On Voronoi's Method of Reducing Positive Definite Quadratic Forms

By

Andrew William Harrell

A.B. (Vanderbilt University) 1970

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Mathematics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:
Edwin
Eyl H. W
Andrew P. Ogg

Committee in Charge

On Voronoi's Method of Reducing Positive Quadratic Forms

Abstract by

Andrew William Harrell

A polyhedron is called a <u>parallelohedron</u> if its parallel translates by the vectors of a lattice just fill up the space it lies in. Centered at the center of symmetry of each such body there is a largest inscribed sphere. In this way we associate a sphere packing with each parallelohedron.

Given a positive definite quadratic form we diagonalize such a form by writing its matrix in the form ${}^{t}PP$ where $P = (\xi_1 \cdots \xi_n)$. Then, the translates of the sphere defined by $\varphi(x) = \xi_1^2 + \cdots + \xi_n^2 \leq M/4$, M = the minimum of φ for integral variable values, by the vectors of the lattice $\xi_1, \cdots \xi_n$ form a sphere packing. The ratio of the volume of the spheres to that of the total space for this packing is just $4\pi \quad (M/2)^3/3\sqrt[2]{D}$ where D is the discriminant of φ. Minkowski [1] proved that no form in dimension 3 not equivalent to that associated with the rhombic dodecahedron gives as large a value as $\pi \ /\sqrt{18}$ for the density of its packing. The word equivalent means equal to a unimodular integral transformation of the original form. We consider such equivalent forms as defining the same "type" of parallelohedron since the lattice remains unchanged.

Voronoi [2] introduced two methods of reducing positive definite quadratic forms. The first method utilized the inner product on the corresponding space of symmetric matrices. The second method involved what he called <u>primitive parallelohedra</u>. A parallelohedron is called a primitive parallelohedron if 1) each vertex belongs to exactly one more edge of the whole collection of translates than the dimension of the space 2) each face coincides with exactly one face of the other bodies. Two primitive parallelohedra are said to be of the same type if they determine the same lattice. Voronoi's idea involved associating a positive definite quadratic form with each primitive parallelohedron. This is done so that the body is the set of points closer to the origin than any other lattice point in the metric defined by the form. Then, generalizing Dirichlet's method of reduction [3], he subdivided the space $\mathcal{P}(n)$ of all positive definite quadratic forms into cones which correspond to a given type of primitive parallelohedron. One such cone is singled out as being the cone of reduced forms.

In short, this thesis is concerned with
1) a clarified presentation of the relation between Voronoi's two methods of reduction in the case $n = 3$.
2) use of this relation and a criterion of Voronoi to prove Minkowski's result mentioned above
3) a geometric proof of the relation between Voronoi's and Minkowski's method of reduction in the case $n = 2$
4) a geometric enumeration of the special automorphisms associated with fixed point stabilizer groups acting under arithmetic equivalence on the boundary of the reduced domain
5) a short discussion of the application of 4) to the problem of desingulariztion of the compactified domain associated with Siegel's modular functions of degree 2 and 3.

[1] Minkowski, H.:"Dichteste Gitterformige Lagerung Kongruenten Korper". Werke, Bd. II S. 1.

[2] Voronoi, G.F.:"Nouvelles Applications des Parametres Continus a ls Theorie des Formes Quadratiques".

On Voronoi's Method of Reducing Positive Quadratic Forms

Abstract by

Andrew William Harrell

A polyhedron is called a parallelohedron if its parallel translates by the vectors of a lattice just fill up the space it lies in. Centered at the center of symmetry of each such body there is a largest inscribed sphere. In this way we associate a sphere packing with each parallelohedron.

Given a positive definite quadratic form we diagonalize such a form by writing its matrix in the form ${}^{t}PP$ where $P = (\xi_1 \cdots \xi_n)$. Then, the translates of the sphere defined by $\varphi(x) = \xi_1^2 + \cdots + \xi_n^2 \leq M/4$, M = the minimum of φ for integral variable values, by the vectors of the lattice $\xi_1, \cdots \xi_n$ form a sphere packing. The ratio of the volume of the spheres to that of the total space for this packing is just $4\pi \ (M/2)^3/3\sqrt[2]{D}$ where D is the discriminant of φ. Minkowski [1] proved that no form in dimension 3 not equivalent to that associated with the rhombic dodecahedron gives as large a value as $\pi/\sqrt{18}$ for the density of its packing. The word equivalent means equal to a unimodular integral transformation of the original form. We consider such equivalent forms as defining the same "type" of parallelohedron since the lattice remains unchanged.

Voronoi [2] introduced two methods of reducing positive definite quadratic forms. The first method utilized the inner product on the corresponding space of symmetric matrices. The second method involved what he called primitive parallelohedra. A parallelohedron is called a primitive parallelohedron if 1) each vertex belongs to exactly one more edge of the whole collection of translates than the dimension of the space 2) each face coincides with exactly one face

of the other bodies. Two primitive parallelohedra are said to be of the same type if they determine the same lattice. Voronoi's idea involved associating a positive definite quadratic form with each primitive parallelohedron. This is done so that the body is the set of points closer to the origin than any other lattice point in the metric defined by the form. Then, generalizing Dirichlet's method of reduction [3], he subdivided the space $\mathcal{P}(n)$ of all positive definite quadratic forms into cones which correspond to a given type of primitive parallelohedron. One such cone is singled out as being the cone of reduced forms.

In short, this thesis is concerned with

1) a clarified presentation of the relation between Voronoi's two methods of reduction in the case n = 3.

2) use of this relation and a criterion of Voronoi to prove Minkowski's result mentioned above

3) a geometric proof of the relation between Voronoi's and Minkowski's method of reduction in the case n = 2

4) a geometric enumeration of the special automorphisms associated with fixed point stabilizer groups acting under arithmetic equivalence on the boundary of the reduced domain

5) a short discussion of the application of 4) to the problem of desingulariztion of the compactified domain associated with Siegel's modular functions of degree 2 and 3.

[1] Minkowski, H.:"Dichteste Gitterformige Lagerung Kongruenten Korper". Werke, Bd. II S. 1.

[2] Voronoi, G.F.:"Nouvelles Applications des Parametres Continus a ls Theorie des Formes Quadratiques".

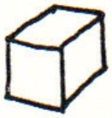
Crelles Journal 133, 134, 136, 1907, 1908 (in three parts).

[3] Dirichlet, L.:"Ueber die Reduction der Positiven Quadratischen Formen mit Drei Unbestimmter Ganzen Zahlen". Werke, Bd. II, S. 41.

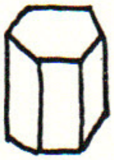
TABLE OF CONTENTS

INTRODUCTION

The theory of positive definite quadratic forms is closely involved with the study of space coverings by "parallelohedra". A polyhedron is called a *parallelohedron* if its parallel translates by the vectors of a lattice just fill up the total space. Fedorov used parallelohedra in his enumeration of crystallographic space groups[6]. He showed that there were only five types in dimension three.

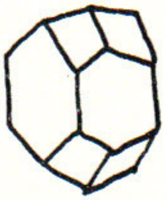


cube

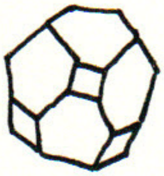


hexagonal prism

rhombic dodecahedron

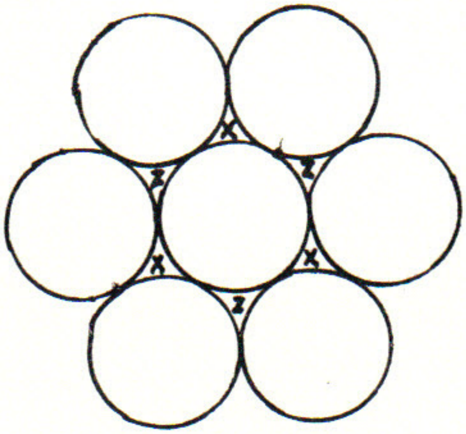


elongated dodecahedron

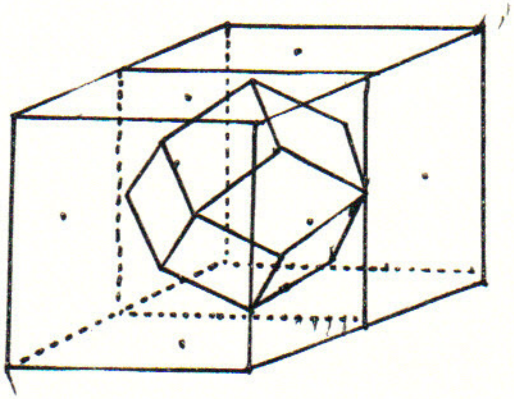


truncated octahedron

A shorter proof of this by Delone is presented in Alexandrov's book[1]. Shortly thereafter, Lord Kelvin encountered the rhombic dodecahedron in his study of closest sphere packings[17]. In two dimensions the closest arrangement of circles corresponds to a hexagonal lattice. Call this lattice Y.



A close-packed three dimensional arrangement can be formed in two ways from Y. A copy of Y with spheres in place of the circles may be laid on top of the 1^{st} layer such that either the spaces marked X or the spaces marked Z are covered. Call the layers formed respectively X and Z. Then, the pattern XYZXYZ of layers leads to the face-centered lattice which is found, for example, in the metals gold, aluminum, and copper. In this example the rhombic dodecahedron arises as the set of all points which are closer to the origin than any other lattice point.

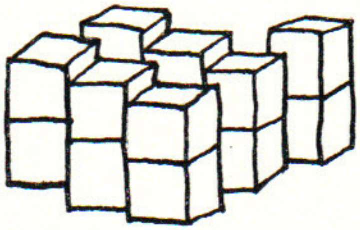


By using the points neighboring the origin as vertices another body can be formed. Call this body the *dual extremal body*. It is the cubo-octahedron. If we choose the second option XYXYXY, we get the hexagonal close-packed arrangement which occurs in the metal cobalt. The dual extremal body in this case does not have parallel sides. And it is easy to see that the centers of the spheres no longer form a lattice.

cubo-octahedron

Minkowski[12]introduced positive definite quadratic forms into these matters. If we diagonalize such a form by writing its matrix in the form ${}^tPP$ where $P = (\xi_1 \cdots \xi_n)$, the translates of the sphere defined by $\varphi(x) = \xi_1^2 + \xi_2^2 + \xi_3^2 \leq M/4$, M = minimum of φ for integral variable values, by the vectors of the lattice with basis $\xi_1 \cdots \xi_n$ form a packing. The ratio of the volume of the spheres to that of the total space for this packing is just $4\pi(M/2)^3/3\sqrt{D}$ where D is the discriminant of φ . The problem of maximizing this is equivalent with that of maximizing $M/\sqrt[3]{D}$. But, Gauss already solved this problem in [7] a review of Seeber's book on the arithmetic reduction of positive definite ternary forms. He proved Seeber's conjecture that we always have $M \leq \sqrt[3]{2D}$. Equality occurs, for instance, in the form $x^2 + y^2 + z^2 + xy + xz + yz$. In this way it was known that the lattice corresponding to the rhombic dodecahedron gives the value $\pi/\sqrt{18}$ for the densest packing of spheres. Minkowski proved that no other form not equivalent by a unimodular integral transformation gives this value. We consider such equivalent forms as defining the same "type" of parallelohedra since the lattice remains unchanged.

In 1908 Voronoi published his work [16] which forms the basis of the following thesis. He introduced two new methods of reducing positive definite quadratic forms. The first method utilized the inner product on the corresponding space of symmetric matrices. The second method involved what he called primitive parallelohedra. A parallelohedron is called a primitive parallelohedron if 1) each vertex belongs to exactly n+1 edges of the whole collection of translates 2) each face coincides with exactly one face of the other bodies. For example, the arrangement below is excluded because a face coincides with more than one other face.



This definition is equivalent to saying that the vertices determine a simplicial subdivision of the space in the manner specified in § 4. Two primitive parallelohedra are said to be of the same type if they determine the same set of simplices. Of the five types of parallelohedra mentioned above, only the truncated octahedron is primitive.

Voronoi's idea involved associating a positive definite quadratic form with each primitive parallelohedron. This is done so that the body is the set of points closer to the origin than any other lattice point in the metric defined by the form. Then, generalizing Dirichlet's method of reduction [5], he subdivided the space $\mathcal{P}(n)$ of all positive definite quadratic forms into cones which correspond to a given type of primitive parallelohedron. We will see that these cones and their images under SL(n,Z) cover the space $\mathcal{P}(n)$ with the exception of their boundary points which correspond to imprimitive parallelohedra. From this it follows that the number of inequivalent cones is the same as the number of types of primitive parallelohedra. The five different types of parallelohedra will be interpreted as limiting cases on the boundary of the "fundamental domain" associated with the truncated octahedron. In order to examine what characterizes the denseness of the sphere packings associated with these parallelohedra, we use the criterion [16] Voronoi showed determines if a form assumes the maximum value of $M / \sqrt[3]{D}$.

By means of a geometric interpretation of the first method of reduction, we shall see the relation of these methods of reduction with Minkowski's method [13] in the case $n = 2$ and some special cases for $n = 3$. The final section mentions an interpretation of this to the desingularizations of the Satake compactification of the Siegel modular variety. The papers of Baily and Borel [2], Igusa [10], Satake [14] [15], Gottschling [8] [9].

witness the large amount of work already done on this problem.

I wish to thank Professor I. Satake for his valuable comments on the exposition of this work.

§1 Voronoi's First Method of Reduction

We denote by $\mathcal{S}(n)$ the space of all real symmetric matrices of degree n. We define an inner product in $\mathcal{S}(n)$ by

$$(A,B) = \mathrm{tr}(AB) = \sum a_{ij}b_{ij} = \sum a_{ij}b_{ji}$$

for $A,B \in \mathcal{S}(n)$. Let $\mathcal{P}(n)$ denote the set of all positive definite symmetric matrices of degree n. Then $\mathcal{P}(n)$ is an open convex cone in $\mathcal{S}(n)$, and is self-dual with respect to the above inner product. Moreover, $\mathcal{P}(n)$ is homogenous under the action $A \mapsto {}^tUAU$ of $GL(n,R)$. With respect to the action of $GL(n,Z)$ on the cone we define a fundamental cone F, using the quadratic form

$$\varphi_0 = x_1{}^2 + x_2{}^2 + \dots + x_1x_2 + x_1x_3 + \dots + x_{n-1}x_n$$

$$A_o = \begin{pmatrix} 1 & \frac{1}{2} & \dots & \frac{1}{2} \\ \frac{1}{2} & 1 & \dots & \frac{1}{2} \\ \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ \frac{1}{2} & \frac{1}{2} & \dots & 1 \end{pmatrix}$$

F is defined to be made up of the matrices Y in $\overline{\mathcal{P}}(n)$ which satisfy the condition

$$(A_o, {}^tUYU) - (A_o, Y) \geq 0 \qquad \text{for all } U \text{ in } GL(n,Z).$$

In order to study F we will define another cone C. The definition of C makes use of the $n(n+1)/2$ vectors which represent the minimum of φ_0. That is, those vectors $u \in Z^n$ such that ${}^tu\,\varphi_0\,u = 1$ and the first non-zero component of u is equal to 1. These vectors are

$$u_1 = (1,0,\dots,0),\ \dots\ ,\ u_{n+1} = (1,-1,0,\dots)\ \dots$$

C is defined to be the set of matrices which correspond to those forms which can be expressed as

$$\sum_{k=1}^{n(n+1)/2} \rho_k (u_k \cdot x)^2\ , \text{ where } \rho_k \text{ is a non-negative real number.}$$

For Y in C we have $(A,Y) - (A_o,Y) =$

$$\Sigma\ \rho_k\ ({}^t u_k A u_k - {}^t u_k A_o u_k) = \Sigma\ \rho_k\ ({}^t u_k A u_k - 1\ )\ .$$

If we assume that A is a positive definite matrix whose corresponding form has half-integral coordinates, then for every point Y in C the above sum is a non-negative integer. In particular, all A of the form ${}^t U A_o U$, where U is in GL(n,Z), satisfy these assumptions. So we see that $C \subsetneq F$.

If $\Sigma\ \rho_k\ (u_k\ x)^2 \longleftrightarrow \left(a_{ij}\right)$, then we must have

$$\begin{aligned} \rho_k &= a_{1k} + a_{2k} + \dots + a_{nk}\ ,\ k = 1,\ \dots\ ,\ n\ , \\ \rho_{ij} &= -a_{ij},\quad k \quad n \quad i \neq j\ . \end{aligned}$$

We now define the normal coordinates of Y in terms of a symmetric matrix of degree n+1. The new matrix has coordinates with $1 \leq i,j \leq n+1$ that are the same as those of Y. The new coordinates Y_{in+1} satisfy the equations

$$\sum_{j=1}^{n+1} Y_{ij} = 0 \qquad i = 1,\ \dots\ ,\ n+1.$$

We call Y_{ij} $(1 \leq i < j \leq n+1)$ the <u>normal coordinates</u> of Y.

These coordinates allow us to write the matrices which belong to C in one simple expression. Let E_{ij} be the matrix of degree n whose ij^{th} normal coordinate is -1, and all of whose other normal coordinates are zero. Inotherwords,

$$\underset{j \neq n+1}{E_{ij}} = \begin{array}{cc} \begin{matrix} i & j \end{matrix} & \\ \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} & \begin{matrix} i \\ j \end{matrix} \end{array} \qquad E_{in+1} = \begin{array}{cc} i & \\ \begin{pmatrix} & & \\ & 1 & \\ & & \end{pmatrix} & i \end{array}$$

with zeroes elsehere.

Using these matrices, we have

$$C = \sum_{1 \leq i < j \leq n+1} R_+ E_{ij} \quad ,$$

where R_+ denotes the non-negative real numbers.

It is known that F = C for n = 3. For a proof of this and other basic facts on F and C we refer the reader to the paper by Igusa [10].

§2 A Theorem of Dirichlet

Theorem 1) Let $ax^2 + 2bxy + cy^2$ be a real positive definite quadratic form. Then, there exist three pairs (ℓ,m), (ℓ',m'), $(\ell'',m'') \in Z^2$ such that for any $(\alpha,\beta) \in R^2$ the inequality

$$(1) \quad ax^2 + 2bxy + cy^2 \pm 2(\alpha x + \beta y) \geq 0$$

holds for all $(x,y) \in Z^2$ if and only if

$$(2) \quad a\ell^2 + 2b\ell m + cm^2 \pm 2(\alpha\ell + \beta m) \geq 0$$

and similarly for (ℓ',m') and (ℓ'',m'').

2) Let H be the hexagon in the (α,β) plane defined by the inequalities (2). Then, the translates of H cover the plane without overlapping.

3) If the form is "reduced", that is, $a-b>0$ $b<0$, $c-b>0$, the (ℓ,m), (ℓ',m'), (ℓ'',m'') may be taken to be $(1,0)$, $(0,1)$, $(1,1)$.

proof)[1] Let $g(x,y) = ax^2 + 2bxy + cy^2$. Define new variables (u_1,u_2) by the equations $-\alpha = au_1 + bu_2$, $-\beta = bu_1 + cu_2$. Then, (1) can be written as

$$2(x(au_1 + bu_2) + y(bu_1 + cu_2)) \leq g(x,y)$$

or $g(u_1,u_2) \leq g(u_1 - x, u_2 - y)$ for all $(x,y) \in Z^2$.

The set in the (u_1,u_2) plane defined by

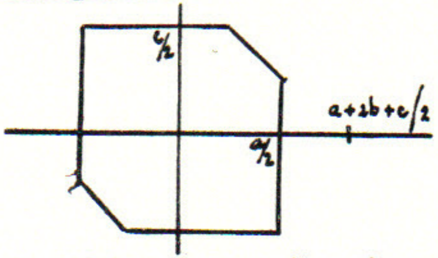
---

[1] The proof is adapted from Cassels [3] to this context. See also Dirichlet [5].

these inequalities is the set of points closer to the origin than to any other lattice point with respect to the metric $g(x,y)^{\frac{1}{2}}$. We call the corresponding set in the (α,β)-plane V. In particular, we have

$$2|\alpha| < a \quad x = \pm 1\,,\ y = 0 \qquad 2|\beta| < c \quad x = 0,\ y = \pm 1$$
$$2|\alpha + \beta| < a + 2b + c \quad x = 1,\ y = 1 \quad x = -1,\ y = -1.$$

First, supposing the form to be "reduced", these inequalities define a figure of area = $(ac - b^2)$.

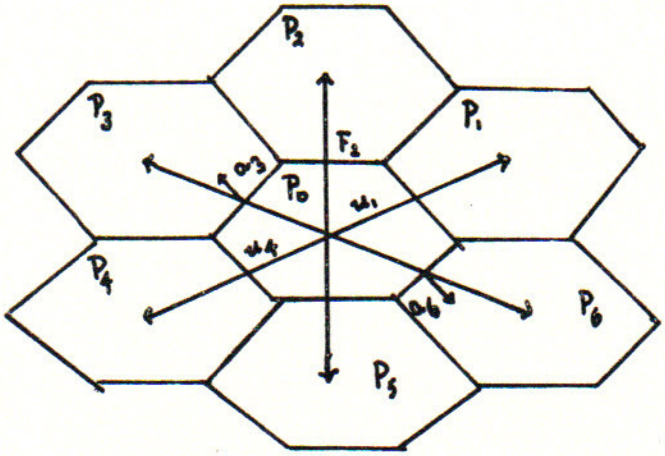


Thus, in this case, the inequalities (2) define a hexagon of area 1 in the new variables (u_1,u_2). H is contained in V. If V were of larger area than H, expanding it by a factor of 2 would give a convex figure containing no lattice points $\neq (0,0)$ and of area larger than 4. This is in contradiction to a well-known theorem of Minkowski [11]. Provided we take into account that both sets are closed, we know that H = V.

In the general case we transform $g(x,y)$ by an integral unimodular transformation into a reduced form.

Remark Note that if c = 0 or b = 0 or a = b the hexagon degenerates into a parallelogram.

§ 3 The Form Which Corresponds to a Given Primitive Parallelohedron [11][16]



We start with a bounded primitive parallelohedron P_0 in n-dimensional Euclidean space with center at the origin. Recall from the Introduction that a primitive parallelohedron is a polyhedron which satisfies the conditions:

1) Its translates by the vectors of a lattice just fill up the space it lies in

2) Each of its vertices meets n+1 edges of adjacent bodies

3) Each face coincides exactly with one face of the translated bodies.

Let the bodies adjacent to P_0 be numbered P_i, and let the face between P_0 and P_i be F_i. The equation which defines the half-space determined by F_i and containing P_i is of the form $a_i \cdot (x - c) < 0$ where c is any point on F_i. An example in two dimensions is shown in the figure above. The u_i denote the vectors from the origin to the centers of the P_i. It is