Proceedings of Symposia in Applied Mathematics
Volume 46, 1992

Number Theory in Computer Graphics

M. DOUGLAS McILROY

ABSTRACT. Computer graphics is geometry on a grid. Hence elementary number theory plays a central role in the design and analysis of basic curve-drawing algorithms. Circles involve matters of representability by sums of squares; straight lines involve continued fractions.

1. Introduction

Computers make drawings by coloring pixels in a bitmap, which may be thought of as the points of a plane integer lattice. People made digital pictures for thousands of years before computers and well before number theory, too (Figure 1). But with computers, where algorithm supplants artistry, the mathematics becomes more important. Drawing a figure becomes a problem in two-dimensional Diophantine approximation: picking points of the lattice to get a good fit.

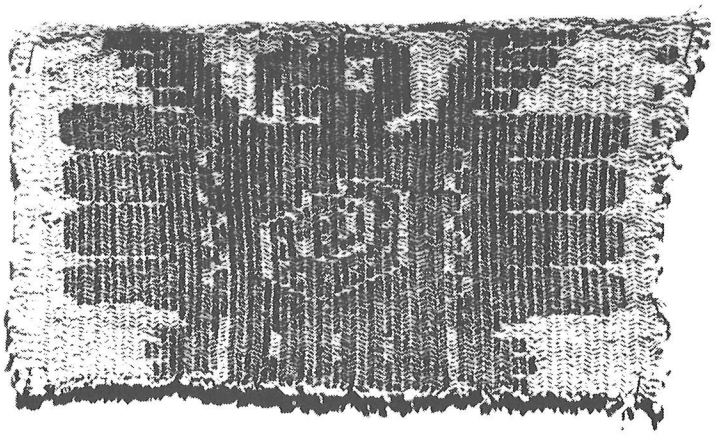

Figure 1. A bitmap image from the Chicama Valley, Peru, circa 2000 BC. American Museum of Natural History. Reprinted with permission from G. G. S. Bushnell, *Ancient Arts of the Americas*, Thames and Hudson, London, 1967.

1991 *Mathematical subject categories:* Primary 11-01; Secondary 68U05, 11J70, 11B57, 11J06.
This paper is in final form. No version of it will be submitted for publication elsewhere.

Resolutions in digital images are typically around one part in a thousand—even coarser on small personal-computer screens. Roundoff becomes visible, and often annoying. Single-pixel errors can have disastrous effects (Figure 2). The details matter, and that is where number theory comes in. Continued fractions, for example, explain the details of the "jaggies" that appear in diagonal lines. Farey series, too, enter into the characterization of digitized lines. The existence and nonexistence of solutions to quadratic Diophantine problems bears on the uniqueness of digital images of circles and ellipses.

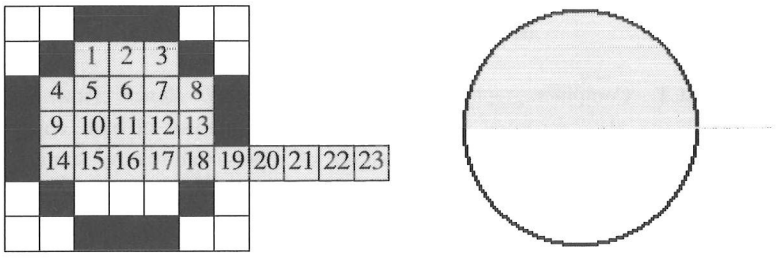

Figure 2. Costly roundoff. (Left) A digital circle of radius 3 with a single-pixel gap, as might result from roundoff error. An attempt to fill it by "painting" the interior pixels in the numbered order fails; the grey paint leaks out. (Right) A similar gap in a digital circle of radius 50 is almost invisible, but leaks the same way.

2. Freeman approximation

In drawing a digital curve, we wish to select pixels that are close to the continuous curve it approximates. Among various criteria for judging closeness [11], one due to Freeman has particularly nice properties [14]. It is invariant under all symmetry operations of the integer lattice: integer translations, half turns, quarter turns, and reflections. And it reduces the two-dimensional problem to a collection of one-dimensional problems.

Definition 0. Unless otherwise specified, *point* means a point of the plane integer lattice.

Definition 1. The *Freeman approximation* to a curve is the set of points (x,y) for which the curve intersects either of the closed unit segments, $H(x,y)$ and $V(x,y)$ centered on (x,y), where

$$H(x,y) = \{(u,y) \mid x-\tfrac{1}{2} \le u \le x+\tfrac{1}{2}\},$$
$$V(x,y) = \{(x,v) \mid y-\tfrac{1}{2} \le v \le y+\tfrac{1}{2}\}.$$

Freeman approximation picks, for each intersection of the curve with a grid line of the lattice, the point nearest to the intersection. Figure 3 shows Freeman points as dots and the segments $H(x,y)$ and $V(x,y)$ as bars.

Ideally a digital curve should look like a curve. It should appear uniformly thin and connected. It should not be too jagged.

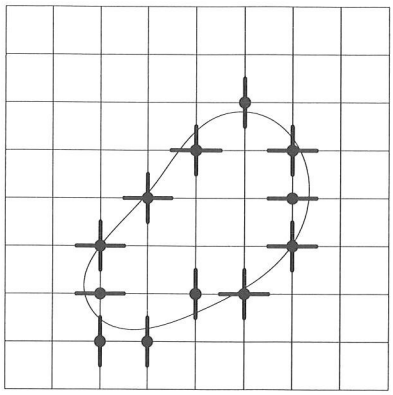

Figure 3. The Freeman approximation to a curve.

Definition 2. A set S of points is *connected* if and only if every pair of points in S can be joined by a path of king moves* within S.

The Freeman approximation to a connected curve is necessarily connected.

Definition 3. A differentiable curve is said to be *predominantly horizontal* if its slope is in the range $[-1,1]$ and *predominantly vertical* if its slope is in the range $[-\infty,-1] \cup [1,\infty]$.

Definition 4. A set of lattice points is *thin* if and only if at most two points of the set are incident on any unit square in the lattice.

The top part of the curve in Figure 4 is predominantly horizontal. It is easy to see that the Freeman approximation to a predominantly horizontal curve must be thin except where it passes exactly halfway between vertically adjacent lattice points.

Thinness is a sometime thing, which can be defeated in several ways, as shown in Figure 4. Some failures of thinness may be defined away. When a curve passes exactly half way between two adjacent lattice points, we may be able to break the tie by an arbitrary choice. When a curve doubles back on itself, we can save the appearances by thinking of the several branches of the curve as lying in different sheets. However, when a curve switches between predominantly horizontal and predominantly vertical there may occur a *square corner*, where the approximation includes three corners of a grid square. This last kind of departure from thinness seems inescapable.

Thinness and connectedness are preserved under the symmetry operations on the lattice.

Half-Freeman approximations are common in computer graphics. A half-Freeman approximation picks points on grid lines that run in only one direction. If the curve is predominantly horizontal, points are picked on vertical grid lines and vice versa (Figure 5). It is easy to see from the figure that the Freeman points on

* In chess, a king move changes one or both coordinates of a point by ± 1.

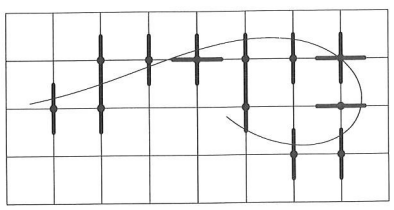

Figure 4 (Left). Thinness is defeated (1) if a curve passes exactly half way between two points as near the left end, (2) if the curve doubles back on itself as near the middle, or in some instances (3) if the curve changes between predominantly horizontal and predominantly vertical, making square corners as at the right side of the figure.

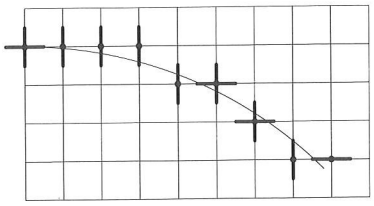

Figure 5 (Right). Vertical bars mark the vertical half-Freeman approximation to a predominantly horizontal octant of a circle. Horizontal bars marking the horizontal half-Freeman approximation agree with the vertical bars except possibly at an end point.

horizontal grid lines coincide with the half-Freeman points on vertical grid lines—except possibly for single outlying points at the ends of the curve.

Half-Freeman approximations to a connected curve are necessarily connected. They are also thin, or can be made so by breaking ties. (To preserve connectedness, ties may have to be broken in a consistent direction.) Thus half-Freeman approximations make visually convincing curves. They are usually easy to compute, too, so they have become ubiquitous in computer graphics. Circles, for example, are usually drawn as eight half-Freeman octants (see box).

To preserve connectivity under half-Freeman approximation, a curve must be split into predominantly horizontal and predominantly vertical sections. The section joins may need special treatment. When a circular quadrant is approximated by octants, there may be one point that is not a half-Freeman point of either octant (Figures 5 and 6).

3. Circles

Circles enter number theory through the study of the Diophantine equation

$$x^2 + y^2 = r^2. \tag{1}$$

In computer graphics, where coordinates are naturally integral, it is customary—and computationally convenient—to limit attention to just such a *standard circle* centered at (0,0) with radius r, or possibly r^2, taken to be integral. The Freeman approximation to the first quadrant is the set of lattice points (x,y) that satisfy either of the inequalities

$$|y - \sqrt{r^2 - x^2}| \le 1/2, \quad 0 \le x \le r \tag{2a}$$

$$|x - \sqrt{r^2 - y^2}| \le 1/2, \quad 0 \le y \le r \tag{2b}$$

How to draw a circle

Formula (3) is the basis of a neat program to trace one octant of a circle [12], [17]. The other octants can be filled in by symmetry.

Start at $(x_0,y_0) = (0,r)$ and then step along the circle computing further points (x_i,y_i) by the scheme

$$x_{i+1} = x_i + 1$$

$$y_{i+1} = \begin{cases} y_i - 1, & \text{if } (y_i - 1/2)^2 + x_{i+1}^2 - r^2 > 0 \\ y_i, & \text{otherwise.} \end{cases}$$

Stop when x_{i+1} would exceed y_{i+1}. The last point drawn may fall beyond the end of the octant. In this event, the point can be shown to be a Freeman point anyway, as in Figure 6.

The scheme avoids any computation of square roots. The single fraction can be eliminated, leaving only integer calculations. Finite differences may be used to update the the quadratic test at each step, thus reducing the calculation of a circular octant to just integer additions, subtractions, and comparisons. The final algorithm is so easy that it comes built into many graphics devices.

Inequality (2a) is satisfied by the greatest integer y such that

$$(y - 1/2)^2 \le r^2 - x^2. \tag{3}$$

The half-Freeman approximation to an octant of a standard circle is necessarily thin because a solution of (3) with $y > 1/2$ is unique. If it were not, then the circle would pass exactly half way between some pair of adjacent lattice points, say (x,y) and $(x,y+1)$, making

$$y + 1/2 = \sqrt{r^2 - x^2}\ .$$

The left side of this equality is rational, but not integral. The right side is the square root of an integer, and therefore either irrational or integral. Hence the situation is impossible and the approximation is thin.

3.1. Other approximation criteria. Freeman circles are remarkably robust, in the sense that different approximating criteria lead to the same approximations. Three natural measures of closeness are

Distance. Euclidean distance from the lattice point to the curve.

Displacement. Distance from the lattice point to the curve measured along the grid line. (This is the Freeman approximation.)

Residual. The absolute value of a defining function $f(x,y)$; the curve is the solution set of $f(x,y) = 0$.

For a standard circle, the minimum-distance approximation picks on each grid line a point (x,y) to minimize $|\sqrt{x^2 + y^2} - r|$. The minimum-displacement approximation minimizes $|y - \sqrt{r^2 - x^2}|$ on vertical grid lines, and a transposed

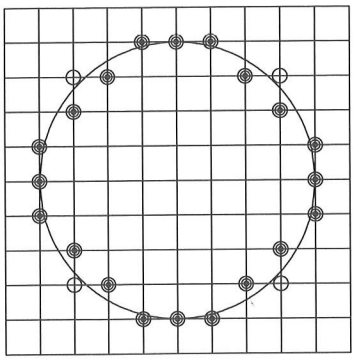

Figure 6. A Freeman circle of radius 4. The black dots mark half-Freeman points of their respective octants. The open dots are Freeman points that are not half-Freeman points of any octant; these points are square corners.

formula on horizontal grid lines. The minimum-residual approximation minimizes $|x^2 + y^2 - r^2|$. Surprisingly, all three approximations are the same [4], [24], [25].

Theorem 1. If r^2 is integral, the minimum-distance, minimum-displacement, and minimum-residual approximations to the circle $x^2 + y^2 = r^2$ are unique and coincide. The approximations are thin except possibly for square corners on the lines $y = \pm x$.

The proof is elementary, but involves considerable case analysis. One case will give some of the flavor. Suppose (x,y) is a minimum-residual point outside the circle and $(x,y-1)$ is a minimum-distance point inside. Both x and y are positive. In analytic terms,

$$r^2 - (x^2 + (y-1)^2) \geq (x^2 + y^2) - r^2,$$
$$r - \sqrt{(x^2 + (y-1)^2} \leq \sqrt{x^2 + y^2} - r.$$

Both sides of both inequalities are positive. Hence the inequalities may be divided to obtain

$$r + \sqrt{x^2 + (y-1)^2} \geq \sqrt{x^2 + y^2} + r,$$

which is absurd. The supposed configuration is impossible.

We originally set out to approximate circles of integer radius. Theorem 1 covers somewhat more: circles with radius equal to the square root of an integer. In particular the three approximations must agree for any circle centered on a lattice point and passing through some lattice point.

It is natural to consider admitting integer diameters and half-integer centers. One could then, for example, cleanly approximate circles inscribed in squares of arbitrary integer dimensions. In this widened class the guarantee of thinness disappears and the approximations may disagree pairwise. Taken together, though, the three approximations always vote unanimously for a unique answer [25].

Theorem 2. If $2x_0$, $2y_0$, *and* $(2r)^2$ *are integral, the intersection of the minimum-distance, minimum-displacement, and minimum-residual approximations to a quadrant of the circle* $(x - x_0)^2 + (y - y_0)^2 = r^2$ *is nonempty and unique on any grid line that intersects the quadrant.*

Theorem 2 is best possible, in the sense that half integers cannot be replaced by any finer rational subdivision of the lattice. If $q > 2$, there exist circles with integral qx_0, qy_0, and $(qr)^2$ for which there is not even majority consensus. The intersection of any two of the three approximations to these circles is disconnected [25].

3.2. Square corners. Just how often in approximating a standard circle of integral radius will we meet square corners as in Figure 6? The answer—about twice between successive powers of 34—depends on the Pell equation [23], [25].

Theorem 3. Square corners appear in the Freeman approximation to a standard circle of integral radius r *if and only if* $r = 4, 11, 134, 373, 4552, 12671, 154634, 430441, \cdots$.

The sequence of radii in Theorem 3 satisfies the linear recurrence $r_{k+2} = 34r_k - r_{k-2}$.

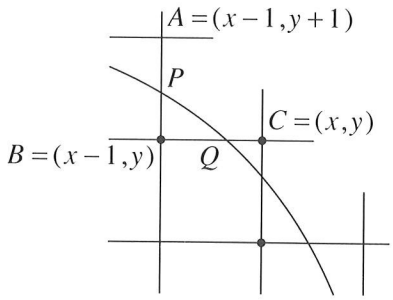

Figure 7.

Figure 7 shows the conditions under which a square corner can happen. The square corner C must be on the diagonal, where $x = y$, and we must have $PB < PA$ and $QC < QB$. By Theorem 1, we may work in terms of residuals, and avoid square roots. Point B is inside the circle; the residual there, $(x - 1)^2 + y^2 - r^2$, is negative. The inequality $PB < PA$ implies that the residuals at B and A satisfy

$$-((x - 1)^2 + y^2 - r^2) < (x - 1)^2 + (y + 1)^2 - r^2,$$

and similarly at C and B,

$$x^2 + y^2 - r^2 < -((x - 1)^2 + y^2 - r^2).$$

Set $x = y$ in these inequalities and simplify, to find

$$4x^2 - 2x + 1 < 2r^2 < 4x^2 - 2x + 3,$$

which, because r is integral, is equivalent to

$$r^2 = 2x^2 - x + 1. \tag{4}$$

Completing the square in (4) and eliminating fractions gives

$$8r^2 = (4x - 1)^2 + 7. \tag{5}$$

The solutions of (5) are solutions of the Pell equation $2p^2 - q^2 = 7$ with p even and $q \equiv 3 \pmod 4$. Standard methods of solution [29] over the unique factorization domain $\mathbf{Z}(\sqrt{2})$ lead to Theorem 3 and the associated recurrence.

One way to read (4) is that the residual at C, namely $x^2 + x^2 - r^2$, is $x - 1$. It follows that the residuals at B and A are $-x$ and $x + 1$. The magnitudes of the three residuals are contiguous integers; the conditions for a square corner are delicate indeed.

3.3. Ellipses, etc. Algorithms for approximating standard ellipses resemble those for circles [12], [27]. Again, as with standard circles, the Freeman approximation to an ellipse with semiaxes of integral length is unique. In other words, it is impossible for the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

with integral a and b to pass exactly half way between adjacent lattice points. To prove it, suppose that the ellipse passes through a point, $(x,z/2)$, where z is an odd integer. We may assume that $\gcd(x,a) = \gcd(z,b) = 1$; if not, we could reduce the fractions in the defining equation

$$\frac{x^2}{a^2} + \frac{z^2}{4b^2} = 1$$

to get a counterexample of the same form where the assumption does hold. The sum of two fractions in lowest terms can be 1 only if their denominators are the same. Hence $a = 2b$. Because x/a is in lowest terms, x must be odd. Consequently (x,z,a) is a Pythagorean triple with parities (odd,odd,even) that satisfies

$$x^2 + z^2 = a^2.$$

As is well known, no such triple exists, for that would imply

$$1 + 1 \equiv x^2 + z^2 \equiv a^2 \equiv 0 \pmod 4.$$

Standard ellipses with integer semiaxes are about the most complicated curves for which a fully satisfactory understanding is available in computer graphics. Other conics, and even nonstandard circles, are customarily handled in an ad hoc way. There's plenty of room to exploit other knowledge from number theory. Here are some questions.

1. If arbitrary centers, (x_0,y_0), and radii are allowed, how many distinct digital circles of radius less than r exist? To what extent does the answer depend on the approximating criterion? Characterize the equivalence classes that digital approximation induces in (x_0,y_0,r) space.
2. Recognize digital circles efficiently.
3. How rough is a digital curve?

4. Describe properties of other digital curves. General 2nd-degree and 3rd-degree polynomials (conics and splines) are of particular interest in computer graphics.

5. The union of digital circles of all integer radii does not cover the integer lattice. Describe the gaps.

4. Straight lines and chain codes

The path of a digital curve is often recorded as a *chain code,* or differential encoding of pixel-to-pixel moves [31]. The intriguing structure of chain codes of straight lines eluded satisfactory explanation until a connection was made with continued fractions.

The chain-code representation of a general digital curve is relatively compact, certainly much more so than a sequence of coordinate pairs. Fortuitously in two dimensions there are 8 directions to nearest neighbors, so general chain codes fit perfectly in 3 bits. Here we shall adopt a limited definition, with just two instead of 8 directions of motion, adequate for the analysis of digital straight lines of positive or zero slope [6].

Definition 5. A *chain code* of a monotone nondecreasing curve, $f(x)$, is a string of 0's and 1's. Let i be an integer abscissa and let $h(i) = \lfloor f(i+1) \rfloor - \lfloor f(i) \rfloor$, where $\lfloor\ \rfloor$ is the integer part function. Then the chain code can be parsed into words, $01^{h(i)}$, made of a single zero followed by $h(i)$ ones.

In the chain code of a predominantly horizontal curve, 1's cannot occur contiguously.

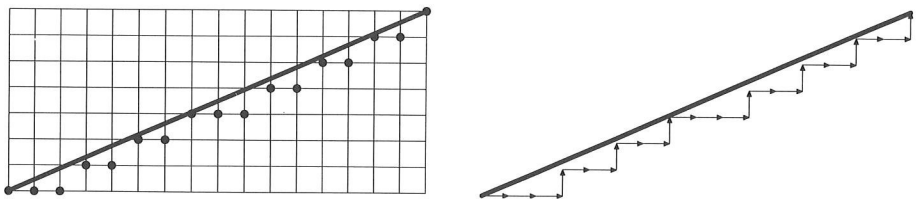

Figure 8. (Left) A digital segment of slope 7/16. (Right) The chain code of the same segment.

A chain code designates a Manhattan path (Figure 8). The digital line segment in Figure 8 has slope $m = 7/16$. If the lower left corner is taken to be $(0,0)$, then the ordinate y of each point (x,y) of the digital segment is the largest integer y that satisfies $y \leq mx$. In other words, the points form a greatest lower bound integer approximation to the line,* sometimes called the *spectrum* of m [15]. The chain code is an infinite repetition of the string

0001001001000100100100 1.

* The greatest lower bound approximation is the same as the Freeman approximation of the shifted line $y = mx - 1/2$, with ties broken in favor of the upper point.

A predominantly horizontal line has a chain code with at least one 0 for every 1. The 1's are isolated and spread as uniformly as possible among the 0's. Thus between every two successive 1's there must be p or possibly $p+1$ zeros, for some integer p. Continued fractions provide the details of the distribution of 1's.

Chain codes by other names have a long history in the geometry of numbers. The "cutting sequence" of grid lines by a line that passes through no lattice points is precisely the chain code, where 0 denotes a vertical grid crossing and 1 a horizontal one. Cutting sequences, presented as the sequence of chain-code exponents, p or $p+1$, also arise in counting the integers between successive terms of the series $\{i\alpha\}$ of integer multiples of an irrational α. In these guises, the study of chain codes goes back at least a century to Christoffel and Markoff [32].

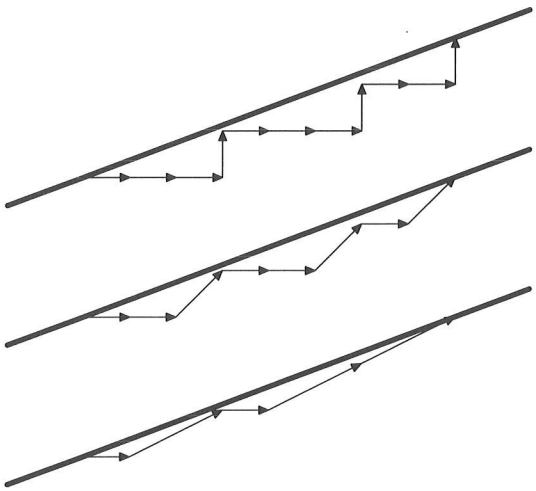

Figure 9. The chain code for a line of slope 3/8 in the standard basis (top) is 00010001001; in the basis transformed by S it is 00100101 (middle), and transformed by S^2 (bottom) it is 01011.

4.1. Chain code transformations and continued fractions. Bruckstein defined a set of invertible transformations among chain codes of digital lines [6]. Two of them, called X and S, are given here.

X. Parse the string into words 0 and 1. Exchange 0 and 1.

S. Parse the string into words 0 and 01. Replace 01 by 1.

Theorem 4. The result of applying either of the transforms X or S to a string of 0's and 1's is the chain code of a digital line if and only if the original string is the chain code of a digital line.

Each transform can be explained as a change of basis in the lattice, and is thus associated with a matrix. Transform X transposes the lattice; it has matrix M_X:

$$M_X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Transform S takes as a new basis the vectors originally coded as 0 and 01 (Figure 9). Code 0, a (1,0) step in original coordinates, maps into code 0 or a (1,0) step in the new. Code 01, a (1,1) step in original coordinates, maps into code 1 or a (0,1) step in the new. The matrix associated with S is thus

$$M_S = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}.$$

When transform S is iterated p times, the short and long words, 0^p1 and $0^{p+1}1$, reduce to 1 and 01 respectively. The resulting string is not parsable into 0's and 01's, so the transform is no longer applicable. The iterated transform has the matrix

$$M_{S^p} = (M_S)^p = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}^p = \begin{bmatrix} 1 & -p \\ 0 & 1 \end{bmatrix}.$$

Continued fractions arise when the transform XS^p is applied repeatedly [5]. (Since S^p changes the predominant symbol from 0 to 1, transform X must be used after each S^p to restore the predominancy of 0.) Table 1 shows the process applied to the chain code for Figure 8. In the table, the operation "shift" simply changes viewpoint on the periodic chain code of an infinite line. The values of p in the table are the partial quotients, [2,3,2], of the continued fraction for the slope [16]:

$$\frac{7}{16} = \cfrac{1}{2 + \cfrac{1}{3 + \cfrac{1}{2}}}.$$

Intuitively, the approximation must rise 7 steps in a run of 16. ("Rise" and "run" are used here in the engineering sense, meaning horizontal and vertical changes.) A rise of 1 in a run of 2 almost does it, so 2 is the first partial quotient. To this degree of approximation, the chain code is 001. But the approximation is too steep, rising 7 steps in a run of 14, not 16 as desired. Thus two out of every seven 001 words must be lengthened. We need one extra 0 in every 3½ words. And that we do by lengthening one word in every 3 (the next partial quotient) to make a superword 0001001001, and sticking in one 001 word after each 2 (the last partial quotient) superwords.

Algebra explains better than English. Since the chain code of a predominantly horizontal line of slope m has one 1 for every p or possibly $p+1$ zeros, we have $p = \lfloor 1/m \rfloor$. Then

$$m = \frac{1}{1/m} = \frac{1}{\lfloor 1/m \rfloor + (\text{number less than } 1)},$$

Thus we see that the iteration exponent p is in fact the partial quotient in the continued fraction for m. The continued fraction for the residual "number less than 1" corresponds to the chain code as transformed by XS^p.

In every case, the transforms must convert invertibly from integer coordinates to integer coordinates. Hence all the matrices must have integer entries and

Table 1. Transforming the chain code for slope 7/16.

Sequence	Applicable transform
0001001001000101001001	$S^p, p = 2$
011101111	X
100010000	Shift
000010001	$S^p, p = 3$
011	X
100	Shift
001	$S^p, p = 2$
1	X
0	

determinant ± 1. Two-by-two matrices of this type form a "general linear group," customarily denoted $GL(2,\mathbf{Z})$. The group elements represent all possible basis changes in the lattice [10]. Transforms X, S, and one other, in particular

$$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{6}$$

generate the group. This group gives a full account of invertible chain-code transformations. As the whole group includes basis changes that turn lines of positive slope into lines of negative slope, it requires a general domain of chain codes, where negative as well as positive steps are admitted. Transform (6) maps chain codes in our limited subdomain into chain codes outside the subdomain, as do higher powers of S than S^p.

4.2. Association of segments and chain codes. By Theorem 4, the method of Table 1 can be used to test whether a periodic or finite sequence of 0's and 1's is the chain code of a digital line or segment. A string of 0's and 1's is a chain code (or period thereof) for a line of rational slope if and only if the the string reduces to all 0's in a finite number of S^p and X steps [33]. The number of such steps is $O(\log n)$, the same as for the continued fraction algorithm [20], so the overall time to decide by this method whether a sequence of n 0's and 1's form the chain code of a straight line is $O(n \log n)$. (This running time is not optimal; the straightness of a chain code can be decided in time $O(n)$ [3].)

The chain code of a given segment is unique. Moreover the chain codes of all segments of a given rational slope with a common projection on the x axis are unique up to a cyclic shift. But many segments may have the same chain code. For any finite chain code C containing n zeros, there is an equivalence class of pairs (m,b) for which the segment

$$y = mx + b, \quad 0 \le x \le n \tag{7}$$

has chain code C.

Let us find the shape of the equivalence sets. If the segment (7) passes between, but does not touch, the lattice points (x,y) and $(x,y+1)$, then b lies in the open interval $-xm + y$ to $-xm + y + 1$. These limits are parallel lines in the m-b

plane with integer slope $(-x)$ and integer intercept. (The x-y and m-b spaces are dual to each other, with points in one corresponding to lines in the other [19].) The network of all such lines for $x = 0,1,\ldots,n$ partitions the m-b plane into *facets* (Figure 10). For every x, every segment (7) with m and b in the interior of one facet will pass between the same two lattice points (x,y) and $(x,y+1)$. Thus, from Definition 5, every such segment has the same chain code. The open facets together with appropriate boundary points are the equivalence classes we seek.

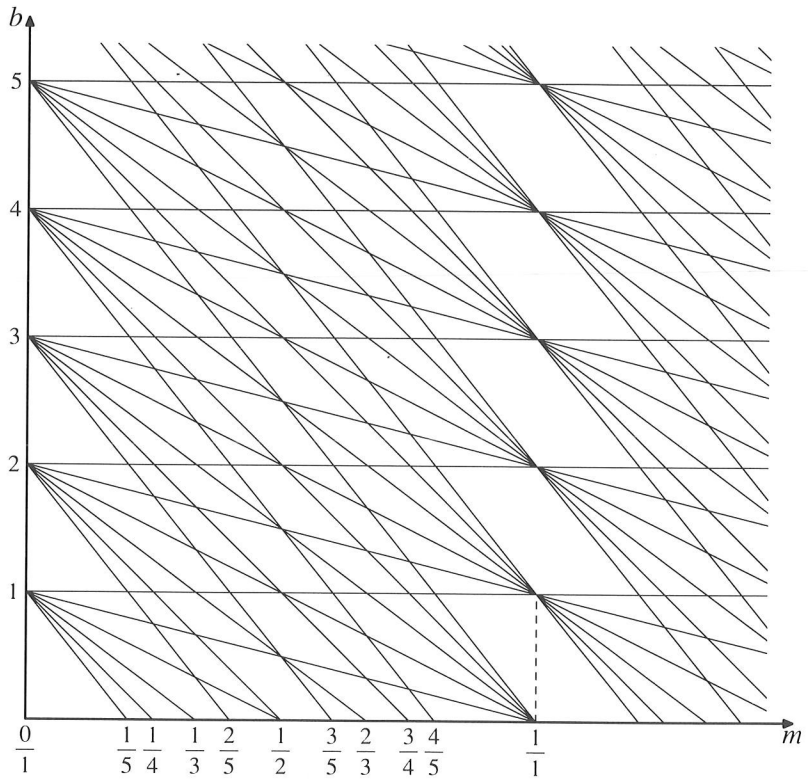

Figure 10. A Farey fan of order 5. Each facet comprises the parameters (m,b) of a family of lines $y = mx+b$ that have the same digital approximation over the interval $0 \le x \le 5$. The dotted line closes the basic rectangle $[0,1]\times[0,1]$.

The facet boundaries, lines of integer slope, fan out from integer points on the y axis. The m-intercepts between 0 and 1 form a Farey series, a well-studied object of number theory.* It may be verified that the ordinates of successive intersections

* A Farey series of order n is the sequence of all proper fractions with denominators less than or equal to n arranged in ascending order. The labels on the m-axis of Figure 10 form a Farey series of order 5.

along each ray also form Farey series of different orders. And the three distinct m-coordinates of each facet are successive terms of some Farey series. Consequently the figure has been dubbed a *Farey fan* [26].

Theorem 5. No facet in a Farey fan has more than four sides [9].

A simple explanation for Theorem 5 is that every segment in an equivalence class is a convex combination of segments in at most four extreme positions, which correspond to the vertices of the facet. Figure 11 exemplifies the four positions.

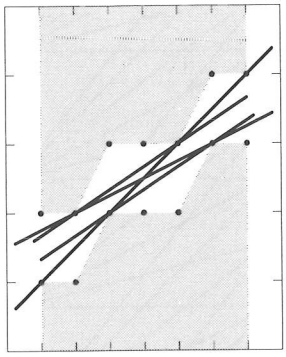

Figure 11. The top and bottom edges of the unit-height channel are 6-unit digital segments of slope 1/3. All lines that pass through the channel without touching the top have the same chain code. Vertical scale exaggerated.

The Farey fan of order n can be built by adding lines of slope $-n$ to a fan of order $n-1$. The values of m where the n new lines cut old lines within the basic rectangle, $0 \le m \le 1$, $0 \le b \le 1$, are precisely a Farey series of order n with the first term, 0/1, deleted. This observation leads to

Theorem 6. The number of distinct chain codes for n-unit digital segments is

$$1 + \sum_{k=1}^{n} (n+1-k)\,\phi(n).$$

As usual, $\phi(n)$ is Euler's totient, the number of positive integers less than n that are relatively prime to n. The number of terms in a Farey series of order n is

$$1 + \sum_{k=1}^{n} \phi(n),$$

from which the theorem follows directly [8]. The first ten values of the formula in the theorem are 2, 4, 8, 14, 24, 36, 54, 76, 104, 136. Asymptotically, the number of distinct n-unit digital segments is n^3/π^2 [2].

The size of a facet in the Farey fan shows the uncertainty inherent in a digital segment. For example, to recover the parameters of a line dropped on an integer lattice, one can read off a bit of chain code, from there determine the facet it lies in, and thence estimate its parameters [8]. The uncertainty is variable; it is greatest for the

big facets at $m = 0$ and $m = 1$. This reflects the fact that an arrow shot along an alignment of trees in a square orchard is likely to go farther when the alignment is in a principal direction; then the arrow's path (a segment) need not be very precisely positioned to fit between the trees.

On average, an n-unit digital segment can be used to estimate the position of a selected point on a straight line to about $1/n$ unit [2]. Such estimates have been used to locate points in digital satellite images to subpixel accuracy [1]. The least uncertainty in estimating the parameters of a digitized line occurs for lines with slope τ^{-1}, the reciprocal of the golden ratio. Because multiples of τ^{-1} are as uniformly distributed modulo 1 as possible [21], the edges of the channel through which the line threads as in Figure 11 cannot be far apart.

For more about digital lines and further pointers to the literature, see Dorst [8] and Bruckstein [6], both of whose work appears in a recent volume of the AMS Contemporary Mathematics series [28]. For a generalization to 3-space, see Forchhammer [13].

5. Conclusion

Simple problems in computer graphics have been tamed by number-theoretic analysis. It is a routine and inexpensive matter to represent straight lines, standard circles, and standard ellipses to the ultimate precision of digital media. Artifacts of the representations (jaggies) are understood in detail. At least where straight lines are involved, it is possible to recover line data from digital images to subpixel accuracy. In another problem, that of drawing circles, the relationship among various approximation criteria has been characterized in some detail.

In computer graphics, more than in most numerical computing, one is vividly confronted by the discrete nature of the pursuit. There is an ultimate, finite level of precision. At that level, numerical analysis merges with number theory, and the imperfection of rounding fades into the exactness of integer computation. Such ultimate precision is occasionally approached in numerical analysis, for example in the exact treatment of Newton's method for the square root [18], in the ellipsoid method for linear programming [30], or in the analysis of floating-point base conversion [7]. It is interesting to speculate about the extent to which numerical computing may one day become explicable as a branch of number theory.

I am grateful to J. A. Reeds for helpful criticism and to J. C. Lagarias for many pointers to the literature.

REFERENCES

1. C. A. Berenstein, L. N. Kanal, D. Lavine, and E. C. Olson, *A geometric approach to subpixel registration accuracy,* Computer Vision, Graphics, and Image Processing **40** (1987), 334-360.
2. C. A. Berenstein and D. Lavine, *On the number of digital straight line segments,* IEEE Transactions on Pattern Analysis and Machine Intelligence **10** (1988), 880-887.
3. M. Boshernitzan and A. S. Fraenkel, *A linear algorithm for nonhomogeneous spectra of numbers,* Journal of Algorithms **5** (1984), 187-198.
4. J. Bresenham, *A linear algorithm for incremental digital display of circular arcs,* Communications of the ACM **20** (1977), 100-106.
5. R. Brons, *Linguistic methods for the description of a straight line on a grid,* Computer Graphics and Image Processing **3** (1974), 48-62.
6. A. M. Bruckstein, *Self-similarity properties of digitized straight lines,* department of Computer Science report #616, Technion – Israel Institute of Technology, March, 1990. Also in Melter, *Vision Geometry.*
7. W. D. Clinger, *How to read floating point numbers accurately,* ACM SIGPLAN '90 Conference on Programming Language (White Plains, 1990) Design and Implementation, Association for Computing Machinery, 1990, pp. 92-101.
8. L. Dorst, *Discrete Straight Line Segments: Parameters, Primitives and Properties,* PhD Thesis, Technische Hogeschool Delft, 1986. Also in Melter, *Vision Geometry.*
9. L. Dorst and A. W. M. Smeulders, *Discrete representation of straight lines,* IEEE Transactions on Pattern Analysis and Machine Intelligence **6** (1984), 450-462.
10. H. Eves, *Elementary Matrix Theory,* Dover, 1966.
11. E. L. Fiume, *The Mathematical Structure of Raster Grpahics,* Academic Press, 1989.
12. J. D. Foley, A. Van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics Principles and Practice,* Addison-Wesley, 1990.
13. S. Forchhammer, *Digital plane and grid point segments,* Computer Vision, Graphics and Image Processing **47** (1989), 373-384.
14. H. Freeman, *Computer processing of line-drawing images,* Computing Surveys **6** (1974), 57-97.
15. R. L. Graham, S. Lin, and C-S. Lin, *Spectra of Numbers,* Mathematics Magazine **51** (1978), 174-176.
16. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers,* Oxford University Press, 1968.
17. B. K. P. Horn, *Circle generators for display devices,* Computer Graphics and Image Processing **5** (1976), 280-288.
18. A. S. Householder, *Principles of Numerical Analysis,* McGraw-Hill, 1953, pp. 10-13.
19. F. Klein, *Elementary Mathematics from an Advanced Standpoint,* Dover, 1945.
20. D. E. Knuth, *The Art of Computer Programming,* Vol. 2 Seminumerical Algorithms, Addison-Wesley, 1971, p. 320.
21. _____, *The Art of Computer Programming,* Vol. 3 Searching and Sorting, Addison-Wesley, 1973, p. 511.

22. J. Koplowitz, M. Lindenbaum, and A. Bruckstein, *The number of digital straight lines on an N×N grid,* IEEE Transactions on Information Theory **36** (1990), 192-197.
23. Z. Kulpa, *On the properties of discrete circles, rings, and disks,* Computer Graphics and Image Processing **10** (1979), 348-365.
24. Z. Kulpa and M. Doros, *Freeman digitization of integer circles minimizes the radial error,* Computer Graphics and Image Processing **17** (1981), 181-184.
25. M. D. McIlroy, *Best approximate circles on integer grids,* ACM Transactions on Graphics **2** (1983), 237-263.
26. _____, *A Note on Discrete Representation of Lines,* AT&T Technical Journal **64** (1985), 481-490.
27. _____, *Getting raster ellipses right,* ACM Transactions on Graphics (to appear).
28. R. A. Melter, A. Rosenfeld, and P. Bhattacharya (eds.), *Vision Geometry.* American Mathematical Society, Providence, 1991. Contemporary Mathematics Series **119**.
29. T. Nagell, *Introduction to Number Theory,* Chelsea, 1964.
30. C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity,* Prentice-Hall, 1982.
31. T. Pavlidis, *Algorithms for Graphics and Image Processing,* Computer Science Press, Rockville, MD, 1982.
32. C. Series, *The geometry of Markoff numbers,* The Mathematical Intelligencer **7**, 3 (1985), 20-29.
33. L. D. Wu, *On the chain code of a line,* IEEE Transactions on Pattern Analysis and Machine Intelligence **4** (1982), 347-353.

SOFTWARE AND SYSTEMS RESEARCH DEPARTMENT, AT&T BELL LABORATORIES, MURRAY HILL, NEW JERSEY, 07974

E-mail: doug@research.att.com