APOLLONIUS' ELLIPSE AND EVOLUTE REVISITED

FREDERICK HARTMANN and ROBERT JANTZEN

1. Introduction

The problem of finding the distance from a point to a curve is a standard exercise in calculus. If the curve is differentiable and has no endpoints, the connecting line segment from the given point to the nearest (farthest) point on the curve must be perpendicular to the tangent line there. One can therefore ask the more general question of how many such points on the curve have this geometric property, i.e. how many "normal line segments" can be drawn from the given point to the curve. If one adopts a parametrization of the curve, then these points correspond to critical points of the distance function.

Since a continuous real-valued function on a closed interval always has a maximum and minimum value, there are at least two such normals for any simple differentiable closed curve like the ellipse. For example, if the curve is a circle and the point is not at the center of the circle, then precisely two normal line segments can be drawn (in the line through the given point and the center), while if the point is at the center of the circle an infinite number of normals can be drawn. Similarly if the curve is a proper ellipse (not a circle) and the point is at the center of the ellipse, then clearly four normals can be drawn to the center from the points on the ellipse lying on its symmetry axes. It is natural to ask how does the number of such normals vary as the given point moves away from the center.

The question of determining the minimum and if it exists, the maximum distance to a conic section was addressed and answered some twenty-two hundred years ago by Apollonius of Perga, "The Great Geometer." Apollonius is most famous for his *Conics* series which originally consisted of eight Books. Only Books I–IV survive in

Date: September 23, 2004.

1991 *Mathematics Subject Classification.* Primary: 51-03, 01-01; Secondary: 01A20.

Key words and phrases. Ellipse, evolute.

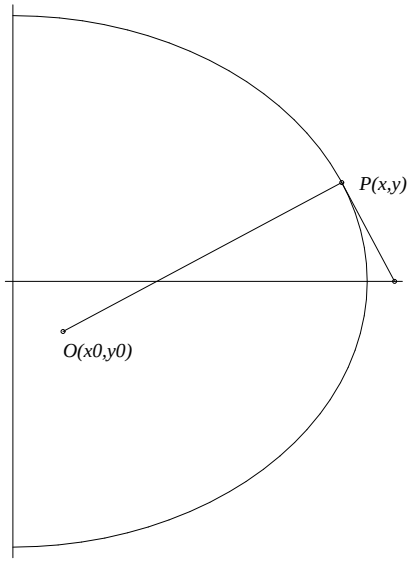

FIGURE 1. The normal line segment from the given point O and the point P together with the tangent line segment from P to a point on the major axis, a configuration whose geometry led Apollonius to the hyperbola whose intersection with the ellipse determines its normal lines.

the original Greek [3], but Books I–VII of the *Conics* exist in an Arabic translation [6]. Apollonius used the "normal" approach in addressing this problem in Book V, and although his original work apparently contains the resolution of the problem, both his proofs and results are very difficult to follow even in their annotated translated form [6, 7]. He lacked even an appropriate mathematical language to discuss quadratic relationships, which makes his results all the more remarkable as well as difficult to translate into modern notation. His approach to the problem relies on the determination of the evolutes of the conics whose equations "can be easily deduced from the results obtained by Apollonius ... and it is a veritable geometric tour de force" [2, p.159].

The purpose of this paper is to develop Apollonius' results using present day mathematics suitable for a second year college mathematics student in a way that demonstrates the geometry through visualization while avoiding excessive formula

manipulation that computer algebra systems are well suited to handle. The approach used here is based on a remark made without detail by Heath in [1, p.cxxvii]. In addition we connect the results to an application of the discriminant of a quartic polynomial, illustrating the no longer well known fact that a familiar and widely used idea for the quadratic case extends to the less familiar cubics and quartics. In the following we restrict our attention to the ellipse and note that a similar approach would work for the remaining conics, leaving some doable problems for similar exploration by interested advanced undergraduates based on our Maple worksheet, where the otherwise tedious details are made manageable even for such students.

2. The Geometry

For a given ellipse

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

and an arbitrary point $O(x_0, y_0)$ in the plane, depicted in Figure 1, the equation of the normal line through this point from the "foot" $P(x, y)$ of that normal line is easily found. By implicit differentiation, the slope of the tangent line is $dy/dx = -(b^2x)/(a^2y)$, whose negative reciprocal is the slope of the normal line

$$y_0 - y = (a^2y)/(b^2x)\,(x_0 - x)\ . \tag{2}$$

Clearing fractions leads to the equation of the hyperbola of Apollonius

$$xy(a^2 - b^2) - x_0a^2y + y_0b^2x = 0\,. \tag{3}$$

Thus the "feet" of the normals are the intersections of the ellipse with a rectangular hyperbola which passes through both the origin and the point $O(x_0, y_0)$ and whose axes are rotated by an angle $\pi/4$ with respect to the coordinate axes. The problem of counting the normals is reduced to finding the number of such intersections as $O(x_0, y_0)$ moves around in the plane. Figure 2 illustrates the situation where O is sufficiently close to the center of the ellipse and in particular inside the evolute, which is shown in the figure and explained in the next section.

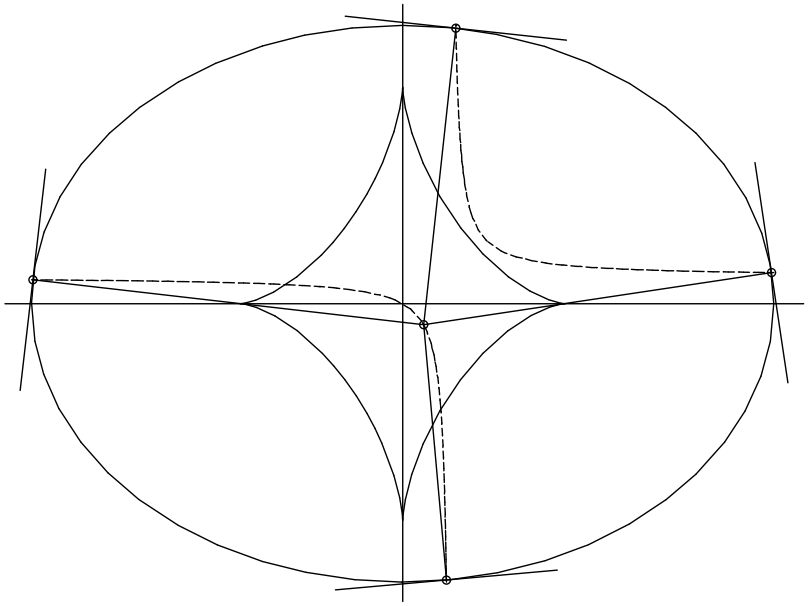

FIGURE 2. For a point just slightly offcenter, the four feet of the normals to the ellipse shift slightly from the center point configuration, where the hyperbola of Apollonius degenerates to its vertical and horizontal asymptotes along the axes of the ellipse.

For simplicity we assume $a > b$ throughout our discussion so that the major axis is the x-axis. Note that the focal distance of the foci from the origin along the x-axis is then $c = (a^2 - b^2)^{1/2}$.

If the point O is at the origin, the hyperbola (3) reduces to the coordinate axes and there are four intersection points as already noted above. If O is on the major axis at a distance from the origin just smaller than the major semi-axis, there are only two intersection points, the endpoints of the major axis of the ellipse. Is there a transition from four intersections to just two?

One may begin to investigate this question by moving the point $O(x_0, y_0)$ along a ray from the origin at a fixed nonzero angle θ by letting $(x_0, y_0) = (r\cos\theta, r\sin\theta)$ and varying the distance r from the origin. For example, in Fig. 2, one can continue moving the point outward from the origin along the line $\theta = -\pi/4$. One can put the equation of the hyperbola (3) into the standard form $XY = \pm a_h^2/2$ by translating its center to the origin through the substitution $(x, y) = (X + x_1, Y + y_1)$ and

setting the linear terms in X and Y to zero to determine its center

$$C(x_1, y_1) = C(a^2 x_0/c^2, -b^2 y_0/c^2) = C(r(a^2/c^2)\cos\theta, -r(b^2/c^2)\sin\theta) \ .$$

Its vertices lie on its major axis $X = \pm Y$, with their distance from the center equalling $\sqrt{2}$ times the absolute value of the coordinates of the intersection with this axis, namely

$$a_h = \frac{ab}{c}|2x_0 y_0|^{1/2} = r\frac{ab}{c}|\sin 2\theta|^{1/2} \ .$$

As r increases, this center moves along the line through the origin on the opposite side of the x-axis from the original ray and whose slope is the ratio of the coefficients of r in the coordinates of the center

$$y = -x\frac{b^2}{a^2}\tan\theta \ ,$$

and therefore makes a smaller angle with the x-axis than the original ray.

Thus the center of the hyperbola moves outward along the second line and its vertices expand outward from it by distances all proportional to r as r increases until the one branch of the hyperbola not passing through the origin moves out of the ellipse and its two intersections with the ellipse degenerate to one and then none. The other branch through the origin must always intersect the ellipse in exactly two points. Thus the four intersections of the hyperbola locating the feet of the normals passing through O degenerate to three when the hyperbola becomes tangent to the ellipse (see Fig. 5), and then two when it moves outside. In the next section we will show exactly where this transition from four to two normals occurs.

3. The evolute

The evolute of an ellipse may be defined in terms of the curvature at a point on the ellipse. Suppose that the ellipse is parameterized by $\vec{r}(t) =< a\cos(t), b\sin(t) >$, $0 \le t < 2\pi$. The curvature $\kappa(t)$ at $\vec{r}(t)$ can be evaluated using the standard calculus formula in parametric form

$$\kappa(t) = \frac{|(-a\cos(t))(b\cos(t)) - (-b\sin(t))(-a\sin(t))|}{[a^2\sin^2(t) + b^2\cos^2(t)]^{3/2}}$$

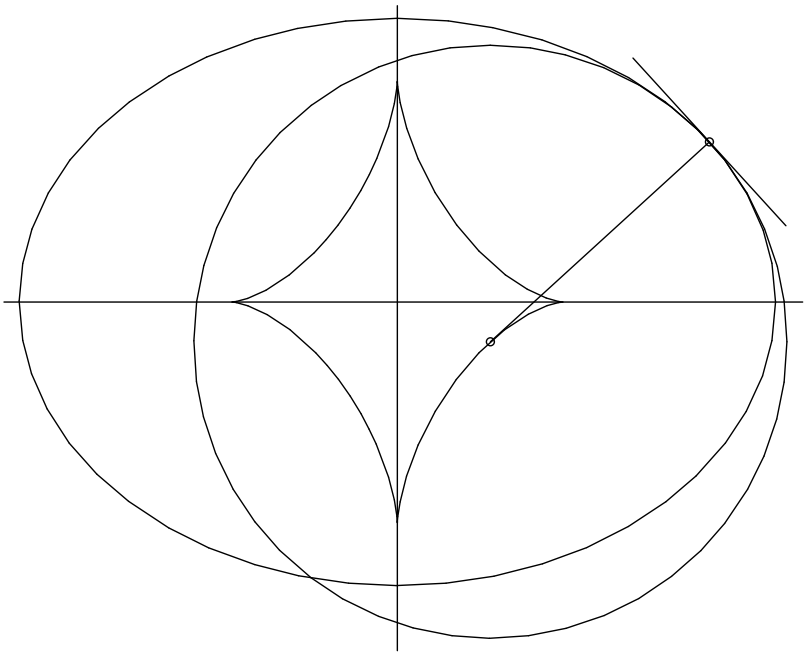

FIGURE 3. The center of the osculating circle traces out the evolute of the ellipse.

and from it one obtains the radius of curvature, $R(t) = 1/\kappa(t)$. By definition the center of the osculating circle is located a distance $R(t)$ along the inward pointing unit normal $\vec{N}(t)$ from its point of origin on the ellipse, so the position vector of this center is simply $\vec{C}(t) = \vec{r}(t) + R(t)\vec{N}(t)$. By direct calculation and simplification one finds

$$\vec{C}(t) =< \frac{a^2 - b^2}{a}\cos^3(t), \frac{b^2 - a^2}{b}\sin^3(t) > \ . \tag{4}$$

If one adopts the definition that the evolute is the locus of the centers of the osculating circles, then for the ellipse it is this parametrized curve $\vec{C}(t)$ as t varies from 0 to 2π. Since the focal distance from the center along the major axis is $c = \sqrt{a^2 - b^2}$, the two cusps of the evolute at a distance $c^2/a = c(c/a) < c < a$ along that axis must fall short of the foci inside the ellipse, while the other two cusps exit the ellipse along the minor axis when $c^2/b > b$ or $a > \sqrt{2}b$.

Figure 3 shows the osculating circle and the normal and tangent lines for a point in the first quadrant. Notice that the normal line to the ellipse is a tangent line to

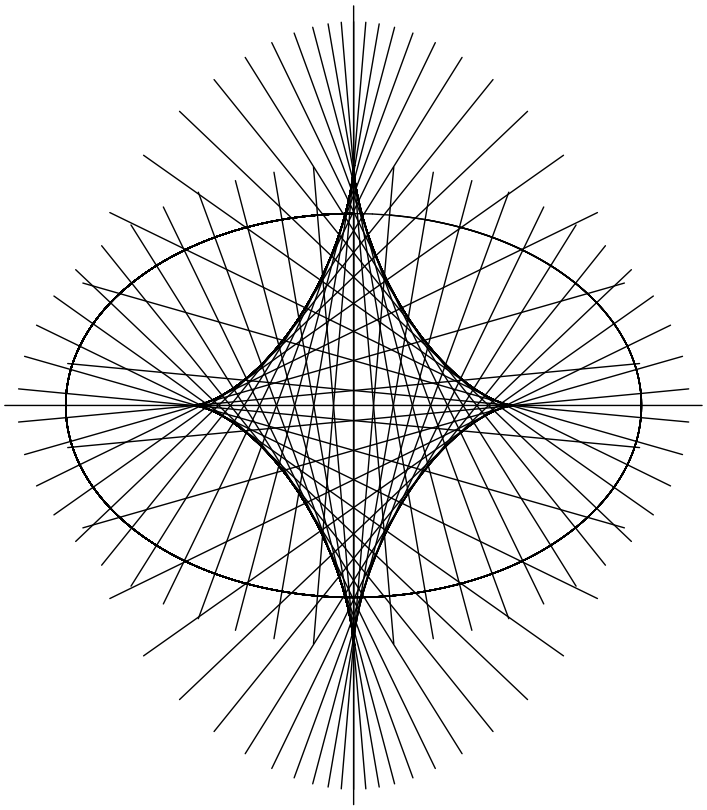

FIGURE 4. The evolute as the envelope of the normals to the ellipse.

its evolute, a property which leads to an alternative way to define the evolute of a curve. In fact the center of the osculating circle may also be thought of in the following way. For a fixed point P on the curve, one may construct two normals to the curve, one at P and another at a nearby point on the curve. The center of the osculating circle is the *limit* of the intersection of these two normals as the nearby point approaches P. The evolute is thus the *envelope* of the curve's normals, illustrated in Figure 4 for a case in which the eccentricity of the ellipse is sufficient to cause the evolute to extend outside the ellipse along the minor axes. Hence if one is looking for the points where two normals coalesce into a single normal leading to a change in the number of normals passing through those points, one would look at the evolute. This was Apollonius' approach to the problem.

One can also derive the equation of the evolute from this alternative idea as the curve traced out by the the limiting intersection point of two successive normals to the curve, which operationally defines the envelope of the normal lines. One can write down the equation of the normal line to the parametrized ellipse, and then its neighbor at a nearby point on the ellipse, find their intersection, and then take the limit as the nearby point approaches the original one. Even without direct computation one easily sees that the unnormalized normal vector $\vec{n}(t) =< b\cos(t), a\sin(t) >$ is orthogonal to the tangent vector $\vec{r}'(t) =< -a\sin(t), b\cos(t) >$ of the parametrized position vector $\vec{r}(t) =< a\cos(t), b\sin(t) >$, so the slope of the normal line is $a\sin(t)/(b\cos(t))$. From the point-slope equation of the normal line with $r(t)$ as the point, one therefore has

$$\frac{y - b\sin(t)}{x - a\cos(t)} = \frac{a\sin(t)}{b\cos(t)}\,.$$

Then finding the intersection of this line and the corresponding line for $t + \Delta t$, and taking the limit as $\Delta t \to 0$ yields exactly the above parametrization (5) of the evolute. The algebra is very tedious and lengthy, but with a computer algebra system, not only can one effortlessly evaluate the result but one can organize the steps in the calculation in such a way that with hindsight one can reconstruct a derivation by hand to understand how this final result comes about.

Let the ellipse and its evolute be parametrized as above and let the point $O(x_0, y_0)$ be on the evolute, i.e.

$$(x_0, y_0) = (\frac{a^2 - b^2}{a}\cos^3(t), \frac{b^2 - a^2}{b}\sin^3(t))\,. \tag{5}$$

Then the equation of the intersecting hyperbola of Apollonius becomes

$$xy - a\cos^3(t)y - b\sin^3(t)x = 0 \tag{6}$$

and it is immediately seen that the corresponding point $P(a\cos(t), b\sin(t))$ on the ellipse lies on this hyperbola. It is straightforward to show that at P, the ellipse and the hyperbola also meet tangentially. (See Figures 5 and 6.) For points $O(x_0, y_0)$ everywhere outside the evolute, only one branch of the hyperbola intersects the ellipse and in exactly two distinct points, so there are only two normals.

To summarize:

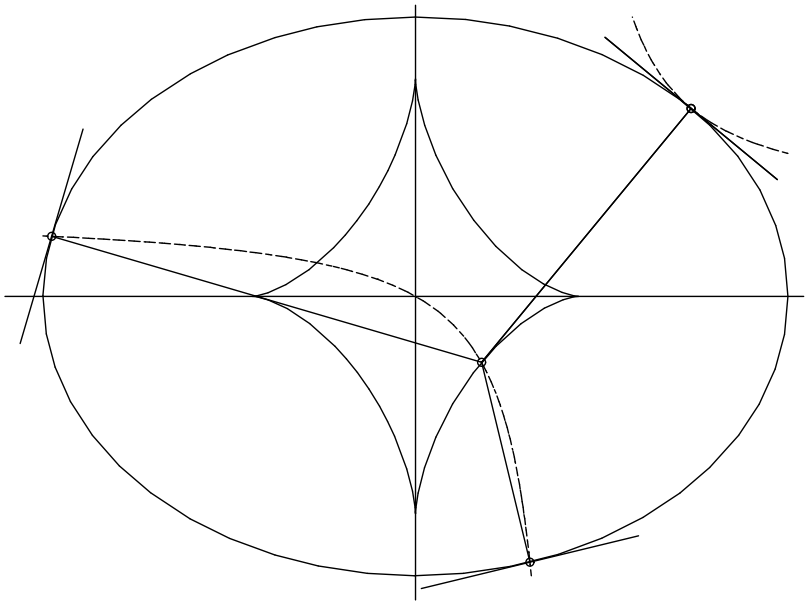

FIGURE 5. As the point approaches the evolute from inside (not at one of its vertices), two of the four normals coallesce as one branch of the hyperbola of Apollonius becomes tangent to the ellipse, in the first quadrant in the configuration shown here.

Theorem. *If a point lies within the evolute of an ellipse, then four distinct normals can be drawn to the ellipse. If the point lies on the evolute, but not at a cusp, then precisely three normals can be drawn, and if the point is at a cusp or outside the evolute only two normals can be drawn.*

4. The Discriminant of the Related Quartic

One may approach the problem in a more algebraic way. If the arbitrary point within the ellipse is $O(x_0, y_0)$ and a point on the ellipse is given parametrically by $P(a\cos(t), b\sin(t))$, then the desired geometric orthogonality condition is that the vector $\overrightarrow{OP}$ be perpendicular to the tangent vector $\vec{T}$ to the ellipse at P or

$$< a\cos(t) - x_0, b\sin(t) - y_0 > \cdot < -a\sin(t), b\cos(t) >= 0\,. \tag{7}$$

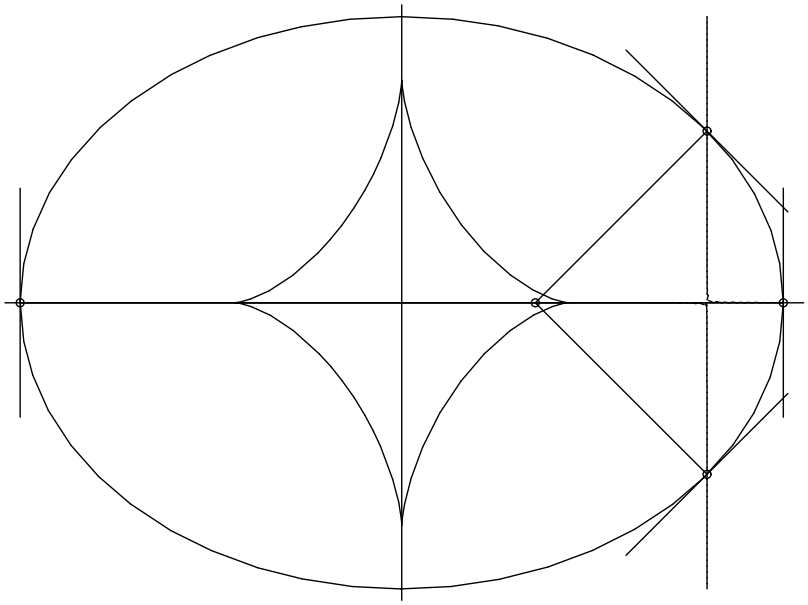

FIGURE 6. As the point approaches one of the axes inside the evolute, the hyperbola of Apollonius degenerates to its pair of horizontal and vertical asympototes, one coinciding with this axis. As the point on the axis approaches a cusp of the evolute from inside along this axis, the other asymptote becomes tangent to the ellipse (the vertical asymptote here), and three of the four normals from the given point (at the right of the ellipse here) coalesce into one (not shown).

Simplifying this condition (7) and eliminating $\sin(t)$ from it gives the location of the normal points as solutions to the quartic equation in $\cos(t)$:

$$(8)\quad (a^2-b^2)^2\cos^4(t) - 2ax_0(a^2-b^2)\cos^3(t) + (a^2x_0^2 + b^2y_0^2 - (a^2-b^2)^2)\cos^2(t) + 2ax_0(a^2-b^2)\cos(t) - a^2x_0^2 = 0\,.$$

This is equivalent to a quartic equation in $x(t) = a\cos(t)$, which in turn is the result of using the equation of the hyperbola of Apollonius to eliminate y from the pair of quadratic equations for the hyperbola and ellipse.

Older books on the theory of equations discuss the nature of the roots of third and fourth degree polynomials, e.g. [4, 5]. They develop the theory of the discriminant of the cubic and quartic, generalizing what every student knows about

quadratic equations, and which is conveniently programmed into modern computer algebra systems. In fact the discriminant can be defined as a product of the squared differences of all the distinct pairs of roots of the polynomial, modulo a normalizing constant, so it is zero precisely when there are multiple roots of the polynomial and nonzero otherwise, while its sign is correlated with the number of real roots.

If $q(x) = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0,\ a_4 \neq 0$ is an arbitrary quartic with real coefficients, then one may define its discriminant in terms of these coefficients, the vanishing of which implies the existence of a multiple root as in the quadratic case. The discriminant of q is given by the formula

$$\begin{aligned}(9)\qquad \Delta(a_4, a_3, a_2, a_1, a_0) &= 4(\frac{a_2^2}{3} + 4a_0a_4 - a_1a_3)^3 \\ &\quad - 27(-a_1^2a_4 - a_0a_3^2 - \frac{2a_2^3}{27} + \frac{a_1a_2a_3}{3} + \frac{8a_0a_2a_4}{3})^2\,.\end{aligned}$$

In our case from (8) we have

$$\begin{aligned}(10)\qquad a_4 &= (a^2 - b^2)^2\,,\ a_3 = 2ax_0(a^2 - b^2) = -a_1\,, \\ a_2 &= a^2x_0^2 + b^2y_0^2 - (a^2 - b^2)^2\,,\ a_0 = -a^2x_0^2\,.\end{aligned}$$

Substituting (10) into (9) and imposing the condition (5) that (x_0, y_0) is on the evolute, one may show that $\Delta = 0$ for all $0 \leq t < 2\pi$. Thus for these points there are either three normals from the point or two when the point is a cusp. One can also show that except for the discriminant vanishing on the axes, where the cosine must have at least a repeated root by symmetry, one has $\Delta < 0$ for points outside the evolute, implying two distinct real roots and hence two normals, and $\Delta > 0$ for points within the evolute, implying either four distinct real roots (and hence four normals) or two distinct pair of complex conjugage roots. However, the latter would imply that there are no normals, but as noted in the introduction there are always at least two so this cannot happen.

The reader interested in seeing more details of this development and a chance to investigate this problem or similar ones may download a Maple worksheet at

http://www.math.villanova.edu/archives/maple/misc/ellipse/
or simply view the worksheet in web page format. This is a beautiful example of how, empowered by a computer algebra system, one can follow one's nose in uncovering elegant mathematical structure that would otherwise be unreachable in practice. It is also a useful lesson on the increasing importance of the use of computer algebra systems in doing mathematics.

Acknowledgements

An anonymous referee is thanked for improving our presentation of these results.

References

[1] Heath, T. L., *Apollonius of Perga: Treatise on Conic Sections*, W. Heffer & Sons Ltd., Cambridge, U.K., 1896

[2] Heath, T. L., *A History of Greek Mathematics*, Vol. II, Clarendon Press, Oxford, U.K., 1921

[3] Hogendijk, J. P., *Arabic Traces of the Lost Works of Apollonius*, Archives for History of Exact Sciences, **35** (1986) 187–253

[4] MacDuffee, C. C., *Theory of Equations*, John Wiley & Sons, New York, 1954

[5] Turnbull, H. W., *Theory of Equations*, Oliver and Boyd, Edinburgh, 1952

[6] Toomer, G. J., *Apollonius: Conics, Books V to VII, The Arabic Translations of the Lost Greek Originals in the Version of Banū Mūsā*, Vol. I, Springer-Verlag, New York, 1990

[7] Zeuthen, H. G., *Die Lehre von der Kegelschnitten im Alterum*, Verlag von Andr. Fred. Höst & Sohn, Kopenhagen, 1886

Department of Mathematical Sciences, Villanova University, Villanova, PA 19085

E-mail address: frederick.hartmann@villanova.edu, robert.jantzen@villanova.edu