectively.

Since C_1, C_2 are basic, the base points of $\varphi(C_1)$, $\varphi(C_2)$ generate the same lattice. Hence

(2) $$\underset{\sim}{c}_1, \underset{\sim}{c}_2 \in Z^2, \qquad \det(\underset{\sim}{c}_1, \underset{\sim}{c}_2) = 1.$$

Now (1), (2) lead to a contradiction, for they imply that for positive integers e_1, f_1, e_2, f_2,

$$\underset{\sim}{c}_1 = (e_1, f_1), \quad \underset{\sim}{c}_2 = (-e_2, f_2),$$

$$\begin{vmatrix} e_1 & f_1 \\ -e_2 & f_2 \end{vmatrix} = 1.$$

But the value of the determinant is $e_1f_1 + e_2f_2$, which is at least 2. This contradiction completes the proof of the lemma.

Note. The condition that C_1, C_2 are both contained in some cone C is essential. For example, if $\underset{\sim}{a}_1$, $\underset{\sim}{a}_2$ were both in the relative interior of C_2, and $C_3 = c(-\underset{\sim}{a}_1, \underset{\sim}{a}_2)$, then C_3, C_2 would not satisfy the condition, nor the conclusion of the lemma.

Corollary. Let C be a 2-dimensional L-cone, and let

$$C = C_1 \cup \ldots \cup C_r,$$

where the C_i are 2-dimensional basic L-cones, and the C_i are all irredundant, that is,

$$C_i \not\subseteq \bigcup_{j \neq i} C_j \qquad \text{for } i = 1,\ldots,r.$$

Then the decomposition $\{C_1,\ldots,C_r\}$ of C is a subdivision of C.

Proof. If

(3) $\qquad\qquad \text{relint } C_i \cap \text{relint } C_j \neq \emptyset,$

then by the lemma, we may suppose that $C_i \subseteq C_j$, hence C_i is redundant. Thus (3) does not hold for any i,j. So $\{C_1,\ldots,C_r\}$ is a dissection, and hence a subdivision of C, since C is 2-dimensional.

The following lemma, which is stated in a form convenient for use in Chapter 5, shows that the irredundancy condition in the above corollary is automatically satisfied if the base points of the C_i are all indecomposable.

Lemma 4A.2. Let C_1, C_2 be basic 1- or 2-dimensional L-cones, and let C be a 2-dimensional L-cone such that

$$C_1,\ C_2 \subseteq C,$$

$$B(C_1),\ B(C_2) \subseteq I(C)$$

(that is, the base points of C_1, C_2 are indecomposable in C). Then $C_1 \cap C_2$ is a face of C_1, C_2.

Proof. The result is trivial unless at least one of C_1, C_2 is 2-dimensional, so suppose C_2 is 2-dimensional. The result will follow if we can show that no base point of C_1 can lie in relint C_2. Suppose, to the contrary, that

$$\underset{\sim}{a} \in B(C_1) \cap \text{relint } C_2.$$

Then

$$\underset{\sim}{a} \in I(C) \cap C_2 \subseteq I(C_2),$$

so $\underset{\sim}{a} \in I(C_2)$. Since C_2 is basic, it follows that

$\underset{\sim}{a} \in B(C_2)$, contradicting the fact that $\underset{\sim}{a} \in$ relint C_2.

The next lemma is an obvious result on arbitrary subdivisions of a 2-dimensional L-cone.

Lemma 4A.3. Let $C = c(\underset{\sim}{a}_1, \underset{\sim}{a}_2)$ be an L-cone, and let $\underset{\sim}{b}_1, \ldots, \underset{\sim}{b}_r$ be r distinct primitive points in C, with $\underset{\sim}{b}_1 = \underset{\sim}{a}_1$, $\underset{\sim}{b}_r = \underset{\sim}{a}_2$, ordered according to increasing angular displacement from $\underset{\sim}{0a}_1$ to $\underset{\sim}{0a}_2$. Then there is a unique subdivision T of C (containing $r - 1$ cones) for which

$$B(T) = \{\underset{\sim}{b}_1, \ldots, \underset{\sim}{b}_r\},$$

namely

$$T = \{D_1, \ldots, D_{r-1}\},$$

where

$$D_i = c(\underset{\sim}{b}_i, \underset{\sim}{b}_{i+1}) \qquad \text{for } i = 1, \ldots, r-1.$$

We recall that by Lemma 3B.5, a 2-dimensional L-cone is basic if and only if it is special.

Suppose that C is a 2-dimensional L-cone of index $m(C) = m$. Since linear transformations do not affect the subdivision problem, we may suppose for the purpose of presenting examples (even though this does not help in the theory of subdivision), that

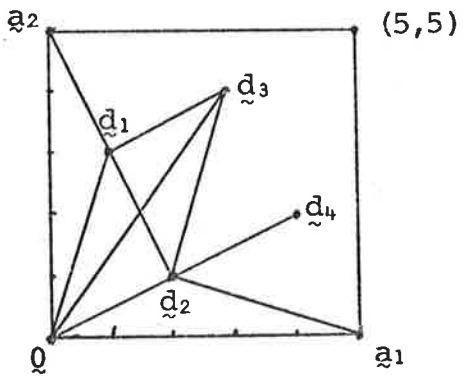
$$C = c(m\underset{\sim}{e}_1, m\underset{\sim}{e}_2), \qquad mZ^2 \subseteq L \subseteq Z^2,$$

$$(L : mZ^2) = (Z^2 : L) = m.$$

We conclude this section by giving a numerical example of basic subdivisions of a 2-dimensional L-cone.

Example. Let $\underset{\sim}{a}_1 = (5,0) = 5\underset{\sim}{e}_1$, $\underset{\sim}{a}_2 = (0,5) = 5\underset{\sim}{e}_2$, $C = c(\underset{\sim}{a}_1, \underset{\sim}{a}_2)$, and let L be the lattice in Z^2 generated by $5Z^2$ and the point $(1,3)$. Then the cone C is the

non-negative orthant E^+ in R^2, and the index of C in L is $m(C) = m = 5$.



The parallelotope $P(\underset{\sim}{a}_1, \underset{\sim}{a}_2)$ contains the following non-zero points of L:

$$\underset{\sim}{d}_1 = (1,3)$$

$$\underset{\sim}{d}_2 = (2,1)$$

$$\underset{\sim}{d}_3 = (3,4)$$

$$\underset{\sim}{d}_4 = (4,2).$$

Of these, $\underset{\sim}{d}_4 = 2\underset{\sim}{d}_2$ is not primitive, and $\underset{\sim}{d}_3 = \underset{\sim}{d}_1 + \underset{\sim}{d}_2$ is decomposable in C. One basic subdivision T_0 of C is the subdivision (see Lemma 4A.3) for which

$$B(T_0) = \{\underset{\sim}{a}_1, \underset{\sim}{d}_2, \underset{\sim}{d}_1, \underset{\sim}{a}_2\} = I(C).$$

Another basic subdivision T_1 of C is given by

$$B(T_1) = \{\underset{\sim}{a}_1, \underset{\sim}{d}_2, \underset{\sim}{d}_3, \underset{\sim}{d}_1, \underset{\sim}{a}_2\},$$

and T_1 is a subdivision of T_0.

(In examples such as this, we can check whether a cone is basic by using determinants and indices, and we can use Cor 3E.2 to prove indecomposability.)

§4B. The minimal basic subdivision of a 2-dimensional L-cone.

Let C be a 2-dimensional L-cone. If T is a basic subdivision of C, then $B(T) = I(T)$, and $I(T) \supseteq I(C)$. In this section I will derive and apply the unique basic subdivision T_0 for which $I(T_0) = I(C)$, or equivalently, for which $B(T_0) = I(C)$. This unique T_0 will be called the minimal basic subdivision of C.

Theorem 4B.1. (Minimal basic subdivision.)

Let $C = c(\underset{\sim}{a}_1, \underset{\sim}{a}_2)$ be an L-cone. Then the minimal special subdivision T_0 of C (see Definition 3D.1) is unique, is basic, and is the unique subdivision satisfying

(1) $$B(T_0) = I(T_0) = I(C) = L \cap V(C),$$

where $V(C)$ is the set of points of $\text{conv}(\dot{L} \cap C)$ which are visible from $\underset{\sim}{0}$.

Proof. By Theorem 3D.2(ii) and Lemmas 3D.1(v) and 3B.5 there is a minimal special, and hence basic, subdivision T_0 of C such that

$$B(T_0) = L \cap V(C) \subseteq I(C).$$

Hence

$$B(T_0) = I(T_0) \supseteq I(C).$$

Thus (1) is proved. The uniqueness of the subdivision satisfying (1) follows from Lemma 4A.3, and the uniqueness of the minimal special subdivision of C follows from this.

The above theorem, which does not depend on Cor 3E.2 leads easily to an alternative proof of that corollary.

A technique for calculating T_0. We now illustrate a technique for determining $I(C)$ and hence T_0. Given C, as in the Example of §4A, we will find a sequence of basic

cones

(2) $\quad c(\underset{\sim}{a}_2,\underset{\sim}{a}),\ c(\underset{\sim}{a},\underset{\sim}{b}),\ c(\underset{\sim}{b},\underset{\sim}{a}_1)$

whose base points belong to $I(C)$. The cones in (2) then form T_0, and the base points $\underset{\sim}{a}_2$, $\underset{\sim}{a}$, $\underset{\sim}{b}$, $\underset{\sim}{a}_1$ comprise $I(C)$.

The point $\underset{\sim}{a}$ is uniquely determined by the conditions that $c(\underset{\sim}{a}_2,\underset{\sim}{a})$ be basic, and $\underset{\sim}{a} \in I(C)$, for these imply that

$$\det(\underset{\sim}{a}_2,\underset{\sim}{a}) = \det\big((0,5),\underset{\sim}{a}\big) = -5,$$

$$\underset{\sim}{a} = (1,3) + k(0,5), \quad k \in Z,$$

where k is least such that

$$(1,3) + k(0,5) \in C = E^+.$$

(The theorem implies that for this k, $\underset{\sim}{a}$ is in fact indecomposable.) Hence $k = 0$, $\underset{\sim}{a} = (1,3)$.

Similarly, $\underset{\sim}{b}$ is uniquely determined by the conditions that $c(\underset{\sim}{a},\underset{\sim}{b})$ be basic, and $\underset{\sim}{b} \in I(C)$. These imply that

$$\underset{\sim}{b} = -\underset{\sim}{a}_2 + h\underset{\sim}{a},$$

where $h \in Z$ is least such that $-\underset{\sim}{a}_2 + h\underset{\sim}{a} \in C$, and this gives $h = 2$, $\underset{\sim}{b} = (2,1)$.

Proceeding in this way we obtain, at the next step, the indecomposable point $\underset{\sim}{a}_1$.

From the technique just illustrated, it is easy to see that the minimal basic subdivision T_0 of C contains all the basic cones whose base points are indecomposable.

A consequence of Theorem 4B.1 is the following theorem on bounds, where $N(S)$, $E(S)$, $H(S)$ are as defined in §3C.

Theorem 4B.2. Let $C = c(\underset{\sim}{a}_1,\underset{\sim}{a}_2)$ be an L-cone of index m. Then there is a basic subdivision T of C

for which

(i) $N(T) \leqslant m$

(ii) $E(T) \leqslant m+1$

(iii) $H(T) \leqslant H(C)$.

Proof. Let T be the minimal basic subdivision T_0 of C, so that $B(T) = I(C)$. But

(3) $$I(C) \subseteq \left(\dot{L} \cap P(C)\right) \cup \{\underset{\sim}{a}_1, \underset{\sim}{a}_2\},$$

and $\dot{L} \cap P(C)$ has $m-1$ elements. Thus

$$E(T) = \# B(T) = \# I(C) = m+1,$$

giving (ii). Hence, by Lemma 4A.3,

$$N(T) = E(T) - 1 = m,$$

that is, (i) holds. From (3) and Cor 3E.2,

$$I(C) \subseteq \operatorname{conv}(\underset{\sim}{0}, \underset{\sim}{a}_1, \underset{\sim}{a}_2),$$

whence (iii) follows.

The above bounds are best possible for cones C such that $\dot{L} \cap P(C)$ is contained in the line segment $\underset{\sim}{a}_1\underset{\sim}{a}_2$.

§4C. Basic subdivisions of a 2-dimensional L-cone.

In this section I will first show the existence of basic subdivisions T of the 2-dimensional L-cone C for certain prescribed $B(T)$. I will then show that any basic subdivision of C can be obtained from the minimal basic subdivision T_0 by repeated barycentric subdivision, where the division points are "mediants".

Theorem 4C.1. Let S be any bounded convex subset of a 2-dimensional L-cone $C = c(\underset{\sim}{a}_1, \underset{\sim}{a}_2)$ such that

$$S \supseteq \operatorname{conv}(\underset{\sim}{0}, \underset{\sim}{a}_1, \underset{\sim}{a}_2).$$

Let A be the set of all primitive points of L which

lie in S. Then there is a basic subdivision T of C such that

$$B(T) = I(T) = A.$$

Proof. Since S is bounded, A is finite. Let T be the unique subdivision of C (see Lemma 4A.3) such that $B(T) = A$. We now show that T is special. Suppose to the contrary that $D \in T$, where the deleted simplex $S^*(D)$ contains a point of L. Then $S^*(D)$ contains a primitive point $\underset{\sim}{a}$ of L. Such a point $\underset{\sim}{a}$ cannot belong to A, since its angular displacement about $\underset{\sim}{0}$ is between those of the base points of D. But S is convex, so $S^*(D) \subseteq S$, and $\underset{\sim}{a} \in A$, a contradiction. Thus T is special, and hence (by Lemma 3B.5) basic, as required.

Any basic subdivision T of a 2-dimensional L-cone C is a subdivision of the minimal basic subdivision T_0, because

$$B(T) \supseteq I(C) = B(T_0).$$

Hence it is trivial that T may be obtained from T_0 by repeated barycentric subdivision. For more detailed information on possible choices of division points, we need the following lemma which we will prove from Lemma 4A.1.

Lemma 4C.2. Let C be a 2-dimensional cone, and let $D = c(\underset{\sim}{a},\underset{\sim}{b})$ and E be basic 2-dimensional L-cones contained in C. If $E \not\supseteq D$, then

$$\underset{\sim}{a} + \underset{\sim}{b} \notin \text{relint } E.$$

Proof. We will prove the contrapositive. Suppose that

(1) $$\underset{\sim}{a} + \underset{\sim}{b} \in \text{relint } E.$$

Consider the basic cones

$$D_1 = c(\underset{\sim}{a}, \underset{\sim}{a}+\underset{\sim}{b}), \quad D_2 = c(\underset{\sim}{a}+\underset{\sim}{b}, \underset{\sim}{b}),$$

which are obtained from D by barycentric subdivision with primitive division point $\underset{\sim}{a}+\underset{\sim}{b}$. It follows from (1), that for $i = 1,2$,

$$\text{relint } E \cap \text{relint } D_i \neq \emptyset, \quad E \not\subseteq D_i.$$

Applying Lemma 4A.1 to the two basic cones E, D_i contained in C, we see that $E \supseteq D_i$, for $i = 1,2$. Hence $E \supseteq D$. This completes the proof of the lemma.

Definition 4C.1. If T is a basic subdivision of the 2-dimensional L-cone C, and $D = c(\underset{\sim}{a}, \underset{\sim}{b}) \in T$, then the sum $\underset{\sim}{a}+\underset{\sim}{b}$ will be called a mediant of T. If U is any basic subdivision of T obtained from T by repeated barycentric subdivision with mediants as division points, we will say that U is obtained from T by constructing mediants.

Notice that a mediant of T cannot be a base point of T. Hence the process of constructing mediants enlarges the set of base points.

Theorem 4C.3. (The mediant construction.) Let T be a basic subdivision of, and T_0 the minimal basic subdivision of the 2-dimensional L-cone C. Then T may be obtained from T_0 by constructing mediants.

Proof. We have that $B(T_0) \subseteq B(T)$, and by Lemma 4A.3, any set of primitive points in C determines a subdivision of C. Thus, it suffices to prove that if S is a basic subdivision of C such that $S \neq T$, and

(2) $$B(S) \subseteq B(T),$$

then some mediant of S lies in T. Since $S \neq T$, there exists D such that

$$D = c(\underset{\sim}{a},\underset{\sim}{b}) \in S, \qquad D \notin T.$$

Now let E be any cone in T. By (2), $E \not\supseteq D$, so by Lemma 4C.2, $\underset{\sim}{a}+\underset{\sim}{b} \notin$ relint E. Thus $\underset{\sim}{a}+\underset{\sim}{b} \in B(T)$, as required.

§4D. Farey sequences, convergents and best approximations.

Farey sequences, convergents and best approximations can be considered in terms of cones.

Let us represent each real number θ by the ray $\underset{\sim}{0}(\theta,1)$ in R^2 emanating from $\underset{\sim}{0}$ and passing through the point $(\theta,1)$, and let us represent the point at infinity by the positive x-axis $\underset{\sim}{0}(1,0)$. Then the set of rays representing $R \cup \{\infty\}$ sweep out the upper half-plane

(1) $$H = \{(x,y) \mid x,y \in R,\ y > 0 \text{ or } (y = 0,\ x \geqslant 0)\}$$

(where the negative x-axis, excluding $\underset{\sim}{0}$, is not included).

If θ is rational, say $\theta = r/s$, where r,s are relatively prime, and $s > 0$, then θ may also be represented by the unique primitive point (r,s) on the ray $\underset{\sim}{0}(\theta,1)$ or $\underset{\sim}{0}(r,s)$, and ∞ may be represented by the point $(1,0)$ on the positive x-axis.

Let

$$D = c(\underset{\sim}{a},\underset{\sim}{b}),$$

where $\underset{\sim}{a} = (a_1,a_2)$, $\underset{\sim}{b} = (b_1,b_2)$ are primitive points of Z^2 lying in H, and $\det(\underset{\sim}{a},\underset{\sim}{b}) < 0$. Let $\underset{\sim}{d} = (d_1,d_2) \in H$. Then

(2) $$\frac{a_1}{a_2} \leqslant \frac{d_1}{d_2} \leqslant \frac{b_1}{b_2} \Leftrightarrow \underset{\sim}{d} \in D.$$

Also,

(3) $\quad \underset{\sim}{d} \in Z^2 \cap D \Leftrightarrow \underset{\sim}{d} = \lambda\underset{\sim}{a} + \mu\underset{\sim}{b},$

where

$$\lambda \geqslant 0, \quad \mu \geqslant 0, \quad \lambda,\mu \in Z.$$

Farey sequences. Let G_n be the sequence

$$\ldots, -\frac{1}{n}, \frac{0}{1}, \frac{1}{n}, \ldots, \frac{1}{1}, \frac{n+1}{n}, \ldots$$

of reduced fractions of denominator not exceeding n, arranged in usual real number order, and let F_n be the subsequence

$$\frac{0}{1}, \frac{1}{n}, \ldots, \frac{1}{1}.$$

Either sequence may be referred to as the Farey sequence of order n. Consider also G'_n which is obtained by taking reduced fractions of height rather than denominator not exceeding n. (The height of a fraction in F_n is equal to its denominator.)

The standard construction of F_n begins with F_1 and proceeds by construction of mediants $(a_1+b_1)/(a_2+b_2)$ of consecutive fractions until there is no pair of consecutive fractions whose mediant has denominator not exceeding n. (It is not actually necessary to pass through F_2, F_3, ... on the way to F_n.)

In terms of cones, the standard construction of F_n, considered as a set of primitive points in H, corresponds to the construction, by constructing mediants (see Definition 4C.1), of the basic subdivision T of the basic cone $c\big((0,1),(1,1)\big)$ for which $B(T) = F_n$. The validity of the construction for F_n is most easily seen from this viewpoint, and is obvious from (2) and (3).

Similarly, G_n may be constructed by subdividing the basic cones $c\big((-1,0),(0,1)\big)$ and $c\big((0,1),(1,0)\big)$, and the same applies for G'_n.

Convergents and best approximations. We will consider the convergents of the real number θ from the cones viewpoint. Let A be the set of all basic Z^2-cones which lie in H and contain $(\theta,1)$ in their relative interiors. There are two cases to consider.

(a) The case when θ is irrational.

We define the convergents of θ to be the base points of cones in A. We will order these convergents and show that, apart from an initial segment, they are the convergents (principal and intermediate) as defined in Lang [7].

Let $C, C' \in A$. From Lemma 4A.1, one of C, C' is contained in the other, so suppose that

$$C' \subseteq C = c(\underset{\sim}{a},\underset{\sim}{b})$$

and let

$$E = c(\underset{\sim}{a},\underset{\sim}{a}+\underset{\sim}{b}), \quad F = c(\underset{\sim}{a}+\underset{\sim}{b},\underset{\sim}{b})$$

be the basic cones obtained from C by the mediant construction. Then from Lemma 4A.1 again it follows that $C' \subseteq E$ or F. Now this E or F belongs to A. Thus under the ordering $\supseteq$, the immediate successor of C is whichever of E, F contains $(\theta,1)$. Consequently, the immediate predecessor of C exists, and is whichever of $c(\underset{\sim}{a},\underset{\sim}{b}-\underset{\sim}{a})$, $c(\underset{\sim}{a}-\underset{\sim}{b},\underset{\sim}{b})$ lies in H. (Notice that exactly one of $\underset{\sim}{a}-\underset{\sim}{b}$, $\underset{\sim}{b}-\underset{\sim}{a}$ lies in H.) Thus the cones in A form a chain under $\supseteq$. From (3), it follows that there are only a finite number of basic cones of A lying between C and C'. Thus the cones of A may be indexed by the

integers, and form a descending sequence

(4) $$C_{-2} \supseteq C_{-1} \supseteq C_0 \supseteq C_1 \supseteq C_2 \supseteq \cdots$$

[From (3), we also see that the height of the base points of C_i tend to infinity as i tends to ∞. Since the C_i are basic, this implies that

$$\bigcap_i C_i = \underset{\sim}{0}(\theta,1).$$

Also, if $t \leq [\theta]$, then $c\big((t,1),(1,0)\big) \in A$. Hence

$$\bigcup_i C_i = H. \quad]$$

We now see that the convergents of θ, defined as the base points of C in A, form a sequence

(5) $$\underset{\sim}{a}_{-\infty} \lessdot \cdots \lessdot \underset{\sim}{a}_{-1} \lessdot \underset{\sim}{a}_0 \lessdot \underset{\sim}{a}_1 \lessdot \underset{\sim}{a}_2 \lessdot \cdots \lessdot \underset{\sim}{a}_i \lessdot \cdots$$

where $\lessdot$ is defined by

(6) $$\underset{\sim}{a} \lessdot \underset{\sim}{b} \Leftrightarrow \underset{\sim}{b} - \underset{\sim}{a} \in H \sim \{\underset{\sim}{0}\},$$

$\underset{\sim}{a}_{-\infty} = (1,0)$, and $\underset{\sim}{a}_i$ is the mediant of C_{i-1}. Let us fix the indices by choosing

$$C_0 = c\big(([\theta],1),(1,0)\big).$$

Writing $\underset{\sim}{a}_0 = ([\theta],1)$, $\underset{\sim}{a}_i = (\alpha_i,\beta_i)$ $\quad (i > 0)$, we see that the α_i/β_i for $i \geqslant 0$ are precisely the convergents to θ, as defined in Lang [7].

Let D_1 denote the cone with apex $\underset{\sim}{0}$ generated by $(\theta,1)$ and $(-1,0)$ and let D_2 denote the cone with apex $\underset{\sim}{0}$ generated by $(\theta,1)$ and $(1,0)$. (These are not Z^2-cones.) Define the linear function L by

(7) $$L(x,y) = y\theta - x.$$

Then for $\underset{\sim}{x}$ not lying on the ray $\underset{\sim}{0}(\theta,1)$,

(8) $$L(\underset{\sim}{x}) \begin{cases} > 0 & \text{if } \underset{\sim}{x} \in D_1 \\ < 0 & \text{if } \underset{\sim}{x} \in D_2 \end{cases}.$$

We define the <u>principal convergents</u> to θ as the convergents $\underset{\sim}{a}_i$ for which $L(\underset{\sim}{a}_i)$, $L(\underset{\sim}{a}_{i+1})$ have opposite signs,

and we define the intermediate convergents to θ as the convergents $\underset{\sim}{a}_i$ for which $L(\underset{\sim}{a}_i)$, $L(\underset{\sim}{a}_{i+1})$ have the same sign. Now if $\underset{\sim}{a}_i$ is any convergent, then from the mediants method of construction of the sequence (5) of convergents, there is a unique j for which

(9) $\qquad c(\underset{\sim}{a}_i,\underset{\sim}{a}_j) \in A, \qquad j < i,$

and for this j, we have

(10) $\qquad L(\underset{\sim}{a}_i)\, L(\underset{\sim}{a}_j) < 0, \qquad \underset{\sim}{a}_i + \underset{\sim}{a}_j = \underset{\sim}{a}_{i+1}$, and

(11) $\qquad \underset{\sim}{a}_k \in \text{relint } c(\underset{\sim}{a}_i,\underset{\sim}{a}_j) \Leftrightarrow k \geqslant i+1.$

Also, with i,j as in (9), and for $h = i,j$, define E_h to be the cone with base $\underset{\sim}{a}_h$ and whichever of $(1,0)$, $(-1,0)$ lies on the same side of $\underset{\sim}{0}(\theta,1)$ as $\underset{\sim}{a}_h$. Then (11) is equivalent to

(12) $\qquad k \leqslant i \Rightarrow \underset{\sim}{a}_k \in E_i \cup E_j,$

and we also have

(13) $\qquad j < k \leqslant i \Rightarrow \underset{\sim}{a}_k \in E_i.$

From (10) the following is immediate.

Proposition 4D.1. Given the convergent $\underset{\sim}{a}_i$, let j be as in (9). Then $\underset{\sim}{a}_i$ is a principal convergent if and only if

$$|L(\underset{\sim}{a}_i)| < |L(\underset{\sim}{a}_j)|.$$

Let θ be an irrational number, and as before let A be the set of all basic Z^2-cones which lie in H and contain $(\theta,1)$. Define $\underset{\sim}{p}$ in $\dot{Z}^2 \cap H$ to be a best approximation to θ if

(14) $\qquad \underset{\sim}{p}' \in \dot{Z}^2 \cap H, \quad \underset{\sim}{p}' \lessdot \underset{\sim}{p} \Rightarrow |L(\underset{\sim}{p})| < |L(\underset{\sim}{p}')|,$

where $\lessdot$ is as defined in (6) and L is as in (7). This agrees with the usual definition except for approximations with denominator 1, and except for the least element $(1,0)$

of $\dot{Z}^2 \cap H$ in the sense of $\preccurlyeq$, which is a best approximation according to our definition.

Proposition 4D.2. The best approximations to θ are just the principal convergents to θ.

Proof. From the mediants construction of the sequence (5) of convergents it is clear that, apart from the ray $\underset{\sim}{Q}(\theta,1)$, the whole of H is covered by (an infinite number of) non-overlapping basic Z^2-cones $c(\underset{\sim}{a},\underset{\sim}{b})$ where $\underset{\sim}{a},\underset{\sim}{b}$ are convergents on the same side of $\underset{\sim}{Q}(\theta,1)$, that is $L(\underset{\sim}{a})L(\underset{\sim}{b}) > 0$. Let $\underset{\sim}{d}$ be a point of $\dot{Z}^2 \cap H$, lying in $c(\underset{\sim}{a},\underset{\sim}{b})$, and such that $\underset{\sim}{d} \neq \underset{\sim}{a},\underset{\sim}{b}$. By (3),

$$\underset{\sim}{d} = \lambda\underset{\sim}{a} + \mu\underset{\sim}{b}$$

where

$$\lambda \geqslant 0, \quad \mu \geqslant 0, \quad \lambda + \mu \geqslant 2.$$

But

$$L(\underset{\sim}{d}) = \lambda L(\underset{\sim}{a}) + \mu L(\underset{\sim}{b}),$$

hence

$$|L(\underset{\sim}{d})| > \min\big(|L(\underset{\sim}{a})|, |L(\underset{\sim}{b})|\big). \tag{15}$$

Also,

$$\underset{\sim}{a} \preccurlyeq \underset{\sim}{d}, \quad \underset{\sim}{b} \preccurlyeq \underset{\sim}{d}. \tag{16}$$

It follows from (15), (16) that $\underset{\sim}{d}$ is not a best approximation. Thus any best approximation is a convergent. It also follows from (15), (16) that for the convergent $\underset{\sim}{p} = \underset{\sim}{a}_i$ to be a best approximation, it is sufficient that (14) be satisfied with the restriction that $\underset{\sim}{p}'$ be a convergent. That is, $\underset{\sim}{a}_i$ is a best approximation if and only if

$$k < i \Rightarrow |L(\underset{\sim}{a}_i)| < |L(\underset{\sim}{a}_k)|. \tag{17}$$

Let $j(< i)$ be as in (9). Then one of $\underset{\sim}{a}_i$, $\underset{\sim}{a}_j$ is in D_1, and the other in D_2. From the mediants construction

of the $\underset{\sim}{a}_k$, and from (8), the $\underset{\sim}{a}_k$ in D_1 have positive $L(\underset{\sim}{a}_k)$ which decrease with k, and the $\underset{\sim}{a}_k$ in D_2 have negative $L(\underset{\sim}{a}_k)$ which increase with k. Hence, in each of D_1 and D_2, and so in each of E_i and E_j (see (12)), $|L(\underset{\sim}{a}_k)|$ decreases with k. Thus

(18) $\quad k < i, \quad L(\underset{\sim}{a}_k)L(\underset{\sim}{a}_i) > 0 \Rightarrow |L(\underset{\sim}{a}_i)| < |L(\underset{\sim}{a}_k)|,$

(19) $\quad k < j, \quad L(\underset{\sim}{a}_k)L(\underset{\sim}{a}_j) > 0 \Rightarrow |L(\underset{\sim}{a}_j)| < |L(\underset{\sim}{a}_k)|.$

From (13) it follows that $k < j$ may be replaced by $k < i$ in (19). From this and (18), we have

(20) $\quad k < i \Rightarrow \min(|L(\underset{\sim}{a}_i)|, |L(\underset{\sim}{a}_j)|) < |L(\underset{\sim}{a}_k)|.$

If $\underset{\sim}{a}_i$ is a principal convergent, then by Proposition 4D.1, $|L(\underset{\sim}{a}_i)| < |L(\underset{\sim}{a}_j)|$, so (20) yields (17), and $\underset{\sim}{a}_i$ is a best approximation. On the other hand, if $\underset{\sim}{a}_i$ is an intermediate convergent, then (17) is false for $k = j$, and so $\underset{\sim}{a}_i$ is not a best approximation.

(b) The case when θ is rational; minimal basic subdivisions.

Let $\theta = r/s$, where (r,s) is a primitive point of Z^2 lying in H. We define the convergents to r/s as for the irrational case, with the addition of the point (r,s) itself. The results are similar to those for the irrational case, but the chain (4) of cones of A terminates with a smallest cone $C_f = c(\underset{\sim}{a},\underset{\sim}{b})$ such that $\underset{\sim}{a} + \underset{\sim}{b} = (r,s)$ and

$$\cap A = \bigcap_i C_i = C_f.$$

We will now state the relation between these convergents and minimal basic subdivisions. (The proof is not hard, and will be omitted.) For simplicity, let us suppose that $r/s > 0$, so that (r,s) lies in the first quadrant

$c\big((0,1),(1,0)\big)$. Let

$$F_1 = c\big((0,1),(r,s)\big), \quad F_2 = c\big((r,s),(1,0)\big).$$

Then for $i = 1,2$ the convergents of r/s which lie in F_i are the indecomposable points of F_i, and so determine the minimal basic subdivision of F_i,

§4E. Note on a lemma of Schinzel.

In this note, we will use Theorem 4B.2(iii) to simplify the proof of, and at the same time improve a constant in Lemma 5 of Schinzel [11], a copy of which is given in the Appendix. We now give the modification to the lemma, and then the modification to the proof.

Modification to the statement of the lemma:

In the bounds for ν_i, μ_i in (40), and for $h(T)$ in (46), make the replacement

(*) "$\big((2d^2)!\big)^2$" by "$(2d^2)!$"

Modification to the proof:

Delete the passage beginning with "Let ξ_1 be" in the second sentence of the second paragraph on page 16 and ending with "$v \geqslant 0$" at the end of the first sentence of the second paragraph on page 17. Apply the replacement (*) to the formula in the second last line of the proof. For the deleted section substitute the following:

"Let ξ be the least positive integer such that $(\xi,0) \in M$, and let η be the least positive integer such that $(0,\eta) \in M$. Then

(47) $$0 < \xi,\eta < (2d^2)!$$

Let N be the sublattice (or submodule) of M with basis $(\xi,0)$, $(0,\eta)$. By Theorem 4B.2(iii) of L. Low's thesis, the

M-cone with base points $(\xi,0)$, $(0,\eta)$, that is, the non-negative orthant in R^2, has a subdivision into basis M-cones whose primitive basis points have height not exceeding $(2d^2)!$ Since the point (n,m) has components which are non-negative integers, it must belong to one of the basic M-cones, say one with base points

$$(\nu_1,\mu_1), \quad (\nu_2,\mu_2).$$

By interchanging the order of these points if necessary, we may arrange that the matrix

$$M = \begin{bmatrix} \nu_1 & \mu_1 \\ \nu_1 & \mu_2 \end{bmatrix}$$

satisfies (41). Also, (40) and (42) hold."

Comparison of the two methods.

We took a sublattice of M with diagonal basis $(\xi,0)$, $(0,\eta)$ and subdivided without increasing the bounds on the height of base points. Schinzel took a triangular basis (ξ_1,η_1), $(0,-\eta_2)$ for M. What he did is equivalent to forming subdivisions of the basic M-cone $c((\xi_1,\eta_1),(0,-\eta_2))$ by the mediants construction to construct the ray $Q(1,0)$. This is equivalent to the continued fraction process for obtaining η_1/η_2. The components of the convergents are also bounded by $(2d^2)!$, and this is where the extra factor of $(2d^2)!$ come in.

APPENDIX TO CHAPTER 4,

TAKEN FROM [11]

Since k_i are integers, $[k_\mu - k_0, k_\nu - k_0] \in \mathfrak{M}$; thus there are integers s, t such that $k_\mu - k_0 = \xi_1 s, k_\nu - k_0 = \eta_1 s + \eta_2 t$.

Putting $\varkappa_i = (\xi_1 D_i' + \eta_1 D_i'')/D$, $\lambda_i = \eta_2 D_i''/D$, we get for $i \leqslant l$

$$k_i - k_0 = \varkappa_i s + \lambda_i t$$

and by (39)

$$|\varkappa_i| \leqslant \frac{\xi_1}{|D|}|D_i'| + \frac{|\eta_1|}{D}|D_i''| \leqslant 2(5c)^{l-1} < (5c)^l,$$

$$|\lambda_i| \leqslant \frac{\eta_2}{|D|}|D_i''| < (5c)^{l-1}.$$

This completes the proof.

Remark. For a given finite linear set, denote by ϱ the number of rationally independent distances and by ϱ_0 the number of rationally independent distances which appear only once. It follows from the lemma that if $\varrho_0 \leqslant 2$, then $\varrho \leqslant 2$. It can easily be found from remark 1 at the end of paper [2] that if $\varrho_0 = 1$ then $\varrho = 1$. The equality $\varrho = \varrho_0$ suggests itself, but I am unable to prove it.

Definition. For a given integral matrix A, $h(A)$ will denote the maximum of absolute values of the elements of A.

Lemma 5. *Let Γ be any given integral matrix 2×2. For arbitrary positive integers d, n, m there exists an integral matrix*

$$M = \begin{bmatrix} \nu_1 & \mu_1 \\ \nu_2 & \mu_2 \end{bmatrix}$$

satisfying the conditions:

$$0 \leqslant \nu_i \leqslant \left((2d^2)!\right)^2, \quad 0 \leqslant \mu_i \leqslant \left((2d^2)!\right)^2 \quad (i = 1, 2), \tag{40}$$

$$|M| > 0, \tag{41}$$

$$[n, m] = [u, v]M, \quad u, v \text{ integers} \geqslant 0, \tag{42}$$

and with the following property. If

$$[n, m]\Gamma = [s, t]\Delta, \tag{43}$$

where s, t are integers, Δ is an integral matrix,

$$|\Delta| \neq 0 \quad \text{and} \quad h(\Delta) \leqslant d, \tag{44}$$

then

$$M\Gamma = T\Delta \quad \text{and} \quad [s, t] = [u, v]T, \tag{45}$$

where T is an integral matrix and

$$h(T) \leqslant 4d\left((2d^2)!\right)^2 h(\Gamma). \tag{46}$$

Proof. Let S be the set of all integral matrices Δ satisfying (43) and (44). Integral vectors $[x, y]$ such that for all $\Delta \in S$ and suitable integers s_Δ, t_Δ, $[x, y]\Gamma = [s_\Delta, t_\Delta]\Delta$ form a module, say $\mathfrak{M}$. By (44) $2d^2 \geqslant |\Delta| \neq 0$, whence $|\Delta|$ divides $(2d^2)!$.

It follows that $[(2d^2)!, 0] \in \mathfrak{M}$ and $[0, (2d^2)!] \in \mathfrak{M}$. Let ξ_1 be the least positive integer such that, for some η_1, $[\xi_1, \eta_1] \in \mathfrak{M}$ and let η_2 be the least positive integer such that $[0, \eta_2] \in \mathfrak{M}$. Clearly $[\xi_1, \eta_1]$ and $[0, \eta_2]$ form a basis for $\mathfrak{M}$ and we may assume without loss of generality that $0 \leqslant \eta_1 < \eta_2$. Hence

$$0 < \xi_1 \leqslant (2d^2)!, \quad 0 \leqslant \eta_1 < \eta_2 \leqslant (2d^2)!. \tag{47}$$

Let

$$\frac{\eta_1}{\eta_2} = \frac{1|}{|b_1} - \frac{1|}{|b_2} - \ldots - \frac{1|}{|b_r}$$

be the expansion of $\frac{\eta_1}{\eta_2}$ into a continued fraction, where b_p are integers > 1 $(1 \leqslant p \leqslant r)$; if $\eta_1 = 0$ let $r = 0$. Put

$$A_{-1} = -1, \quad B_{-1} = 0; \quad A_0 = 0, \quad B_0 = 1;$$
$$A_{p+1} = b_p A_p - A_{p-1}, \quad B_{p+1} = b_p B_p - B_{p-1} \quad (0 \leqslant p < r).$$

It follows that the sequences A_p, B_p are increasing and for $p \leqslant r$

$$A_p B_{p-1} - B_p A_{p-1} = 1, \tag{48}$$

$$0 \leqslant A_p \leqslant \eta_1, \quad 0 < B_p \leqslant \eta_2, \tag{49}$$

$$A_p/B_p < A_r/B_r = \eta_1/\eta_2. \tag{50}$$

Since m, n are > 0, we have

$$\frac{\eta_1}{\eta_2} - \frac{m}{n} \cdot \frac{\xi_1}{\eta_2} < \frac{A_r}{B_r}.$$

Let q be the least non-negative integer which can be substitued for r in the last inequality. Assuming $A_{-1}/B_{-1} = -\infty$ we have therefore

$$\frac{A_{q-1}}{B_{q-1}} \leqslant \frac{\eta_1}{\eta_2} - \frac{m}{n} \cdot \frac{\xi_1}{\eta_2} < \frac{A_q}{B_q}. \tag{51}$$

Let us put

$$M = \begin{bmatrix} \nu_1 & \mu_1 \\ \nu_2 & \mu_2 \end{bmatrix} = \begin{bmatrix} B_q & -A_q \\ B_{q-1} & -A_{q-1} \end{bmatrix} \begin{bmatrix} \xi_1 & \eta_1 \\ 0 & \eta_2 \end{bmatrix} = \begin{bmatrix} B_q \xi_1 & B_q \eta_1 - A_q \eta_2 \\ B_{q-1} \xi_1 & B_{q-1} \eta_1 - A_{q-1} \eta_2 \end{bmatrix}.$$

Inequalities (47), (49) and (50) imply (40). By (48)

$$|M| = \begin{vmatrix} \xi_1 & \eta_1 \\ 0 & \eta_2 \end{vmatrix} = \xi_1\eta_2 > 0.$$

Moreover, the vectors $[\nu_1, \mu_1]$, $[\nu_2, \mu_2]$ form a basis for $\mathfrak{M}$. Since $[n, m] \in \mathfrak{M}$, there are integers u, v satisfying (42). We have

$$[u, v] = [n, m]M^{-1} = \frac{1}{\xi_1\eta_2}[n, m]\begin{bmatrix} B_{q-1}\eta_1 - A_{q-1}\eta_2 & -B\eta_1 + A_q\eta_2 \\ -B_{q-1}\xi_1 & B_q\xi_1 \end{bmatrix}$$

$$= \frac{1}{\xi_1\eta_2}[B_{q-1}(n\eta_1 - m\xi_1) - A_{q-1}\eta_2,\ A_q\eta_2 - B_q(n\eta_1 - m\xi_1)].$$

It follows from (51) that $u \geqslant 0$, $v \geqslant 0$. In order to prove the last statement of the lemma suppose that for some integral matrix Δ (43) and (44) hold. Thus $\Delta \in S$ and since $[\nu_i, \mu_i] \in \mathfrak{M}$ $(i = 1, 2)$ there are integers σ_i, τ_i such that $[\nu_i, \mu_i]\Gamma = [\sigma_i, \tau_i]\Delta$ $(i = 1, 2)$. Putting

$$T = \begin{bmatrix} \sigma_1 & \tau_1 \\ \sigma_2 & \tau_2 \end{bmatrix} \tag{52}$$

we get

$$M\Gamma = T\Delta.$$

On the other hand, (42) and (43) imply

$$[u, v]M\Gamma = [s, t]\Delta. \tag{53}$$

Since $|\Delta| \neq 0$ by (44), we get (45) from (52) and (53). Finally, by (52), (40) and (44)

$$h(T) = h(M\Gamma\Delta^{-1}) \leqslant 4h(M)h(\Gamma)h(\Delta) \leqslant 4d\left((2d^2)!\right)^2 h(\Gamma).$$

This completes the proof.

LEMMA 6. *Let $f(x)$ be an irreducible polynomial not dividing $x^\delta - x$ $(\delta > 1)$, α, β integers, $\alpha > 0$ or $\beta > 0$. For arbitrary positive integers n, m such that $an + \beta m > 0$ there exists an integral matrix*

$$M = \begin{bmatrix} \nu_1 & \mu_1 \\ \nu_2 & \mu_2 \end{bmatrix}$$

satisfying the conditions

$$0 \leqslant \nu_i \leqslant C(f, \alpha, \beta), \quad 0 \leqslant \mu_i \leqslant C(f, \alpha, \beta) \quad (i = 1, 2), \tag{54}$$

$$|M| > 0, \tag{55}$$

$$[n, m] = [u, v]M, \quad u, v \text{ integers} \geqslant 0, \tag{56}$$

$$\alpha\nu_i + \beta\mu_i \geqslant 0 \quad (i = 1, 2), \tag{57}$$

and having the following property:

CHAPTER 5

SUBDIVISION OF 3-DIMENSIONAL LATTICE CONES

§5A. Introduction.

Let L be a lattice in R^d. In this chapter I will consider the subdivision problem for 3-dimensional L-cones. As in Chapter 3, a subdivision into basic L-cones of dimension 3 (see Note 3A.1) will be called a basic subdivision, and the terms special subdivision, simplicial subdivision, special complex, etc., will be used similarly.

The argument will hinge on the basic subdivision of special cones (see Definition 3B.9). I shall characterise 3-dimensional special cones in §5B and apply this characterisation to the basic subdivision of special cones in §5C. Then in §5D I shall obtain the main result of this chapter, namely, that any complex S of 3-dimensional L-cones has a basic subdivision T whose base points are indecomposable in S, so that

$$B(T) = I(T) = I(S).$$

The results of §5C, §5D will lead to improvements for the bounds in Theorem 3C.2, in the 3-dimensional case, and these will be given explicitly for a complex of special cones in Theorem 5C.4. A further application of the characterisation of special lattices will be given in §5E.

We conclude this section by settling the question of equivalence of basic subdivisions and basic dissections (see §3A.3.) for the purposes of this chapter.

Proposition 5A.1. Let D be a 3-dimensional L-cone in R^d, and suppose that D_1, D_2 are 3-dimensional basic L-cones contained in D with base points indecomposable in D, and with disjoint relative interiors. Then the intersection $F = D_1 \cap D_2$ is a face of D_1, and of D_2.

Proof. By Lemma 1A.1, it follows that no point of F can belong to relint D_1 or to relint D_2. If $F = \{\underset{\sim}{0}\}$, or F is an edge common to D_1, D_2, then the result holds. The remaining two possibilities are

(i) F is an edge of one of D_1, D_2 lying in a facet of the other, or

(ii) $F = C_1 \cap C_2$, where C_1, C_2 are facets of D_1, D_2 respectively, $\text{lin } C_1 = \text{lin } C_2$ (and D_1, D_2 lie on opposite sides of F in $\text{lin } D$).

In cases (i), (ii) the result follows easily from Lemma 4A.2 (concerning cones in two dimensions). For example, in case (ii), take $C = D \cap \text{lin } C_1$. Then $C_1, C_2 \subseteq C$. Since D_1, D_2 are basic, so are C_1, C_2. The base points of C_1, C_2, being base points of D_1, D_2 respectively, are indecomposable in D, and therefore in C, so the conditions of the lemma are satisfied.

The following result is immediate.

Corollary 5A.2. Suppose that $\mathcal{D}$ is a dissection of a 3-dimensional L-cone C into basic L-cones whose base points are indecomposable in C. Then $\mathcal{D}$ is a subdivision of C.

Since we will be concerned with basic L-cones D in an L-cone C such that the base points of D are indecomposable in C, there is thus no loss of generality in considering subdivisions of C, rather than dissections.

§5B. Characterisation of 3-dimensional special L-cones.

The main result of this section is the characterisation of 3-dimensional special L-cones (Theorem 5B.2). The proof uses Proposition 2.6 from Chapter 2. We begin with two examples.

Example 1. Let $\underset{\sim}{a}_1 = (1,0,0)$, $\underset{\sim}{a}_2 = (0,1,0)$, $\underset{\sim}{a}_3 = (0,0,1)$ in R^3, and let L be the lattice in R^3 consisting of all points $\frac{1}{2}(\alpha,\beta,\gamma)$, where α, β, γ in Z are all even, or all odd. Then $C = E^+ = c(\underset{\sim}{a}_1, \underset{\sim}{a}_2, \underset{\sim}{a}_3)$ is a special L-cone of index 2. The lattice L has basis

$$\underset{\sim}{a}_1, \underset{\sim}{a}_2, \quad \underset{\sim}{a} = \tfrac{1}{2}(\underset{\sim}{a}_1+\underset{\sim}{a}_2+\underset{\sim}{a}_3) = (\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2}),$$

and $\underset{\sim}{a}$ is the only non-zero point of $L \cap P(C)$.

Example 2. Let $\underset{\sim}{u}$, $\underset{\sim}{v}$, $\underset{\sim}{w}$ be 3 linearly independent points in R^d. Let m, b, c be positive integers such that

(1) $\qquad m > 1, \qquad b + c = m, \qquad (b,m) = (c,m) = 1.$

Let L be the lattice generated by $\underset{\sim}{u}$, $\underset{\sim}{v}$, $\underset{\sim}{w}$ and the point

(2) $$\underset{\sim}{a} = \frac{1}{m}\,(\underset{\sim}{u} + b\underset{\sim}{v} + c\underset{\sim}{w}).$$

That is,

(3) $$L \cap P(C) = \left\{\frac{1}{m}\,(r\underset{\sim}{u} + \overline{rb}\ \underset{\sim}{v} + \overline{rc}\ \underset{\sim}{w}) \,\middle|\, 0 \leq r \leq m-1\right\},$$

where $\bar{x}$ denotes the least non-negative residue of x mod m. Then the cone

$$C = c(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$$

is special of index m.

Proof. The index of C is clearly (by Proposition 3B.2) equal to the number of points in $L \cap P(C)$, that is, m. For any point in $L \cap P(C)$, the sum of the coordinates with respect to basis $\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w}$ is, by (1),

$$(4) \qquad \frac{r}{m} + \frac{\overline{rb}}{m} + \frac{\overline{rc}}{m} = \frac{r}{m} + 1 \qquad (1 \leqslant r \leqslant m-1)$$

$$> 1.$$

Hence C is special of index m.

It will turn out that all special cones of index greater than 1 are of this type. We therefore introduce the following definition.

Definition 5B.1. An L-cone $C = c(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$ is $\underset{\sim}{u}$-special if it is basic or if there exist positive integers m, b, c such that (1), (3) hold. The point (2) will be called the minimal point of C.

Lemma 5B.1. Let $C = c(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$ be an L-cone, and let S be the set of coordinate vectors with respect to $\underset{\sim}{u}$, $\underset{\sim}{v}$, $\underset{\sim}{w}$ of the lattice points in the parallelotope of C, that is,

$$(5) \qquad S = \{(\lambda,\mu,\nu) \mid \lambda\underset{\sim}{u} + \mu\underset{\sim}{v} + \nu\underset{\sim}{w} \in L \cap P(C)\}.$$

Let m be an integer greater than 1. Then C is $\underset{\sim}{u}$-, $\underset{\sim}{v}$- or $\underset{\sim}{w}$-special of index m if and only if for some a, b, c in Z,

$$(6) \qquad \begin{cases} S = \left\{\frac{1}{m}\,(\overline{ra},\overline{rb},\overline{rc}) \mid 0 \leqslant r \leqslant m-1\right\}, \\ 1 \leqslant a,\ b,\ c \leqslant m-1, \\ \overline{ra} + \overline{rb} + \overline{rc} = m + r \quad \text{for} \quad r = 1,\ldots,m-1. \end{cases}$$

Proof. Suppose that C is $\underset{\sim}{u}$-special of index m. Then by Definition 5B.1, (1), (3) hold for suitable b,c. Thus (4), and hence (6) holds with a = 1. Similarly, (6) holds if C is $\underset{\sim}{v}$- or $\underset{\sim}{w}$-special.

Conversely, suppose that (6) holds. Then the index of C, being the number of elements in S, is equal to m.

Also, by Proposition 2.6 with $k = 3$, we have that a, b, c are all relatively prime to m, one is equal to 1, and the other two add up to m. In the case $a = 1$, (1), (3) follow, and C is $\underset{\sim}{u}$-special. In the cases $b = 1$, $c = 1$, C is $\underset{\sim}{v}$-special or $\underset{\sim}{w}$-special respectively. This completes the proof of the lemma.

We are now in a position to characterise 3-dimensional special cones.

Theorem 5B.2. Let $C = c(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$ be an L-cone. Then C is special if and only if C is $\underset{\sim}{u}$-, $\underset{\sim}{v}$- or $\underset{\sim}{w}$-special.

Proof. The proof of Example 2 shows that if C is $\underset{\sim}{u}$-, $\underset{\sim}{v}$- or $\underset{\sim}{w}$-special then it is special. For the converse, we may suppose that C is special and of index $m > 1$. Let S be defined in (5) of Lemma 5B.1. Then S has m members, all of the form $\frac{1}{m}(\alpha,\beta,\gamma)$, where α,β,γ are non-negative integers less than m. We must show that (6) holds. First we give two preliminary results about S.

(i) Suppose $\frac{1}{m}(\alpha,\beta,\gamma)$ is a non-zero element of S. Then

$$(7) \qquad \begin{cases} m+1 \leqslant \alpha+\beta+\gamma \leqslant 2m-1 \\ 1 \leqslant \alpha,\beta,\gamma \leqslant m-1. \end{cases}$$

These inequalities are easily deduced from the fact that C is special by considering points such as $(1,1,1) - \frac{1}{m}(\alpha,\beta,\gamma)$, $(0,1,1) - \frac{1}{m}(\alpha,\beta,\gamma)$ as well as (α,β,γ).

(ii) Suppose that $\underset{\sim}{a}_1, \underset{\sim}{a}_2 \in S$, where $\underset{\sim}{a}_1 \neq \underset{\sim}{a}_2$, and for $i = 1,2$, $\underset{\sim}{a}_i = \frac{1}{m}(\alpha_i,\beta_i,\gamma_i)$. Then $\alpha_1+\beta_1+\gamma_1 \neq \alpha_2+\beta_2+\gamma_2$.

Proof of (ii). Suppose to the contrary that $\alpha_1 + \beta_1 + \gamma_1 = \alpha_2 + \beta_2 + \gamma_2$. Since $\underset{\sim}{a}_1 \neq \underset{\sim}{a}_2$, we may suppose without loss of generality, that $\alpha_1 > \alpha_2$, $\beta_1 \geqslant \beta_2$, $\gamma_1 < \gamma_2$. Let $\alpha_3 = \alpha_1 - \alpha_2$, $\beta_3 = \beta_1 - \beta_2$, $\gamma_3 = \gamma_1 - \gamma_2 + m$, and let $\underset{\sim}{a}_3 = \frac{1}{m}(\alpha_3, \beta_3, \gamma_3)$. Then $\underset{\sim}{a}_3 = \underset{\sim}{a}_1 - \underset{\sim}{a}_2 + (0,0,1)$ is a non-zero point of S, and $\alpha_3 + \beta_3 + \gamma_3 = m$, contradicting the fact that C is special (the first inequality in (7)).

From (i), (ii) we see that for non-zero elements $\frac{1}{m}(\alpha,\beta,\gamma)$ of S, $\alpha + \beta + \gamma$ takes all the values $m+1, \ldots, 2m-1$. Let $\frac{1}{m}(a,b,c)$ be the member of S for which $a + b + c = m + 1$. Then by (7), we obtain (6), as required.

Corollary 5B.3. Let $C = c(\underset{\sim}{u}, \underset{\sim}{v}, \underset{\sim}{w})$ be a special L-cone. Then C is $\underset{\sim}{u}$-special if and only if

(8) $\quad \mu + \nu = 1$ for all $\underset{\sim}{a} = \lambda\underset{\sim}{u} + \mu\underset{\sim}{v} + \nu\underset{\sim}{w}$ in $\dot{L} \cap P(C)$.

Proof. If C is $\underset{\sim}{u}$-special, then it is immediate that (8) holds. Conversely, suppose that (8) holds, and suppose that C is $\underset{\sim}{v}$-special. Then for all $\underset{\sim}{a} = \lambda\underset{\sim}{u} + \mu\underset{\sim}{v} + \nu\underset{\sim}{w}$ in $\dot{L} \cap P(C)$, we have $\lambda + \nu = 1$, and by (8), $\mu + \nu = 1$, so $\lambda = \mu$. Hence C is also $\underset{\sim}{u}$-special. Similarly if C is $\underset{\sim}{w}$-special.

The following corollary is useful in calculations with 3-dimensional L-cones (for example, in checking whether a given complex is special). The denominator q of a non-zero rational number $\lambda = \frac{p}{q}, (p,q) = 1$, will be denoted by $\delta(\lambda)$.

Corollary 5B.4. Let $C = c(\underset{\sim}{u}, \underset{\sim}{v}, \underset{\sim}{w})$ be an L-cone and let (λ,μ,ν) denote the coordinate vector of $\underset{\sim}{a}$ with

respect to basis $\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w}$. If C is $\underset{\sim}{u}$-special, then

(i) $\mu + \nu \in Z$ for all $\underset{\sim}{a} \in \dot{L}$

(ii) $\delta(\lambda) = \delta(\mu) = \delta(\nu)$ for all $a \in \dot{L}$

(iii) The index of C is

$$m(C) = \max_{\underset{\sim}{a} \in \dot{L}} \delta(\lambda)$$

(iv) $\dot{L} \cap P(C) \subseteq$ relint C.

(Hence, by Theorem 5B.2, (ii)-(iv) hold if C is special.) Conversely, if (i), (ii) hold, then C is $\underset{\sim}{u}$-special.

Proof. Let C be $\underset{\sim}{u}$-special. Then (i), (ii), (iii) clearly hold. We now prove that (ii) implies (iv). Let $\underset{\sim}{a} \in \dot{L} \cap P(C)$. Then

(9) $$\underset{\sim}{a} = \lambda\underset{\sim}{u} + \mu\underset{\sim}{v} + \nu\underset{\sim}{w}, \qquad 0 \leqslant \lambda,\mu,\nu < 1.$$

If $\lambda = 0$, then $\delta(\lambda) = 1$. Hence by (ii), $\delta(\mu) = \delta(\nu) = 1$, and so $\mu = \nu = 0$, so $\underset{\sim}{a} = \underset{\sim}{0}$, a contradiction. Thus $\lambda > 0$. Similarly $\mu,\nu > 0$. Thus

(10) $$\lambda,\mu,\nu > 0,$$

and this is equivalent to (iv).

To prove the converse part, let (i), (ii) hold. Let $\underset{\sim}{a} \in \dot{L} \cap P(C)$. Then (9), (10) hold. By (i), (9), $\mu + \nu$ must be 0 or 1. By (10),

$$\mu + \nu = 1, \qquad \lambda + \mu + \nu > 1.$$

Thus C is special, and the condition of Corollary 5B.3 is satisfied. It follows that C is $\underset{\sim}{u}$-special.

Corollary 5B.5. Let $C = c(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$ be a special L-cone of index m. Then

(11) $$I(C) = (\dot{L} \cap P(C)) \cup \{\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w}\}.$$

Thus C has $m+2$ indecomposable points, of which

$\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w}$ are on the relative boundary of C, and the other $m-1$ points, comprising $\dot{L} \cap P(C)$, are in the relative interior of C.

Proof. Clearly that left hand side is contained in the right hand side. For the reverse inclusion we need only show that

$$\dot{L} \cap P(C) \subseteq I(C). \tag{12}$$

We will use Lemma 3D.1 (iv) with $F = \text{conv}(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$. By definition of $\underset{\sim}{n} = \underset{\sim}{n}(F)$, $\underset{\sim}{n} \cdot \underset{\sim}{x} = 1$ for all $\underset{\sim}{x}$ in F. Hence, if

$$\underset{\sim}{a} = \frac{1}{m}(\alpha\underset{\sim}{u}+\beta\underset{\sim}{v}+\gamma\underset{\sim}{w}) \in \dot{L} \cap P(C),$$

then

$$\underset{\sim}{n} \cdot \underset{\sim}{a} = \frac{1}{m}(\alpha+\beta+\gamma)$$

$$\leqslant 2 - \frac{1}{m} < 2 \quad \text{by Theorem 5B.2,}$$

and it follows, from the lemma, that $\underset{\sim}{a} \in I(C)$. This completes the proof of (12), and hence of (11). The last part of the corollary follows from Corollary 5B.4 (iv) (and Theorem 5B.2).

[The converse does not quite hold. If (12) holds, then from Corollaries 3E.2,3 we can show that if $\underset{\sim}{a} \in \dot{L} \cap P(C)$, then

$$\underset{\sim}{a} = \frac{1}{m}(\alpha\underset{\sim}{u}+\beta\underset{\sim}{v}+\gamma\underset{\sim}{w}),$$

$$m \leqslant \alpha+\beta+\gamma \leqslant 2m-1.$$

But we cannot show that C is special, for, as is easily shown, the left hand bound cannot be improved to $m+1$.]

The following corollary of Corollary 5B.5 relating the number of indecomposable points of a special complex S of 3-dimensional L-cones to the inner index $m_i(S)$ of

S (defined in Definition 3D.1) will be a useful adjunct to the theorem (Theorem 5C.4) on bounds obtained later. If S is a complex of 3-dimensional L-cones, then as in §3C, we let

$$N(S) = \#\{C \mid C \in S\},$$

$$E(S) = \#\{\underset{\sim}{b} \mid \underset{\sim}{b} \text{ is a base point of } S\},$$

$$H(S) = \max\{H(\underset{\sim}{b}) \mid \underset{\sim}{b} \text{ is a base point of } S\},$$

and if $\cup S$ is 3-dimensional, we also let

$$E_b(S) = \#\{\underset{\sim}{b} \mid \underset{\sim}{b} \text{ is a base point of } S \text{ on the relative boundary of } \cup S\},$$

and

$$E_i(S) = \#\{\underset{\sim}{b} \mid \underset{\sim}{b} \text{ is a base point of } S \text{ in the relative interior of } \cup S\},$$

so that

$$E(S) = E_i(S) + E_b(S).$$

Corollary 5B.6. Let S be a special complex of 3-dimensional L-cones. Then, in terms of the above notation,

(i) $\# I(S) = m_i(S) - N(S) + E(S)$

(ii) If S is a subdivision of an L-cone, or $\cup S$ lies in R^3 and is simply connected, then

$$\# I(S) = m_i(S) + 2 - E_i(S).$$

Proof. (i) By Corollary 5B.5,

$$\# I(S) = \sum_{C \in S} (m(C)-1) + E(S)$$

$$= m_i(S) - N(S) + E(S)$$

(ii) By Lemma 1C.1.

$$2E_i(S) + E_b(S) = N(S) + 2. \tag{13}$$

Hence (ii) follows from (i).

The following two results are useful in dealing with $\underset{\sim}{u}$-special cones, and are immediate consequences of the definition.

Proposition 5B.7. If $C = c(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$ is a $\underset{\sim}{u}$-special L-cone, then the $m+1$ indecomposable points of $\dot{L} \cap C$ not equal to $\underset{\sim}{u}$ all lie on the plane through $\underset{\sim}{v},\underset{\sim}{w}$ and parallel to $\underset{\sim}{o}\underset{\sim}{u}$.

Proposition 5B.8. Let $C = c(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$ be an L-cone. Then C is $\underset{\sim}{u}$-special if and only if there is an integer c_1 such that

$$(c_1,m) = 1,$$

$$(\dot{L} \cap P(C)) \cup \{\underset{\sim}{v},\underset{\sim}{w}\} = \{\underset{\sim}{b}_r \mid r = 0,1,\ldots,m\},$$

$$\underset{\sim}{b}_r = \frac{1}{m}\left(\overline{rc_1}\ \underset{\sim}{u} + (m-r)\underset{\sim}{v} + r\underset{\sim}{w}\right).$$

§5C. Subdivision of 3-dimensional special L-cones, and complexes.

The main result in this section (Theorem 5C.3) is that every special 3-dimensional complex S of L-cones has a unique basic subdivision T whose base points are indecomposable points of S, so that

$$B(T) = I(T) = I(S)$$

(See Definition 3E.1 and Note 3E.5. The first equality holds because T is basic.)

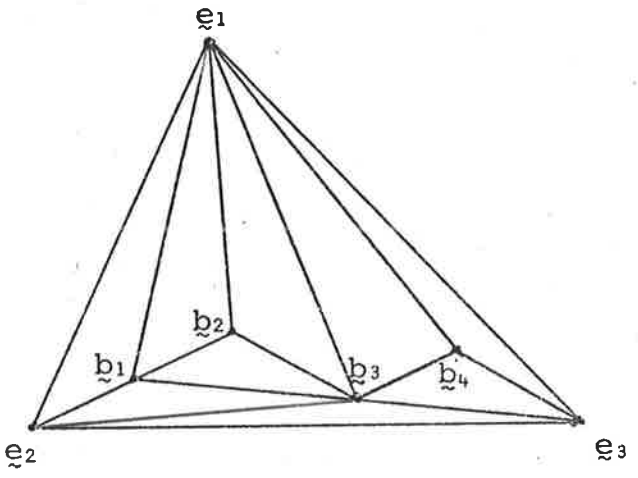
Example 1. Let $C = E^+ = c(\underset{\sim}{e}_1, \underset{\sim}{e}_2, \underset{\sim}{e}_3)$, and let L be given by $(\dot{L} \cap P(C)) \cup \{\underset{\sim}{e}_2, \underset{\sim}{e}_3\} = \{\underset{\sim}{b}_i \mid i = 0, \ldots, 5\}$, where

$$\underset{\sim}{b}_0 = \tfrac{1}{5}(0,5,0) = \underset{\sim}{e}_2$$
$$\underset{\sim}{b}_1 = \tfrac{1}{5}(2,4,1)$$
$$\underset{\sim}{b}_2 = \tfrac{1}{5}(4,3,2)$$
$$\underset{\sim}{b}_3 = \tfrac{1}{5}(1,2,3)$$
$$\underset{\sim}{b}_4 = \tfrac{1}{5}(3,1,4)$$
$$\underset{\sim}{b}_5 = \tfrac{1}{5}(0,0,5) = \underset{\sim}{e}_3$$

(See Proposition 5B.8.)

Then C is $\underset{\sim}{e}_1$-special of index 5, and the diagram below represents a basic subdivision T with indecomposable base points (plane cross-sections being shown).

Subdivision T:



All the 9 (= 2.5 - 1) cones have determinant $1/5 = 1/5|\det(\underset{\sim}{e}_1,\underset{\sim}{e}_2,\underset{\sim}{e}_3)|$, and so are basic. Barycentric subdivision of C with division point $\underset{\sim}{b}_3$ produces three $\underset{\sim}{e}_1$-special cones (this could be verified numerically, but will follow from the theory) and further barycentric subdivisions with division points $\underset{\sim}{b}_2,\underset{\sim}{b}_1,\underset{\sim}{b}_4$ results in the subdivision T as shown.

Lemma 5C.1. Let S be a complex of 3-dimensional special cones, and let T be the complex obtained from S by barycentric subdivision with division point $\underset{\sim}{t}$ in C, where $C \in S$. If $\underset{\sim}{t} = \underset{\sim}{t}_0$, the minimal point of C (as a $\underset{\sim}{u}$-, $\underset{\sim}{v}$-, or $\underset{\sim}{w}$-special cone; see Theorem 5B.2 and Definition 5B.1), then T is special, and

$$I(T) = I(S).$$

Proof. Suppose $\underset{\sim}{t} = \underset{\sim}{t}_0$, and (using Lemma 5B.1 with $r=1$) let

$$\text{(1)} \quad \begin{cases} C = c(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w}), \quad C_{\underset{\sim}{u}} = c(\underset{\sim}{t}_0,\underset{\sim}{v},\underset{\sim}{w}), \\ C_{\underset{\sim}{v}} = c(\underset{\sim}{u},\underset{\sim}{t}_0,\underset{\sim}{w}), \quad C_{\underset{\sim}{w}} = c(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{t}_0), \\ m = m(C) \quad (= \text{index of } C), \\ \underset{\sim}{t}_0 = \frac{1}{m}(a\underset{\sim}{u} + b\underset{\sim}{v} + c\underset{\sim}{w}), \\ \text{so } a+b+c = m+1. \end{cases}$$

Then T is obtained from S by replacing C by the three cones $C_{\underset{\sim}{u}}$, $C_{\underset{\sim}{v}}$, $C_{\underset{\sim}{w}}$, and it is sufficient to show that taking D to be any of these three cones, D is special, and

$$\text{(2)} \qquad I(D) = D \cap I(C).$$

If $S^*(D)$ is the deleted simplex of D, and

$$\underset{\sim}{d} = \lambda\underset{\sim}{u} + \mu\underset{\sim}{v} + \nu\underset{\sim}{w} \in L,$$

then

$$\lambda + \mu + \nu < 1 + \frac{1}{m} \quad \text{for} \quad \underset{\sim}{d} \in L \cap S^*(D), \quad \text{by (1)},$$

$$\geqslant 1 + \frac{2}{m} \quad \text{for} \quad \underset{\sim}{d} \in \dot{L} \cap C,$$

$$\underset{\sim}{d} \neq \underset{\sim}{t}_0, \underset{\sim}{u}, \underset{\sim}{v}, \underset{\sim}{w}.$$

Hence $L \cap S^*(D) = \emptyset$, and D is special.

To prove (2), we note that the right hand side is trivially contained in the left hand side. We will complete the proof of (2) by a counting argument. By Definition 3B.4 and Proposition 3B.1, the indices of the cones $C_{\underset{\sim}{u}}$, $C_{\underset{\sim}{v}}$, $C_{\underset{\sim}{w}}$ are

$$m(C_{\underset{\sim}{u}}) = a, \quad m(C_{\underset{\sim}{v}}) = b, \quad m(C_{\underset{\sim}{w}}) = c.$$

Let

$$a' = \#\big((C_{\underset{\sim}{u}} \cap I(C)) \sim \{\underset{\sim}{t}_0, \underset{\sim}{v}, \underset{\sim}{w}\}\big),$$

and similarly define b', c'. Now

$$a - 1 = \#\big(I(C_{\underset{\sim}{u}}) \sim \{\underset{\sim}{t}_0, \underset{\sim}{v}, \underset{\sim}{w}\}\big)$$

by Corollary 5B.5, and similar results hold for $b-1$, $c-1$. Thus, by the trivial part of (2), we obtain

(3) $$a' \leqslant a - 1, \quad b' \leqslant b - 1, \quad c' \leqslant c - 1.$$

Thus

(4) $$m + 2 = a + b + c + 1 \geqslant a' + b' + c' + 4 \geqslant \#\big(I(C)\big)$$

$$= m + 2 \quad \text{by Corollary 5B.5.}$$

Hence equality holds throughout (4) and (3), and therefore (2) holds for $D = C_{\underset{\sim}{u}}, C_{\underset{\sim}{v}}, C_{\underset{\sim}{w}}$. This completes the proof of Lemma 5C.1.

Note. It is easily shown, using Corollary 5B.3 that if C is $\underset{\sim}{u}$-special, and $\underset{\sim}{t} = \underset{\sim}{t}_0$, then $C_{\underset{\sim}{u}}$ is basic and $C_{\underset{\sim}{v}}$, $C_{\underset{\sim}{w}}$ are $\underset{\sim}{u}$-special.

Note. If $\underset{\sim}{t} \neq \underset{\sim}{t}_0$, $\underset{\sim}{u}$, $\underset{\sim}{v}$, $\underset{\sim}{w}$, then it is easily shown that T is not special, and it can be shown from the uniqueness part of Theorem 5C.3 below, that $I(T) \neq I(S)$.

The following lemma partially states the uniqueness part of Theorem 5C.3 and will be used to prove it. It is well illustrated by Example 1 above.

Lemma 5C.2. Let $C = c(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$ be a $\underset{\sim}{u}$-special L-cone of index m, and let

$$\underset{\sim}{b}_r = \frac{1}{m}\,(\overline{rc}_1\underset{\sim}{u} + (m-r)\underset{\sim}{v} + r\underset{\sim}{w})$$

for $r = 0,1,\ldots,m$, and $(c_1,m) = 1$, as in Proposition 5B.8. Let T be a basic subdivision of C for which $I(T) = I(C)$. Then the cones in T containing the point $\underset{\sim}{u}$ are precisely the $D_r = c(\underset{\sim}{u},\underset{\sim}{b}_{r-1},\underset{\sim}{b}_r)$, $r = 1,\ldots,m$.

Proof. If $0 \leqslant s < r \leqslant m$, then $\det(\underset{\sim}{u},\underset{\sim}{b}_s,\underset{\sim}{b}_r) = (r-s)/m$, as is easily checked, so that $c(\underset{\sim}{u},\underset{\sim}{b}_s,\underset{\sim}{b}_r)$ has index $r-s$. The subdivision T must contain some $c(\underset{\sim}{u},\underset{\sim}{b}_0,\underset{\sim}{b}_r)$, $1 \leqslant r \leqslant m$, but this r must be 1 for the cone to be basic. Hence T must contain some $c(\underset{\sim}{u},\underset{\sim}{b}_1,\underset{\sim}{b}_r)$ such that $\det(\underset{\sim}{u},\underset{\sim}{b}_1,\underset{\sim}{b}_r) = +1$ and this r must be 2 for the cone to be basic. Proceeding this way, the result follows.

Theorem 5C.3. Let S be a special complex of 3-dimensional L-cones in R^d. Then S has a unique basic subdivision T whose base points are the indecomposable points of S. That is,

$$B(T) = I(T) = I(S)$$

(where the first equality holds because T is basic).

The subdivision T may be obtained from S by repeated barycentric subdivision, where each division point is the minimal point (see Definition 5B.1) of the

relevant cone.

Proof. Existence. This follows from Lemma 5C.1, as the process must terminate (with basic T), because for any subdivision V obtained in the process, $B(V) \subseteq I(S)$, and each barycentric subdivision enlarges $B(V)$ by one point.

Uniqueness. It is sufficient to prove the result in the case when S consists of a single $\underset{\sim}{u}$-special cone $C = c(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$ of index m greater than 1. Suppose that T is a basic subdivision of C such that $I(T) = I(C)$. Since C is special, it follows from Corollary 5B.5 that the facet $c(\underset{\sim}{v},\underset{\sim}{w},)$ contains no point of $I(C)$ apart from $\underset{\sim}{v},\underset{\sim}{w}$. Hence for some point $\underset{\sim}{a}$ in $I(C)$, we have $c(\underset{\sim}{v},\underset{\sim}{w},\underset{\sim}{a}) \in T$. By Definition 5B.1, since $c(\underset{\sim}{v},\underset{\sim}{w},\underset{\sim}{a})$ is basic, we must have $\underset{\sim}{a} = \frac{1}{m}(\underset{\sim}{u} + b\underset{\sim}{v} + c\underset{\sim}{w}) = \underset{\sim}{b}_c$ (in the notation of Lemma 5C.2). Thus the cones $c(\underset{\sim}{v},\underset{\sim}{b}_c)$, $c(\underset{\sim}{w},\underset{\sim}{b}_c)$ lie on the relative boundary of a member of T. By Lemma 5C.2, and in the notation of that lemma, $D_c \in T$, so $c(\underset{\sim}{u},\underset{\sim}{b}_c)$ also lies on the relative boundary of a member of T. Thus, none of these three 2-dimensional cones can contain a point of the relative interior of a member of T (by Lemma 1A.1, as T is a dissection of C). It follows that if $D \in T$, then relint D lies in one of $C_{\underset{\sim}{u}}$, $C_{\underset{\sim}{v}}$, $C_{\underset{\sim}{w}}$ (in the notation (1) of Lemma 5C.1), and so D lies in that cone. Thus $T = T_{\underset{\sim}{u}} \cup T_{\underset{\sim}{v}} \cup T_{\underset{\sim}{w}}$, where $T_{\underset{\sim}{u}},T_{\underset{\sim}{v}},T_{\underset{\sim}{w}}$ are basic subdivisions of $C_{\underset{\sim}{u}},C_{\underset{\sim}{v}},C_{\underset{\sim}{w}}$ respectively, and $I(T_{\underset{\sim}{u}}) = I(C_{\underset{\sim}{u}})$, and similarly for $C_{\underset{\sim}{v}},C_{\underset{\sim}{w}}$. Since $C_{\underset{\sim}{u}},C_{\underset{\sim}{v}},C_{\underset{\sim}{w}}$ have smaller indices than C, the proof of the uniqueness result can be carried out by induction. Thus the theorem is proved.

We now give the consequent improvements to Theorem 3C.2 for the particular 3-dimensional case, as in Theorem 5C.3. The notation $N(S)$, $E(S)$, $H(S)$, and so on, will be as in §3C, or as in §5B. The proof of the following theorem will use a combinatorial result (Lemma 1C.1) of Chapter 1.

Theorem 5C.4. Let S be a special complex of 3-dimensional L-cones in R^d, and let T be the basic subdivision of S such that $B(T) = I(T) = I(S)$. Then in the case when $S = \{C\}$, and the index of C is $m(C) = m$, we have

(i) $N(T) = 2m - 1$

(ii) $E(T) = m + 2$

(iii) $H(T) < 2H(C)$, and

$H(T) = H(C)$ if $C = E^+$ (the non-negative orthant).

In the general case, we have

(i)' $N(T) = 2m_i(S) - N(S)$

(ii)' $E(T) = \#I(S) = m_i(S) - N(S) + E(S)$

(iii)' $H(T) < 2H(S)$

Proof. In (ii)' the first equality is by data, and the second is Corollary 5B.6(i) proved earlier from Corollary 5B.5. Part (ii) is the special case, and is proved similarly, or deduced from (ii)'. Part (i) follows from Corollary 5B.5 and Lemma 1C.1. (The lemma gives §5B (13).) Part (i)' is obtained from result (i) by summation. Part (iii) follows from Definition 5B.1 or from §5B (4). Part (iii)' follows from part (iii).

Notice that from Lemma 1C.1, any enlargement of the set $I(T)$, where T is an arbitrary simplicial subdivision of the special complex S, increases $N(T)$. Hence the unique basic subdivision is most economical, not only from the point of view of number of base points, but also from the point of view of number of cones.

§5D Subdivision of 3-dimensional L-cones and complexes.

In this section we obtain results on general 3-dimensional complexes of L-cones from the results in §5C on special 3-dimensional complexes, and from the results in §3D on minimal special subdivisions which were defined in Definition 3D.1, and exist by Theorem 3D.2.

Theorem 5D.1. Let S be a complex of 3-dimensional L-cones, and let S_1 be any minimal special subdivision of S. Then $I(S_1) = I(S)$.

Proof. The proof is similar to that for Corollary 5B.5. If $\underset{\sim}{a} = \lambda\underset{\sim}{u} + \mu\underset{\sim}{v} + \nu\underset{\sim}{w} \in I(C)$ for $C = c(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$ in S_1, then $\lambda + \mu + \nu < 2$, so $\underset{\sim}{a} \in I(S)$ by Lemma 3D.1 (iv). Hence $I(S_1) \subseteq I(S)$, and the reverse inclusion is of course trivial.

Theorem 5D.2. Any complex S of 3-dimensional L-cones has a basic subdivision T such that $I(T) = I(S)$.

Proof. From Theorem 5D.1 and Theorem 3D.2 there is a (minimal) special subdivision S_1 of S such that $I(S_1) = I(S)$. Applying the existence part of Theorem 5C.3 to the special complex S_1 we obtain a basic subdivision T of S_1, and hence of S, such that $I(T) = I(S_1) = I(S)$. This completes the proof of Theorem 5D.2.

(Results on $N(T)$, $E(T)$, $H(T)$ would also follow as in Theorem 5C.4.)

§5E. Applications of §5B.

In this section we will discuss an application of the characterisation of special cones to the characterisation of lattice octahedra in R^3. Another application is given later in the proof of Theorem 6C.3.

Let $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n$ be linearly independent points in R^n, and consider the convex hull

(1) $$K = \text{conv}(\pm\underset{\sim}{a}_1,\ldots,\pm\underset{\sim}{a}_n).$$

If L is a lattice in R^n which contains $\underset{\sim}{a}_1,\ldots,\underset{\sim}{a}_n$ but contains no other points of K, then K is called a lattice octahedron (with respect to lattice L). The problem of characterising lattice octahedra is the problem of determining for given K, those lattices L for which K is a lattice octahedron with respect to L. Mordell [10] discusses this problem, and gives simple proofs of known results for $n = 3,4$.

We will now obtain a simple proof of the result for $n = 3$ from our theory of 3-dimensional special cones of §5B.

Theorem 5E.1. Let K, given by (1), be a lattice octahedron with respect to L. Then L has basis $\underset{\sim}{a}_1, \underset{\sim}{a}_2, \underset{\sim}{a}_3$, or has basis $\underset{\sim}{a}_1, \underset{\sim}{a}_2, \frac{1}{2}(\underset{\sim}{a}_1+\underset{\sim}{a}_2+\underset{\sim}{a}_3)$.

Proof. The eight cones $c(\pm\underset{\sim}{a}_1,\pm\underset{\sim}{a}_2,\pm\underset{\sim}{a}_3)$ are all special L-cones, and each is thus special with respect to a base point. Let m be their common index. If $m = 1$, then $\underset{\sim}{a}_1, \underset{\sim}{a}_2, \underset{\sim}{a}_3$ are a basis for L, so suppose $m > 1$. Without loss of generality, we may suppose that $C_1 = c(\underset{\sim}{a}_1,\underset{\sim}{a}_2,\underset{\sim}{a}_3)$ is $\underset{\sim}{a}_1$-special. Let the minimal point of C_1 be

$$\frac{1}{m}\,(\underset{\sim}{a}_1+h\underset{\sim}{a}_2+k\underset{\sim}{a}_3),$$

$$(h,k) = 1,\quad 1 \leqslant h,\ k \leqslant m-1,\quad h+k = m.$$

Without loss of generality, suppose that $h \leqslant k$, so that $k \geqslant \frac{m}{2}$. The cone $C_2 = c(\underset{\sim}{a}_1,-\underset{\sim}{a}_2,\underset{\sim}{a}_3)$ contains the lattice point $\frac{1}{m}\,(\underset{\sim}{a}_1+k(-\underset{\sim}{a}_2)+k\underset{\sim}{a}_3)$, and is special. If C_2 is $\underset{\sim}{a}_1$-special, then $k+k = m$ (by the easy part of Corollary 5B.3), so $h = k = 1$, $m = 2$, the minimal point of C_1 is $\frac{1}{2}(\underset{\sim}{a}_1+\underset{\sim}{a}_2+\underset{\sim}{a}_3)$, and $\underset{\sim}{a}_1$, $\underset{\sim}{a}_2$, $\frac{1}{2}(\underset{\sim}{a}_1+\underset{\sim}{a}_2+\underset{\sim}{a}_3)$ are a basis for L. If, on the other hand, C_2 is $(-\underset{\sim}{a}_2)$-special, or $\underset{\sim}{a}_3$-special, then $1+k = m$. But then the cone $C_3 = c(\underset{\sim}{a}_1,\underset{\sim}{a}_2,-\underset{\sim}{a}_3)$ contains the lattice point $\frac{1}{m}\,(\underset{\sim}{a}_1+\underset{\sim}{a}_2+(-\underset{\sim}{a}_3))$, so since C_3 is special, $m = 2$, and it follows again that $\underset{\sim}{a}_1$, $\underset{\sim}{a}_2$, $\frac{1}{2}(\underset{\sim}{a}_1+\underset{\sim}{a}_2+\underset{\sim}{a}_3)$ are a basis for L. This completes the proof of the theorem.

CHAPTER 6

QUADRATIC FAREY SEQUENCES

Introduction.

This chapter concerns the quadratic analogue of Farey sequences and the analogous determinant properties. The work in this chapter provides some applications of some of the ideas and results of the earlier chapters, and this work was in fact the source of my initial interest in the investigation of the cone subdivision problem.

In §6A I shall introduce quadratic Farey sequences, and then in §6B I shall develop some basic results and notation and outline the plan of the remainder of the chapter.

§6A. The quadratic Farey sequence K_n.

We recall that in §4D we considered the classical Farey sequence F_n of order n and also the sequence G'_n which consists of all reduced fractions of height not exceeding n. We recall that G'_n may be generated from $-1/0$, $0/1$ and $1/0$ by the construction of mediants. From this construction it is immediate that G'_n has the following property:

Determinant property of G'_n. If r/s, r'/s' are any two consecutive members of G'_n $(r/s < r'/s')$, then

$$rs' - r's = -1.$$

We shall often identify a polynomial

$$f(x) = ax^2 + bx + c$$

with the triple

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

and we shall denote its height by

$$|f| = \max(|a|,|b|,|c|).$$

Recall that the height of an algebraic number is defined as the height of its primitive minimal polynomial.

We now define a quadratic analogue of G'_n.

Definition. The quadratic Farey sequence K_n of order n is the sequence of all real quadratic or rational numbers of height less than or equal to n ordered according to the natural order of the real numbers.

Let

(1) $\theta_i \in K_n \qquad (i = 1,2,3)$

be consecutive in K_n, and let

(2) $f_i = (a_i,b_i,c_i) \qquad (i = 1,2,3)$

be the corresponding primitive minimal polynomials, so that by definition of K_n,

$$|f_i| \leqslant n \qquad (i = 1,2,3).$$

The f_i are of course unique except for sign. Let

(3) $$D = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}.$$

We shall refer to D as the determinant of the f_i or as the determinant of the θ_i. It is natural to ask whether D has some property analogous to the determinant property of G'_n.

Brown and Mahler [2] introduced the above definition of quadratic Farey sequences, obtained some numerical information (a table of information for $n = 5$ is given in [2]), and made some observations and conjectures (without any proofs). In particular, they conjectured that if $\theta_1,\theta_2,\theta_3$ are as in (1), and D is as defined by (2), (3),

then

(4) $\qquad D = 0,\ 1 \text{ or } -1$

provided θ_2 is irrational. We will refer to (4) as the determinant property, so that the conjecture was that the determinant property always holds when θ_2 is irrational.

However, it was pointed out by A. Hesterman (private communication) that the result is false for $n = 7$, and I have verified Hesterman's calculations. The sequence K_7 has 1311 members (of which 655 are positive) and the determinant property for irrational θ_2 holds everywhere except for one set of irrationals $\theta_1,\theta_2,\theta_3$ in the interval (0,1) (between 1/5 and 1/4) and the related sets $-\theta_3,-\theta_2,-\theta_1$; $\theta_3^{-1},\theta_2^{-1},\theta_1^{-1}$; $-\theta_1^{-1},-\theta_2^{-1},-\theta_3^{-1}$ which lie respectively in the intervals $(-1,0)$, $(1,\infty)$, $(-\infty,1)$. The exception involves the 42nd, 43rd and 44th positive members of K_7, for which $D = -2$ (when the a_i in (2) are chosen to be positive) :

θ_i	Approximate value	f_i
θ_1	.2265	(7,-6, 1)
θ_2	.2287	(7, 3,-1)
θ_3	.2319	(7, 7,-2)

Despite this, some results about the determinant D of three consecutive elements of K_n will be obtained in this chapter. These will include, in particular, a way of eliminating the most obvious discrepancy when θ_2 is rational (see §6F). Before going further, we need some preliminary results and notation.

§6B. Preliminary results for K_n : elements in $(r/s, r'/s')$.

In this chapter we will be considering three consecutive members $\theta_1, \theta_2, \theta_3$ of K_n such that θ_2 is irrational, and as far as the determinant property in K_n is concerned, we will be able to suppose that the θ_i all lie in some closed Farey interval $[r/s, r'/s']$, that is an interval $[r/s, r'/s']$, where r/s, r'/s' are consecutive elements of G'_n. For if $\theta_i \geqslant n$ for $i = 1,2,3$, we may consider the θ_i^{-1}, which lie in $[0,1/n]$, and for $\theta_i \leqslant -n$ we may consider $[-1/n,0]$. (We may also suppose without loss of generality, when convenient, that the θ_i are non-negative.)

Lemma 6B.1. Let r/s, r'/s' be consecutive elements of G'_n.

(i) The elements of K_n in the open interval $(r/s, r'/s')$ are all irrational.

(ii) If f is a quadratic polynomial such that $|f| \leqslant n$, and $f(\theta) = 0$, where $\theta \in (r/s, r'/s')$, then $\theta \in K_n$, and f is an integral multiple of the primitive minimal polynomial of θ.

Proof. (i) This is immediate from the definitions. (ii) It is easily checked that any integral linear factor of a reducible integral polynomial of height not exceeding n itself has height not exceeding n. Result (ii) easily follows from this and from (i).

Lemma 6B.2. Let r/s, r'/s' be consecutive elements of G'_n, let $\theta \in K_n \cap (r/s, r'/s')$, let f be the primitive minimal polynomial of θ, and let θ' be the conjugate of θ. Then

(i) $\theta' \notin (r/s, r'/s')$, and

(ii) we may "normalise" f by multiplying by ± 1 so that either

$$(1) \qquad \begin{cases} f(x) < 0 & \text{for } x \text{ in } (r/s, \theta) \\ f(x) > 0 & \text{for } x \text{ in } (\theta, r'/s'), \end{cases}$$

or (1) holds with the inequality symbols interchanged.

Proof. (i) (Notice that Lemma 6B.1(i) implies that f is quadratic.) Suppose to the contrary that θ, θ' lie in the same Farey interval $(r/s, r'/s')$. Then the conjugate elements $-\theta', -\theta$ lie in the same Farey interval $(-r'/s', -r/s)$, and similarly for $\theta'^{-1}, \theta^{-1}$ and for $-\theta^{-1}, -\theta'^{-1}$. Thus we may suppose that

$$0 \leqslant r/s < \theta, \theta' < r'/s' \leqslant 1,$$

$$s \geqslant 1, \ s' \geqslant 1, \ r \geqslant 0, \ r' \geqslant 1,$$

$$s + s' \geqslant n + 1.$$

Then

$$\frac{r'}{s'} - \frac{r}{s} = \frac{1}{ss'} \leqslant \frac{1}{s+s'-1} \leqslant \frac{1}{n},$$

contradicting the fact that

$$|\theta - \theta'| = \frac{\sqrt{b^2-4ac}}{|a|} > \frac{1}{n}.$$

(ii) This is immediate from (i).

Remark. The lemma gives a one-one correspondence between the θ in K_n in a given Farey interval, and their quadratics f. Hence we will sometimes refer to the order of quadratics f, when we mean the order of their zeros in $(r/s, r'/s')$.

If $f(x) = ax^2 + bx + c$ is a quadratic, it will sometimes be convenient to consider the quadratic form

$$(2) \qquad f(x,y) = ax^2 + bxy + cy^2 = y^2 f(x/y).$$

Definition 6B.1. A quadratic f given by (2) has form Δ at r/s (where r,s are relatively prime) if

$$f(r,s) = \Delta.$$

An element θ in K_n has form Δ if its primitive minimal polynomial has form $\pm\Delta$.

The plan of the remaining sections of this chapter is as follows. In §6C I shall derive some general results concerning pairs and triples of consecutive elements of K_n. In §6D I shall specify the elements of K_n in $[0,1/n]$ completely and show that the determinant property holds in this interval. In §6E I shall obtain corresponding results for $\left[\frac{n-1}{n}, 1\right]$. In §6F I shall show how to obtain the smallest element of K_n in $(r/s, r'/s')$ for consecutive elements r/s, r'/s' of G_n', I will show that it has form 1 at r/s, and using this, I will show how to eliminate the discrepancy in the determinant when the middle element is rational. In §6G I will give results on the second smallest element of K_n in $(r/s, r'/s')$, without going into all the details of the proof. In §6H I shall investigate the set S_1 of elements of K_n in $(r/s, r'/s')$ with form 1 at r/s, and show that the determinant property holds on S_1 (that is, for triples of consecutive elements in S_1).

We remark that we may view many of the problems concerning the elements of K_n in terms of lattice points (a point in Z^3 corresponds to a polynomial $f = (a,b,c)$) and lattice cones. Thus we will use a result on 3-dimensional special cones from Chapter 5 in §6C, and we will use element-

ary results on 2-dimensional lattice cones in §6H.

§6C: General results for 2 or 3 consecutive members of K_n.

Two consecutive quadratics.

Theorem 6C.1. Let r/s, r'/s' be consecutive elements of G'_n, and let

$$f_i = (a_i,b_i,c_i) \qquad (i = 1,2)$$

be the primitive minimal polynomials of two consecutive elements θ_1,θ_2 of K_n in $[r/s, r'/s']$. Then

(1) $$\ell(f_1,f_2) = Z^3 \cap \text{lin}\,(f_1,f_2),$$

that is, f_1,f_2 are a basis for the 2-dimensional sublattice of Z^3 in the plane of f_1,f_2.

Proof. In the terminology of §3A.1, we must prove that the Z^3-cone

$$C = c(f_1,f_2)$$

is basic. We may suppose that $\theta_1 < \theta_2$ and that f_1,f_2 are normalised according to §6B(1), so that

$$f_1(\theta_1) = 0, \quad f_1(\theta_2) > 0, \quad |f_1| \leqslant n,$$

$$f_2(\theta_1) < 0, \quad f_2(\theta_2) = 0, \quad |f_2| \leqslant n.$$

Suppose that C is not basic. Then by Lemma 3B.5, C is not special, so there exist λ_1,λ_2 such that

$$f_3 = \lambda_1 f_1 + \lambda_2 f_2 \in Z^3 \cap C$$

$$\lambda_1 + \lambda_2 \leqslant 1, \quad 0 < \lambda_i < 1 \quad \text{for} \quad i = 1,2.$$

It follows that

$$|f_3| \leqslant n, \quad f_3(\theta_1) < 0, \quad f_3(\theta_2) > 0,$$

so

$$f_3(\theta) = 0 \quad \text{for some} \quad \theta \in (\theta_1,\theta_2).$$

By Lemma 6B.1(ii), $\theta \in K_n$, contradicting the consecutiveness of θ_1,θ_2. Thus C is basic.

Corollary 6C.2. Under the conditions of Theorem 6C.1, the 2×2 minors of the matrix

$$\begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{pmatrix}$$

have no common factor.

Proof. By Lemma 2 of Chapter I of Cassels [3], (1) is equivalent to the condition on the minors.

Three consecutive quadratics.

Theorem 6C.3. Let r/s, r'/s' be consecutive elements of G_n', let $\theta_1,\theta_2,\theta_3$ $(\theta_1 < \theta_2 < \theta_3)$ be three consecutive members of K_n in $[r/s, r'/s']$, and let

$$f_i = (a_i,b_i,c_i) \qquad (i = 1,2,3)$$

be their primitive minimal polynomials. Let

$$D = \det(f_1,f_2,f_3) = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} .$$

Then

$$|D| \leqslant n.$$

Proof. We may suppose that the f_i are normalised according to §6B(1) so that

$$\begin{aligned} f_1(\theta_1) &= 0, & f_1(\theta_2) &> 0, & f_1(\theta_3) &> 0 \\ f_2(\theta_1) &< 0, & f_2(\theta_2) &= 0, & f_2(\theta_3) &> 0 \\ f_3(\theta_1) &< 0, & f_3(\theta_2) &< 0, & f_3(\theta_3) &= 0 . \end{aligned}$$

Suppose that $|D| > n$. Then the f_i are linearly independent, and the Z^3-cone

$$C = c(f_1,f_2,f_3)$$

has index $|D| > 1$.

We will show that C is special. Suppose not. Then there are numbers $\lambda_1,\lambda_2,\lambda_3$ in $[0,1)$ no two of which are zero, such that

$$\lambda_1 + \lambda_2 + \lambda_3 \leqslant 1$$

and such that

$$f_4 = \lambda_1 f_1 + \lambda_2 f_2 + \lambda_3 f_3 \in Z^3.$$

It follows that

$$|f_4| \leqslant n, \quad f_4(\theta_1) < 0, \quad f_4(\theta_3) > 0,$$

so

$$f_4(\theta) = 0 \quad \text{for some} \quad \theta \in (\theta_1, \theta_3).$$

By Lemma 6B.1(ii), $\theta \in K_n$, and f_4 is an integral multiple of the primitive minimal polynomial of θ. Since f_4 is not a multiple of f_2, $\theta \neq \theta_2$. This contradicts the consecutiveness of $\theta_1, \theta_2, \theta_3$. Thus we have proved that C is special.

By Theorem 5B.2, C is $\underset{\sim}{f}_1$-, $\underset{\sim}{f}_2$- or $\underset{\sim}{f}_3$-special. Hence, by Lemma 5B.1 (taking $r = 1$ in §5B(6)), there exist positive μ_1, μ_2, μ_3 such that

$$f_5 = \mu_1 f_1 + \mu_2 f_2 + \mu_3 f_3 \in Z^3,$$

$$\mu_1 + \mu_2 + \mu_3 = 1 + \frac{1}{|D|}.$$

Thus

$$|f_5| \leqslant n\left(1 + \frac{1}{|D|}\right) < n + 1,$$

and so $|f_5| \leqslant n$.

As before, this contradicts the consecutiveness of $\theta_1, \theta_2, \theta_3$. Thus we have proved that $|D| \leqslant n$.

In the next two sections we will study two rational Farey intervals within which the determinant D of 3 consecutive elements of K_n always satisfies

$$|D| \leqslant 1.$$

§6D Elements of K_n in [0/1, 1/n].

The elements of K_n in $(0,1/n)$ may be obtained by taking reciprocals of elements of K_n greater than n, so suppose that θ is a zero greater than n of the integral quadratic

$$f(x) = ax^2 + bx + c \qquad (a > 0, |f| \leqslant n).$$

Then

$$\theta = \frac{-b}{2a} \pm \sqrt{\left(\frac{b}{2a}\right)^2 - \frac{c}{a}} \quad .$$

We may then deduce the following results in the order given:

$$\left|\frac{b}{2a}\right| \leqslant \frac{n}{2} \quad ,$$

the + sign must hold,

$$\left|\sqrt{\left(\frac{b}{2a}\right)^2 - \frac{c}{a}}\right| > \frac{n}{2} \quad .$$

$$c < 0,$$

$$a = 1, \quad \left(a \geqslant 2 \Rightarrow \theta \leqslant \frac{n}{4} + \frac{1}{4}\sqrt{n^2+8n}\right\{ \begin{array}{l} < \frac{n}{2} + 1 \leqslant n \text{ if } n \geqslant 2 \\ = n \text{ if } n = 1 \end{array}$$

$$b = -n \quad \left(b \geqslant -n+1 \Rightarrow \theta \leqslant \frac{n-1}{2} + \frac{1}{2}\sqrt{(n-1)^2+4n} = n\right),$$

$$f = (1, -n, c),$$

$$\theta = \frac{n + \sqrt{n^2 + 4|c|}}{2} \quad .$$

Hence the possible quadratics $f(x)$ are, in order of ascending zero greater than n,

$$(1, -n, -1)$$

$$(1, -n, -2)$$

$$\cdots\cdots$$

$$(1, -n, -n) \quad .$$

For each of these quadratics, the positive zero lies between n and $n+1$ exclusive. We may now deduce (using Lemma 6B.1(ii)) that the elements of K_n between 0 and

$1/n$ have primitive minimal polynomials, which when arranged in ascending order, with the reducible polynomials $(1,0,0)$, $(n,-1,0)$ inserted, give the table:

quadratic			determinant
1	0	0	
0	1	0	
n	n	-1	-1
n-1	n	-1	1
.	.	.	0
.	.	.	⋮
2	n	-1	0
1	n	-1	
n	-1	0	-1
0	n	-1	1 .

We also have

$$\begin{vmatrix} 2 & n & -1 \\ 1 & n & -1 \\ 0 & n & -1 \end{vmatrix} = 0 \quad .$$

Thus we have verified that $|D| \leqslant 1$ for any three consecutive members of K_n in $[0/1, 1/n]$, as well as verifying this determinant property in the slightly modified form of the above table. The first and last figures in the last column of the table illustrate a general result proved later, in §6F, that the least and greatest elements of K_n in $(r/s, r'/s')$, f_1, f_2, say, when normalised according to 6B(1), satisfy

$$f_1(r,s) = -1, \quad f_2(r',s') = 1.$$

§6E Elements of K_n in $\left[\frac{n-1}{n}, \frac{1}{1}\right]$.

By Lemma 6B.1(ii) the elements of K_n in $\left(\frac{n-1}{n}, 1\right)$ correspond to the primitive quadratics

(1) $f(x) = ax^2 + bx + c,$

for which

(2) $$|f| \leqslant n,$$

(3) $$\begin{cases} f(n-1,n) = n^2 f\left(\frac{n-1}{n}\right) = a(n-1)^2 + b(n-1)n + cn^2 < 0, \\ \Delta \quad = f(1,1) = f(1) = a+b+c > 0. \end{cases}$$

Taylor's Theorem gives

(4) $$\begin{cases} f(x) = f(1) + (x-1)\left(f'(1) + \frac{x-1}{2} f''(1)\right) \\ \quad = \Delta + (x-1)\left(2a+b+\frac{x-1}{2}2a\right) . \end{cases}$$

Since

$$\frac{-1}{2n} < \frac{x-1}{2} < 0 \quad \text{for} \quad x \in \left(\frac{n-1}{n}, 1\right),$$

it follows from (1), (2) that the graphs of the $f(x)$, for fixed Δ, do not cross in $\left(\frac{n-1}{n}, 1\right)$.

We now show that when (3) is satisfied too, then $\Delta = 1$, with at most 2 exceptions where $\Delta = 2$. The determinant

$$\begin{vmatrix} n & -(n-1) & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{vmatrix} = 1 .$$

Hence an arbitrary integer triple is an integral linear combination of the rows. The quadratics corresponding to these rows have, respectively, $\Delta = 1,0,0$. Thus an arbitrary integer quadratic may be written as

(5) $f = (a,b,c) = \Delta(n,-n+1,0) + \alpha(1,-1,0) + \beta(0,1,-1)$,

where α,β are integral, and $\Delta = a+b+c$. It follows that

(6) $$f(n-1,n) = -\alpha(n-1) - \beta n.$$

From (5), (6) it follows that (2), (3) can only hold if $\Delta \leqslant 2$.

The case $\Delta = 2$ yields at most two primitive quadratics:

$$f_0 = (n-1,\ 3,\ -n) \qquad (\text{when } n \geqslant 3)$$

for which

$$f_0(n-1,\ n) = -1, \quad \tfrac{1}{2}f_0(1) = 1$$

and

$$g = (n,\ 2,\ -n) \qquad (\text{when } n \geqslant 3 \text{ and is odd})$$

for which

$$g(n-1,\ n) = -n, \quad \tfrac{1}{2}g(1) = 1.$$

(When n is even, g is not primitive.)

The case $\Delta = 1$ yields $\frac{n^2+3n-2}{2}$ primitive quadratics, with zeros in $\left(\frac{n-1}{n}\ ,\ 1\right)$. Their order is determined by (4). For $n = 1$, there is only the quadratic $(1,1,-1)$ and $n = 1$ is already covered by §6D. For $n \geqslant 2$, we have:

	a	b	c	α	β	f(n-1,n)
2a + b = n + 1	n-1	-n+3	-1	-1	1	-1
	n-2	-n+5	-2	-2	2	-2
	.	.	.	.	.	.
	1	n-1	-n+1	-n+1	n-1	-n+1
2a + b = n + 2	n	-n+2	-1	0	1	-n
	n-1	-n+4	-2	-1	2	-n-1
	.	.	.	.	.	.
	1	n	-n	-n+1	n	-2n+1
2a + b = n + 3	n	-n+3	-2	0	2	-2n
	n-1	-n+5	-3	-1	3	-2n-1
	.	.	.	.	.	.
	2	n-1	-n	-n+2	n	-3n+2
2a + b = n + 4	n	-n+4	-3	0	3	-3n
	n-1	-n+6	-4	-1	4	-3n-1
	.	.	.	.	.	.
	3	n-2	-n	-n+3	n	-4n+3

	a	b	c	α	β	f(n-1,n)
$2a+b=n+5$	.	.	.	.	.	.
. . . . .	.	.	.	.	.	.
$2a+b=2n$	n	0	-n+1	0	n-1	$-n(n-1)$
	n-1	2	-n	-1	n	$-n(n-1)-1$
$2a+b=2n+1$	n	1	-n	0	n	$-n^2$.

Notice that

$$a + b + c = 1$$

$$\alpha + \beta = 2a + b - (n+1).$$

The general rule for the ordering is as follows:

Let the quadratics be

$$f_i = (a_i, b_i, c_i) \quad \left(i = 1, \ldots, N, \quad N = \frac{n^2+3n-2}{2}\right),$$

and let the corresponding zeros in $\left(\frac{n-1}{n}, 1\right)$ be

$$\theta_1 < \theta_2 < \ldots < \theta_N.$$

Then (from (4))

$$\theta_i < \theta_j \Leftrightarrow 2a_i + b_i < 2a_j + b_j \quad \text{or}$$

$$2a_i + b_i = 2a_j + b_j \quad \text{and} \quad a_i > a_j.$$

Let us insert into this table the quadratics for which $\Delta = 2$. These were obtained above, and are:

$$f_0 = (a_0, b_0, c_0) = (n-1,\ 3,\ -n) \qquad (n \geqslant 3)$$

$$f_0(\theta_0) = 0, \quad \frac{n-1}{n} < \theta_0 < 1, \quad \text{and}$$

$$g = (a', b', c') = (n,\ 2,\ -n) \qquad (n \geqslant 3,\ n \text{ odd})$$

$$g(\phi) = 0, \quad \frac{n-1}{n} < \phi < 1.$$

Consider the graph of $\frac{1}{2}f_0$. Since

$$\tfrac{1}{2}(a_0 + b_0 + c_0) = 1,$$

$$\tfrac{1}{2}(2a_0 + b_0) = n + \tfrac{1}{2}, \quad \text{and for} \quad i \geqslant 1,$$

$$\frac{x-1}{2}(\tfrac{1}{2}a_0 - a_i) > -\tfrac{1}{2} \quad \text{in} \quad \left(\frac{n-1}{n}, 1\right),$$

it follows from (4) that the graph of $\frac{1}{2}f_0$ lies above that of any f_i $(i \geqslant 1)$, so that

$$\theta_0 \leqslant \theta_i \qquad \text{for} \quad i \geqslant 1.$$

As for $\frac{1}{2}g$, we have

$$\tfrac{1}{2}(a'+b'+c') = 1,$$

$$\tfrac{1}{2}(2a'+b') = n+1,$$

so the graph of $\frac{1}{2}g$ lies in $\left(\frac{n-1}{n},\ 1\right)$ between the graphs of two quadratics in the first "block" of $n-1$ quadratics for which $2a+b = n+1$, namely between

$$\left(\frac{n}{2}+\tfrac{1}{2},\ 0,\ \frac{-n}{2}+\tfrac{1}{2}\right) \quad \text{and} \quad \left(\frac{n}{2}-\tfrac{1}{2},\ 2,\ \frac{-n}{2}-\tfrac{1}{2}\right)$$

and ϕ lies between the corresponding zeros. Thus ϕ lies strictly between the first and last quadratics in that block.

We are now able to calculate the determinants of 3 consecutive quadratic irrationals in $\left(\frac{n-1}{n},\ 1\right)$. Let us initially insert the "normalised" quadratics $\frac{1}{2}f_0$, $\frac{1}{2}g$ into the appropriate positions. Then the quadratic polynomials fall into "blocks" according to the value of $2a+b$. When $n=2$, there are 3 blocks, but when $n \geqslant 3$, there are $n+2$ blocks, the extra block being formed by $\frac{1}{2}f_0$, for which

$$2a + b = \tfrac{1}{2}(2a_0+b_0) = n + \tfrac{1}{2}\ .$$

Suppose that

$$h_i = (p_i, q_i, r_i) \qquad (i = 1,2,3)$$

are consecutive, and let

$$D = \begin{vmatrix} p_1 & q_1 & r_1 \\ p_2 & q_2 & r_2 \\ p_3 & q_3 & r_3 \end{vmatrix}.$$

Since $\Delta = p + q + r = 1$ for normalised quadratics,

$$D = \begin{vmatrix} p_1 - p_2 & q_1 - q_2 \\ p_2 - p_3 & q_2 - q_3 \end{vmatrix}.$$

(i) If the h_i are from the same block, then they differ by multiples of $(-1, 2, -1)$, so $D = 0$.

If the h_i are not all from the same block, then two equations must come from the same block, and the other from an adjoining block, since all blocks have at least two equations, except for the last block $(2a + b = 2n + 1)$ and the first block $(2a + b = n + \frac{1}{2}$ if $n \geqslant 3$, and $2a + b = n + 1$ if $n = 2)$ which have one equation each.

Suppose that h_1, h_2 are from one block, and h_3 from the next. Then

(ii) if $h_1 \neq \frac{1}{2}g$, then

$$D = \begin{vmatrix} 1 & -2 \\ k & -2k-1 \end{vmatrix} = -1 \qquad (k \in Z),$$

(iii) if $h = \frac{1}{2}g$, then

$$D = \begin{vmatrix} \frac{1}{2} & -1 \\ k & -2k-1 \end{vmatrix} = -\frac{1}{2} \qquad (k \in Z).$$

Now suppose that h_1 is from one block, and h_2, h_3 are from the next. Then

(iv) if $h_1 \neq \frac{1}{2}f_0$, then

$$D = \begin{vmatrix} k & -2k-1 \\ 1 & -2 \end{vmatrix} = 1 \qquad (k \in Z),$$

(v) if $h_1 = \frac{1}{2}f_0$, and $h_3 \neq \frac{1}{2}g$, then

$$D = \begin{vmatrix} k & -2k-\frac{1}{2} \\ 1 & -2 \end{vmatrix} = \frac{1}{2} \qquad (k \in Z),$$

(vi) if $h_1 = \frac{1}{2}f_0$, and $h_3 = \frac{1}{2}g$ (this can only happen if $n = 3$), then

$$D = \begin{vmatrix} k & -2k-\frac{1}{2} \\ \frac{1}{2} & -1 \end{vmatrix} = \frac{1}{4} \qquad (\text{in fact } k = 1).$$

Thus, in all cases, on reverting where necessary from the normalised forms $\frac{1}{2}f_0$, $\frac{1}{2}g$ to the primitive forms f_0, g, the determinant of 3 consecutive quadratic irrationals in $\left(\frac{n-1}{n}, 1\right)$ is $0, \pm 1$.

It may be checked that if we add one or both of $(n, -n+1, 0)$, $(0, n, -n+1)$ to the beginning of the table, and if we add one or both of $(1, -1, 0)$, $(0, 1, -1)$ to the end of the table, then the additional determinants obtained are all equal to $0, \pm 1$. It follows in particular that the determinant property holds for K_n in the closed interval $\left[\frac{n-1}{n}, \frac{1}{1}\right]$.

§6F. The smallest element of K_n in $(r/s, r'/s')$.

Throughout this section we let r/s, r'/s' be consecutive elements of G_n' $(r/s < r'/s')$.

We will show that the least element of K_n in $(r/s, r'/s')$ has form 1 at r/s. (See Definition 6B.1.) Notice that results about the least element of K_n in $(r/s, r'/s')$ may be converted into results about the greatest element of K_n in the Farey interval $(s'/r', s/r)$ by considering the mapping $\theta \to \theta^{-1}$ $(\theta \in K_n)$. For example, the result stated above implies that the greatest element of K_n in a Farey interval $(r/s, r'/s')$ has form 1 at r'/s'.

Also, results for $r/s \leqslant 0$ may be deduced from results for $r/s \geqslant 0$ by considering the mapping $\theta \to -\theta$ $(\theta \in K_n)$. Thus we will often be able to assume that $r/s \geqslant 0$ without loss of generality. Notice that the case $r/s = 0$ has been considered fully in §6D.

Incidentally, it is an easy matter to show that there do exist elements of K_n in $(r/s, r'/s')$. For example, there exists such an element corresponding to one of the quadratics

$$(s, -r, 0) \pm (0, s', -r').$$

This is easily checked using the result

(1) $$r's - s'r = 1.$$

We prove the following stronger assertion.

Lemma 6F.1. Let r/s, r'/s' be consecutive elements of G_n'. Then there is an element of K_n in $(r/s, r'/s')$ having form 1 at r/s.

Proof. When $r/s = 0$, the quadratic $(-n,-n, 1)$ determines an element of form 1 at 0 in $(0, 1/n)$. Thus we may suppose that $r/s \neq 0$, and without loss of generality, $r/s > 0$. Thus $r \geqslant 1$, $s \geqslant 1$.

Since g.c.d. $(r,s) = 1$, the equation

(2) $$f(r,s) = ar^2 + brs + cs^2 = 1$$

has an integral solution

(3) $$f = f_0 = (a_0,b_0,c_0), \quad f_0(r,s) = 1.$$

(In fact, $f = (s'^2, -2r's', r'^2)$ is a solution by (1).) As (3) can be written

(4) $$\begin{vmatrix} a_0 & b_0 & c_0 \\ s & -r & 0 \\ 0 & s & -r \end{vmatrix} = 1,$$

it follows that the rows of the matrix are a basis for Z^3, and hence the general solution of (2) is

(5) $$f = f_0 + \alpha(s, -r, 0) + \beta(0, s, -r) \qquad (\alpha,\beta \in Z).$$

Let $f = f_1$ be given by (3), (5) and

(6) $$f = f_1 = (a_1,b_1,c_1) \Leftrightarrow \alpha = \left[\frac{-a_0}{s}\right], \quad \beta = \left[\frac{c_0}{r}\right].$$

Then, denoting the fractional part of x by $\{x\}$, it follows from (5), (6) that

$$(7) \qquad a_1 = -s\left\{\frac{-a_0}{s}\right\}, \quad c_1 = r\left\{\frac{c_0}{r}\right\}.$$

By using (5), (6) and (3) it is also easily checked that

$$(8) \qquad b_1 = r\left\{\frac{-a_0}{s}\right\} - s\left\{\frac{c_0}{r}\right\} + \frac{1}{rs}.$$

From (7), (8), $|f_1|$ is at most $\max(r,s)$. However, a furhter adjustment may be necessary to the solution (6) of (2), since $f_1(r',s')$ might not be negative, and by Lemma 6B.1(ii') this is a necessary and sufficient condition for f_1 to give an element of K_n.

From (5), it follows that

$$(9) \qquad f(r',s') = f_0(r',s') + \alpha r' + \beta s'.$$

Hence from (6), it may be verified, using (1), (3), that

$$(10) \qquad f_1(r',s') = -r'\left\{\frac{-a_0}{s}\right\} - s'\left\{\frac{c_0}{r}\right\} + \frac{r's'}{rs}.$$

Let r'', s'' be the integers determined uniquely by

$$(11) \qquad r''s - rs'' = 1, \quad 0 < r'' \leqslant r, \quad 0 \leqslant s'' < s.$$

(The first and third conditions imply the second, as $r > 0$.) Then

$$\begin{bmatrix} r' \\ s' \end{bmatrix} = \begin{bmatrix} r'' \\ s'' \end{bmatrix} + k\begin{bmatrix} r \\ s \end{bmatrix}, \qquad k \in \mathbb{Z}, \quad k \geqslant 0$$

(and $k > 0$ if $s'' = 0$). We have

$$(12) \qquad \left[\frac{s'}{s}\right] = k; \quad \left[\frac{r'}{r}\right] = k \text{ if } r > 1, \text{ and}$$

$$(13) \qquad \frac{r'}{r} = k + 1 \text{ if } r = 1.$$

Since r/s, r'/s' are consecutive in G_n', k may also be described as the largest non-negative integer such that (for given r, s, r'', s'', n)

$$(14) \qquad r' = kr + r'' \leqslant n, \quad s' = ks + s'' \leqslant n$$

both hold.

In the remainder of the proof, there are three cases to consider.

Case $r \geqslant 2$, $s \geqslant 2$. In this case, (10) gives

$$f_1(r',s') \leqslant -\frac{r'}{s} - \frac{s'}{r} + \frac{r's'}{rs} .$$

Since, by (1),

$$\frac{r'}{r} = \frac{s'}{s} + \frac{1}{rs} > \frac{s'}{s} ,$$

it follows that

$$f_1(r',s') < \frac{-s'}{s}\left(\frac{r}{s} + \frac{s}{r} - \frac{r'}{r}\right) , \quad \text{so}$$

$$(15) \qquad f_1(r',s') < \frac{s'}{s}\left(\frac{r'}{r} - 2\right).$$

Now define $f = f_2$ by (5), substituting f_1, α_1, β_1 for f_0, α, β respectively. Then (9) gives

$$(16) \qquad f_2(r',s') = f_1(r',s') + \alpha_1 r' + \beta_1 s' .$$

Choose

$$(\alpha_1,\beta_1) = \begin{cases} (0,0) & \text{if } k = 0,1 \\ (-(k-1),-(k-1)) & \text{if } k \geqslant 2. \end{cases}$$

Then from (12), (15), (16), $f_2(r',s') < 0$, and from (7), (8), (14), $|f_2| \leqslant n$. Thus f_2 is the primitive (all polynomials of form 1 are primitive) minimal polynomial of an element of K_n in $(r/s, r'/s')$ having form 1 at r/s.

Case $r = 1$. Here, $r'' = 1$, $s'' = s - 1$. Also, $a_0 \equiv 1(s)$. Thus

$$(17) \qquad f_1 = (-s+1, 1, 0) = (-s'', r'', 0).$$

From (10), or directly from (17),

$$f_1(r',s') = kr' .$$

Define f_2 as before, choosing

$$(\alpha_1,\beta_1) = \begin{cases} (0, -1) & \text{if } k = 0 \\ (-k, -k) & \text{if } k \geqslant 1. \end{cases}$$

Then f_2 gives an element of K_n as required.

Case $s = 1$. Here, we proceed as in the previous case to get

$$f_1 = (0, 0, 1) = (-s'', s'', r''),$$

$$f_1(r', s') = s'^2 = ks',$$

choosing

$$(\alpha_1, \beta_1) = \begin{cases} (-1, 0) & \text{if } k = 0 \\ (-k, -k) & \text{if } k \geqslant 1. \end{cases}$$

Then f_2 gives the required element of K_n. This completes the proof of Lemma 6F.1.

The following corollary is immediate.

Corollary 6F.2. Let r/s, r'/s' be consecutive elements of G_n'. Then there is an element of K_n in $(r/s, r'/s')$ having form 1 at r'/s'.

Minimal quadratics. In order to determine the least element of K_n greater than r/s (for fixed n) we need to introduce the concept of a minimal quadratic at r/s.

Let r/s be a non-negative element of G_n' $(r \geqslant 0,\ s > 0)$, and let Δ be a positive integer. By (4) above, there is a solution $f = f_1$ in Z^3 of

(18) $$f = (a,b,c), \quad f(r,s) = \Delta,$$

and f satisfies (18) if and only if

(19) $$f - f_1 = \alpha(s,-r,0) + \beta(0,s,-r),$$

where $\alpha, \beta \in Z$. (Of course (19) is valid for arbitrary r/s in G_n'.) From this, and the fact that $\Delta > 0$, it is easy to check that (18) has a solution f such that

(20) $$a \geqslant -n, \quad |b| \leqslant n, \quad |c| \leqslant n,$$

(even with $a > 0$). If we begin with such an f and subtract $(s, -r, 0)$ or $(0, s, -r)$ repeatedly, subject to the condition that the f obtained always satisfies (20) then the

process eventually stops, and then at least one of the following holds:

$$(21)\quad \begin{cases} \text{(i)} & a < -n + s, \quad b < -n + s \\ \text{(ii)} & a < -n + s, \quad c > n - r \\ \text{(iii)} & b > n - r, \quad c > n - r. \end{cases}$$

Definition 6F.1. Let $r/s \geqslant 0$, $r/s \in G_n'$, and let Δ be a positive integer. Then the polynomial f satisfying (18), (20), and at least one part of (21) will be called the minimal quadratic of form Δ (at r/s, for given n), will be denoted by g_Δ, and will be said to be of type (i), (ii), or (iii) if (21)(i), (ii), or (iii) holds respectively.

Remarks. Despite its name, a minimal quadratic is not assumed to have degree 2. For any g_Δ of type (i) or type (ii), $|g_\Delta| \leqslant n$. However, a g_Δ of type (ii) may have $|g_\Delta| > n$, that is, $a > n$, though not for small Δ. For example, if $r/s \neq 0,1$, then $r^2 + rs + s^2 \geqslant 7$, so if $\Delta \leqslant 6$, then $a \leqslant 0$, so certainly $|g_\Delta| \leqslant n$. The following lemma is also easily proved from the inequalities in (21).

Lemma 6F.3. Let $r/s \geqslant 0$, $r/s \in G_n'$, and let $g_1 = (a^{(1)}, b^{(1)}, c^{(1)})$ be minimal of form 1. Then $a^{(1)} < 0$ and $c^{(1)} > 0$.

We have shown that g_Δ exists for each positive integer Δ. Also, g_Δ is unique. This is immediate from the first part of the following lemma.

Lemma 6F.4. Let $r/s \geqslant 0$, $r/s \in G_n'$, and let Δ be a positive integer.

(i) If f is any integral polynomial satisfying (18), (20), then

(22) $f = g_\Delta + \alpha(s, -r, 0) + \beta(0, s, -r)$,

where α, β are non-negative integers.

(ii) If, in addition, $|f| \leqslant n$, then $|g_\Delta| \leqslant n$.

Proof. (i) is easy to check, and (ii) is immediate from (i).

Theorem 6F.5. Let r/s, r'/s' be consecutive elements of G_n' with $0 \leqslant r/s < r'/s'$, and let Δ be a positive integer. Then

(i) $|g_1| \leqslant n$, and g_1 has a zero θ_1 in K_n, and in $(r/s, r'/s')$,

(ii) $g_\Delta = \Delta g_1 + \alpha(s, -r, 0) + \beta(0, s, -r)$, where α, β are non-negative integers,

(iii) if there is an f satisfying the conditions of Lemma 6F.4, and having a zero θ in $(r/s, r'/s')$, then g_Δ has a zero θ_Δ in $(r/s, r'/s')$, and

$$\theta_1 \leqslant \theta_\Delta \leqslant \theta,$$

with equality only when the quadratics are equal up to a constant factor.

Proof. Part (i) follows from Lemma 6F.1, and Lemma 6F.4. (Actually, the fact that $|g_1| \leqslant n$ is also easily seen by a direct argument similar to that in the Remarks above.)

(ii) follows from the inequalities (21).

(iii) $\theta_\Delta \leqslant \theta$ follows from Lemma 6F.4 for if α, β are not both zero in (21), then the graph of g_Δ lies below that of f for $x > r/s$. Similarly, $\theta_1 \leqslant \theta_\Delta$ follows from (ii).

The following corollary is immediate:

Corollary 6F.6. Let r/s, r'/s' be consecutive elements of G_n'. Then the least element of K_n in $(r/s, r'/s')$ has form 1 at r/s, and the greatest element of K_n in $(r/s, r'/s')$ has form 1 at r'/s'.

This result suggests an easy way of eliminating the most apparent discrepancy (mentioned in §6A) in the determinant corresponding to $\theta_1, \theta_2, \theta_3$ when θ_2 is rational. We simply eliminate these determinants by inserting an extra (reducible) quadratic $(s, -r, 0)$ at the rational point r/s just before $(0, s, -r)$. To say that g_1 has form 1 at r/s is to say that on writing

$$g_1 = (a^{(1)}, b^{(1)}, c^{(1)}),$$

we have

$$\begin{vmatrix} s & -r & 0 \\ 0 & s & -r \\ a^{(1)} & b^{(1)} & c^{(1)} \end{vmatrix} = 1 .$$

This is one of the two new determinants created by inserting $(s, -r, 0)$. In the case where $r/s \neq 0$, if (c', b', a') is the primitive minimal polynomial of form 1 at s/r of the least element of K_n greater than s/r, then (a', b', c') is the primitive minimal polynomial of form 1 at r/s of the greatest element of K_n less than r/s, and the other new determinant created is

$$\begin{vmatrix} a' & b' & c' \\ s & -r & 0 \\ 0 & s & -r \end{vmatrix} = \begin{vmatrix} r & -s & 0 \\ 0 & r & -s \\ c' & b' & a' \end{vmatrix} = 1.$$

(The case $r/s = 0$ can be handled separately, the least positive element of K_n being given by $(a^{(1)}, b^{(1)}, c^{(1)}) =$

$(-n, -n, 1)$ and the greatest negative element being given by $(a',b',c') = (-n, n, 1)$.)

The eliminated determinant was

$$D = \begin{vmatrix} a' & b' & c' \\ 0 & s & -r \\ a^{(1)} & b^{(1)} & c^{(1)} \end{vmatrix} .$$

Now by Lemma 6F.4(i),

(23) $$(a',b',c') = (a^{(1)},b^{(1)},c^{(1)}) + \alpha(s, -r, 0) + \beta(0, s, -r),$$

for non-negative integers α,β, not both zero. Hence $D = \alpha$. Also, when $r/s > 0$, it follows from Lemma 6F.3 applied to g_1, and to (c',b',a'), that $a^{(1)} < 0$ and $a' > 0$. Then from (23), D is positive, as was observed in the numerical data of Brown and Mahler [2] in the reference to Table 2. (Their determinants were actually negative, as their quadratics had non-negative leading coefficients.)

§6G. The second smallest element of K_n in $(r/s, r'/s')$.

In this section, we will suppose that r/s, r'/s' are consecutive elements of G_n' with $0 \leqslant r/s < r'/s'$, although the results may be easily translated into results for negative r/s. We will let

$$f^{(i)} = (a^{(i)},b^{(i)},c^{(i)}) \qquad (i = 1,2)$$

denote the primitive quadratics of the smallest and second smallest elements $\theta^{(1)}$, $\theta^{(2)}$ in $(r/s, r'/s')$. Thus, Corollary 6F.6 shows that $f^{(1)} = g_1$. I have also investigated the quadratic $f^{(2)}$ by methods which include further study of the minimal quadratics of §6F. As the details are lengthy I will not include them here, but

I shall describe the results and show how they can be used in connection with the determinant

$$\det\big((0,\ s,\ -r),\ f^{(1)},f^{(2)}\big)\ .$$

It will be convenient to express an arbitrary integral quadratic f such that

$$f(r,s) = \Delta > 0$$

in the notation

$$\begin{aligned} f &= [\Delta,\alpha,\beta] \\ &= \Delta f^{(1)} + \alpha(s,\ -r,\ 0) + \beta(0,\ s,\ -r) \end{aligned}$$

where by Lemma 6F.4(i), α,β are non-negative integers if $|f| \leqslant n$. This notation is useful not only in determining, of two given quadratics, which has the least zero greater than r/s, but also in applying various techniques of obtaining from a given quadratic, a quadratic with a lesser zero in $(r/s,\ r'/s')$.

For example, if

$$f = [\Delta,\alpha,\beta]\ ,$$

where $\beta = 0$, $\alpha \geqslant 2$, is the primitive minimal polynomial of an element θ_f of K_n in $(r/s,\ r'/s')$, and $[x]^+$ denotes the least integer not less than x, and

$$h = \left[\left[\frac{\Delta}{\alpha}\right]^+,\ 1,\ 0\right]\ ,$$

then $|h| \leqslant n$, and h has a zero θ_h in $(r/s,\ r'/s')$, such that

$$\theta^{(1)} < \theta_h < \theta_f\ ,$$

so $\theta_f \neq \theta^{(2)}$, and $f \neq f^{(2)}$.

Let us write in square bracket notation,

$$f^{(2)} = [\Delta^{(2)},\alpha^{(2)},\beta^{(2)}]\ .$$

Then it is easily checked that

(1) $$D = \det\big((0,\ s,\ -r),\ f^{(1)},f^{(2)}\big) = \alpha^{(2)}\ .$$

I have proved by detailed considerations, that

(2) $\quad f^{(2)} = [\Delta,1,0],\ [\Delta,0,1],\ [\Delta,1,1]\ ,$

$[3,1,2],\ \text{or}\ [3,2,1]\ ,$

and that the case $f^{(2)} = [3,2,1]$ can only occur when $r/s > 1$ and $b^{(1)} + s > n$. It follows that the determinant in (1) satisfies

(3) $\quad D = 0,1,2,$ and $D = 0,1$ if $0 \leqslant r/s \leqslant 1.$

(Negative determinants can occur if we normalise the quadratics to get non-negative leading coefficients.)

The possibility $f^{(2)} = [3,1,2]$ is interesting. An example where this holds is when $n = 11$, $r/s = 8/11$, $r'/s' = 3/4$, when it can be checked by considering minimal quadratics in the sense of §6F and using the various reduction techniques alluded to in the present section, that

$$f^{(1)} = [1,0,0] = (-6,-8,9),$$

$$f^{(2)} = [3,1,2] = (-7,-10,11).$$

A consequence of this example is that at $11/8$ in K_{11} we have the two consecutive quadratics preceding $11/8$, forming with $11/8$, the determinant,

$$\begin{vmatrix} 11 & -10 & -7 \\ 9 & -8 & -6 \\ 0 & 8 & -11 \end{vmatrix} = -2 \qquad (= -\beta^{(2)} \text{at } 8/11)$$

Thus the determinant property for three consecutive elements of K_n does not hold in this example where the largest element is rational (providing another counter-example to a conjecture of Brown and Mahler), although the determinant property does hold (at this point) in the modified scheme (explained after Corollary 6F.5) where we place $(8,-11,0)$ before $(0,8,-11)$, for then the determinant becomes

$$\begin{vmatrix} 11 & -10 & -7 \\ 9 & -8 & -6 \\ 8 & -11 & 0 \end{vmatrix} = -1 \quad (= -\alpha^{(2)} \text{ at } 8/11).$$

I have not determined whether the case $f^{(2)} = [3,2,1]$ can actually occur (when $r/s > 1$). If it cannot, then it would follow that the determinant property for three consecutive elements

$$f, \quad g, \quad (s, -r, 0)$$

in our modified scheme also holds (for arbitrary $r/s \in G_n'$).

§6H. The quadratics of form 1 at r/s.

In this section I will prove the determinant property for quadratics of form 1. That is, I will prove the following theorem.

Theorem 6H.1. Let r/s, r'/s' be consecutive elements of G_n'. Then the primitive quadratics of elements of K_n in $(r/s, r'/s')$ of form 1 have the property that the determinant of any 3 consecutive quadratics is $0, \pm 1$.

Proof. Let S be the set of all integral quadratics

$$f(x) = ax^2 + bx + c$$

such that

$$f(r,s) = 1, \quad |f| \leq n,$$

and let S_1 be the subset of S consisting of the quadratics in S satisfying

$$f(r',s') < 0.$$

By Lemmas 6B.1(ii) and 6B.2(i), there is a one-one correspondence between the elements of K_n in $(r/s, r'/s')$ and the quadratics in S_1. We will prove the determinant result about S_1 by proving a determinant result about S.

As we found in §6F(19) (which is of course valid for arbitrary r/s in G_n'), the elements of S may be written as

(1) $\qquad f = f_0 + \alpha(s, -r, 0) + \beta(0, s, -r), \quad \alpha,\beta \in Z,$

where f_0 is a particular quadratic of form 1 at r/s. For $i = 1,2$ and elements f_i of S such that

(2) $\qquad f_i = f_0 + \alpha_i(s, -r, 0) + \beta_i(0, s, -r)$

we have

(3) $\qquad f_1(x) - f_2(x) = (sx - r)\big((\alpha_1 - \alpha_2)x + (\beta_1 - \beta_2)\big).$

Thus the graphs of $f_1(x)$, $f_2(x)$ cross at r/s, and if $\alpha_1 \neq \alpha_2$, at $-(\beta_1-\beta_2)/(\alpha_1-\alpha_2)$. Thus the graphs cross in $(r/s, r'/s')$ if and only if

(4) $$-\frac{\beta_1 - \beta_2}{\alpha_1 - \alpha_2} \in (r/s, r'/s') .$$

We now show that this is not possible.

Without loss of generality, suppose that $r/s \geqslant 0$, $s > 0$. Since r/s, r'/s' are consecutive in K_n, the larger of $r+r'$, $s+s'$ exceeds n. Suppose that (4) holds. Then

$$|\beta_1 - \beta_2| \geqslant r+r' , \quad |\alpha_1 - \alpha_2| \geqslant s+s' .$$

If $r \geqslant s$, then $r' > s'$, $r+r' > s+s'$, so $r+r' > n$, $|\beta_1 - \beta_2| > n$. By considering the third components of f_1, f_2, we see that $r \leqslant 1$, $s = 1$, $r/s = 1/1$. On the other hand, if $r < s$, then $r' \leqslant s'$, $r+r' < s+s'$, $s+s' > n$, $|\alpha_1 - \alpha_2| > n$, and from the first components of f_1, f_2, $s \leqslant 1$, so $r/s = 0/1$. Thus $(r/s, r'/s') = \left(\frac{0}{1}, \frac{1}{n}\right)$ or $\left(\frac{1}{1}, \frac{n}{n-1}\right)$. But it is readily seen from §6D, §6E respectively that (4) is false in these cases. (In fact the determinant problem has been disposed of for these intervals.) Thus we have shown that (4) does not hold, that is the graphs of the

quadratics f in S do not cross in $(r/s, r'/s')$, and so an ordering on the quadratics in S may be defined as follows: If $f_1, f_2 \in S$, then

(5) $\qquad f_1 < f_2 \Leftrightarrow f_1(x) < f_2(x) \qquad \forall\, x \in (r/s, r'/s')$.

Clearly, this ordering on S agrees with the ordering on S_1 corresponding to the order of the zeros of the quadratics in S_1, and S_1 is an initial segment of S in the sense that there is an element f_N in S_1 such that

$$S_1 = \{f \in S \mid f \leqslant f_N\}.$$

Thus, we need only prove the corresponding determinant result for S.

It is also clear from (3) that if f_1, f_2 in S are given by (2), then

(6) $\qquad f_1 < f_2 \Leftrightarrow \alpha_1 r + \beta_1 s < \alpha_2 r + \beta_2 s$

$$\text{or} \begin{cases} \alpha_1 r + \beta_1 s = \alpha_2 r + \beta_2 s \\ \text{and} \quad \alpha_1 < \alpha_2 \quad . \end{cases}$$

Thus, the ordering $<$ on S is a "linear ordering". One final observation about S which will make our proof work is that S consists of the points (a,b,c) of the lattice Z^3 which lie in a 2-dimensional convex set, namely that given by the inequalities

$$\begin{cases} -n \leqslant a = a_0 + \alpha s \leqslant n \\ -n \leqslant b = b_0 - \alpha r + \beta s \leqslant n \\ -n \leqslant c = c_0 - \beta r \leqslant n \ , \end{cases}$$

or, in terms of α, β, the point $\underset{\sim}{\alpha} = (\alpha, \beta)$ lies in the convex set K in R^2 given by the inequalities

$$\begin{cases} (-n-a_0)/s \leqslant \alpha \leqslant (n-a_0)/s \\ -n - b_0 \leqslant -\alpha r + \beta s \leqslant n - b_0 \\ (-n-c_0)/r \leqslant \beta \leqslant (n-c_0)/r \ . \end{cases}$$

Let f_1, f_2, f_3 be consecutive elements of S given by (2) with $i = 1,2,3$ and write also

(7) $\qquad f_i = (a_i, b_i, c_i) \qquad (i = 1,2,3).$

Then the relevant determinant is

$$D = \det \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix}$$

$$= \det \left(\begin{bmatrix} 1 & \alpha_1 & \beta_1 \\ 1 & \alpha_2 & \beta_2 \\ 1 & \alpha_3 & \beta_3 \end{bmatrix} \begin{bmatrix} a_0 & b_0 & c_0 \\ s & -r & 0 \\ 0 & s & -r \end{bmatrix} \right)$$

$$= \det \begin{bmatrix} 1 & \alpha_1 & \beta_1 \\ 1 & \alpha_2 & \beta_2 \\ 1 & \alpha_3 & \beta_3 \end{bmatrix}$$

$$= \det \begin{bmatrix} \alpha_2-\alpha_1 & \beta_2-\beta_1 \\ \alpha_3-\alpha_2 & \beta_3-\beta_2 \end{bmatrix} .$$

$$= \det (\underset{\sim}{\alpha}_2-\underset{\sim}{\alpha}_1, \underset{\sim}{\alpha}_3-\underset{\sim}{\alpha}_2),$$

where $\underset{\sim}{\alpha}_i = (\alpha_i, \beta_i)$ for $i = 1,2,3$.

Suppose that $|D| > 1$. Then by Lemma 3B.5, or by Theorem 34 of Hardy and Wright [6], there is a point $\underset{\sim}{\alpha}_4 = (\alpha_4, \beta_4)$ in Z^2 lying within the triangle with vertices $\underset{\sim}{\alpha}_1, \underset{\sim}{\alpha}_2, \underset{\sim}{\alpha}_3$ or on its boundary, but distinct from $\underset{\sim}{\alpha}_i$ for $i \neq 4$. Since $\underset{\sim}{\alpha}_4 \in K$, it follows from (6) that the quadratic f_4 of S, given by

$$f_4 = f_0 + \alpha_4(s, -r, 0) + \beta_4(0, s, -r),$$

lies between f_1 and f_3. But $f_4 \neq f_2$, and this contradicts the consecutiveness of f_1, f_2, f_3. Thus $|D| \leq 1$, as required. This completes the proof of Theorem 6H.1.

CHAPTER 7

AN INDEX PROBLEM OF CASSELS

§7A Introduction.

Throughout this chapter, M will denote a 3-dimensional lattice, and L a 2-dimensional lattice in R^3, and the octants in R^3 will be denoted in the obvious way by A_{+++}, A_{-++}, and so on. For example

$$A_{-++} = \{\underset{\sim}{x} | x_1 \leqslant 0,\ x_2 \geqslant 0,\ x_3 \geqslant 0\}.$$

Similarly the quadrants in R^2 will be denoted by $\pm A_{++}$, $\pm A_{-+}$.

Throughout this chapter, we will also let F denote the convex distance function in R^3 given by

$$F(\underset{\sim}{x}) = F(x_1,x_2,x_3) = |x_1| + |x_2| + |x_3|. \tag{1}$$

Now let M be a 3-dimensional lattice, and A an octant in R^3. Define a minimal point of M in A with respect to F as a point $\underset{\sim}{u}$ in $\dot{M} \cap A$ with minimal $F(\underset{\sim}{u})$.

For simplicity, suppose that $\dot{M}$ contains no points on the coordinate planes, and let X be a set of 4 minimal points, one for each pair of opposite octants $\pm A$. Then we have the following problem, which is the subject of the present chapter.

Index problem: What can be said about the index $(M:\ell(X))$ of the sublattice of M generated by X?

This problem was raised by Professor J.W.S. Cassels in the Arithmetic Seminar in Cambridge in 1974. Professor Cassels suggested that one ought to be able to show that the index is at most 6, or perhaps at most 2, or perhaps even always equal to 1. The existence of a small bound for the index is suggested by results on the classical successive minima. For

example, by VIII.4.3 Corollary of Cassels [3], the index of the associated minimal vectors in R^n does not exceed $n!$ In fact, for $n = 3$, Bantegnie [1] has shown that the index of these minimal vectors does not exceed 4. (See also Lekkerkerker [8] p.56.)

For the problem as stated above, the index need in fact not be 1. In §7B we will give an example where the index is 2, and to retain the possibility that the index be always equal to 1, we will introduce one modified 'minimal' lattice point. We will also discuss briefly the connection with successive minima.

Then in §7C, I will prove index results for 2-dimensional lattices in R^2 or R^3. Sections §7D, §7E concern the two main ideas which we will use to tackle the index problem of Cassels. Finally, in §7F I will obtain a bound for this index.

§7B. The set of minimal points of a 3-dimensional lattice, notation, examples.

Let M be a 3-dimensional lattice in R^3. Suppose that $\dot{M}$ has no points on the coordinate planes. Then each point $\underset{\sim}{u}$ in $\dot{M}$ belongs to a uniquely determined octant $A(\underset{\sim}{u})$. Let F be as in §7A(1).

Let the term minimal point of M in A (or of $\dot{M} \cap A$) be as defined in §7A, and let X be a set of minimal points as in §7A. Suppose that

(1) $$X = \{\underset{\sim}{a}, \underset{\sim}{b}, \underset{\sim}{c}, \underset{\sim}{d}\},$$

where

(2) $$F(\underset{\sim}{a}) \leqslant F(\underset{\sim}{b}) \leqslant F(\underset{\sim}{c}), \quad \text{and} \quad F(\underset{\sim}{b}) \leqslant F(\underset{\sim}{d}).$$

Since $\dot{M}$ has no points on the coordinate planes, $\underset{\sim}{a}, \underset{\sim}{b}$ are

linearly independent. Let $\underset{\sim}{c}'$, $\underset{\sim}{d}'$ be points such that

$$(3)\quad \begin{cases} \underset{\sim}{c}' \in M \cap A(\underset{\sim}{c}) \sim \lin(\underset{\sim}{a},\underset{\sim}{b}) = V \text{ say}, \\ \underset{\sim}{d}' \in M \cap A(\underset{\sim}{d}) \sim \lin(\underset{\sim}{a},\underset{\sim}{b}) = W \text{ say}, \\ F(\underset{\sim}{c}') = \min\{F(\underset{\sim}{u}) \mid \underset{\sim}{u} \in V\}, \\ F(\underset{\sim}{d}') = \min\{F(\underset{\sim}{u}) \mid \underset{\sim}{u} \in W\}, \end{cases}$$

and let

$$X' = \{\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{c}',\underset{\sim}{d}'\}.$$

In this chapter, we shall investigate both indices

$$(M:\ell(X)), \quad (M:\ell(X')).$$

Since a plane in R^3 through $\underset{\sim}{0}$ contains interior points of at most three octant pairs $\pm A$, it follows that at most one of $\underset{\sim}{c}$ and $\underset{\sim}{d}$ lies in $\lin(\underset{\sim}{a},\underset{\sim}{b})$. Hence we may suppose that $\underset{\sim}{c} \notin \lin(\underset{\sim}{a},\underset{\sim}{b})$, so that

$$(4)\quad \begin{aligned} &\underset{\sim}{c}' = \underset{\sim}{c}, \quad X' = \{\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{c},\underset{\sim}{d}'\}, \\ &X' = X \Leftrightarrow \underset{\sim}{d} \notin \lin(\underset{\sim}{a},\underset{\sim}{b}). \end{aligned}$$

We now give an example where Cassels's index $(M:\ell(X))$ is 2 but the modified index $(M:\ell(X'))$ is 1.

Example 1.

We note first that if $\underset{\sim}{x}$ belongs to the lattice with basis

$$(0,1,0), \quad (0,0,1), \quad (1,\tfrac{1}{2},\tfrac{1}{2})$$

and $\underset{\sim}{x}$ is not a linear combination of $(0,1,0)$ and $(0,0,1)$, then $F(\underset{\sim}{x}) \geqslant 2$. We consider a lattice M which is a slight deformation of the above lattice, namely the lattice M with basis

$$\underset{\sim}{u} = (-\varepsilon_1,\ 1,\ \varepsilon_2),$$

$$\underset{\sim}{v} = (-\varepsilon_3,\ -\varepsilon_4,\ 1),$$

$$\underset{\sim}{w} = (1,\ \tfrac{1}{2}+\varepsilon_5,\ \tfrac{1}{2}+\varepsilon_6)$$

where the ε_i are small positive numbers chosen to satisfy

conditions to be stated. The ε_i are to satisfy the appropriate linear independence conditions over the rationals, so that $\dot{M}$ contains no point on a coordinate plane. The ε_i are to be small enough so that $F(\underset{\sim}{x}) > 1\frac{1}{2}$ say for all $\underset{\sim}{x}$ in $M \sim \ell(\underset{\sim}{u},\underset{\sim}{v})$. It then follows that one of the minimal points $\underset{\sim}{a}$, $\underset{\sim}{b}$ (see (2)) is $\underset{\sim}{u}$ in A_{-++}, and the other is $\underset{\sim}{v}$ in A_{--+}, and that these are unique. (All the while we are assuming that the ε_i are small enough.)

Now let $\underset{\sim}{c}$, $\underset{\sim}{d}$ be minimal points of M in A_{+++} and A_{+-+} respectively. Consider the vector

$$\underset{\sim}{u} \times \underset{\sim}{v} = (1 + \varepsilon_2\varepsilon_4,\ -\varepsilon_2\varepsilon_3 + \varepsilon_1,\ \varepsilon_1\varepsilon_4 + \varepsilon_3),$$

normal to $\mathrm{lin}(\underset{\sim}{u},\underset{\sim}{v})$, and consider the lattice point

$$\underset{\sim}{v} - \underset{\sim}{u} = (-\varepsilon_3 + \varepsilon_1,\ -\varepsilon_4 - 1,\ 1 - \varepsilon_2).$$

Suppose that

$$\varepsilon_1 > \varepsilon_3, \quad \varepsilon_2 < 1. \tag{5}$$

Then

$$\underset{\sim}{u} \times \underset{\sim}{v} \in A_{+++}, \quad \underset{\sim}{v} - \underset{\sim}{u} \in A_{+-+},$$

and $\mathrm{lin}(\underset{\sim}{u},\underset{\sim}{v})$ contains no point of $\dot{M} \cap A_{+++}$. Hence

$$\underset{\sim}{c} \notin \mathrm{lin}(\underset{\sim}{u},\underset{\sim}{v}) = \mathrm{lin}(\underset{\sim}{a},\underset{\sim}{b}),$$

$$\underset{\sim}{c}' = \underset{\sim}{c},$$

where $\underset{\sim}{c}'$ is as in (3).

Let us consider the minimal values of $F(\underset{\sim}{x})$, where $\underset{\sim}{x}$ lies in a coset

$$\begin{cases} \underset{\sim}{x} \in h\underset{\sim}{w} + \ell(\underset{\sim}{u},\underset{\sim}{v}) & (h \in Z), \\ \underset{\sim}{x} \neq \pm \underset{\sim}{u},\ \pm \underset{\sim}{v},\ \underset{\sim}{0} & \\ \underset{\sim}{x} \in A_{+++} \cup A_{+-+}. & \end{cases} \tag{6}$$

For $h = 0$, we have

$$\underset{\sim}{v} - \underset{\sim}{u} \in A_{+-+}, \quad F(\underset{\sim}{v}-\underset{\sim}{u}) = 2 + \varepsilon_1 - \varepsilon_3 + \varepsilon_4 - \varepsilon_2, \tag{7}$$

and

$$F(\underset{\sim}{x}) > 2\tfrac{1}{2} \quad \text{for other} \quad \underset{\sim}{x} \quad \text{with} \quad h = 0.$$

For $h = 1$, we have

$$\underset{\sim}{w} \in A_{+++} , \quad F(\underset{\sim}{w}) = 2 + \varepsilon_5 + \varepsilon_6 . \tag{8}$$

We also have

$$\begin{cases} \underset{\sim}{w} - \underset{\sim}{u} = (1 + \varepsilon_1, \ -\tfrac{1}{2} + \varepsilon_5, \ \tfrac{1}{2} + \varepsilon_6 - \varepsilon_2) \\ \qquad \in A_{+-+} , \\ F(\underset{\sim}{w}-\underset{\sim}{u}) = 2 + \varepsilon_1 - \varepsilon_5 + \varepsilon_6 - \varepsilon_2 , \end{cases} \tag{9}$$

and

$$F(\underset{\sim}{x}) > 2\tfrac{1}{2} \quad \text{for other} \quad \underset{\sim}{x} \quad \text{with} \quad h = 1.$$

For $h = 2$, we have

$$2\underset{\sim}{w} - \underset{\sim}{u} - \underset{\sim}{v} = (2 + \varepsilon_1 + \varepsilon_3, \ 2\varepsilon_5 + \varepsilon_4, \ 2\varepsilon_6 - \varepsilon_2)$$

which lies in A_{+++} on assuming that

$$2\varepsilon_6 > \varepsilon_2 , \tag{10}$$

and we have

$$F(2\underset{\sim}{w}-\underset{\sim}{u}-\underset{\sim}{v}) = 2 + \varepsilon_1 + \varepsilon_3 + 2\varepsilon_5 + \varepsilon_4 + (2\varepsilon_6-\varepsilon_2) \tag{11}$$

and

$$F(\underset{\sim}{x}) > 2\tfrac{1}{2} \quad \text{for other} \quad \underset{\sim}{x} \quad \text{with} \quad h = 2.$$

For $h \geqslant 3$, we have $F(\underset{\sim}{x}) > 3\tfrac{1}{2}$, and for $h \leqslant 0$, $\underset{\sim}{x}$ clearly cannot lie in $A_{+++} \cup A_{+-+}$ unless $F(\underset{\sim}{x})$ is rather large, certainly greater than $3\tfrac{1}{2}$ (always assuming that the ε_i are small enough).

Thus, in considering the conditions (6), we find that the minimal point $c = \underset{\sim}{c}'$ in A_{+++} must be one of the points $\underset{\sim}{w}$, $2\underset{\sim}{w} - \underset{\sim}{u} - \underset{\sim}{v}$ which satisfy (8), (10) and (11), and which have $h = 1,2$ respectively. We can ensure that

$$\underset{\sim}{c} = 2\underset{\sim}{w} - \underset{\sim}{u} - \underset{\sim}{v} \tag{12}$$

by choosing the ε_i so that

$$F(2\underset{\sim}{w} - \underset{\sim}{u} - \underset{\sim}{v}) < F(\underset{\sim}{w}) ,$$

that is, to satisfy

(13) $$\varepsilon_1 + \varepsilon_3 + \varepsilon_5 + \varepsilon_4 + \varepsilon_6 > \varepsilon_2 ,$$

and (13) is certainly compatible with (5) and (10).

Equation (12) implies that

(14) $$\big(M: \ell(\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{c})\big) = 2.$$

From the above consideration of the conditions (6), we also conclude that the point $\underset{\sim}{d}'$ in A_{+-+} (see (3)) is

(15) $$\underset{\sim}{d}' = \underset{\sim}{w} - \underset{\sim}{u}$$

calculated in (9), so that with X' as in (4),

$$\big(M: \ell(X')\big) = 1 ,$$

as claimed.

We next prove that the minimal point $\underset{\sim}{d}$ in A_{+-+} is

(16) $$\underset{\sim}{d} = \underset{\sim}{v} - \underset{\sim}{u} ,$$

(see (7)).

Inequalities (10), (13) imply that

$$\varepsilon_6 > \varepsilon_1 + \varepsilon_3 + \varepsilon_5 + \varepsilon_4$$

$$> \varepsilon_5 + \varepsilon_4 - \varepsilon_3 .$$

Hence by (7), (9),

$$F(\underset{\sim}{v}-\underset{\sim}{u}) < F(\underset{\sim}{w}-\underset{\sim}{u}) ,$$

and (16) follows. From this, we see that with X as in (1),

$$\ell(\underset{\sim}{X}) = \ell(\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{c},\underset{\sim}{d})$$

$$= \ell(\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{c}),$$

so by (14), we have

$$\big(M: \ell(X)\big) = 2,$$

as we claimed.

Connection with successive minima.

Let $\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{c},\underset{\sim}{d}$ be as in (1), (2). It is clear that the first minimal point $\underset{\sim}{a}$ is just a first successive minimal point of the lattice M with respect to the distance function

F. (See Cassels [3], p.201.) Further, given a lattice M, it may well turn out that the successive minimal points are $\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{c}$ or $\underset{\sim}{a}, \underset{\sim}{b}, \underset{\sim}{d}$. However, we now give a simple example which illustrates that the second successive minimal point need not even lie in $\text{lin}(\underset{\sim}{a},\underset{\sim}{b})$.

Example 2.

Let M be the lattice in R^3 with basis

$$u = (1,\varepsilon_1,\varepsilon_2),$$

$$v = (\varepsilon_3,1,\varepsilon_4),$$

$$w = (\varepsilon_5,\varepsilon_6,-1\tfrac{1}{2}),$$

where $\varepsilon_1 + \varepsilon_2 < \varepsilon_3 + \varepsilon_4,$

$0 < \varepsilon_i < \frac{1}{8}$ for all i.

Then $\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w}$ are successive minimal points, and $\underset{\sim}{a} = \underset{\sim}{u}$, but $\underset{\sim}{b} = \underset{\sim}{w}$, so

$$\text{lin}(\underset{\sim}{a},\underset{\sim}{b}) \neq \text{lin}(\underset{\sim}{u},\underset{\sim}{v}).$$

Let M be a 3-dimensional lattice in R^3, let λ_1,λ_2 denote the first two successive minima of M with respect to F, and let $\underset{\sim}{a},\underset{\sim}{b}$ denote the first two minimal points as in (1), (2). Also, let

(17) $$F(\underset{\sim}{a}) = \mu_1, \quad F(\underset{\sim}{b}) = \mu_2.$$

Then

(18) $$\lambda_1 = \mu_1, \quad \mu_2 - \lambda_1 \leqslant \lambda_2 \leqslant \mu_2.$$

This is trivial, except for the left hand inequality, which will be proved in §7D.

In the case when the (classical) second successive minimal point $\underset{\sim}{m}_2$ of M lies in $\text{lin}(\underset{\sim}{a},\underset{\sim}{b})$ it may be proved that, re-choosing $\underset{\sim}{b}$ if necessary,

(19) $$\underset{\sim}{b} = \underset{\sim}{m}_2 \quad \text{or} \quad \underset{\sim}{m}_2 - \underset{\sim}{a}$$

(the proof would use Theorems 7C.1, 7C.3.), and in this case, (18) obviously follows from (19).

§7C. Minimal points of a 2-dimensional lattice.

In this section, we let L be a general 2-dimensional lattice in R^3 which may have points on the coordinate planes or axes. We will prove if $\underset{\sim}{a},\underset{\sim}{b}$ are the minimal points as defined below, then

$$(L : \ell(\underset{\sim}{a},\underset{\sim}{b})) = 1,$$

except in certain cases when L contains points on the coordinate axes, when the index is 2. It will then follow that the minimal points $\underset{\sim}{a},\underset{\sim}{b}$ of §7B (1), (2) are a basis for $M \cap \text{lin}(\underset{\sim}{a},\underset{\sim}{b})$.

Let P be an arbitrary plane in R^3 passing through the origin. There are three possibilities:

(i) P is a coordinate plane,

(ii) P is not a coordinate plane but passes through a coordinate axis.

(iii) P contains no coordinate axis.

These cases correspond to the cases when the Cartesian equation for P has two, one or none of its coefficients equal to zero. Consider the intersections $P \cap A$, where A is an octant in R^3. When non-degenerate, $P \cap A$ is a 2-dimensional cone, which we will refer to as a coordinate <u>sector</u> in P. In cases (i), (ii) P has four distinct sectors. (In case (ii), exactly two of the lines of intersection of P with a coordinate plane coincide.) In case (iii), P has six distinct sectors, these being determined by the three distinct lines in which the coordinate planes intersect P.

Now consider the convex distance function F of §7A (1). That is,

(1) $$F(\underset{\sim}{x}) = |x_1| + |x_2| + |x_3|.$$

We will be using the fact that F is linear in each octant and hence in each sector in P. We will also use the fact that the inequality

(2) $$F(\underset{\sim}{x}+\underset{\sim}{y}) < F(\underset{\sim}{x}) + F(\underset{\sim}{y})$$

holds whenever no octant contains both $\underset{\sim}{x}$ and $\underset{\sim}{y}$, and so holds for any two points of P which do not lie in a common sector.

The minimal points $\underset{\sim}{a}$, $\underset{\sim}{b}$ of L.

Let L be an arbitrary 2-dimensional lattice in the plane P in R^3. Choose $\underset{\sim}{a}$ in L to satisfy

(3) $$\underset{\sim}{a} \in \dot{L}, \quad F(\underset{\sim}{a}) = \min_{\underset{\sim}{u}\in\dot{L}} F(\underset{\sim}{u}).$$

Choose a sector A in P such that $\underset{\sim}{a} \in A$, and choose $\underset{\sim}{b}$ in L such that

(4) $$\underset{\sim}{b} \in L \sim (\text{relint } A) \sim (-\text{ relint } A) \sim \ell(\underset{\sim}{a}) = T, \text{ say,}$$

(5) $$F(\underset{\sim}{b}) = \min_{\underset{\sim}{u}\in T} F(\underset{\sim}{u}).$$

We will refer to $\underset{\sim}{a}$, $\underset{\sim}{b}$ as the first and second minimal points of L with respect to F. (An analogous general definition could be given for the first two minimal points of an arbitrary 3-dimensional lattice, though we have omitted this for the sake of simplicity.) If M is a 3-dimensional lattice in R^3 such that $\dot{M}$ has no points on the coordinate planes, if $\underset{\sim}{a}$, $\underset{\sim}{b}$ are taken as the first two minimal points as in §7B, and if $L = M \cap \text{lin}(\underset{\sim}{a},\underset{\sim}{b})$, then $\underset{\sim}{a}$, $\underset{\sim}{b}$ are minimal points of L in the above sense, as is easily seen.

Theorem 7C.1. Let L be a 2-dimensional lattice in the plane P in R^3, and let $\underset{\sim}{a}$, $\underset{\sim}{b}$ be minimal points with respect to F as given by (1), (3), (4), (5) above. Then

$$(L : \ell(\underset{\sim}{a},\underset{\sim}{b})) = 1$$

except when $F(\underset{\sim}{a}) = F(\underset{\sim}{b})$ and L has basis $\underset{\sim}{a}$, $\frac{1}{2}(\underset{\sim}{a}+\underset{\sim}{b})$ (in which case the index is 2). The exceptional case can only occur when P has just four sectors, one of which contains $\underset{\sim}{a}$, $\underset{\sim}{b}$ on its bounding rays.

Note 1: Thus, in the exceptional case, one of $\underset{\sim}{a}$, $\underset{\sim}{b}$ lies on a coordinate axis, and the other lies on a coordinate plane.

Note 2: If $P = \{\underset{\sim}{x} | x_3 = 0\}$ we may identify F with the corresponding distance function on R^2, and infer the analogous result for 2-dimensional lattices in R^2.

Proof of Theorem. Let F_A denote the linear function that F coincides with on A. Replacing $\underset{\sim}{b}$ by $-\underset{\sim}{b}$ if necessary we may suppose that the following (strict) inequality holds

$$F_A(\underset{\sim}{b}) < F(\underset{\sim}{a}) \tag{6}$$

(that is, $\underset{\sim}{0}$, $\underset{\sim}{b}$ lie on the same side of the line $F_A(\underset{\sim}{x}) = F(\underset{\sim}{a})$ in P). From (3), we have

$$F(\underset{\sim}{b}) \geqslant F(\underset{\sim}{a}). \tag{7}$$

We suppose the required index is not 1. Then by Lemma 3B.5 the triangle $\underset{\sim}{0}\ \underset{\sim}{a}\ \underset{\sim}{b}$ contains a point $\underset{\sim}{y}$ of L other than $\underset{\sim}{0}$, $\underset{\sim}{a}$, $\underset{\sim}{b}$; any such point is of the form

$$\underset{\sim}{y} = \lambda\underset{\sim}{a} + \mu\underset{\sim}{b}, \quad 0 \leqslant \lambda < 1, \quad 0 \leqslant \mu < 1, \quad 0 < \lambda + \mu \leqslant 1.$$

Now since $\underset{\sim}{b} \notin -$ relint A, $\underset{\sim}{a} \in A$, we have that $\underset{\sim}{y} \notin$ relint A. Thus it follows from (3), (4), (6) and the linearity of F_A that $\underset{\sim}{y}$ must be in T as defined by (4). If $F(\underset{\sim}{a}) < F(\underset{\sim}{b})$, or $\underset{\sim}{a}$ and $\underset{\sim}{b}$ are not contained in a common sector of P, or $\lambda + \mu < 1$, then by (2), (7) and the convexity of F we obtain $F(\underset{\sim}{y}) < F(\underset{\sim}{b})$, a contradiction to (5). Hence we must have

(i) $F(\underset{\sim}{b}) = F(\underset{\sim}{a})$,

(ii) some sector in P contains both $\underset{\sim}{a}$ and $\underset{\sim}{b}$, and

(iii) $\underset{\sim}{y}$ is on the line-segment $\underset{\sim}{a}\,\underset{\sim}{b}$.

It follows from (iii) that $\underset{\sim}{b} - \underset{\sim}{a}$ is not primitive, so let h be the largest integer such that

$$\underset{\sim}{a} - \underset{\sim}{b} = h\underset{\sim}{u} \qquad (h \geqslant 2) \tag{8}$$

for some $\underset{\sim}{u}$ in L. By (3), we have

$$F(\underset{\sim}{a}-\underset{\sim}{b}) \geqslant hF(\underset{\sim}{a}).$$

But

$$\begin{aligned} F(\underset{\sim}{a}-\underset{\sim}{b}) &\leqslant F(\underset{\sim}{a}) + F(-\underset{\sim}{b}) \\ &= 2F(\underset{\sim}{a}) \quad \text{by (i)}. \end{aligned}$$

Hence

$$F\big(\underset{\sim}{a}+(-\underset{\sim}{b})\big) = F(\underset{\sim}{a}) + F(-\underset{\sim}{b}), \tag{9}$$

and

$$h = 2. \tag{10}$$

From (9) (10) we have that

(iv) some sector in P contains both $\underset{\sim}{a}$ and $-\underset{\sim}{b}$.

Now conditions (ii), (iv) imply that there are two adjacent sectors in P which both contain $\underset{\sim}{a}$, and which each contain one of $\underset{\sim}{b}$, $-\underset{\sim}{b}$. This can only happen when P has only four sectors, and $\underset{\sim}{a}$, $\underset{\sim}{b}$ lie on the bounding rays of one sector. Moreover, (iii) and (10) show that $\underset{\sim}{y} = \frac{1}{2}(\underset{\sim}{a}+\underset{\sim}{b})$, and that $\underset{\sim}{a}$, $\underset{\sim}{y}$ form a basis for L. Thus if the index is not 1 then the exceptional case specified in the theorem must occur.

We now have the following immediate corollary.

Theorem 7C.2. Let M be a 3-dimensional lattice in R^3 having no non-zero points on the coordinate planes, and let $\underset{\sim}{a}$, $\underset{\sim}{b}$, $\underset{\sim}{c}$, $\underset{\sim}{d}$, X be as given in §7B (1), (2), (3), (4). Then

(i) $\quad M \cap \mathrm{lin}(\underset{\sim}{a},\underset{\sim}{b}) = \ell(\underset{\sim}{a},\underset{\sim}{b})$,

(ii) $\quad$ If $\underset{\sim}{d} \in \mathrm{lin}(\underset{\sim}{a},\underset{\sim}{b})$, then $\ell(X) = \ell(\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{c})$.

The following result can be proved by a proof similar to that of Theorem 7C.1, but simpler, and we omit the details.

Theorem 7C.3. Let M be a 3-dimensional lattice in R^3 with first two (classical) successive minimal points $\underset{\sim}{a}$, $\underset{\sim}{b}$ with respect to F. Then the index

$$\left(M \cap \mathrm{lin}(\underset{\sim}{a},\underset{\sim}{b}) : \ell(\underset{\sim}{a},\underset{\sim}{b})\right)$$

is only greater than 1 in the same exceptional cases as in Theorem 7C.1 with $L = \mathrm{lin}(\underset{\sim}{a},\underset{\sim}{b})$.

In Theorem 7C.3, the exceptional cases of course cannot occur if $\dot{M}$ has no points on the coordinate planes.

§7D. Distance of lattice planes from the origin.

In this section, we will let M be a 3-dimensional lattice in R^3, and although not absolutely necessary, we will assume, without explicit mention, that $\dot{M}$ has no points on the coordinate planes. We will consider a lattice plane not through the origin, and parallel to $\mathrm{lin}(\underset{\sim}{a},\underset{\sim}{b})$ where $\underset{\sim}{a}$, $\underset{\sim}{b}$ are as in §7B, and find a lower bound for its distance from $\underset{\sim}{0}$ with respect to the distance function F of §7A(1).

Consider a plane P in R^3, with equation

(1) $$\gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 = t, \quad |\gamma_1| = \max_i |\gamma_i|.$$

By the distance τ of the plane P from $\underset{\sim}{0}$ with respect to F we will mean the minimum value of F on P. We have

(2) $$\tau = F(t/\gamma_1,0,0) = |t/\gamma_1| .$$

This is easily seen from the fact that F is non-negative and linear on each octant A, and so achieves a minimum value at a vertex of one of the polygonal regions $P \cap A$.

The following lemma establishes (18) of §7B.

Lemma 7D.1. Let M be a 3-dimensional lattice in R^3, let $\underset{\sim}{a}$ and $\underset{\sim}{b}$ be as in §7B(1), (2), $\mu_1 = F(\underset{\sim}{a})$, $\mu_2 = F(\underset{\sim}{b})$ and let λ_1 $(= \mu_1)$, λ_2 be the first two successive minima as in §7B(17), (18). Then

$$\lambda_2 \geqslant \max(\mu_2-\mu_1, \mu_1). \tag{3}$$

Proof. The only non-trivial part to prove is

$$\lambda_2 \geqslant \mu_2 - \mu_1 . \tag{4}$$

Let $\underset{\sim}{u}$ be a point of $\dot{M}$ such that

$$F(\underset{\sim}{u}) < \mu_2 - \mu_1 . \tag{5}$$

We must prove that

$$\underset{\sim}{u} \in \ell(\underset{\sim}{a}) . \tag{6}$$

Let A be the octant such that $\underset{\sim}{a} \in A$.

By definition, we have

$$F(\underset{\sim}{a}) = \mu_1 = \min_{v \in \dot{M}} F(\underset{\sim}{v}) \tag{7}$$

and

$$\underset{\sim}{v} \in M, \quad F(\underset{\sim}{v}) < \mu_2 \Leftrightarrow \underset{\sim}{v} \in \pm A . \tag{8}$$

By (5), (8),

$$\underset{\sim}{u} \in \pm A .$$

Take $\underset{\sim}{u}' = \pm \underset{\sim}{u}$ so that

$$\underset{\sim}{u}' \neq \underset{\sim}{0}, \quad \underset{\sim}{u}' \in A, \quad F(\underset{\sim}{u}') < \mu_2 - \mu_1 . \tag{9}$$

Let F_A be the linear function which F coincides with on A. Then

$$F_A(\underset{\sim}{u}') = F(\underset{\sim}{u}') \geqslant F(\underset{\sim}{a}) \quad \text{by (7), since } \underset{\sim}{u}' \neq \underset{\sim}{0},$$
$$= F_A(\underset{\sim}{a}).$$

Hence

$$0 \leqslant F_A(\underset{\sim}{u}'-\underset{\sim}{a}) < F_A(\underset{\sim}{u}') . \tag{10}$$

Now (5), (7), (8), (10) imply

(11) $\qquad \underset{\sim}{u}' - \underset{\sim}{a} \in A, \quad F(\underset{\sim}{u}'-\underset{\sim}{a}) < F(\underset{\sim}{u}') < \mu_2 - \mu_1 .$

Thus (9) implies (11). Repeated use of this fact shows that $\underset{\sim}{u}'$ is a positive integral multiple of $\underset{\sim}{a}$. This establishes (6), and completes the proof of the lemma.

Lemma 7D.2. Let M be a 3-dimensional lattice in R^3. Let $\underset{\sim}{u}$, $\underset{\sim}{v}$ be linearly independent points in M, and let $\underset{\sim}{f}$ be a point of R^3 such that

(12) $$\begin{cases} F(\underset{\sim}{u}) = \alpha, \quad F(\underset{\sim}{v}) = \beta, \quad F(\underset{\sim}{f}) = \varepsilon, \\ \underset{\sim}{f} \notin \mathrm{lin}(\underset{\sim}{u},\underset{\sim}{v}). \end{cases}$$

Let k be a number such that

(13) $\qquad F(\underset{\sim}{x}) \geqslant k \quad$ for all $\underset{\sim}{x}$ in $M \sim \mathrm{lin}(\underset{\sim}{u},\underset{\sim}{v})$.

Let P be the plane

(14) $$P = \underset{\sim}{f} + \mathrm{lin}(\underset{\sim}{u},\underset{\sim}{v}).$$

Suppose that

$$M \cap P \neq \emptyset .$$

Then

(15) $$\varepsilon \geqslant \tfrac{1}{4}\left(3k-(\alpha+\beta)\right).$$

Proof. Let $\underset{\sim}{w}$ be a point in $M \cap P$. Replacing $\underset{\sim}{w}$ by $\underset{\sim}{w} - \underset{\sim}{u}$, $\underset{\sim}{w} - \underset{\sim}{v}$, or $\underset{\sim}{w} - \underset{\sim}{u} - \underset{\sim}{v}$ if necessary, and then replacing $\underset{\sim}{u}$ by $-\underset{\sim}{u}$ and/or $\underset{\sim}{v}$ by $-\underset{\sim}{v}$ if necessary, we may suppose that

$$\underset{\sim}{w} = \underset{\sim}{f} + \lambda\underset{\sim}{u} + \mu\underset{\sim}{v} ,$$

where

$$0 \leqslant \lambda \leqslant \tfrac{1}{2}, \quad 0 \leqslant \mu \leqslant \tfrac{1}{2} .$$

Applying (13) to the point $\underset{\sim}{w}$, and to the point

$$2\underset{\sim}{w} - \underset{\sim}{u} - \underset{\sim}{v} = 2\underset{\sim}{f} + (2\lambda-1)\underset{\sim}{u} + (2\mu-1)\underset{\sim}{v} ,$$

and using the convexity of F, we obtain

$$\begin{cases} \varepsilon + \lambda\alpha + \mu\beta \geqslant k \\ 2\varepsilon + (1-2\lambda)\alpha + (1-2\mu)\beta \geqslant k, \end{cases}$$

which immediately gives (15).

Theorem 7D.3. Let M be a 3-dimensional lattice in R^3, and let $\underset{\sim}{a}$, $\underset{\sim}{b}$ be minimal points as in Lemma 7D.1. Suppose (changing scale if necessary) that

(16) $$F(\underset{\sim}{a}) = 1 (= \mu_1), \quad F(\underset{\sim}{b}) = m (= \mu_2).$$

Let P be a lattice plane parallel to $\mathrm{lin}(\underset{\sim}{a},\underset{\sim}{b})$, but not through $\underset{\sim}{0}$. Then with notation as in (1), (2) (re-naming coordinates if necessary to satisfy (1)), the distance from P to $\underset{\sim}{0}$ with respect to F satisfies

(17) $$\tau \geqslant \max\left\{\frac{m-2}{2}, \frac{2-m}{4}, \frac{1}{6m}\right\}$$

Proof. Apply Lemma 7D.2 with

$$\underset{\sim}{u} = \underset{\sim}{a}, \quad \underset{\sim}{v} = \underset{\sim}{b}, \quad \underset{\sim}{f} = (t/\gamma_1, 0, 0) .$$

Since (2) states that

$$\tau = |t/\gamma_1| ,$$

the lemma gives

(18) $$\tau \geqslant \tfrac{1}{4}\big(3k-(1+m)\big).$$

By Lemma 7D.1, condition (13) of Lemma 7D.2 will be satisfied if $k = m-1$ or $k = 1$. Substituting these values of k into (18) gives the first two inequalities of (17).

The third inequality in (17) is obtained by a simple application of Minkowski's convex body theorem. The parallelepiped determined by $\underset{\sim}{a}$, $\underset{\sim}{b}$, $\underset{\sim}{f}$ contains a fundamental domain for M. Hence

$$\begin{aligned} 1.m.\tau &= F(\underset{\sim}{a})F(\underset{\sim}{b})F(\underset{\sim}{f}) \\ &\geqslant |\underset{\sim}{a}||\underset{\sim}{b}||\underset{\sim}{f}| \\ &\geqslant \det M \\ &\geqslant 2^{-3}. \text{ (Volume of } F(\underset{\sim}{x}) < 1), \text{ by Minkowski} \\ &= \frac{1}{6} , \text{ as required.} \end{aligned}$$

Note 1. The third inequality in (17), and its proof, are valid if $\underset{\sim}{a},\underset{\sim}{b}$ are arbitrary linearly independent points of M with

$$F(\underset{\sim}{a}) = 1, \quad F(\underset{\sim}{b}) = m.$$

Note 2. Application of Lemma 7D.2 to the classical successive minimal points (instead of the minimal points of §7B) and corresponding distance τ' say, yields

(19) $$\tau' \leqslant \tfrac{1}{2}\lambda_2 - \tfrac{1}{4}$$

(from (18) with τ, k, m replaced by τ', λ_2, λ_2, respectively). Thus, if the minimal point $\underset{\sim}{b}$ happened to be a second successive minimal point, then (19) would be an improvement on (17). In the case when $k = m = 1$, (17) gives

$$\tau \geqslant \tfrac{1}{4}$$

which is easily seen to be best possible, and similarly, if $\lambda_2 = 1$ $(= \lambda_1)$, then (19) gives $\tau' \geqslant \frac{1}{4}$ which is best possible.

§7E. Covering domains.

Let L be a 2-dimensional lattice in the plane P through the origin in R^3. We shall call a set $\mathcal{D}$ a covering domain for L if

(1) $$P = \bigcup_{u \in L} (\mathcal{D} + \underset{\sim}{u}),$$

that is, if L is a covering lattice of $\mathcal{D}$.

In this section, we will find covering domains $\mathcal{D}$ in which F is small, where F is as in §7A(1), and hence derive an auxiliary result needed for the index problem.

Suppose that P has equation

$$\gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 = 0 , \tag{2}$$

where

$$|\gamma_1| = \max_i |\gamma_i| > 0. \tag{3}$$

It follows that if $\underset{\sim}{x}$ is in P then

$$|x_1| = |(\gamma_2 x_2 + \gamma_3 x_3)/\gamma_1| \leqslant |x_2| + |x_3| . \tag{4}$$

As explained in §7C, P has four or six coordinate sectors. We will be concerned with the quadrants $\pm A$, $\pm B$ in P, where

$$\begin{cases} A = \{\underset{\sim}{x} | \underset{\sim}{x} \in P,\ x_2 \geqslant 0,\ x_3 \geqslant 0\} , \\ B = \{\underset{\sim}{x} | \underset{\sim}{x} \in P,\ x_2 \leqslant 0,\ x_3 \geqslant 0\} , \end{cases} \tag{5}$$

which are non-degenerate since $\gamma_1 \neq 0$.

We observe the following immediate consequence of (2), (3) and (4).

Lemma 7E.1. If $\underset{\sim}{x}$ and $\underset{\sim}{y}$ are in P, then

$$x_2 = y_2 \text{ and } |x_3| < |y_3| \Rightarrow F(\underset{\sim}{x}) < F(\underset{\sim}{y}),$$

$$x_3 = y_3 \text{ and } |x_2| < |y_2| \Rightarrow F(\underset{\sim}{x}) < F(\underset{\sim}{y}).$$

We shall use this observation in the proof of the following lemma.

Lemma 7E.2. Let $\underset{\sim}{u}$, $\underset{\sim}{v}$ be linearly independent points of L such that

$$\underset{\sim}{u}, \underset{\sim}{v} \in A, \quad \underset{\sim}{v} - \underset{\sim}{u} \in B.$$

Let

$$C = \{\underset{\sim}{x} | \underset{\sim}{x} \in P,\ -u_2 \leqslant x_2 \leqslant u_2,\ \ 0 \leqslant x_3 \leqslant v_3\} ,$$

$$\mathcal{D}_1 = C \cap A,$$

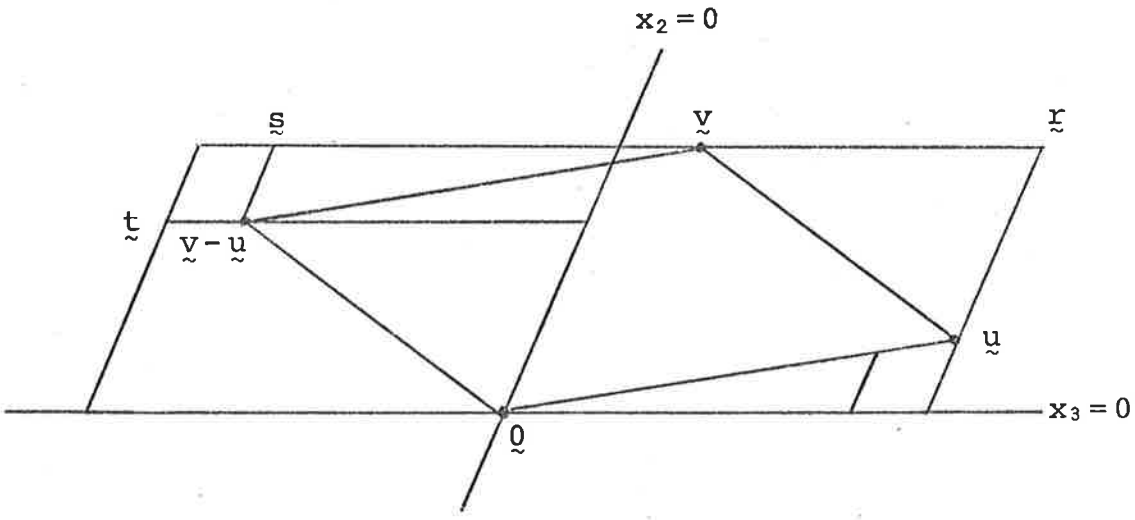
$$\mathcal{D}_2 = C \cap B \sim \{\underset{\sim}{x} | x_2 < v_2 - u_2,\ x_3 > v_3 - u_3\} .$$

Then

(i) $\mathcal{D}_1$ and $\mathcal{D}_2$ are covering domains for L, and

(ii) for all $\underset{\sim}{x}$ in $\mathcal{D}_1$ or $\mathcal{D}_2$

$$F(\underset{\sim}{x}) \leqslant \min\{F(\underset{\sim}{u}) + F(\underset{\sim}{v}),\ F(\underset{\sim}{u}) + F(\underset{\sim}{v}-\underset{\sim}{u}),\ F(\underset{\sim}{v}) + F(\underset{\sim}{v}-\underset{\sim}{u})\}\ .$$



Proof. (i) The parallelogram F with vertices $\underset{\sim}{0}$, $\underset{\sim}{u}$, $\underset{\sim}{v}$, $\underset{\sim}{v}-\underset{\sim}{u}$ is certainly a covering domain for L, and if $\underset{\sim}{x}$ is any point in F then either $\underset{\sim}{x}$ itself or $\underset{\sim}{x}+\underset{\sim}{u}$ or $\underset{\sim}{x}+\underset{\sim}{u}-\underset{\sim}{v}$ is in $\mathcal{D}_1$, as illustrated in the diagram. Hence $\mathcal{D}_1$ is a covering domain for L, and the proof for $\mathcal{D}_2$ is similar.

(ii) Let $\underset{\sim}{r}$, $\underset{\sim}{s}$, $\underset{\sim}{t}$ be the points in P such that

$$r_2 = u_2,\ r_3 = v_3;\quad s_2 = v_2 - u_2,\ s_3 = v_3;\quad t_2 = -u_2,\ t_3 = v_3 - u_3.$$

Then by Lemma 7E.1 the maximum values of F on $\mathcal{D}_1$ and $\mathcal{D}_2$ are attained at $\underset{\sim}{r}$ and $\underset{\sim}{s}$ or $\underset{\sim}{t}$. The result now follows by using Lemma 7E.1 together with the convexity of F. For example,

$$F(\underset{\sim}{r}) \leqslant F(\underset{\sim}{u}+\underset{\sim}{v}) \leqslant F(\underset{\sim}{u}) + F(\underset{\sim}{v}),$$

$$F(\underset{\sim}{r}) \leqslant F(\underset{\sim}{v}) + F(\underset{\sim}{r}-\underset{\sim}{v}) \leqslant F(\underset{\sim}{v}) + F(\underset{\sim}{u}-\underset{\sim}{v}).$$

We have the following immediate corollary.

Corollary 7E.3. Let L, P, A, B be as in Lemma 7E.2, and let $\underset{\sim}{u}$, $\underset{\sim}{v}$, $\underset{\sim}{v}-\underset{\sim}{u}$ be points of L not all lying in $\pm A$, and not all lying in $\pm B$. Let $\underset{\sim}{\delta}$ be any point of P. Then in each quadrant $\pm A$, $\pm B$ there is a point $\underset{\sim}{x}$ such that

$$\underset{\sim}{x} \in L + \underset{\sim}{\delta}, \quad F(\underset{\sim}{x}) \leqslant F(\underset{\sim}{u}) + F(\underset{\sim}{v}) \ . \tag{6}$$

We now introduce further notation in the case when there are six sectors in P as in (2), (3), that is, when $\gamma_2\gamma_3 \neq 0$. We then suppose, by mapping (x_1,x_2,x_3) to $(x_1,x_2,-x_3)$ if necessary, that

$$\gamma_2\gamma_3 > 0. \tag{7}$$

In this case the line $x_1 = 0$ in P divides the quadrant B into two sectors, B_1, B_2, say, so that

$$B = B_1 \cup B_2. \tag{8}$$

For each sector E, let F_E denote the linear function that coincides with F on E. We then have the following lemma, which is easily checked using (4).

Lemma 7E.4. Suppose the plane P given by (2) has six coordinate sectors and (3), (5), (7) and (8) hold. Then

$$F_{B_1}(\underset{\sim}{x}) > 0 \quad \text{for all } \underset{\sim}{x} \text{ in } B_2,$$

except possibly when $\underset{\sim}{x}$ lies on the common boundary of $\pm A$ and $\pm B$, and a similar result holds with B_1 and B_2 interchanged.

The lemma is equivalent to the statement that the lines in P on which F_{B_1}, F_{B_2} vanish must lie in $\pm A$, possibly on its boundary. We now use this lemma together with Corollary 7E.3 to prove the following auxiliary result for our index problem.

Theorem 7E.5. Let L be a 2-dimensional lattice in the plane P in R^3 such that (2), (3) hold. Let $\underset{\sim}{a}$, $\underset{\sim}{b}$, be

minimal points of L as defined in §7C (3), (4) and let

$$F(\underset{\sim}{a}) = 1, \quad F(\underset{\sim}{b}) = m.$$

Let A, B be the quadrants in P as defined in (5), and let $\underset{\sim}{\delta}$ be any point of P. Then in each of the four quadrants $\pm A$, $\pm B$ there is a point $\underset{\sim}{x}$ such that

$$\underset{\sim}{x} \in L + \underset{\sim}{\delta}, \quad F(\underset{\sim}{x}) \leqslant m + 1.$$

Proof. By Corollary 7E.3, it is sufficient to show that the minimal points $\underset{\sim}{a}$, $\underset{\sim}{b}$ chosen appropriately if not unique, satisfy the condition that $\underset{\sim}{a}$, $\underset{\sim}{b}$, $\underset{\sim}{b} - \underset{\sim}{a}$ do not all lie in $\pm A$, or all in $\pm B$. We will not use the fact that $m \geqslant 1$, so that $\underset{\sim}{a}$, $\underset{\sim}{b}$ are actually interchangeable in the argument.

Case 1. When $\underset{\sim}{a} \in \pm A$, and $\underset{\sim}{b} \in \pm B$ or vice versa, the only way for $\underset{\sim}{a}$, $\underset{\sim}{b}$ to lie in the same pair $\pm A$ or $\pm B$ is for $\underset{\sim}{a}$, $\underset{\sim}{b}$ to lie on the two bounding rays of one quadrant, in which case neither of $\pm (\underset{\sim}{b}-\underset{\sim}{a})$ can lie in the same quadrant.

Case 2. When Case 1 does not hold, both $\underset{\sim}{a}$ and $\underset{\sim}{b}$ must lie in $\pm B$ but in different sector pairs $\pm B_i$, and we may suppose that $\underset{\sim}{a} \in B_1$, $\underset{\sim}{b} \in B_2$. If $F_{B_2}(\underset{\sim}{a}) = 0$, we replace $\underset{\sim}{b}$ by $\underset{\sim}{b} - h\underset{\sim}{a}$ where $h \in Z$, $h \geqslant 0$ so that the new $\underset{\sim}{b}$ still satisfies $\underset{\sim}{b} \in B_2$, $F(\underset{\sim}{b}) = m$, but also $\underset{\sim}{b} - \underset{\sim}{a} \notin B_2$. If $F_{B_2}(\underset{\sim}{a}) \neq 0$, then by Lemma 7E.4, $F_{B_2}(\underset{\sim}{a}) > 0$, so $F_{B_2}(\underset{\sim}{b}-\underset{\sim}{a}) < F(\underset{\sim}{b})$, which would contradict the minimality of $F(\underset{\sim}{b})$ in $B_2 \sim \ell(\underset{\sim}{a})$ if we had $\underset{\sim}{b} - \underset{\sim}{a} \in B_2$. Hence $\underset{\sim}{b} - \underset{\sim}{a} \notin B_2$. Similarly, we may suppose that $\underset{\sim}{a} - \underset{\sim}{b} \notin B_1$, and so $\underset{\sim}{b} - \underset{\sim}{a} \notin -B_1$.

Now if $\underset{\sim}{b} - \underset{\sim}{a} \in B_1$, then $\underset{\sim}{b} = (\underset{\sim}{b}-\underset{\sim}{a}) + \underset{\sim}{a}$ would belong to B_1, so $\underset{\sim}{b}$ and hence all three of $\underset{\sim}{b}$, $\underset{\sim}{b} - \underset{\sim}{a}$, $\underset{\sim}{a}$ would lie on the common boundary of B_1 and B_2, a contradiction since $\underset{\sim}{b} \notin \ell(\underset{\sim}{a})$. Thus $\underset{\sim}{b} - \underset{\sim}{a} \notin B_1$. Similarly, $\underset{\sim}{a} - \underset{\sim}{b} \notin B_2$, so

$\underset{\sim}{b} - \underset{\sim}{a} \notin -B_2$. Thus, we have that $\underset{\sim}{b} - \underset{\sim}{a} \notin \pm B_1, \pm B_2$, so $\underset{\sim}{b} - \underset{\sim}{a} \notin B$, as required.

§7F. Bounds for the index.

Throughout this section, M is a 3-dimensional lattice in R^3 having no non-zero points on the coordinate planes, and F is as in §7A(1). Let

(1) $$X = \{\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{c},\underset{\sim}{d}\}\ , \quad X' = \{\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{c},\underset{\sim}{d}'\}$$

be as in §7B (1), (4) with

(2) $$F(\underset{\sim}{a}) = 1, \quad F(\underset{\sim}{b}) = m \geqslant 1.$$

We will use the results of §7D, 7E to obtain bounds for the indices of the sublattices $\ell(X)$ and $\ell(X')$ in M.

Let the plane

(3) $$P = \mathrm{lin}(\underset{\sim}{a},\underset{\sim}{b})$$

have equation

(4) $$\gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 = 0,$$

where

(5) $$-\gamma_1 \geqslant \gamma_2 \geqslant \gamma_3 \geqslant 0.$$

Condition (5) implies that for all $\underset{\sim}{x}$ in P

(6) $$|x_1| \leqslant \frac{\gamma_2}{-\gamma_1 + \gamma_2}\ F(\underset{\sim}{x}) \leqslant \tfrac{1}{2}\, F(\underset{\sim}{x}).$$

Since (5) also implies that (4) of §7E holds, we will be able to use the results obtained there on the quadrants $\pm A$, $\pm B$ of P (bounded by the lines $x_2 = 0$, $x_3 = 0$ in P). Using the notation A_{+++}, etc., for the octants as at the beginning of §7A, we note that

(7) $$P \cap A_{-++} = \{\underset{\sim}{0}\}\ ,$$

which follows from (5). Since M has no point on the coordinate planes, we have $\gamma_2 \neq 0$. But γ_3 may be zero, in which case we also have $P \cap A_{-+-} = \{\underset{\sim}{0}\}$.

For $\xi \in R$, let $P(\xi)$ denote the plane

(8) $$P(\xi) = (\xi,0,0) + P,$$

and recall that by §7D (1), (2),

(9) $$\underset{\sim}{u} \in P(\xi) \Rightarrow F(\underset{\sim}{u}) \geqslant |\xi|.$$

We define

(10) $$\tau_0 = \min\{|\xi| \mid \xi \neq 0, P(\xi) \text{ contains a point of } M\}.$$

Thus, by §7D (1), (2), τ_0 is the distance with respect to F between adjacent lattice planes parallel to P. We recall that by Theorem 7D.3

(11) $$\tau_0 \geqslant \max\left\{\frac{m-2}{2}, \frac{2-m}{4}, \frac{1}{6m}\right\}.$$

We shall consider the index of $\ell(\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{u})$ in M for certain vectors $\underset{\sim}{u}$. By Theorem 7C.2 we have

$$\ell(\underset{\sim}{a},\underset{\sim}{b}) = M \cap P,$$

and it therefore follows that, for any $\underset{\sim}{u}$ in M and integral h,

(12) $$\underset{\sim}{u} \in M \cap P(h\tau_0) \Leftrightarrow \left(M:\ell(\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{u})\right) = |h|.$$

The following lemma produces special points in $M \cap P(h\tau_0)$ for suitable h.

Lemma 7F.1. Let M, $\underset{\sim}{a},\underset{\sim}{b}$, P, τ_0 be as above, satisfying (2) to (5) and (10), and let $h \in Z$.

(i) If

(13) $$h\tau_0 \geqslant \frac{m+1}{2}$$

then each of the octants A_{+++}, A_{+-+}, A_{+--}, A_{++-} contains a point $\underset{\sim}{w}$ such that

(14) $$\begin{cases} \underset{\sim}{w} \in M \cap P(h\tau_0), \\ F(\underset{\sim}{w}) \leqslant h\tau_0 + m + 1. \end{cases}$$

(ii) If $h \geqslant 0$ the octant A_{+++} contains a point $\underset{\sim}{w}$ such that

$$\underset{\sim}{w} = \underset{\sim}{v} + (h\tau_0, 0, 0) \in M \cap P(h\tau_0),$$

where

$$\underset{\sim}{v} \in A_{+++}, \quad F(\underset{\sim}{v}) \leqslant m + 1.$$

Proof. Let $\underset{\sim}{u}$ be any point of $M \cap P(h\tau_0)$, and let

$$\underset{\sim}{\delta} = \underset{\sim}{u} - (h\tau_0, 0, 0).$$

By applying Theorem 7E.5 with this $\underset{\sim}{\delta}$, we see that each quadrant E $(= \pm A, \pm B)$ contains a point $\underset{\sim}{v}$ such that

$$\underset{\sim}{v} \in E, \quad F(\underset{\sim}{v}) \leqslant m + 1,$$

$$\underset{\sim}{w} = \underset{\sim}{v} + (h\tau_0, 0, 0) \in M \cap P(h\tau_0).$$

Since, by (6), each $\underset{\sim}{v}$ has $|v_1| \leqslant \frac{1}{2}(m+1)$, the condition (13) ensures that each $\underset{\sim}{w}$ has $w_1 \geqslant 0$, and this gives the points required in (i). In the case of the octant A_{+++}, (ii) follows from the above and (7).

We can now give a result on indices.

Lemma 7F.2. Let M, $\underset{\sim}{a}$, $\underset{\sim}{b}$, P, τ_0 be as above, satisfying (2) to (5) and (10). Let G be any octant in R^3, and let $\underset{\sim}{u}$ be a point of $G \sim P$ with minimal $F(\underset{\sim}{x})$. Then

(i) $$\big(M{:}\,\ell(\underset{\sim}{a}, \underset{\sim}{b}, \underset{\sim}{u})\big) < \frac{3}{2}\,\frac{m+1}{\tau_0} + 1,$$

and

(ii) if $G = \pm A_{+++}$, then

$$\big(M{:}\,\ell(\underset{\sim}{a}, \underset{\sim}{b}, \underset{\sim}{u})\big) \leqslant \frac{m+1}{\tau_0} + 1.$$

Proof. Without loss of generality, we may suppose that G is one of the octants A_{+++}, A_{++-}, A_{+--}, A_{+-+}. Apply Lemma 7F.1 with h as the least integer greater than or equal to $(m+1)/2\tau_0$, so that

(15) $$1 \leqslant h < \frac{m+1}{2\tau_0} + 1,$$

to obtain a point $\underset{\sim}{w}$ in the same octant as $\underset{\sim}{u}$ and satisfying (14), so that $\underset{\sim}{w} \in M \sim P$, and

(16) $$F(\underset{\sim}{w}) \leqslant h\tau_0 + (m+1).$$

Let k be the required index, so that, by (12),

$$\underset{\sim}{u} \in M \cap P(k\tau_0),$$

hence by (9), we have

(17) $$F(\underset{\sim}{u}) \geqslant k\tau_0.$$

Suppose that (i) does not hold. Then

$$k\tau_0 \geqslant \frac{m+1}{2} + \tau_0 + (m+1) > 0,$$

so by (15), (17),

$$F(\underset{\sim}{u}) > h\tau_0 + (m+1), \quad \underset{\sim}{u} \notin P.$$

By (16), this contradicts the minimality of $\underset{\sim}{u}$. Thus (i) holds. The result (ii) is proved similarly, using (ii) of Lemma 7F.1 with $h = 1$.

We now combine our results to give our final index estimates.

Theorem 7F.3. Let M be a 3-dimensional lattice in R^3 having no non-zero points on the coordinate planes. Let X, X' be as in (1) above and in §7B (1), (4), and let

$$m = F(\underset{\sim}{b})/F(\underset{\sim}{a}) \geqslant 1,$$

$$\omega = \max\left\{\frac{m-2}{2}, \frac{2-m}{4}, \frac{1}{6m}\right\}.$$

(i) Each of the indices $(M:\ell(X))$, $(M:\ell(X'))$ is positive and less than

$$\frac{3}{2}\ \frac{m+1}{\omega} + 1.$$

(ii) Let P be as in (3), (4), (5). If neither $\underset{\sim}{a}$ nor $\underset{\sim}{b}$ lies in the quadrant

$$A = \{\underset{\sim}{x} \mid \underset{\sim}{x} \in P,\ x_2 \geqslant 0,\ x_3 \geqslant 0\},$$

then

$$(M:\ell(X')) \leqslant \frac{m+1}{\omega} + 1.$$

Corollary 7F.4. For all m, the indices in (i) are at most 62 and if $m \geqslant 11$ these indices are at most 4. For all

m, the index in (ii) is at most 41 and if $m \geqslant 8$ this index is at most 3.

Note. The above bounds are rather crude for small m. It should be possible to drastically improve them by obtaining an improved version of (17) of Theorem 7D.3, in particular for the case when m is near 2.

Proof of Theorem 7F.3. The indices $(M:\ell(X))$ and $(M:\ell(X'))$ are just the greatest common divisors of $(M:\ell(\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{c}))$ with $(M:\ell(\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{d}))$ and $(M:\ell(a,b,d'))$ respectively. Hence to to prove (i) it is sufficient to prove that

$$(18) \qquad 0 < (M:\ell(\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{u})) < \frac{3}{2}\ \frac{m+1}{\omega} + 1$$

holds when $\underset{\sim}{u} = \underset{\sim}{c}$, and to prove (ii) it is sufficient to prove that under the assumptions on $\underset{\sim}{a}$, $\underset{\sim}{b}$ in (ii), we have

$$(19) \qquad (M:\ell(\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{u})) < \frac{m+1}{\omega} + 1$$

when $\underset{\sim}{u}$ is one of $\underset{\sim}{c}$, $\underset{\sim}{d}'$. Now if $\underset{\sim}{u}$ is any of the points $\pm \underset{\sim}{c}$, $\pm \underset{\sim}{d}'$ and G is the octant containing $\underset{\sim}{u}$, then $\underset{\sim}{u}$ is the minimal point of $G \sim \mathrm{lin}(\underset{\sim}{a},\underset{\sim}{b})$. Assume without loss of generality that P is as in (ii). Then on applying Lemma 7F.2(i) and using (11), we obtain (18), in particular with $\underset{\sim}{u} = \underset{\sim}{c}$.

If neither of $\underset{\sim}{a}$, $\underset{\sim}{b}$ lies in A, then $\pm a$, $\pm b$ must lie in A_{++-}, A_{+-+} (or vice versa) and one of $\pm \underset{\sim}{c}$, $\pm \underset{\sim}{d}'$ must lie in A_{+++}. Taking this point as $\underset{\sim}{u}$ in Lemma 7F.2 (ii), and using (11), we obtain (19). This concludes the proof of the theorem.

BIBLIOGRAPHY

[1] Bantegnie, R., "A propos d'un problème de Mordell sur les octaedres latticiels", J. London Math. Soc., 37 (1962), 320-328.

[2] Brown, H. and Mahler, K., "A generalization of Farey sequences: Some exploration via the computer", J. Number Theory 3 (1971), 364-370.

[3] Cassels, J.W.S., "An introduction to the geometry of numbers", Springer-Verlag, Berlin-Göttingen-Heidelberg, 1959.

[4] Gaskell, R.W., Klamkin, M.S. and Watson, P., "Triangulations and Pick's Theorem", Math. Mag. 49 (1976), 35-37.

[5] Grünbaum, B., "Convex polytopes", Interscience, London, 1967.

[6] Hardy, G.H. and Wright, E.M., "An introduction to the theory of numbers", Clarendon Press, Oxford, 1960.

[7] Lang, S., "Introduction to diophantine approximations", Addison-Wesley, Reading, Mass., 1966.

[8] Lekkerkerker, Cornelis Gerrit, "Geometry of numbers", Wolters-Noordhoff, Groningen; North-Holland, Amsterdam-London, 1969.

[9] Low, L., "A problem of Schinzel on lattice points", Acta Arith. 31 (1976), 385-388.

[10] Mordell, L.J., "Lattice octahedra", Canad. J. Math. 12 (1960), 297-302.

[11] Schinzel, A., "On the reducibility of polynomials and in particular of trinomials", Acta Arith. 11 (1965), 1-34.

[12] Schmidt, W.M., "A problem of Schinzel on lattice points", Acta Arith. 15 (1969), 199-203.