Drawing ellipses, hyperbolas or parabolas with a fixed number of points and maximum inscribed area

L. B. Smith*
CERN, Geneva, Switzerland

In interactive graphical work one may want to represent curves by connecting a fixed number of points on the curves by straight lines. Parametric representations are given which lead to efficient algorithms for computing piecewise linear representations of ellipses, hyperbolas and parabolas. It is proved that the representations give inscribed polygons with maximum area in all three cases.

(Received September 1969)

1. Introduction

Pitteway (1967) describes an algorithm for drawing ellipses or hyperbolas with a digital plotter. The method involves choosing a sequence of pen movements in terms of the various available directions and the smallest pen increment. The sequence is chosen so that the straight line segments traced by the pen deviate from the desired curve as little as possible. For an off-line plotting device which moves a pen in basic increments of certain fixed directions and lengths, Pitteway's algorithm for ellipses and hyperbolas is effective and efficient. Another similar algorithm for drawing ellipses and hyperbolas is given by Partridge (1968). Botting and Pitteway (1968) consider Pitteway's original method for conic sections and give an extension to cubic curves.

When using a cathode-ray tube (CRT) for on-line computer display, it may be useful to represent a curve by a fixed number of points. This limitation may be dictated by the hardware involved or by programming considerations. For example, CRT display units often have an associated buffer of limited capacity from which the picture is generated. A picture requiring more storage space than that available in the buffer will flicker. Thus a buffer size may dictate limitations on the number of points that can be displayed without flicker. On the other hand some graphics software systems have facilities for saving sub-pictures for later re-display. Such systems may store the sub-picture in the computer memory as a sequence of graphic commands which create a display of the points (vectors) involved. In this case a limitation on the number of points representing a sub-picture (a curve) might be dictated by storage requirements of the graphics system. A fixed number of points can, of course, also be useful in limiting the computation time involved in representing a curve. Pitteway's algorithm has no provision for using a fixed number of points since it was designed for a plotting device which uses a given fixed increment size until the curve is completely traced.

The algorithms given here assume that a fixed number of points are available to represent an ellipse (or hyperbola or parabola). Those points are then distributed along the curve so that a faithful representation of the curve is given when the points are joined by straight lines. Thus, when the curvature is great the points must be closely spaced and when the curvature is small the points must be widely spaced.

Visitor from Stanford Linear Accelerator Center, Stanford, USA

2. Plotting ellipses

Given a fixed number of points there is an obvious distribution of those points to represent graphically a circle. Equal angle increments with a point on the perimeter for each angle will nicely depict the circle (by approximating it with a regular polygon). An ellipse, however, is a different matter, particularly if the eccentricity is large (near 1). Equal angle increments used to pick points for display can easily fail to represent the small ends of an ellipse as shown in Fig. 1 even though the sides appear quite smooth. This is because at the ends, the points for plotting are too far apart with respect to the curvature.

A seemingly better method would be to use equal perimeter lengths to separate the representative points. However, this would use more points than necessary on the sides in order to have points close enough together on the ends. Another drawback to the use of equal perimeter lengths is the fact that their computation involves an elliptic integral, an undesirably complex operation. One would like the calculation to be efficient, if at all possible.

Another scheme which produces faithful representations of ellipses (including those with large eccentricity) with a relatively small amount of computation can be devised. Consider an ellipse centred at the origin with major axis $2a$ and minor axis $2b$. The radial line from the origin to the point (a, b) intersects the ellipse at the point $(a/\sqrt{2}, b/\sqrt{2})$. If we are given N points with which to represent the ellipse at hand, the scheme is as follows (see Fig. 2) for the representation in the first quadrant. The other three quadrants will be handled similarly.

(*a*) Compute

$$dx = \left(\frac{a}{\sqrt{2}}\right) \bigg/ \left(\frac{N}{8}\right)$$

$$dy = \left(\frac{b}{\sqrt{2}}\right) \bigg/ \left(\frac{N}{8}\right)$$

(*b*) set $x = a$, $y = 0$

(*c*) plot (x, y)

(*d*) if $y < b/\sqrt{2}$, set $y = y + dy$, compute x, go to (*c*)

(*e*) if $y \geqslant b/\sqrt{2}$, set $x = x - dx$,

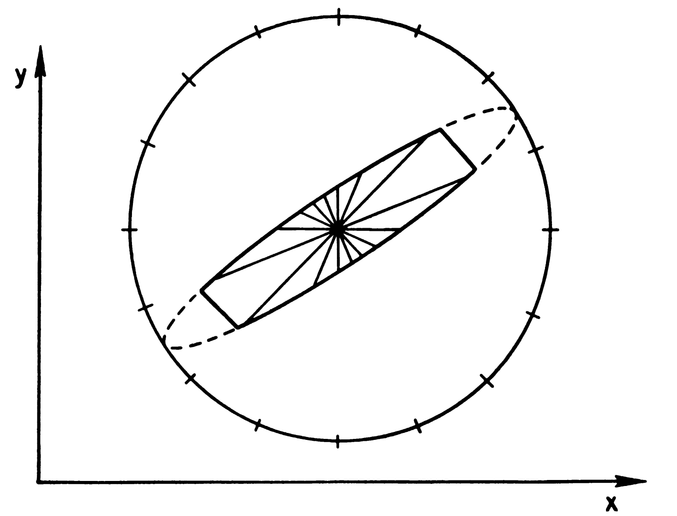

Fig. 1. Equal angle representation of an ellipse. For high eccentricity ellipses, the ends are not well represented.

if $x \geqslant 0$, compute y, go to (c)
if $x < 0$, go to (f)
(f) that is the end of the first quadrant.

A fourth (and best known to the author) scheme for graphical representation of ellipses by a fixed number of points uses the parametric representation of an ellipse centred at the origin with axes $2a$ and $2b$,

$$x = a \cos \phi, \tag{1a}$$

$$y = b \sin \phi. \tag{1b}$$

Using (1), as the parameter ϕ is incremented from 0 to 2π, the points (x, y) trace the desired ellipse. It is easy to use a fixed number of points, say N, by using increments of $2\pi/(N-1)$ to vary ϕ over the interval $[0, 2\pi]$.

We see that this method automatically gives the desired changes in perimeter increments, namely relatively small changes at the ends and relatively large changes along the sides which are necessary for ellipses with eccentricity near one. We have

$$dy = b \cos \phi \, d\phi \tag{2a}$$

$$dx = -a \sin \phi \, d\phi \tag{2b}$$

so that for ϕ near zero (or near π) we have $|dy| \approx b d\phi$ and dx near zero. For ϕ near $\pi/2$ or $3\pi/2$ we have $|dx| \approx a d\phi$ and dy near zero. Thus the ratio of perimeter increment size at the small ends to that on the sides is approximately b/a as desired for a faithful representation of an ellipse with eccentricity near one ($b/a \ll 1$). We also see that this method gives equal perimeter increments (a regular polygon) if $a = b$ for the best possible representation in the case of a circle. In the next section we prove that the inscribed area is maximum.

Another advantage of the use of the parametric equations for plotting an ellipse is the saving in computer time to calculate the points. The previous method, using equal y increments along the sides involves a SIN, a COS and a SQRT calculation plus considerable testing to determine each point to be plotted. The parametric method requires only a few arithmetic operations to calculate each point since we can use the multiple angle trigonometric identities as we are incrementing ϕ by an equal amount for each point. Given (x_c, y_c) the centre of the ellipse, ϑ, the tilt of the major axis, and the major and minor semi-axes, a and b we compute the points to be plotted $\{(x_n, y_n)\}_{n=1}^{N}$ as follows

Initialise

$$\begin{cases} d\phi = 2\pi/(N-1) \\ CT = \cos(\vartheta) \\ ST = \sin(\vartheta) \\ CDP = \cos(d\phi) \\ SDP = \sin(d\phi) \\ CNDP = 1{\cdot}0 \\ SNDP = 0{\cdot}0 \end{cases}$$

Then repeat the following for $n = 1, 2, \ldots, N$

$$\begin{cases} x' = a * CNDP \\ y' = b * SNDP \\ \\ x_n = x_c + x' * CT - y' * ST \\ y_n = y_c + x' * ST + y' * CT \\ \\ TEMP = CNDP * CDP - SNDP * SDP \\ SNDP = SNDP * CDP + CNDP * SDP \\ CNDP = TEMP \end{cases}$$

The above algorithm is a straightforward coding of the numerical processes involved. However, it can be made much more efficient by the following equivalent algorithm which reduces the number of multiplications and additions in the inner loop to four each.

Initialise:

$$\begin{aligned} d\phi &= 2\pi/(N-1) \\ CT &= \cos(\theta);\ ST = \sin(\theta) \\ CDP &= \cos(d\phi);\ SDP = \sin(d\phi) \\ A &= CDP + SDP * ST * CT * (a/b - b/a) \\ B &= -\ SDP(\ (b * ST) **2 + (a * CT) **2)/(a * b) \\ C &= SDP\ (\ (b * CT) **2 + (a * ST) **2\ )/(a * b) \\ D &= CDP + SDP * ST * CT * (b/a - a/b) \\ D &= D - (C * B)/A \\ C &= C/A \\ x &= a * CT;\ y = a * ST. \end{aligned}$$

Then repeat the following for $n = 1, 2, \ldots, N$.

$$\begin{aligned} x_n &= x_c + x \\ y_n &= y_c + y \\ x &= A * x + B * y \\ y &= C * x + D * y. \end{aligned}$$

The parametric method of ellipse plotting has been used in an interactive graphical environment. The author gives elsewhere (Smith, 1969) some discussion of fitting empirical elliptical data by the method of least squares. The ellipses

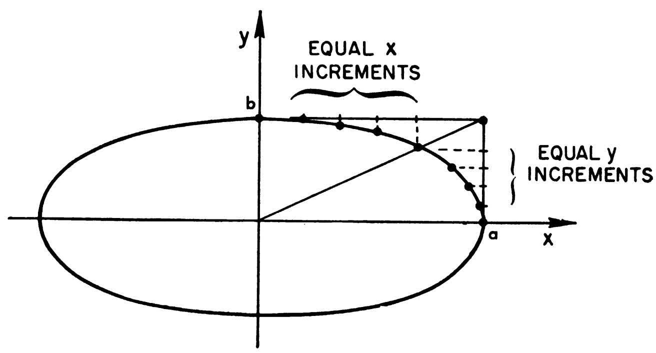

Fig. 2. A successful ellipse plotting scheme. This scheme works, but it is not as efficient as the parametric representation.

encountered in that work often had major to minor axis ratios of near 60 to 1. Such elongated (high eccentricity) ellipses require special treatment in order to be displayed accurately. Fig. 3 shows an ellipse plot shown on-line during a session with the least squares data-fitting program described by the author (Smith, 1969). The ellipse shown in Fig. 3 has a major to minor axis ratio of 14·4 to 0·164. The abscissa and ordinate scale factors have been made unequal for the display, thus to some extent belying such a high eccentricity.

3. Parametric ellipse gives maximum inscribed area

We have just seen that the parametric representation of an ellipse is efficient and we have shown that it is intuitively accurate in that the points are spaced widely when the curvature is small and closely when the curvature is great. Here we shall prove that the area inscribed by the polygon formed by the points given by the parametric representation is maximum. This shows that in addition to being efficient, the algorithm for plotting ellipses is the most accurate (in the sense of maximum area).

The maximum inscribed area criterion used to examine the conic section representations discussed here is somewhat, but not entirely, arbitrarily chosen. Other criteria such as minimising the maximum error between the polygon and the curve could be used and would be more appropriate in some circumstances. However, the criterion used here involves points calculated *on the curve* itself and satisfies the requirement that point spacing be inversely proportional to the curvature. The fact that the points are on the curve make the representation exact if only the points are displayed. The inscribed polygon, although farther from a convex curve than for example a minimax polygon, gives a representation that is quite satisfactory for on-line viewing.

Consider the N-sided polygon formed by connecting the N points (connect point N to point 1)

$$\left.\begin{aligned} x_n &= a\cos\phi_n \\ y_n &= b\sin\phi_n \end{aligned}\right\} n = 1, 2, \ldots, N \tag{3}$$

where

$$\phi_n = (n-1)\frac{2\pi}{N}\ .$$

Assume $N \geqslant 3$.

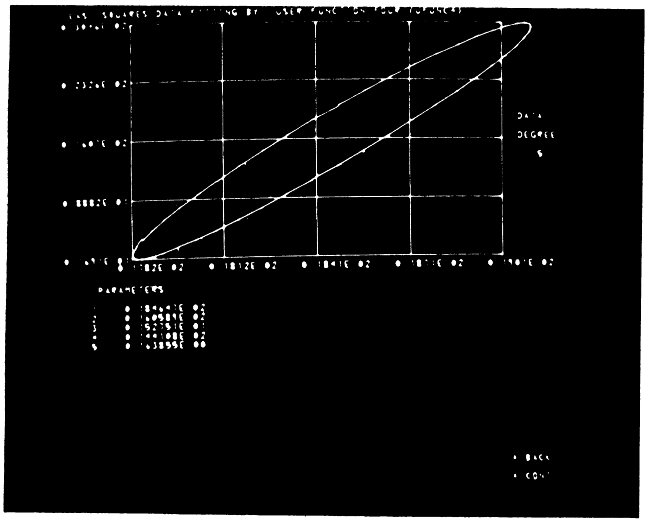

Fig. 3. Display of data, computed ellipse, and current parameter values during on-line least square ellipse fitting.

If the area of this polygon is not maximal, for N-sided polygons inscribed in the given ellipse, then by moving one of the points (x_n, y_n) for some n, the new area will be larger. Thus, if we can show that movement of a point (x_n, y_n) for any n causes the area to decrease, we have proved that the area is maximal for the given representation.

Without loss of generality, let us assume that the ellipse is centred at the origin, has its major axis of length $2a$ parallel to the x-axis and that the minor axis has length $2b$. Now let us perturb the nth point by a small amount and examine the change in area of the inscribed polygon. The change in area of the inscribed polygon due to moving the point (x_n, y_n) by a small amount (still between the $n-1$st and $n+1$st points) will be given by the change in area of the triangle formed by the $n-1$st, nth and $n+1$st points. Considering the base of the triangle as the chord connecting the $n-1$st and $n+1$st points the area of this triangle will be maximal when the perpendicular distance from the base to the nth point is maximal. Since the ellipse is convex, the position of the nth point giving maximum height to the triangle will occur where the tangent to the ellipse is parallel to the base. If this condition is satisfied at the nth point defined by (3), then a perturbation of the nth point will decrease the inscribed area proving that the chosen representation gives maximum inscribed area.

The slope of the chord between the $n-1$st and $n+1$st point is given by

$$\frac{y_{n+1}-y_{n-1}}{x_{n+1}-x_{n-1}} = \frac{b(\sin\phi_{n+1}-\sin\phi_{n-1})}{a(\cos\phi_{n+1}-\cos\phi_{n-1})}\ . \tag{4}$$

Noting that $\phi_{n=1} = \phi_n \pm 2\pi/N$ we see that the slope of the chord can be reduced to

$$\frac{y_{n+1}-y_{n-1}}{x_{n+1}-x_n} = -\frac{b\cos\phi_n}{a\sin\phi_n}, \tag{5}$$

but this is equal to the tangent at the nth point which is

$$\left.\frac{dy}{dx}\right|_{\phi_n} = -\frac{b\cos\phi_n}{a\sin\phi_n}\ . \tag{6}$$

This proves that the inscribed area is maximal for the chosen representation.

This proof that the inscribed area of an ellipse by the parametrically defined polygon is maximum involves a lemma which can be stated more generally as follows:

Lemma:
A polygon inscribed in a convex curve, contains maximum area if and only if for every three points adjacent in the polygon, the tangent at the middle point parallels the chord between the other two points.

4. Plotting hyperbolas and parabolas

The use of a parametric representation for on-line plotting can be carried over to hyperbolas and parabolas. We will consider these two curves separately.

Hyperbola
For simplicity, assume we would like to plot a hyperbola which is centred at the origin and whose axis is collinear with the x-axis as shown in Fig. 4. Furthermore, assume that the distance from the centre to a vertex is a and that the slopes of the asymptotes are $\pm b/a$. The rectangular coordinate representation of the hyperbola is then

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1\ . \tag{7}$$

A parametric representation of (11) is

$$x = \pm a \sec \phi \tag{8a}$$

$$y = \pm b \tan \phi. \tag{8b}$$

As ϕ varies continuously from 0 to $\pi/2$ the points given by (8) will trace the desired hyperbola.

As shown by the author, Smith (1969a), this representation does not satisfy the maximum inscribed area criterion as did the parametric ellipse representation. However, it is also shown by the author, Smith (1969a), that the area is 'nearly' maximum and that the algorithm to calculate the representative points is efficient in that the multiple angle trigonometric identities can be used to advantage.

A second parametric representation of (7) is (in the first quadrant)

$$x = a \cosh \phi \tag{9a}$$

$$y = b \sinh \phi. \tag{9b}$$

As ϕ varies continuously from 0 to ∞ the points given by equations (9) will also trace the desired hyperbola. This representation can also be efficiently calculated by making use of multiple argument identities for the hyperbolic functions.

The representation, equations (9), is preferable to equation (8) in that the maximum inscribed area criterion is satisfied. To see this, we apply the Lemma given above since the hyperbola is a convex curve. For equal increments, $\delta\phi$, of the parameter ϕ, three adjacent points are given by $\phi_n - \delta\phi$, ϕ_n and $\phi_n + \delta\phi$. The tangent at ϕ_n is given by

$$\left.\frac{dy}{dx}\right|_{\phi_n} = \frac{b \cosh \phi_n}{a \sinh \phi_n}. \tag{10}$$

The slope of the chord joining the other two points is given by

$$\frac{y_{n+1} - y_{n-1}}{x_{n+1} - x_{n-1}} = \frac{b}{a}\left[\frac{\sinh(\phi_n + \delta\phi) - \sinh(\phi_n - \delta\phi)}{\cosh(\phi_n + \delta\phi) - \cosh(\phi_n - \delta\phi)}\right]$$

$$= \frac{b \cosh \phi_n}{a \sinh \phi_n} \tag{11}$$

Since the tangent at ϕ_n and the slope of the chord are equal, we have by the Lemma that the inscribed area is maximal.

The hyperbola differs from the ellipse, however, in that it is not a closed curve of finite length. Therefore a decision must be made as to how much of the hyperbola is to be displayed. If this decision can be made in terms of the desired range of one coordinate, then an upper limit on ϕ for plotting purposes can be specified. For example, if we consider the branch of the hyperbola in the first and fourth quadrants and we would like to plot for values of x in the interval $[a, a + c]$, the upper limit, ϕ_{max}, on the parameter ϕ, would be $\phi_{max} = \cosh^{-1}[(a + c)/a]$. Similarly, other limits could be specified.

Once an upper limit, ϕ_{max}, has been specified, the calculation of the representative points, $\{x_n, y_n\}_{n=1}^{N}$, in the first quadrant can proceed as follows:

Initialise:

$$\begin{aligned} d\phi &= \phi_{max}/(N - 1) \\ A &= \cosh(d\phi) \\ B &= (a/b) * \sinh(d\phi) \\ C &= (b/a) * \sinh(d\phi) \\ x_1 &= a \\ y_1 &= 0. \end{aligned}$$

Repeat the following for $n = 2, 3, \ldots, N$.

$$x_n = A * x_{n-1} + B * y_{n-1}$$

$$y_n = C * x_{n-1} + A * y_{n-1}.$$

This will give N points representing the portion of the hyperbola appearing in the first quadrant. The other points may be obtained from these by appropriate changes of sign due to the symmetry involved. For hyperbolas with centres displaced from the origin, and/or other than horizontal axes, a simple change of coordinates will give the desired results.

Parabola

For plotting parabolas we can again make use of a parametric representation. For simplicity, consider the parabola shown in Fig. 5, centred at the origin and with horizontal axis.

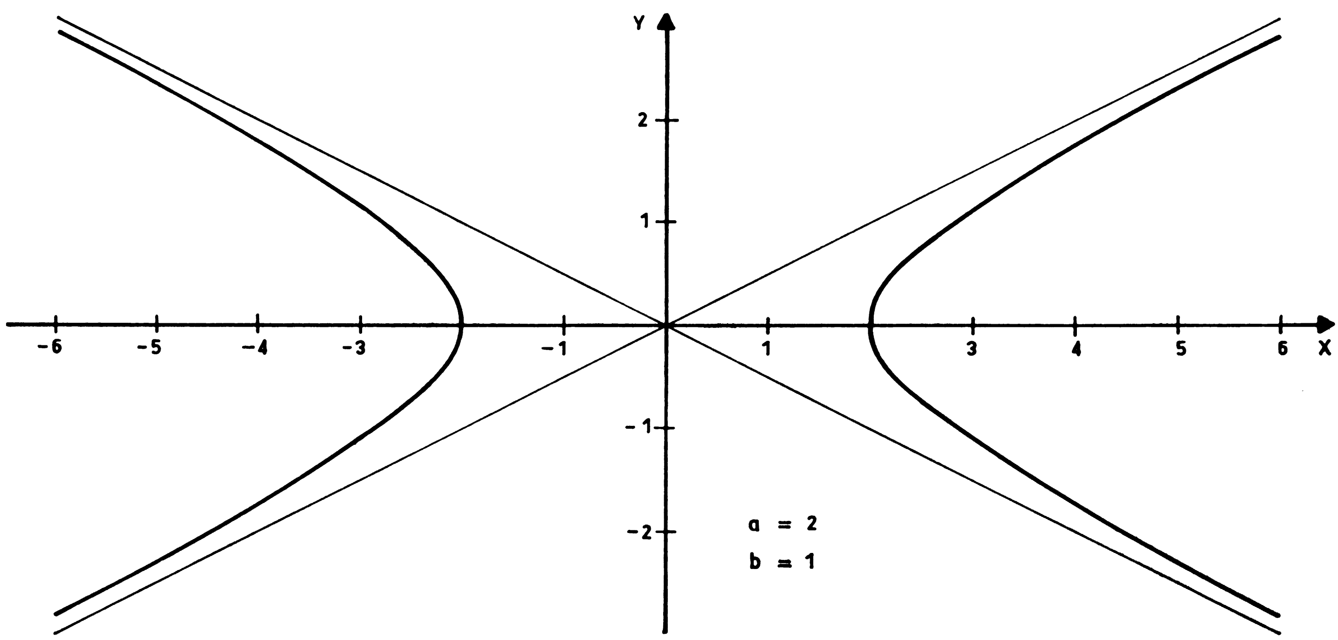

Fig. 4. Hyperbola centred at origin with horizontal axis.

A parabola such as that shown in Fig. 5 can be expressed in rectangular coordinates as

$$y^2 = 4ax. \tag{12}$$

A parametric representation of (12) is

$$x = \tan^2 \phi \tag{13a}$$

$$y = \pm 2\sqrt{a} \tan \phi. \tag{13b}$$

As ϕ varies continuously from 0 to $\pi/2$ the points given by equations (13) will trace the desired parabola.

As shown by the author, Smith (1969a), this representation leads to a 'nearly' maximal inscribed area and an efficient computational algorithm. A second parametric representation, however, can be used to give maximal inscribed area and be even more computationally efficient. This second representation is

$$x = a\phi^2 \tag{14a}$$

$$y = 2a\phi. \tag{14b}$$

As ϕ varies continuously from 0 to ∞ the points given by equations (14) will trace the desired parabola.

To see that the representation given by equations (14) satisfies the maximum inscribed area criterion we can again make use of the Lemma since the parabola is a convex curve. Assuming equal increments, $\delta\phi$, of the parameter, three adjacent points are given by $\phi_n - \delta\phi$, ϕ_n, and $\delta_n + \phi$. The tangent at ϕ_n is given by

$$\left.\frac{dy}{dx}\right|_{\phi_n} = \frac{2a}{2a\phi_n} = \frac{1}{\phi_n}\,. \tag{15}$$

The slope of the chord joining the other two points is given by

$$\frac{y_{n+1} - y_{n-1}}{x_{n+1} - x_{n-1}} = \frac{2a(\phi_n + \delta\phi - \phi_n + \delta\phi)}{a[(\phi_n + \delta\phi)^2 - (\phi_n - \delta\phi)^2]} = \frac{1}{\phi_n}\,. \tag{16}$$

Since the tangent at ϕ_n and the slope of the chord are equal, we have by the Lemma that the inscribed area is maximal.

To calculate N representative points for the first quadrant we must, as we did for the hyperbola, establish an upper limit, ϕ_{max}, on ϕ. Once ϕ_{max} has been given, the algorithm to compute N points for the first quadrant proceeds as follows:

Initialise:

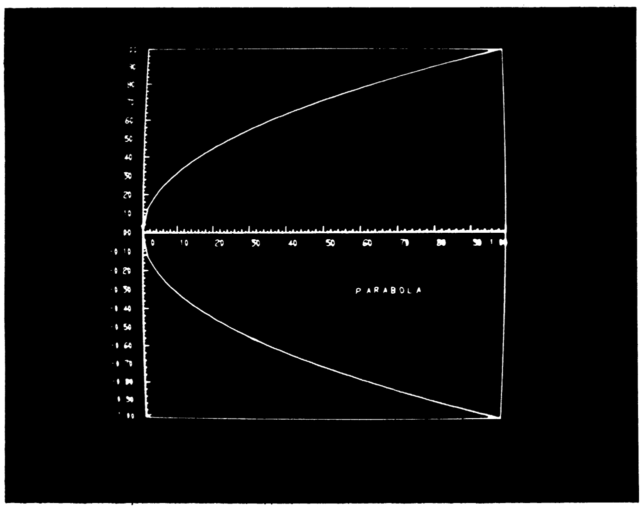

Fig. 5. Parabola centred at origin with horizontal axis.

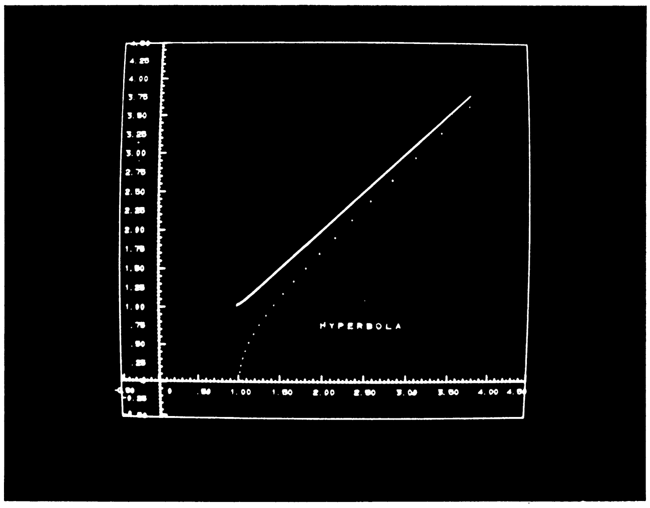

Fig. 6. Hyperbola by parametric representation.

$$
\begin{aligned}
d\phi &= \phi_{max}/(N - 1)\\
x_1 &= 0\\
y_1 &= 0\\
A &= a * d\phi * d\phi\\
B &= 2 * a * d\phi.
\end{aligned}
$$

Repeat the following for $n = 2, 3, \ldots, N$.

$$
\begin{aligned}
x_n &= A + x_{n-1} + d\phi * y_{n-1}\\
y_n &= B + y_{n-1}.
\end{aligned}
$$

This will give N points in the first quadrant. The fourth quadrant points can be obtained by negating the values of $y_n(n = 1, \ldots, N)$. Again, parabolas with displaced centres and other than horizontal axes can be handled by a change of coordinates.

Fig. 6 and Fig. 7 are on-line point plots of a hyperbola and a parabola (first quadrant) using the methods recommended in this report. These pictures were created by using the GAMMA* (*G*raphically *A*ided *M*athematical *MA*chine)

*GAMMA is based on the Culler-Fried system currently in use at the University of California at Santa Barbara.

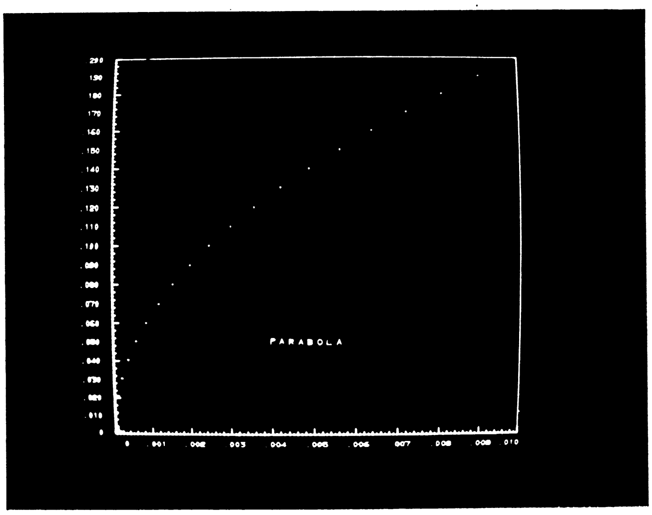

Fig. 7. Parabola by parametric representation.

on-line graphical mathematical system at CERN (see Vandoni, 1969).

5. Summary

Efficient algorithms for calculating a fixed number of representative points for ellipses, hyperbolas, and parabolas have been given. These are based on parametric representations of the curves with the parameters being varied by equal increments. The nature of the representation, and the use of equal increments in the parameter, lead to very efficient algorithms.

For the ellipse we have proved that the parametric representation is best in the sense of maximum inscribed area. This led us to a Lemma which was used to prove that for the hyperbola and parabola the parametric representations are also best in the same sense.

6. Acknowledgements

An earlier version of this paper, Smith (1969a), was based on the trigonometric parametric representations mentioned here. D. Wiskott (of CERN) has pointed out that these representations are, in fact, all related to the trigonometric representation of a circle by linear projective transformations. That paper also contained a complicated proof of the maximum inscribed area property of the parametric ellipse representation. The referee pointed out that by using the Lemma, given here in Section 3, the proof could be quite simple. The referee also pointed out the parametric representations for the hyperbola and ellipse which satisfy the maximum inscribed area criterion. For the Lemma, the hyperbola and ellipse representations, and his other valuable comments, my grateful thanks go to the referee.

References

Botting, R. J., and Pitteway, M. L. V. (1968). Algorithm for drawing ellipses or hyperbolas with a digital plotter (Letter to the Editor), *The Computer Journal*, Vol. 11, p. 120.

Partridge, M. F. (1968). Algorithm for drawing ellipses or hyperbolas with a digital plotter (Letter to the Editor), *The Computer Journal*, Vol. 11, pp. 119-120.

Pitteway, M. L. V. (1967). Algorithm for drawing ellipses or hyperbolas with a digital plotter, *The Computer Journal*, Vol. 10, pp. 282-289.

Smith, L. B. (1969). The use of man-machine interaction in data-fitting problems, (Ph.D. thesis), SLAC report No. 96, Stanford Linear Accelerator Center, Stanford, California.

Smith, L. B. (1969a). Drawing ellipses, hyperbolas or parabolas with a fixed number of points, CERN/DD/DH/69/9, CERN, DD Division, Geneva, Switzerland.

Vandoni, C. E. (1969). GAMMA User's Manual. CERN, DD Division, Geneva, Switzerland.

Correspondence

To the Editor
The Computer Journal

Sir,

'This subprogram operates on a dissimilarity coefficient to generate the clusters of the numerically stratified hierarchy (dendrogram) specified by the single link method.'

'Providing alternatives to the go to statement is one of the distinguishing features of high level programming languages.'

Both these quotations were taken from Vol. 13, No. 3 of *The Computer Journal*. Neither of them is written in English. The first, I submit, and regret as it comes from Cambridge, is undisciplined jargonese. The second is transformed into English more simply, by putting the words 'go to' in, say, inverted commas or italics.

This edition of the *Journal* provides for authors' notes on submission of papers. Nowhere therein is to be found any reference to quality of writing. I accept, of course, that specialised vocabularies must develop, and that to avoid them entirely would be possible only at the cost of excessive demands on space. Nevertheless, I cannot accept that vocabularies and jargon should be allowed to become so undisciplined and over-specialised that communication between intelligent people working in broadly the same areas should reach the point of total inhibition; or so time-consuming for others specialised elsewhere that it is barely attempted, and we reach at last such fragmentation that only a handful of people can convey meaning, one to another. Mathematical authors will find by going back to the original works of the great masters, like Euler, that they were capable of expressing themselves clearly, and took no great pride in avoidable obscurity. Bertrand Russell dealt with very complex philosophical concepts, yet his writings are always models of clarity.

I charge the Editorial Board with failing in their duty. Admiral of the Fleet, Earl Mountbatten of Burma, has made broadly similar comments. Can we not arrest this drift now?

Yours faithfully,

R. L. Allen (Brigadier)

Inventory Systems Development Wing
Headquarters, Base Organisation RAOC
Vauxhall Barracks
Didcot, Berkshire
20 August 1970