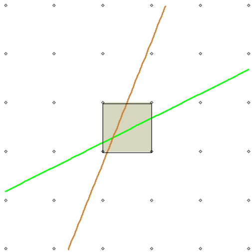
Resurrecting the divided cell algorithm for inhomogeneous Diophantine approximation

by

Richard T. Bumby and Mary E. Flahive

January 14, 2006

A divided cell



The bold lines are the lines where the given expression is zero. The cell is **divided** by having one vertex in each quadrant bounded by the lines.

Two coordinate systems

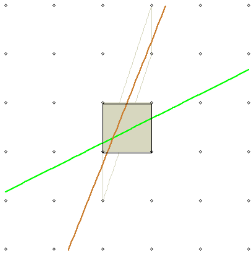
The picture shows two coordinate systems related by an **affine** transformation. One is the basis of the integer lattice; the other is a basis whose **axes** are the lines. If the equations of the lines are $a_i x + b_i y + c_i = 0$ for $i = 0$ (crossing the first generator) and $i = 1$ (crossing the second generator), the matrix

$$M = \begin{bmatrix} a_0 & b_0 & c_0 \\ a_1 & b_1 & c_1 \\ 0 & 0 & 1 \end{bmatrix}$$

defines the relation between the coordinate systems.

The divided cell step

The **next** divided cell in a **chain** is formed by extending the vertical sides until they cross the other line.



The lattice basis

The basis of the lattice has changed, so the fundamental domain is no longer a square. To put the cell in the position of the original picture, both the **basis** of lattice and the location of origin as the **lower left corner** of fundamental domain (where the expressions defining the lines are negative when $a_0 > 0$ and $b_1 > 0$) must be changed. This includes an arbitrary choice of **base vertex** for the cell. It will require that a cell be distinguished from its reflection. Such a choice was usually **implicit** in classical work, but we now know that it is safer to describe the automorphisms that identify equivalent objects before ignoring them.

The step matrix

When the lines are expressed in terms of the new lattice basis, the matrix M is right multiplied by the matrix of an affine transformation from the plane with the new basis to the plane with the old basis. The third column is the old name for the new origin and the first two columns are the old names for the directions crossing the given lines (in the given order). This matrix has one of two forms depending on the sign of the slope of b_0/a_0 (which is the sign of b_0 since $a_0 > 0$, by convention).

Positive slope

The entries in the first row have opposite signs and the transition matrix has the form

$$\begin{bmatrix} 0 & 1 & 0 \\ -1 & n & -k \\ 0 & 0 & 1 \end{bmatrix}$$

with $k \geq 0$.

Negative slope

The entries in the first row have the same sign and the transition matrix has the form

$$\begin{bmatrix} 0 & -1 & 1 \\ 1 & n & -k \\ 0 & 0 & 1 \end{bmatrix}$$

with $k \geq 0$.

Three coordinate systems

There is a third coordinate system: the one used for **viewing images**. Thus, one has **two** changes of coordinates.

$$\text{lattice} \longrightarrow \text{lines} \longrightarrow \text{view}$$

It would simplify the presentation if the lines $\longrightarrow$ view map were always a simple (positive) rescaling of axes. This convention will be used in the remainder of this presentation.

The relative approach

Instead of requiring that a scaling of the axes preserve the **size** of the lattice (as measured, for example, by the area of the fundamental domain), one could use an **arbitrary** scaling, and divide any values found by an appropriate measure of size to obtain something invariant under scaling. This approach is now standard in the study of the (homogeneous) **Markoff Spectrum**, where it is also customary to invert the values of forms to get quantities that are more easily identified in the continued fraction.

Other views

If a product of two linear forms is found to be a multiple of an indefinite quadratic form with integer coefficients, the algebra can be emphasized by considering only forms with relatively prime integer coefficients. In this case, the **discriminant** of the form is usually preserved. This involves the use of scalings so that reductions are represented by ideals in fixed order.

What killed the divided cell algorithm?

- The **Markoff Spectrum** had not yet been studied systematically;

- the main motivating problem — the Euclidean algorithm (for the norm) in quadratic fields — was settled;

- ...

. . . and calculation

The **Brunviga** was used for numerical work.

The chain of divided cells contains all minima

We are now considering the expression xy on a lattice. The points in each quadrant that are farther from **both** lines can be excluded from consideration in a search for the minima of $|xy|$. Furthermore, there are no lattice points in a strip bounded by two parallel sides of a divided cell, and the divided cell step includes a vertex on each of the **next nearest** lines in those directions. Repeating this observation in both directions on the chain shows that only the vertices of the divided cells need be considered in a search for minima.

The cell and the box

If we have a fundamental parallelogram of the lattice, we can ask which **translations** of that figure can be a divided cell. The set of all possible **origins** of the axis coordinate system form a rectangle which we call the **inner box** of the cell.

In the axis coordinate system, it is the **columns** of matrix M (introduced on an early slide) that are significant. The third column (with entries c_i) is the location of the base vertex and the other two columns are vectors generating the lattice.

Reduced forms

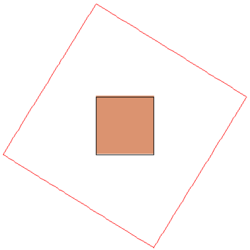
The lattice bases for which this box is nonempty were called I-reduced by Barnes. He showed that there are only finitely many I-reduced bases belonging to ideals of quadratic orders with fixed discriminant. This reduces the problem of finding the behavior of all inhomogeneous problems involving the same quadratic part **with rational coefficients** to the study of a **finite set of cells** and the relations between them determined by the divided cell step.

Characterizing reduced forms

For the vertices of the cell to lie in different quadrants with the base vertex in the third quadrant requires that the entries of M satisfy: $c_0 \leq 0$, $c_0 + a_0 \geq 0$, $c_0 + b_0 \leq 0$, $c_0 + a_0 + b_0 \geq 0$. Existence of a suitable c_0 requires $a_0 \geq |b_0|$. A similar collection of inequalities for the elements in the second row of M gives $b_1 \geq |a_1|$. This is (essentially) the characterization given by Barnes. Note that these give **closed** intervals of admissible values of the off-diagonal terms since we do not wish to exclude forms that represent zero.

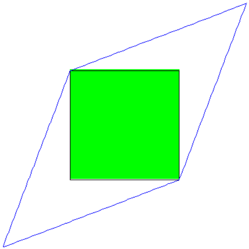
A picture of a cell and its box

There are two types of I-reduced cells. One is Gaussian reduced. Here is an example with its inner box.



Another picture of a cell and its box

Here is a non-Gaussian example with its inner box.



Existence of divided cells

If you have one divided cell, there is an algorithm for producing a **chain** of divided cells, but the question of **existence** of a divided cell has been avoided so far in this presentation. The existence of a divided cell for every expression was proved by Delone, but a better proof was a by-product of the thesis of Jane Pitman (to be shown presently).

If the quadratic part is fixed, it is natural to consider the **chain of reduced forms** given by the **continued fraction** of the quadratic part of the expression.

The three boxes

The cell of a Gaussian reduced form found in the usual continued fraction chain of forms has two I-reduced neighbors obtained by multiplying its matrix on the right by

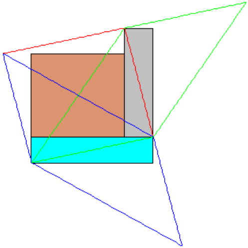
$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

Only 2 by 2 matrices are shown because we are interested in the shape of the cell and not its position.

The three box theorem

Pitman's theorem asserts that the union of the inner boxes of these three cells is (apart from edges) a **fundamental domain** of the lattice. In particular, every possible **origin** belongs to one of these three inner boxes, and the corresponding **cell** is divided by the axes.

A picture of the three box theorem

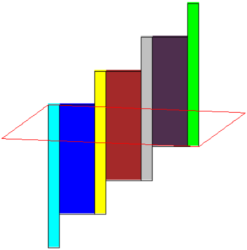


Technicalities

The strict alternation of orientation in the ordinary continued fraction is not present in the chain of divided cells. Indeed, a **consistent** orientation has been enforced. The signs of the off-diagonal terms distinguish two views of the same cell. Thus, any matrix with $a_1 b_0 < 0$, not just those with $a_1 < 0$, should be considered Gaussian, and this will modify the rules for constructing its neighbors in the 3 box theorem. There is a more subtle pattern of orientation here (that will be suppressed in this report, except to note that the **parity of partial quotients** plays a major role).

Axis view of successors

Here are the inner boxes for all divided cell successors of one cell.



Finding all divided cells

The I-reduced forms have $a_0 \geq |b_0|$ and $b_1 \geq |a_1|$. If the cell is non-Gaussian with $a_1 \geq 0$ and $b_0 \geq 0$, and $a_0 \geq 2b_0$, then

$$M_- = \begin{bmatrix} a_0 - b_0 & b_0 \\ a_1 - b_1 & b_1 \end{bmatrix}$$

is also I-reduced and is the second type of Gaussian cell.

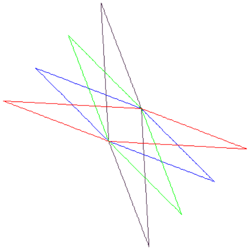
If $2a_1 \geq a_0 \geq a_1$, then the slopes of M_- are in the intervals $[0, 1]$ and $[-1, 0]$. Right multiplication by a matrix that inverts these quantities, adjusts the signs of the diagonal and translates to make $-b_1 \leq a_1 \leq 0$ gives a Gaussian reduced cell.

Restriction to the 3-box cells

The first type of non-Gaussian cell is one of the neighboring cells in the 3-box picture, but the second may not be. However, a study of divided cell steps shows both how they arise in the divided cell algorithm and why only those in the 3-box picture need to be considered. A picture will illustrate this result.

The superfluous cells

Here are some successive cells. All have the same box, so two vertices are shared by all cells. The remaining vertices lie on a fixed pair of lines, and in a pair of opposite quadrants.



What does the picture show?

Only the **unshared** vertices closest to the axes need to be considered. The cells having these vertices can be shown to belong to 3-box pictures. Thus, **it is not necessary to strictly follow the divided cell algorithm** — one can use the ordinary continued fraction and the neighbors of its reduced cells to list the **essential** divided cells.

The Markoff viewpoint

Studies of the Diophantine properties of quadratic forms and their related inhomogeneous expressions now look at the actual infimum of values of the expression over a chain of equivalent expressions. In the homogeneous case, these infima lead to the **Markoff Spectrum**. Earlier work blurred the distinction between chains of equivalent forms and sequences of best approximants of numbers associated with the factors of the form, which leads to what is now known as the **Lagrange Spectrum**.

Zero is a number

The Lagrange approach assumes that an isolated small value is an accident, and uses a limit to identify **essential** properties of rational approximations to a number.

By contrast, the Markoff approach will accept an actual minimum, however atypical. In particular, a single lattice point where the expression is zero allows one to ignore everything else about the expression.

Expressions based on zero forms

The expressions

$$(ax + y + c_0)(y + c_1)$$

for integer a give families that are valid objects of study although they were largely ignored in earlier work. Using the coefficients of the factors to define the matrix M and interpreting the columns of M tells us that this is equivalent to considering translates of the lattice generated by $(a, 0)$ and $(1, 1)$. The fundamental domain of this lattice has area a, so all values of the expression need to be divided by a to get comparable quantities.

An extreme case

For $a = 1$, this is equivalent to the integer lattice. Here, the cell with vertices $(0,0)$, $(1,0)$, $(1,1)$, $(0,1)$ is **the only** divided cell whose box doesn't reduce to an interval. For $-1 \leq c \leq 0$, there is an integer x with $0 \leq |x + c| \leq 1/2$, and these bounds are attained. The other factor is similar, so the absolute value of the expression takes all values between 0 and $1/4$.

The next case

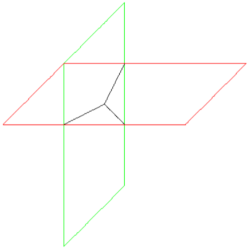
For $a = 2$, the given lattice generators may be interpreted as being Gaussian reduced or not. In the latter interpretation, it is a neighbor of a Gaussian reduced cell with vertices $(0, 0)$, $(1, -1)$, $(2, 0)$, $(1, 1)$, whose box reduces to the single point $(1, 0)$. The non-Gaussian cells in this picture are the only cells that need to be considered. The point $(3/2, 1/2)$ has largest **minimum distance to cell vertex** since the other diagonal of the box separates the points closer to $(2, 0)$ from those close to $(1, 1)$. The values at those vertices are equal on this line and maximal at the center.

Something new

The case $a = 3$ does not seem to have been considered previously. The "largest minimum" is $4/9$ attained at $(5/3, 1/3)$ To prove this, one considers the intersection of two boxes for successive divided cells for the lattice spanned by $(1, 1)$ and $(3, 0)$. The other lattice, spanned by $(2, 1)$ and $(3, 0)$ doesn't allow minima greater than $1/4$.

A picture of the proof

The critical point is found by considering the lines shown that give equal values at two vertices. Other lattice points give larger values.



It’s not always so easy

To compare vertices in adjacent quadrants, where the expression has opposite signs, the locus of equal absolute values is an arc of a hyperbola, so the pictures aren’t as neat. However, it is still possible to find critical points by hand computation.