4.B. Tangent and normal lines to conics

Apollonius' work on conics includes a study of tangent and normal lines to these curves. The purpose of this document is to relate his approaches to the modern viewpoints based upon analytic geometry and differential calculus.

Classification of conics up to congruence

Whenever one works with coordinate geometry, it is clearly useful to choose coordinate axes so that the algebraic equations defining the objects of study are as simple as possible. Geometrically, this essentially corresponds to the physical idea of moving the objects without changing their sizes or shapes. Linear algebra provides a powerful method for interpreting rigid motions in terms of coordinates. Specifically, if we define an abstract *isometry* of the coordinate plane $\mathbb{R}^2$ as a transformation $\mathbf{T}$ of the form

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \;=\; \begin{pmatrix} \cos\theta & -\varepsilon\sin\theta \\ \sin\theta & \varepsilon\cos\theta \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;+\; \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

where θ is some real number (often we assume $0 \leq \theta \leq 2\pi$), $\varepsilon = \pm 1$, and (b_1, b_2) is a fixed vector in $\mathbb{R}^2$, then two plane figures A and B are said to be *congruent* if there is an isometry $\mathbf{T}$ which maps A onto B. Additional comments on this concept appear on pages 66–73 (= document pages 10–17) of the following online document:

`http://math.ucr.edu/~res/math133/geometrynotes2b.pdf`

We are particularly interested in finding a short list of conics so that every such curve defined by an equation of the form

$$Ax_1^2 \;+\; 2Bx_1x_2 \;+\; Cx_2^2 \;+\; Dx_1 \;+\; Ex_2 \;+\; F \;=\; 0$$

is congruent to one of the curves on the list. We need to impose some restrictions on the coefficients in order to eliminate unwanted examples; for example, if $A = B = C = 0$, then we have a first degree equation, and clearly we should find a way to eliminate such cases. If we do this, then basic results from a second linear algebra course imply that every curve defined by such an equation is congruent to an example with one of the following forms in which all coefficients except possibly F are nonzero. In each case, we might as well assume that A is positive.

$$Ax_1^2 \;+\; F \;=\; 0$$

$$Ax_1^2 \;+\; Cx_2^2 \;+\; F \;=\; 0$$

$$Ax_1^2 \;+\; Ex_2 \;=\; 0$$

This is just a notational reformulation of a result (the Rigid Change of Coordinates Theorem) on page 80–81 of the following document:

`http://math.ucr.edu/~res/math132/linalgnotes.pdf`

As indicated in the table on page 82 of the same document, the first equation defines either a line or a pair of lines, and it is of no interest to us here. The third equation always defines a parabola. However, the solutions to an equation of the second type depend upon the signs of the various coefficients:

(*a*) If C is positive and $F \leq 0$, then the equation either defines a single point when $F = 0$ (the origin) or no points at all when $F < 0$, so this case is also of no interest to us.

(*b*) If C is positive and $F > 0$, then the equation defines and ellipse when $C \neq A$ and a circle when $C = A$.

(*c*) If C is negative and $F \neq 0$, then the equation defines a hyperbola.

(*d*) If C is negative and $F = 0$, then the equation defines a pair of lines which intersect at the origin, so this case is also of no interest to us.

Thus we see that every conic of interest to us is congruent to either a parabola, an ellipse (or circle), or a hyperbola given by a fairly standard equation, and as usual we may rewrite these equations as follows:

$$(\textbf{ELLIPSE}) \qquad \frac{x^2}{a^2} \, + \, \frac{y^2}{b^2} \;\; = \;\; 1 \qquad (a, \; b \; > \; 0)$$

$$(\textbf{HYPERBOLA}) \qquad \frac{x^2}{a^2} \, - \, \frac{y^2}{b^2} \;\; = \;\; 1 \qquad (a, \; b \; > \; 0)$$

$$(\textbf{PARABOLA}) \qquad y \;\; = \;\; ax^2 \qquad (a \neq 0)$$

Parametrizations of conics

The first point to note is that ellipses, hyperbolas and parabolas can all be represented by parametrized curves of the form

$$\mathbf{s}(t) \;\; = \;\; \big(x(t), \, y(t) \big)$$

where $x(t)$ and $y(t)$ are differentiable functions; strictly speaking, a hyperbola is given by two parametrized curves corresponding to its two branches, and we shall have to be a bit careful about this as we proceed. In any case, the tangent line to the curve at the point $\mathbf{s}(t_0)$ is represented parametrically by $\mathbf{s}(t_0) \, + \, u \, \mathbf{s}'(t_0)$, where $\mathbf{s}'(t_0)$ is given by taking the derivatives of the coordinate functions where $x(t)$ and $y(t)$. This parametric representation will define a line if $\mathbf{s}'(t_0) \neq (0,0)$, and we claim this is true for suitably defined parametrizations of ellipses, hyperbolas and parabolas.

Examples. **1.** For the ellipse given by

$$\frac{x^2}{a^2} \, + \, \frac{y^2}{b^2} \;\; = \;\; 1 \qquad (a, \; b \; > \; 0)$$

one has the parametrization $(a \cos t, b \sin t)$, and for each t_0 we have $\mathbf{s}'(t_0) = (-a \sin t_0, b \cos t_0)$. The right hand expression is never the zero vector because the sine and cosine functions are never simultaneously zero.

2. If we now consider the hyperbola given by

$$\frac{x^2}{a^2} \, - \, \frac{y^2}{b^2} \;\; = \;\; 1 \qquad (a, \; b \; > \; 0)$$

one has the parametrization $(\pm \, a \cosh t, b \sinh t)$ in terms of the hyperbolic functions $\sinh u$ and $\cosh u$; as noted before, this curve has two branches, and the $\pm$ signs correspond to parametrizations of the pieces of the curve in the half-planes $x \geq 0$ and $x \leq 0$. For each t_0 we have $\mathbf{s}'(t_0) =$

$(\pm\, a \sinh t_0, b \cosh t_0)$. The right hand expression is never the zero vector because the hyperbolic cosine function is never zero (recall that $\cosh^2 u - \sinh^2 u = 1$).

3. Finally, for the the parabola $y = ax^2$ (where $a \neq 0$) we have the graph parametrization $\mathbf{s}(t) = (t, at^2)$; in this case $\mathbf{s}'(t_0) = (1, 2at_0)$ and the latter is never zero because its first coordinate is always nonzero.

Tangent lines to conics

Apollonius' approach to tangent lines for conics is typical of Greek geometry, and it is based upon the fact that such lines meet the curves only at the point of contact. This viewpoint is evident in elementary geometry, where one views a tangent line to a circle as a line which meets the circle in exactly one point. Experimetation with ellipses, parabolas and hyperbolas suggests that the same is true for these curves. The purpose of this section is to prove this property using the definition of tangent lines from calculus.

Our discussion of tangent lines will require the concept of the two open half-planes determined by the line (less formally, the two **sides** of the line). Specifically, if the line is given by an equation of the form $f(x, y) = Lx + My + n = 0$ where $(L, M) \neq (0, 0)$, then two points (u_0, v_0) and (u_1, v_1) not on the line are said to lie on the same half-plane (or side or the line) provided $f(u_0, v_0)$ and $f(u_1, v_1)$ are either both positive or both negative; the points are said to lie on opposite half-planes or sides if one is positive and the other is negative. For example, the two sides of the x-axis are defined by the inequalities $y > 0$ and $y < 0$.

In our discussion below, we shall need to know that an isometry has the following two properties:

(1) If $\mathbf{T}$ is an isometry and $\mathcal{L}$ is a line which determines half-planes $\mathcal{H}_+$ and $\mathcal{H}_-$, then the image of $\mathcal{L}$ under $\mathbf{T}$ is a line, say $\mathcal{M}$, and $\mathbf{T}$ maps $\mathcal{H}_+$ and $\mathcal{H}_-$ onto the two half-planes determined by $\mathcal{M}$.

(2) If $\mathbf{T}$ is an isometry and $\mathbf{s}(t)$ is a continuous parametrized curve such that $\mathbf{s}'(t_0)$ is defined, then the composite $\mathbf{T} \circ \mathbf{s}(t)$ also has these properties, and in fact $(\mathbf{T} \circ \mathbf{s})'(t_0)$ is equal to $\mathbf{T}\big(\mathbf{s}'(t_0)\big)$.

Both proofs are fairly straighforward and left to the reader.∎

Here is the result implying that all for each point $\mathbf{p}$ of a conic curve, all the remaining points lie on one side of the tangent line at $\mathbf{p}$; strictly speaking, one must modify this slightly in the hyperbolic case so that the conclusion only implies to the unique branch of the hyperbola on which $\mathbf{p}$ lies.

THEOREM. *Suppose that Γ is either a parabola, a circle, an ellipse, or a branch of a hyperbola. Let $\mathbf{p}$ be a point of Γ, and let $\mathcal{L}$ be the tangent line to Γ at $\mathbf{p}$. Then all points of Γ except $\mathbf{p}$ lie on* ONE *of the half-planes determined by $\mathcal{L}$.*

Proof. The argument splits into cases depending upon the type of the conic. Since tangents and half-planes are preserved under isometries, it is enough to prove the results for the standard examples of conics described above.

Suppose first that Γ is a parabola; we shall assume that in the equation $y = ax^2$ the coefficient a is positive; the other case will follow by the change of variables $u = -x,\ v = y$. Let (c, ac^2) be a point on the parabola, so that the tangent line is defined by the equation $y - ac^2 = 2ac(x - c)$. We claim that all other points on the parabola belong to the half-plane defined by the inequality

$y - ac^2 > 2ac(x - c)$. Let t be an arbitrary real parameter value; then the question is whether $at^2 - ac^2$ is greater than $2ac(t - c)$ for $t \neq c$. But

$$(at^2 - ac^2) \ - \ 2ac(t - c) \quad = \quad at^2 \ - \ 2act \ + \ ac^2 \quad = \quad a(t - c)^2$$

and the latter is clearly positive if $t \neq c$.

Now suppose that Γ is a circle or an ellipse with the parametrization $(a \cos t, b \sin t)$. We shall use the following result to describe the tangent line to the ellipse at a typical point:

> *If we are given a curve which satisfies an equation $f(u, v) = 0$ where $f(u_0, v_0) = 0$ but $\nabla f(u_0, v_0) \neq \mathbf{0}$, then the tangent line to the curve at $\mathbf{p} = (u_0, v_0)$ is defined by the vector equation $\nabla f(\mathbf{p}) \cdot (\mathbf{x} - \mathbf{p}) = 0$, and the normal direction at $\mathbf{p}$ is determined by $\nabla f(\mathbf{p})$.*

This can be found in nearly any calculus book which covers partial derivatives.

Applying the preceding result to the ellipse, we see that the normal direction to the ellipse at a point (x_0, y_0) is given by $(2x/a^2,\ 2y/b^2)$, and therefore the tangent line at the point $\mathbf{p} = (a \cos \theta, b \sin \theta)$ is defined by the following equation:

$$\left(\frac{2a \cos \theta}{a^2}\right) \ (u - a \cos \theta) \ + \ \left(\frac{2b \sin \theta}{b^2}\right) \ (v - b \sin \theta) \quad = \quad 0$$

We claim that the expression on the right hand side is negative for all points (u, v) on the ellipse except for $\mathbf{p}$. But if we make the substitution $(u, v) = (a \cos t, b \sin t)$ in the expression on the left hand side of the equation and simplify by cancellations of the form $c^2/c^2 = 1$, the left hand side reduces to

$$(\cos \theta \cos t - \cos^2 \theta) \ + \ (\sin \theta \sin t - \sin^2 \theta) \quad = \quad \cos(t - \theta) \ - \ 1 \ .$$

This expression is always nonnegative, and it is zero if and only if t is an integral multiple of 2π; since $(u, v) = \mathbf{p}$ if t is an integral of 2π, it follows that all other points of the ellipse lie on the same side of the tangent line at $\mathbf{p}$.

A similar analysis applies to the branches of the hyperbola. Since the change of variables $u = -x, \ v = y$ interchanges the two branches, by symmetry it will suffice to consider the right hand branch on which $x > 0$ and the branch is parametrized by the curve $(a \cosh t, b \sinh t)$. Suppose that $\mathbf{p}$ has coordinates $(a \cosh t_0, b \sinh t_0)$. Then as in the elliptical case the normal direction to the hyperbola branch at a point (x_0, y_0) is given by the gradient expression $(2x/a^2,\ -2y/b^2)$, and therefore the tangent line at the point $\mathbf{p} = (a \cosh t_0, b \sinh t_0)$ is defined by the following equation:

$$\left(\frac{2a \cosh t_0}{a^2}\right) \ (u - a \cosh t_0) \ - \ \left(\frac{2b \sinh t_0}{b^2}\right) \ (v - b \sinh t_0) \quad = \quad 0$$

If we now make the substitution $(u, v) = (a \cosh t, b \sinh t)$ and simplify once again, we see that the expression on the left side is equal to

$$(\cosh t_0 \cosh t - \cosh^2 t_0) \ - \ (\sinh t_0 \sinh t - \sinh^2 t_0) \quad = \quad \cosh(t - \theta) \ - \ 1 \ .$$

Since $\cosh w \geq 1$ with equality if and only if $w = 0$, it follows as before that all points on the branch of the hyperbola except for $\mathbf{p}$ must lie on the same side of the tangent line at $\mathbf{p}$.∎

Differential calculus and normal lines to curves

By definition, a normal line to a curve at a given point $\mathbf{p}$ on the curve is a line through $\mathbf{p}$ which is perpendicular to the tangent line at $\mathbf{p}$. Apollonius' work on conics includes a study of normal lines to these curves; it is based upon the observation that the shortest or longest distance from an external point to a curve is a given by a line segment which is perpendicular to the tangent line to the conic at the point of contact. For example, if the curve is a circle, then the line segment lies in the line joining the external point to the center of the circle; by elementary geometry, we know that this line is perpendicular to the tangent line at the point where the lines meet the circle. Our goal here is to give a simple proof of the general fact using calculus and parametric equations and vector representations for curves.

The relationship between normal lines and maximum or minimum distances is then given by the following result:

PROPOSITION. *Suppose that* $\mathbf{s}(t) \;=\; \big(\, x(t),\, y(t) \,\big)$ *is a parametrized curve such that the coordinate functions* $x(t), y(t)$ *are differentiable everywhere and for each value of* t *at least one of* $x'(t), y'(t)$ *is nonzero. Let* $\mathbf{p} = (c, d)$ *be a point not on the curve, and suppose that* t_0 *is such that the distance from* $\mathbf{s}(t_0)$ *to* $\mathbf{p}$ *is minimized or maximized. Then the direction of the line joining these two points is perpendicular to* $\mathbf{s}'(t_0)$.

Proof. We shall need the following vector differentiation identity from first year calculus:

$$(\mathbf{a} \cdot \mathbf{b})' \;=\; \mathbf{a}' \cdot \mathbf{b} \;+\; \mathbf{a} \cdot \mathbf{b}'$$

Since the distance between two points is nonnegative, the problem of maximizing or minimizing the distance is equivalent to maximizing or minimizing the **square** of the distance, and using vectors we can write the square of the distance between $\mathbf{p}$ and a point $\mathbf{s}(t)$ on the curve as

$$\delta_2(t) \;=\; (\mathbf{s}(t) - \mathbf{p}) \cdot (\mathbf{s}(t) - \mathbf{p}) \;.$$

We can now use differential calculus to search for a minimum by setting $\delta_2'(t) = 0$, and if we apply the previously stated differentiation identity to the expression for δ_2 we obtain

$$0 \;=\; \delta_2'(t) \;=\; 2\,\mathbf{s}'(t) \cdot (\mathbf{s}(t) - \mathbf{p}) \;.$$

Since $\mathbf{s}(t) - \mathbf{p}$ is the direction of the line joining $\mathbf{s}(t)$ and $\mathbf{p}$, this is exactly the conclusion that we want.∎

The proof of the proposition has the following simple implication.

PROPOSITION. *Suppose that* $\mathbf{s}(t) \;=\; \big(\, x(t),\, y(t) \,\big)$ *is a parametrized curve such that the coordinate functions* $x(t), y(t)$ *are differentiable everywhere and for each value of* t *at least one of* $x'(t), y'(t)$ *is nonzero. Let* $\mathbf{p} = (c, d)$ *be a point not on the curve, and let* $\delta_2(t)$ *denote the square of the distance from* $\mathbf{p}$ *to* $\mathbf{s}(t)$. *Then* t_0 *is a critical point of* δ_2 *if andn only if the direction of the line joining* $\mathbf{p}$ *to* $\mathbf{s}(t_0)$ *is perpendicular to* $\mathbf{s}'(t_0)$.

This is an immediate consequence of the formula for $\delta_2'(t)$.∎

Critical points of squared distance functions

Let Γ be a nonsingular conic, and let $\mathbf{p}$ be a point in the plane. The first goal is to show that the squared distance function δ_2 always has an absolute minimum and therefore always has at least one critical point.

The case of an ellipse. In this case the distance squared function δ^2 satisfies the periodicity identity $\delta_2(t + 2\pi) = \delta_2(t)$, so over the interval $[0, 2\pi]$ it has both maximum and minimum values. Since δ_2 is periodic, these are maximum and minimum values for the function over the entire real line.∎

The case of a parabola. In this case $\delta_2(t)$ is a fourth degree polynomial whose highest degree term has the form a^4t^4 where $a \neq 0$. Every such polynomial function has an absolute minimum value over the entire real line; in particular, we have $\lim t \to \pm\infty\ \delta_2(t) = +\infty$ (the same is true for every even degree polynomial function for which the coefficient of the highest degree term is positive).∎

The case of a branch of a hyperbola. We shall only consider the branch of the standard hyperbola which lies in the half-plane $x > 0$; symmetry conditions will imply the same sorts of conclusions for the other branch. — If $\mathbf{p} = (c, d)$, then $\delta_2(t) = (\cosh t - c)^2 + (\sinh t - d)^2 \leq (\cosh t - c)^2$, and since the right hand side goes to infinity as $t \to \pm\infty$, we can argue as in the parabolic case that δ_2 has an absolute minimum over the real line: Specifically, let v be some value that δ_2 takes, and using the limit property choose $M > 0$ so large that $\delta_2(t) \geq v + 1$ for $|t| \geq M$. Then δ_2 takes an absolute minimum value on the interval $[-M, M]$, and by choice of M this value is taken at an interior point of the interval. Since this minimum value is no greater than v and $\delta_2(t) > v$ if $|t| \geq M$, this minimum must be the minimum value of the function over the entire real line.∎

Counting normals from a point to a conic. Having shown that there is always at least one normal to a conic through a given point in the plane, it is natural to ask two further questions:

(1) Given a point $\mathbf{p}$ and a conic Γ, how many normals to Γ pass through the point $\mathbf{p}$?

(2) Given a point $\mathbf{p}$ on the conic Γ and a normal line $\mathcal{L}$ to Γ through $\mathbf{p}$, are there other points $\mathbf{q}$ on $\mathcal{L} \cap \Gamma$ such that $\mathcal{L}$ is also a normal to Γ at $\mathbf{q}$?

There are simple answers to both questions if the conic is a circle.

PROPOSITION. *Let Γ be a circle, and let $\mathbf{p}$ be a point in the coordinate plane.*

(i) If $\mathbf{p}$ is the center of Γ, then every line through $\mathbf{p}$ is a normal to Γ.

(ii) If $\mathbf{p}$ is not the center of Γ, then there is exactly one normal line to Γ through $\mathbf{p}$, and it contains the center of the circle.

In either case, if $\mathcal{L}$ is a line through $\mathbf{p}$ which is normal to Γ, then $\mathcal{L}$ meets Γ in a second point $\mathbf{q}$ and is also a normal to Γ at $\mathbf{q}$.

Proof. The final statement follows from the preceding two because every the latter show that every normal line through Γ passes through the center of the circle, so that every such line meets the circle at two points and is normal to the circle at each of these points.

Statement (i) follows because tangent lines to a circle are perpendicular to radial lines at their points of contact. To prove (ii), we first use an isometry to find a congruent circle centered at the origin and note that it will suffice to prove the result for the resulting circle, which has an equation of the form $x^2 + y^2 = r^2$ for some $r > 0$. We are assuming that $\mathbf{p}$ is not the center, so we may set $\mathbf{p}$ equal to $(b\cos\alpha, b\sin\alpha)$ where $b > 0$ and α is some real number.

There are two cases, depending upon whether or not $\mathbf{p}$ lies on the given circle. If $\mathbf{p}$ lies on the circle, then the radial line through $\mathbf{p}$ is the unique line through $\mathbf{p}$ which is perpendicular to the tangent line at $\mathbf{p}$, and since it passes through the center of the circle the conclusion of (ii) is true in this case.

Suppose now that $\mathbf{p}$ does not lie on the given circle, so that $b \neq r$. For each point $(r\cos\theta, r\sin\theta)$ on the circle, the direction of the normal line is determined by the vector $(\sin\theta, -\cos\theta)$, and therefore the line joining $\mathbf{p}$ to $(r\cos\theta, r\sin\theta)$ is perpendicular to the circle at the latter point if and only if

$$0 \;=\; (\sin\theta, -\cos\theta) \cdot (b\cos\alpha - r\cos\theta, b\sin\alpha - r\sin\theta) \; .$$

Since $(\cos\theta, \sin\theta)$ and $(\sin\theta, -\cos\theta)$ are perpendicular, this equation reduces to

$$0 \;=\; b\cos\alpha\sin\theta \;-\; b\sin\alpha\cos\theta \;=\; b\sin(\theta - \alpha)$$

and since $b \neq 0$ the latter reduces to $\sin(\theta - \alpha) = 0$, which is true if and only if $\theta - \alpha$ is an integral multiple of π. Since $\sin(\alpha + \pi) = -\sin\alpha$ and $\cos(\alpha + \pi) = -\cos\alpha$, it follows that a normal line to the circle through $\mathbf{p} = (b\cos\alpha, b\sin\alpha)$ meets the circle at either $(r\cos\alpha, r\sin\alpha)$ or $(-r\cos\alpha, -r\sin\alpha)$. But the line through $\mathbf{p}$ and one of these points also goes through the other and through the center of the circle, which implies the conclusion of (ii).∎

For the other conics, the answer to the first question is more complicated. If the conic Γ is not a circle, then there are always at most four normals to the conic from a given point $\mathbf{p}$, but the exact number of normals varies in each case and depends upon the position of $\mathbf{p}$. In one of the exercises for this unit, this question is studied when Γ is the standard parabola $y = x^2$ and $\mathbf{p}$ lies on the y-axis; for some choices of $\mathbf{p}$ there are three normals to the parabola, but for others there is only one. Here are some additional references:

`http://mathworld.wolfram.com/ParabolaEvolute.html`

`http://mathworld.wolfram.com/EllipseEvolute.html`

`http://mathworld.wolfram.com/HyperbolaEvolute.html`

`http://www3.villanova.edu/maple/misc/ellipse/Apollonius2004.pdf`

Finally, we address the question of whether a single line can be a normal line to several different points on a conic which is not a circle. The next result implies that a line can be a normal to at most two points of a conic.

PROPOSITION. *Let Γ be a nonsingular conic, and let $\mathcal{L}$ be a line. Then $\mathcal{L}$ and Γ have at most two points in common.*

As in other cases, it suffices to verify this result when Γ is one of the standard examples. If $\mathcal{L}$ is given by the algebraic equation $px + qy = k$ where $(p, q) \neq (0, 0)$ this amounts to showing that for each of the quadratic equations in two variables defining the standard examples

$$y \;=\; ax^2, \quad b^2x^2 + a^2y^2 \;=\; a^2b^2 \; , \quad b^2x^2 - a^2y^2 \;=\; a^2b^2$$

there are at most two simultaneous solutions of the system determined by the quadratic equation and the equation of the line. We can solve $px + qy = k$ for x in terms of y or vice versa; it will be convenient to call the solved-for variable the *constrained variable* and the other variable the *free variable*. Substitution of for the constrained variable in terms of the free variable yields a quadratic equation in the free variable, which has at most two solutions. For each of these solutions there is a unique corresponding value of the constrained variable, and thus there are at most two simultaneous solutions for the given system(s) of equations.∎

Here is the result on lines which are normal at two points of a conic.

PROPOSITION. *Let Γ be a conic which is not a circle, and let $\mathcal{L}$ be a line. Then $\mathcal{L}$ is a normal line to Γ at two distinct points if and only if Γ is an ellipse or hyperbola and $\mathcal{L}$ is one of its axes.*

In particular, there are two such lines for an ellipse (which is not a circle) and one for a hyperbola, but none for a parabola.

Proof. The argument breaks down into cases depending upon whether Γ, is an ellipse, a hperbola, or a parabola. In each case we shall look for conditions under which two points on the curve have the same normal direction (up to sign as usual), and in each case where this happens we shall show that the normal lines at the two points must be different.

The case of a parabola. The curve's equation is $y = ax^2$ where $a > 0$. We shall show that no two points on the curve have the same normal direction. The tangent vector at a typical point (t, at^2) has the form $(1, 2at)$, so the normal direction at that point is given by $(2at, -1)$. If (s, as^2) is a second point on the curve (so $s \neq t$), then (t, at^2) and (s, as^2) will have the same normal direction if and only if $(2at, -1)$ and $(2as, -1)$ are nonzero scalar multiples of each other. If c is a scalar such that the first vector is c times the second, then we have $(2at, -1) = (2cas, -c)$ and by equating coefficients we get $c = 1$. But this means that $2as = 2at$, which implies $s = t$, contradicting our choice of $s \neq t$. Therefore different points on the parabola have different normal directions, so no line can be normal to the parabola at two distinct points.

The case of an ellipse. The curve is given in parametric form by $(a\cos t, b\sin t)$ where $a, b > 0$ and $a \neq b$ since we are excluding the case of a circle. In this case the normal direction is determined by $(b\cos t, a\sin t)$. If t is an integral multiple of $\frac{1}{2}\pi$, then the normal line is one of the coordinate axes, so that the point is a vertex of the ellipse and the normal goes through the opposite vertex, so we knows what happens in such cases and henceforth we shall only consider cases where t is not an integral multiple of $\frac{1}{2}\pi$. This implies that both $\cos t$ and $\sin t$ are nonzero.

If u is a now point such that $(b\cos u, a\sin u) = c(b\cos t, a\sin t)$ for some nonzero scalar k, then it follows that $\cos u = k\cos t$ and $\sin u = k\sin t$; using the identity $\sin^2 + \cos^2 = 1$ we can conclude that $1 = k^2$, so that the only nontrivial possibility is $k = -1$, and in fact this is realized when $u = t + \pi$. To conclude the proof in this case, it is enough to show that the normal line to the ellipse at $(a\cos u, b\sin u) = (-a\cos t, -b\sin t)$. If it did, then there would be a scalar s such that

$$(a\cos t, b\sin t) \ + \ s \cdot (b\cos t, a\sin t) \ = \ (-a\cos t, -b\sin t)$$

so that $a + sb = -a$ and $b + sa = -b$; in deriving the last two equations we use the fact that both $\sin t$ and $\cos t$ are nonzero. If we solve the pair of equations we obtain

$$s \ = \ -\frac{2a}{b} \ = \ -\frac{2b}{a}$$

and if we multiply both sides of the last equation by $-ab/2$ we find that $a^2 = b^2$. Since we specifically assumed that $a^2 \neq b^2$, this means that there is no value of s which solves the equation

$$(a\cos t, b\sin t) \ + \ s \cdot (b\cos t, a\sin t) \ = \ (-a\cos t, -b\sin t)$$

and therefore the normal lines to the curve at parameter values t and $t + \pi$ have no points in common. Therefore, if a normal line to an ellipse is not an axis, then it is not normal to the ellipse at any other point on the curve.

The case of a hyperbola. This is similar to the preceding case, but now the curve's two branches are given in parametric form by $(\varepsilon a\cosh t, b\sinh t)$ where $a, b > 0$ and $\varepsilon = \pm 1$. In this case

the direction of the normal line is determined by $(\varepsilon b \cosh t, -a \sinh t)$, and once again the question is to see if there is a second point at which the normal direction is the same; we shall write this point as $(\eta b \cosh u, -a \sinh u)$, where $\eta = \pm 1$ does not depend upon ε. If k is the proportionality constant, then as before we must have

$$\varepsilon \cosh t \;=\; k\eta \cosh u \;, \qquad \sinh t = k \sinh u \;.$$

As in the elliptical case, it is useful to dispose of cases where the points lie on the x-axis (note that no points lie on the y-axis). We know that the x-axis meets the hyperbola normally at the two vertex points $(\pm a, 0)$, so this case is understood. Therefore we shall assume that $\sinh t > 0$ for the rest of the discussion; since $\cosh^2 - \sinh^2 = 1$ and $\cosh \geq 1$, this also yields $\cosh t > 1$.

Also as in the elliptical case, given a point $(\varepsilon a \cosh t, b \sinh t)$ with $t \neq 0$, it follows that the same normal direction arises for a second point on the hyperbola of the form

$$(-\varepsilon a \cosh -t, b \sinh -t) \;\;=\;\; -(\varepsilon a \cosh t, b \sinh t)$$

since $\cosh t$ is even and $\sinh t$ is odd; note that this point lies on the other branch of the hyperbola. We would like to imitate the preceding argument and show that (i) the same normal direction does not arise for any other point on either branch of the hyperbola, (ii) the normal to the hyperbola at $(\varepsilon a \cosh t, b \sinh t)$ does not contain the point $-(\varepsilon a \cosh t, b \sinh t)$.

The sign factors ε and η introduce complications not present in the elliptical case, but we can reduce everything to the case where $\varepsilon = 1$ because the hyperbola is symmetric with respect to the isometry (reflection about the y-axis) sending (x, y) to $(-x, y)$. Thus we shall also assume that $\varepsilon = +1$ for the remainder of this argument.

To dispose of (i), we need to show that if k, η and u are such that

$$(b \cosh t, -a \sinh t) \;\;=\;\; (k\eta b \cosh u, -ka \sinh u)$$

then $k = \pm 1$, $u = kt$ and $\eta = k$. Set $p = \cosh t$ and $q = \sinh t$, so that $p > 1$ and $q > 0$, with $p^2 = 1 + q^2$. Since a and b are positive, equating the coefficients on both sides of the displayed equation yields $p = k\eta \cosh u$ and $q = k \sinh u$, so that

$$p^2 \;\;=\;\; k^2 \cosh^2 u \;\;=\;\; k^2 \;+\; k^2 \sinh^2 u \;\;=\;\; k^2 \;+\; q^2$$

and if we subtract $p^2 = 1 + q^2$ from this we get $k^2 - 1 = 0$, or $k = \pm 1$. Since $\sinh ku = k \sinh u$ for $k = \pm 1$ and sinh is strictly increasing, if $k = -1$ then the coordinate equation $-a \sinh t = -ka \sinh u = -a \sinh -u$ implies that $u = -t$ (recall that we have reduced things to cases where $\sinh t, \sinh u \neq 0$), and $b \cosh t = k\eta b \cosh -t = k\eta b \cosh t = -\eta b \cosh t$ implies $\eta = -1 = k$ (since $\cosh t, \cosh u > 1$). — Now suppose that $k = 1$. Then the coordinate equations yield $-a \sinh t = -a \sinh u$, so that $t = u$, and $b \cosh t = \eta b \cosh t$, so that $\eta = 1 = k$. This completes the verification of (i).

The normal line at $(a \cosh t, b \sinh t)$ is given by the parametric form

$$(a \cosh t, b \sinh t) \;+\; s \cdot (b \cosh t, -a \sinh t)$$

and the verification of (ii) amounts to saying there is no choice of s for which this expression equals $(-a \cosh t, -b \sinh t)$. This argument is parallel to the elliptical case. If we can find a scalar s so that

$$(a \cosh t, b \sinh t) \;+\; s \cdot (b \cosh t, -a \sinh t) \;\;=\;\; (-a \cosh t, -b \sinh t)$$

so that $a + sb = -a$ and $b - sa = -b$; in deriving the last two equations we use the fact that both $\sinh t$ and $\cosh t$ are nonzero. If we solve the pair of equations we obtain

$$s \;=\; -\frac{2a}{b} \;=\; \frac{2b}{a}$$

and if we multiply both sides of the last equation by $-ab/2$ we find that $a^2 = -b^2$. Since $a, b > 0$ this is impossible, and hence there is no value of s which solves the equation

$$(a\cosh t, b\sinh t) \;+\; s \cdot (b\cosh t, -a\sinh t) \;=\; (-a\cosh t, -b\sinh t)$$

and therefore the normal to the hyperbola at the point $(\varepsilon a\cosh t, b\sinh t)$ does not contain the point $-(\varepsilon a\cosh t, b\sinh t)$. To sum up the entire discussion for hyperbolas, if a normal line to a hyperbola is not an axis, then it is not normal to the hyperbola at any other point on the curve.∎

Addendum: Intersections of two conics

In `history04Y.pdf` we mentioned that two conics have at most four points in common. For a given pair of examples, it is usually fairly straightforward to do this using standard algebraic techniques from precalculus and earlier courses, but for the sake of completeness we note that a very abstract and general result of this type appears in Exercise 11 on pages 175–175 (= document pages 33–34) of the following online document:

`http://math.ucr.edu/~res/progeom/pgnotes07.pdf`