# The Cartesian ovals

Rida T. Farouki
Department of Mechanical and Aerospace Engineering,
University of California, Davis, CA 95616, USA
e–mail: farouki@ucdavis.edu

## 1 Introduction

In 1637 René Descartes published his treatise *La Géométrie* [5] — which first systematically expounded the use of coordinates to permit the investigation of geometrical problems by algebraic methods, and thereby launched the field of *analytic geometry.* Descartes insisted that only loci specified by algebraic equations (e.g., conics) are "geometrical curves" and those arising from, for example, the rolling motion of a circle (e.g., cycloids) are "mechanical curves" (we now call them *algebraic* and *transcendental* curves). Descartes' insistence that only those curves that admit algebraic equations are truly "geometrical" was subsequently criticized by Leibniz and Newton, among others.

*La Géométrie* appeared, together with *La Dioptrique* and *Les Météores*, as appendices to Descartes' famous philosophical work *Discours de la méthode pour bien conduire sa raison, et chercher la vérité dans les sciences* — which contains his famous credo "I think, therefore I am." Descartes' keen interest in both geometry and optics is exemplified by his introduction, in Book II of *La Géométrie*, of a family of "oval" curves that arise from the refraction of light at a smooth interface between two different media.

Although these *Cartesian ovals* are a natural generalization of the ellipse and hyperbola, and have many fascinating properties and applications, they have sadly settled into relative obscurity over the past century. They receive cursory treatment (perhaps more from a desire for completeness than their intrinsic merit) in "catalogs of plane curves" [19, 20] although others [26] neglect to mention them. Several classic works, concerned with the geometry of plane curves and theory of functions, discuss the Cartesian ovals in some depth — e.g., Harkness and Morley [14], Salmon [24], Steiner [29], Williamson [31], and Zwikker [32] — but date mostly from the mid–19th to the mid–20th centuries. The most comprehensive treatment from this era may be found in Francisco Gomes Teixeira's 1905 *Tratado de las Curvas Especiales Notables* [12], which has been translated into French [13] but not English. The impetus for this treatise was a prize, proposed by the Royal Academy of Sciences of Madrid, for "An orderly list of all the curves of every kind to which definite names have been assigned, accompanying each with a succinct exposition of its form, equations and general properties, and with a statement of the books in which, or the authors by whom, it was first made known" [27]. A contemporary counterpart to Gomes Teixeira's *Tratado* is the *Encyclopedia of Remarkable Mathematical Forms* website [33] — recipient of the 2008 Anatole Decerf Prize — which gives a comprehensive treatment of the properties of Cartesian ovals, and illustrative animations of their morphology.

In this article, we attempt to revive interest in the Cartesian ovals among modern readers through an easily–accessible introduction to their fascinating geometry, their algebraic properties, and their connections and applications — with particular focus on their intimate relationship to the algebra of point sets in the complex plane and classical geometrical optics.

## 2 Descartes' oval constructions

Figure 1, from Book II of *La Géométrie*, shows an example of Descartes' oval constructions, which he states are "very useful in the theory of catoptrics and dioptrics" [5], i.e., in problems of reflection and refraction of light. Descartes explains the construction shown in Figure 1 — in a style that, unfortunately, is rather lacking in both clarity and motivation — as follows.

He considers a point A on the straight line FG, at which an inclined line is drawn, such that the ratio of lengths FA:AG is a given value $k$ (representing a ratio of refractive indices). A circle $C_1$ with center F is drawn, that meets the line FG at the point labelled 5. The line 56 is then drawn such that the ratio of lengths A6:A5 is also equal to $k$. Along the inclined line, the point R is identified such the lengths AR and AG are equal, and a circle $C_2$ is drawn with center at G and radius equal to the length[1] 6R. The circles $C_1$ and $C_2$ intersect at two positions labelled 1 — above and below the line FG — which

[1] That is, from the point 6 to the point R in the diagram (Descartes confusingly employs both letters and numbers to label points).

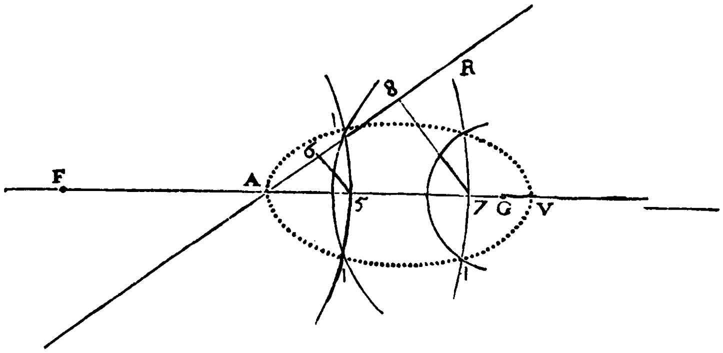

Figure 1: Descartes' geometrical construction of one of his ovals.

are points on the desired oval. By repeating this process for circles $C_1$ with centers at F and different radii (e.g, F7), the entire oval can be traced.

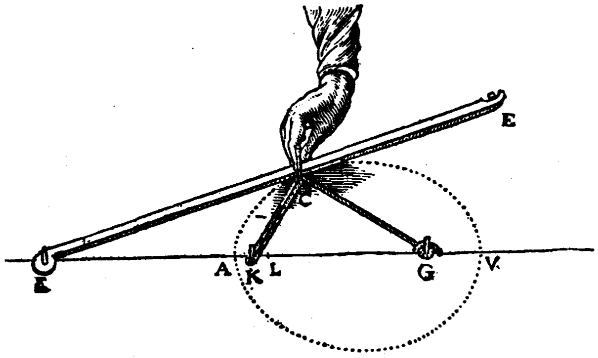

Figure 2: Descartes' mechanism to construct one of his ovals.

Figure 2 shows a mechanism proposed by Descartes [5] to construct one of his ovals, in the particular case FA=AG. Attaching one end of a length of string to point K, it is wrapped around a pulley at C, around another pulley at G, and the other end is attached to point C. The point C then traces out the oval with extreme points A and V as the ruler FE rotates about point F.

Although bipolar coordinates are implicit in Descartes' oval constructions, his approach is difficult to follow and notably does *not* result in an explicitly algebraic formulation. Boyer [3] observes that Newton, in his *Artis analyticae specimina vel geometria analytica* (which was not published until 1779 — 52 years after Newton's death) had criticized Descartes for describing the ovals "in a very prolix manner" and he states that "Newton therefore seems to have been the originator of bipolar coordinates in the strict sense of the word."

## 3 Bipolar coordinates

A well–known simple mechanism — the "gardener's method" — can be used to accurately draw an ellipse. Two pins are pressed into a sheet of paper to identify the foci $\mathbf{p}_1, \mathbf{p}_2$ of the ellipse, and a loop of string with total length $\ell > 2\,|\mathbf{p}_2 - \mathbf{p}_1|$ is then wrapped around them. The point $\mathbf{p}$ of a pencil that keeps the loop taut as it moves will then trace out an ellipse. In terms of the distances $r_1, r_2$ of $\mathbf{p}$ from $\mathbf{p}_1, \mathbf{p}_2$ the equation of the ellipse can be written as

$$r_1 \,+\, r_2 \,=\, k\,, \tag{1}$$

where $k = \ell - |\mathbf{p}_2 - \mathbf{p}_1|$. The values $(r_1, r_2)$ are the *bipolar coordinates* of $\mathbf{p}$ with respect to the *poles* $\mathbf{p}_1, \mathbf{p}_2$. Without loss of generality, we may choose $\mathbf{p}_1 = (+1, 0)$ and $\mathbf{p}_2 = (-1, 0)$ — then $(r_1, r_2)$ are non–negative values that must (see Figure 3) satisfy

$$r_1 + r_2 \geq 2 \qquad \text{and} \qquad |r_1 - r_2| \leq 2\,. \tag{2}$$

The constraints (2) are satisfied with equality by points on the $x$–axis (the former for $-1 \leq x \leq +1$, and the latter for $x \leq -1$ or $x \geq +1$).

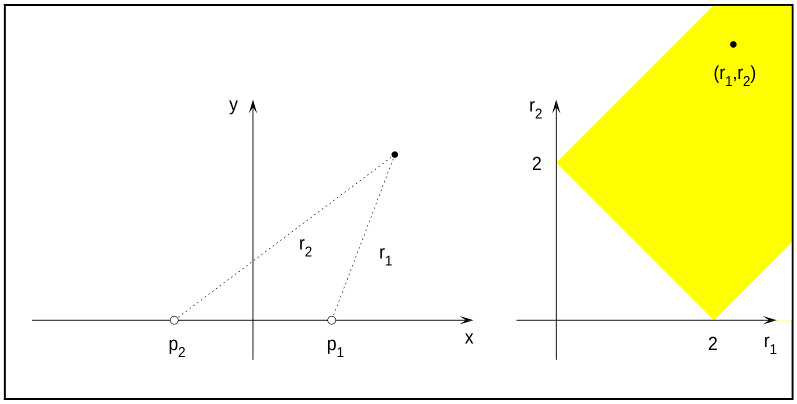

Figure 3: Bipolar coordinates $(r_1, r_2)$ with respect to poles $\mathbf{p}_1 = (+1, 0)$ and $\mathbf{p}_2 = (-1, 0)$: the region of valid $(r_1, r_2)$ values is shown shaded on the right.

Note that the bipolar coordinates $(r_1, r_2)$ actually identify *two* points, that are images of each other under a reflection in the $x$–axis. A curve specified in bipolar coordinates is therefore symmetric about the $x$–axis, unless separate equations are defined for the lower and upper half–planes. For a point $\mathbf{p}$ with Cartesian coordinates $(x, y)$ the bipolar coordinates are $r_1 = \sqrt{(x-1)^2 + y^2}$, $r_2 = \sqrt{(x+1)^2 + y^2}$. Conversely, the bipolar coordinates are given in terms of the Cartesian coordinates by

$$(x, y) \;=\; \frac{1}{4}\left(r_2^2 - r_1^2, \pm\sqrt{8\,(r_1^2 + r_2^2 - 2) - (r_2^2 - r_1^2)^2}\right) .$$

We can convert (1) into a polynomial equation in $(x, y)$ by re–arranging terms and squaring twice, to obtain

$$(r_1^2 - r_2^2)^2 \;-\; 2\,k^2(r_1^2 + r_2^2) \;+\; k^4 \;=\; 0\,. \tag{3}$$

This equation describes not only the ellipse (1), but all the loci defined (with independent sign choices) by

$$r_1 \;\pm\; r_2 \;=\; \pm\,k\,. \tag{4}$$

For $k < |\mathbf{p}_2 - \mathbf{p}_1|$, the loci $r_1 - r_2 = \pm\,k$ define the two branches of a hyperbola with foci $\mathbf{p}_1, \mathbf{p}_2$. However, the equation $r_1 + r_2 = -\,k$ defines a vacuous real locus for any $k > 0$. Substituting $r_1^2 = (x-1)^2 + y^2$, $r_2^2 = (x+1)^2 + y^2$ into (3) and simplifying, we obtain

$$\frac{4\,x^2}{k^2} \;+\; \frac{4\,y^2}{k^2 - 4} \;=\; 1\,. \tag{5}$$

This defines an ellipse or a hyperbola, with semi–axis $k/2$ and $\sqrt{|(k/2)^2 - 1|}$ according to whether $k > 2$ or $k < 2$ (with $k = 2$ being the degenerate case of a doubly–traced line between $\mathbf{p}_1$ and $\mathbf{p}_2$) — see Figure 4.

# 4 The Cartesian ovals

A simple generalization of the bipolar ellipse equation (1) can be obtained by multiplying $r_1, r_2$ with positive constants $mk, nk$ (where $m \neq n$) and dividing out $k$ to obtain the equation

$$m\,r_1 \;+\; n\,r_2 \;=\; 1\,, \tag{6}$$

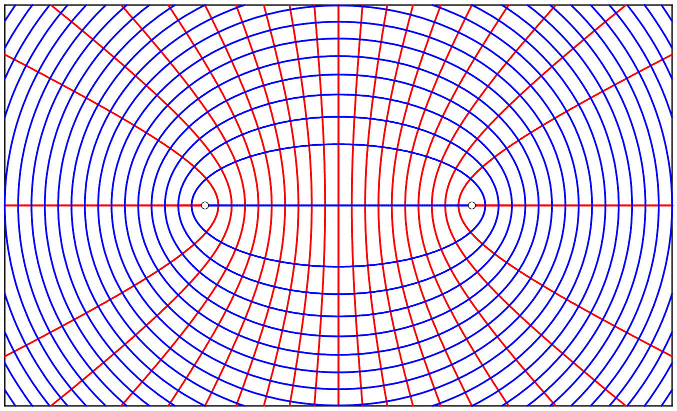

Figure 4: The confocal family of conics (ellipses $r_1 + r_2 = k$ with $k > 2$, and hyperbolas $r_1 - r_2 = \pm k$ with $k < 2$) defined by the bipolar equations (4).

which defines a *Cartesian oval*. By squaring twice and simplifying, we can formulate (6) as an equation involving only $r_1^2$ and $r_2^2$, namely

$$(m^2 r_1^2 - n^2 r_2^2)^2 \,-\, 2\,(m^2 r_1^2 + n^2 r_2^2) \,+\, 1 \,=\, 0\,.$$

This equation describes not only the locus (6), but all the curves defined (for independent sign choices) by

$$m\,r_1 \,\pm\, n\,r_2 \,=\, \pm 1\,. \tag{7}$$

In the $(r_1, r_2)$ plane, equations (7) identify four lines that pass through the points $(\pm 1/m, 0)$ and $(0, \pm 1/n)$. Since $m$ and $n$ are positive, $m\,r_1 + n\,r_2 = -1$ obviously defines a vacuous locus. Of the remaining three lines, one can easily see (Figure 5) that only two possess segments within the valid domain (2) for bipolar coordinates. In general, the Cartesian oval comprises two nested loops — depending upon $m$ and $n$, the equations from (7) that individually define these loops are identified in Table 1.

On substituting $r_1 = \sqrt{(x+1)^2 + y^2}$ and $r_2 = \sqrt{(x-1)^2 + y^2}$ into (6) and simplifying, we obtain the equation of the Cartesian oval as an irreducible quartic algebraic curve,

$$(\alpha x^2 + \alpha y^2 - 2\,\beta x + \alpha)^2 \,-\, 2\,(\beta x^2 + \beta y^2 - 2\,\alpha x + \beta) \,+\, 1 \,=\, 0\,, \tag{8}$$

where $\alpha = m^2 - n^2$, $\beta = m^2 + n^2$. For $\alpha = 0$ (i.e., $m = \pm n$) this reduces to the equation (5) with $\beta = 2/k^2$ for an ellipse ($\beta < \frac{1}{2}$) or a hyperbola ($\beta > \frac{1}{2}$). Figure 6 illustrates some examples of the quartic curves defined by (8).

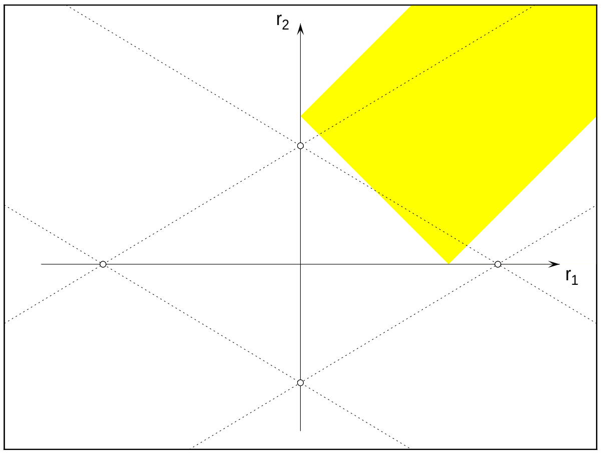

Figure 5: The four lines in the $(r_1, r_2)$ plane defined by (7), of which only two cross the valid domain (shaded region) for bipolar coordinates when $m \neq n$. These two line segments determine the two nested loops of a Cartesian oval.

| $m$ and $n$ | Cartesian oval equations |
|---|---|
| $m < n < \frac{1}{2}$ or $m < \frac{1}{2} < n$ | $m\,r_1 + n\,r_2 = +1,\;\; m\,r_1 - n\,r_2 = -1$ |
| $n < m < \frac{1}{2}$ or $n < \frac{1}{2} < m$ | $m\,r_1 + n\,r_2 = +1,\;\; m\,r_1 - n\,r_2 = +1$ |
| $m > \frac{1}{2}$ and $n > \frac{1}{2}$ | $m\,r_1 - n\,r_2 = +1,\;\; m\,r_1 - n\,r_2 = -1$ |

Table 1: The appropriate members from equations (7) defining the two loops of a Cartesian oval for various $m$ and $n$ values (with $m \neq n$ and $m, n \neq \frac{1}{2}$).

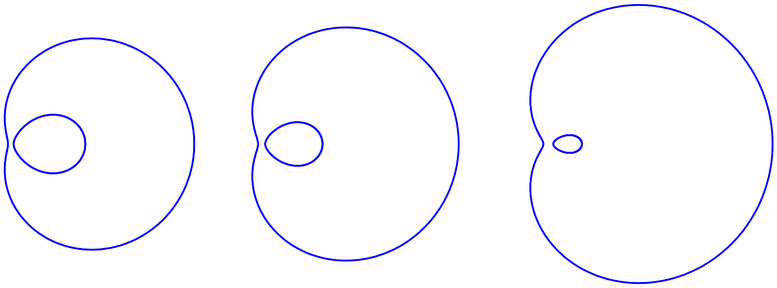

Figure 6: Examples of the Cartesian ovals defined by equation (8).

Invoking homogeneous coordinates $(W, X, Y)$ with $(x, y) = (X/W, Y/W)$ equation (8) becomes

$$\begin{aligned} &[\,\alpha(X^2+Y^2+W^2)+2\,\beta\,WX\,]^2 \\ &\qquad -\,2\,W^2\,[\,\beta(X^2+Y^2+W^2)-2\,\alpha\,WX\,]+W^4 \;=\; 0\,, \end{aligned}$$

and the behavior at infinity is identified by setting $W = 0$ to obtain

$$\alpha\,(X^2+Y^2)^2 \;=\; 0\,.$$

Thus, a Cartesian oval with $\alpha = m^2 - n^2 \neq 0$ is a "bicircular quartic" curve, since the *circular points at infinity* $(W, X, Y) = (0, 1, \pm\mathrm{i}\,)$ are double points on it. In general, it has no other singular points, and is therefore of genus 1. However, if $m = \frac{1}{2}$ or $n = \frac{1}{2}$, one of the poles is also a double point — this special (rational) form is called a *limaçon of Pascal*.

There appears to be no consensus on whether to refer to the curve defined by equation (8) in the singular or the plural. Although the equation defines a single irreducible quartic curve, its generic real locus consists of two disjoint nested loops. It seems unnecessary to be too pedantic on this point.

# 5 Minkowski geometric algebra

The *Minkowski sum* [21] of point sets $A, B \subset \mathbb{R}^n$ is the set of the vector sums of all pairs of points $\mathbf{a}$ and $\mathbf{b}$ selected independently from $A$ and $B$,

$$A \oplus B \;=\; \{\,\mathbf{a}+\mathbf{b} \mid \mathbf{a} \in A \text{ and } \mathbf{b} \in B\,\}\,. \tag{9}$$

Now the vector sum in $\mathbb{R}^2$ is equivalent to complex number addition in $\mathbb{C}$, and since the complex–number product is commutative, we can also introduce a *Minkowski product* of sets $A, B \subset \mathbb{C}$ as

$$A \otimes B \;=\; \{\,\mathbf{a}\,\mathbf{b} \mid \mathbf{a} \in A \text{ and } \mathbf{b} \in B\,\}\,. \tag{10}$$

These operations define a *Minkowski geometric algebra of complex sets* [10]. It is also possible to define the difference and quotient sets $A \ominus B$ and $A \oslash B$, but it is important to note that $\ominus$ and $\oslash$ are *not* inverses to $\oplus$ and $\otimes$. Note also that (9) and (10) obey the *subdistributive* rule

$$(\mathcal{A} \oplus \mathcal{B}) \otimes \mathcal{C} \;\subset\; (\mathcal{A} \otimes \mathcal{C}) \oplus (\mathcal{B} \otimes \mathcal{C})\,.$$

The operations (9)–(10) specify an algebra of point sets that yields rich geometric structures. Consider, for example, $A$ and $B$ as circles with centers at the point 1 on the real axis,[2] as seen on the left in Figure 7. When we plot the products of a large number of pairs of complex values randomly selected from these circles, we obtain the distribution shown on the right in Figure 7, and we immediately recognize that the set $A \otimes B$ of these products occupies the region bounded by the inner and outer loops of a Caretsian oval! When $A$ and $B$ are the *disks* bounded by the two circles, $A \otimes B$ occupies the entire area bounded by the outer loop [10] — the inner loop is no longer empty.

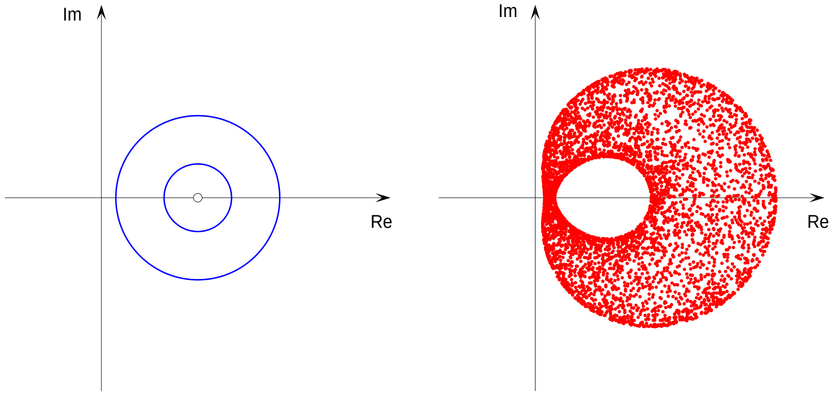

Figure 7: The products of complex values sampled randomly from two circles of different radii centered at the point 1 on the real axis in the complex plane populate the region between the inner and outer loops of a Cartesian oval.

Now since complex number multiplication correponds to a scaling/rotation operation — dubbed an *amplitwist* by Needham [23] — we can interpret the Cartesian oval seen in Figure 7 as the envelope of a one–parameter family of circles, specified by the instances of one circle scaled/rotated by the complex points of the other circle, as illustrated in Figure 8. Hence, we can say that "a Cartesian oval is the boundary of the Minkowski product of two circles," a more elegant and succinct characterization than the description by Gomes Teixeira [13] in terms of real geometry:

> "L'enveloppe d'un cercle variable dont le centre parcourt la circonférence d'un autre cercle donné et dont le rayon varie

[2] We consider only circles or disks centered at the point 1, since the results for arbitrary (non–zero) centers are simply scaled/rotated versions of the resulting Minkowski products.

> proportionnellement à la distance de son centre à un point fixe
> est un couple d'ovales de Descartes."

or "The envelope of a variable circle whose center lies on the circumference of another circle and whose radius is proportional to the distance of its center from a fixed point is a pair of ovals of Descartes."

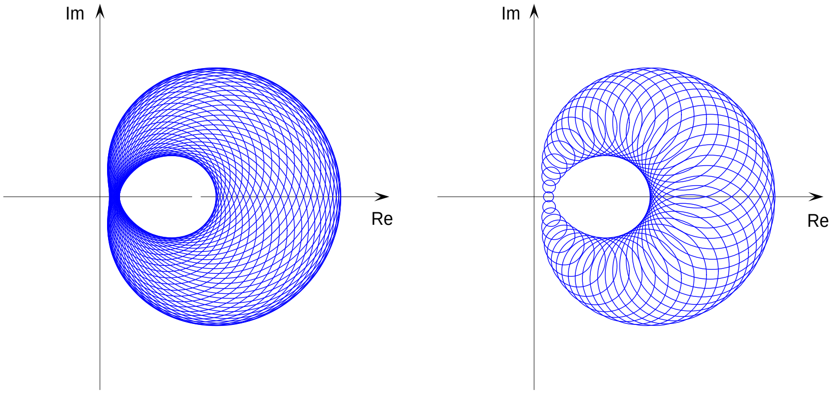

Figure 8: The Cartesian oval in Figure 7 as the envelopes of one–parameter families of circles, generated by a sequence of scalings/rotations of the larger circle by the complex points of the smaller circle (left), and vice–versa (right).

Degenerate forms of the Cartesian oval arise if either or both of the circles in Figure 7 has radius equal to 1. As is evident in Figure 9, if only one circle passes through the origin, the two loops merge into a single, self–intersecting loop (the limaçon of Pascal). When both the circles pass through the origin, we obtain a cusped single loop (the cardioid).

The Minkowski geometric algebra may be regarded as a generalization of *interval arithmetic* [22] from real–number to complex–number sets. Although circles and disks are perhaps the simplest operands, algorithms to compute the Minkowski products of more general complex sets can be developed [9]. Other operations may also be formulated within this algebra. For example, for fixed circular discs $A$ and $B$, the generic solution $X$ to the simple equation

$$A \otimes X \,=\, B\,,$$

(which exists if and only if the radius of $A$ is less than that of $B$) is the region bounded by the *inner* loop of a Cartesian oval [8]. The $n^{\rm th}$ Minkowski power

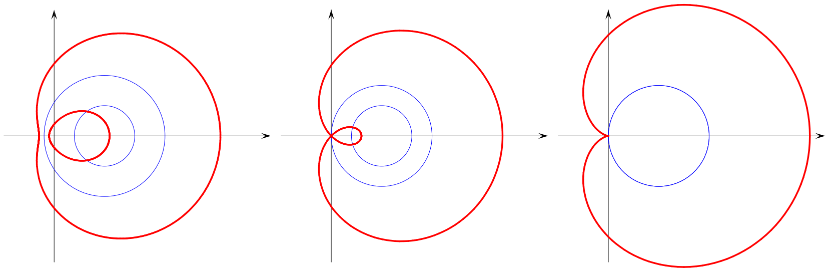

Figure 9: Two circles with centers at the point 1 in the complex plane will generate a generic Cartesian oval, a limaçon of Pascal, or a cardioid according to whether neither, one, or both of the circles pass through the origin.

$\bigotimes^n A$ of a set $A$ is simply the product of $n$ instances of $A$, and correspondingly one can define the $n^{\text{th}}$ root $\bigotimes^{1/n} A$ by the property

$$\{\, \mathbf{z}_1\mathbf{z}_2\cdots\mathbf{z}_n \mid \mathbf{z}_i \in \bigotimes^{1/n} A \ \text{ for } \ i = 1,\ldots,n \,\} \;=\; A\,,$$

for complex values $\mathbf{z}_1,\ldots,\mathbf{z}_n$ chosen independently from $\bigotimes^{1/n} A$. Such roots do not always exist, but for a circular disk $A$ that does not contain the origin it can be shown [7] that $\bigotimes^{1/2} A$ is the region bounded by a single loop of the *ovals of Cassini* defined by a bipolar equation of the form $r_1 r_2 = b^2$, if $a > b$ where $a$ is the distance between the poles (the situation is more complicated when $A$ contains the origin). This can be extended to the $n^{\text{th}}$ root $\bigotimes^{1/n} A$ by an $n$–polar generalization of the Cassini ovals with equation $r_1 r_2 \cdots r_n = b^n$, employing the $n^{\text{th}}$ roots of unity as poles [7].

It is also possible to obtain an exact description of the Minkowski product $A_1 \otimes \cdots \otimes A_N$ of $N$ complex disks as a generalized Cartesian oval [11]. The key observation is that points on the disk boundaries $\partial A_1,\ldots,\partial A_N$, whose products may contribute to the product boundary $\partial(A_1 \otimes \cdots \otimes A_N)$, are identified by the intersections of a system of coaxal circles, through the points 0 and 1 on the real axis, with the circles $\partial A_1,\ldots,\partial A_N$ — see Figure 11.

# 6 Inversion in circles

It is a remarkable fact [13] that any Cartesian oval, regarded as the Minkowski product of two circles with center 1 and radii $\rho_1$ and $\rho_2$, may be described in terms of bipolar coordinates with respect to *three different pairs of poles.* To

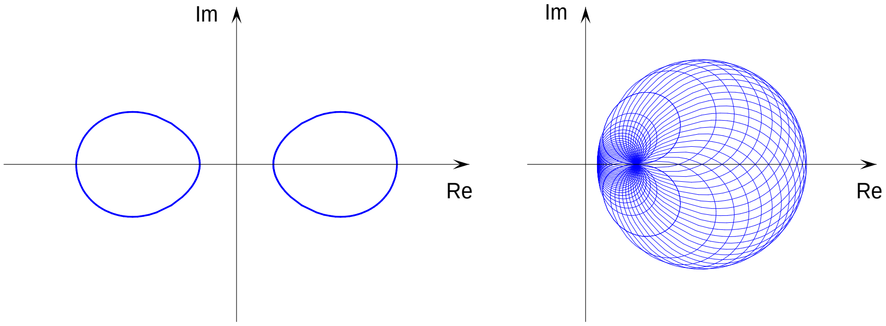

Figure 10: Either loop of the ovals of Cassini (left) is a Minkowski square root $\bigoplus^{1/2} A$ of a given circular disk $A$ (right) not encompassing the origin.

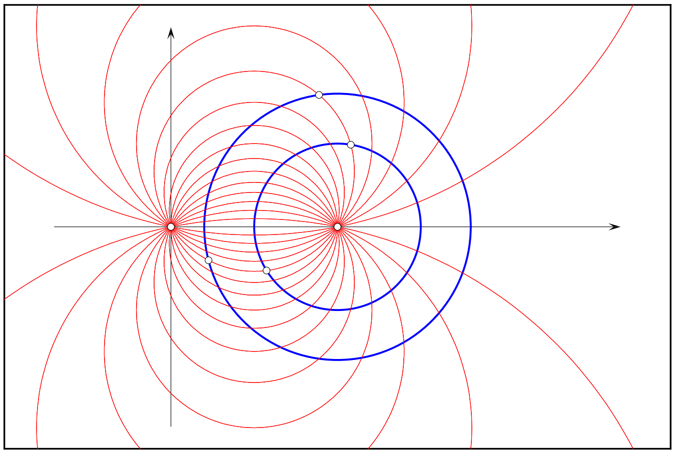

Figure 11: The points on two circles (blue) with center 1 that may contribute to the boundary of their Minkowski product correspond to their intersections with a system of coaxal circles (red) that pass through the points 0 and 1 on the real axis. This characterization generalizes to the product of $N$ circles.

see this, it is convenient to choose the three points 0, $a_1 = 1-\rho_1^2$, $a_2 = 1-\rho_2^2$ on the real axis as poles $\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2$. Then, writing

$$r_0 \;=\; \sqrt{x^2+y^2}\,, \;\;\; r_1 \;=\; \sqrt{(x-a_1)^2+y^2}\,, \;\;\; r_2 \;=\; \sqrt{(x-a_2)^2+y^2}$$

as the distances of a point $x+\mathrm{i}y$ from these poles, the three bipolar equations

$$\rho_1 r_0 \,\pm\, r_1 \;=\; \pm\, a_1\rho_2\,, \;\;\; \rho_2 r_0 \,\pm\, r_2 \;=\; \pm\, a_2\rho_1\,, \;\;\; \rho_2 r_1 \,\pm\, \rho_1 r_2 \;=\; \pm(a_1-a_2)$$

describe the *same* Cartesian oval. By squaring to eliminate radicals, one can verify that these bipolar equations all define the same quartic curve,

$$(x^2+y^2-2x+a_1a_2)^2 \,-\, 4\,\rho_1^2\rho_2^2(x^2+y^2) \;=\; 0\,. \qquad (11)$$

The limaçon of Pascal ($\rho_1 = 1$ or $\rho_2 = 1$) corresponds to the case where $\mathbf{p}_1$ or $\mathbf{p}_2$ coincides with $\mathbf{p}_0$, and the cardioid with the case $\rho_1 = \rho_2 = 1$ where both coincide with $\mathbf{p}_0$. Table 2 lists the appropriate signs in $\rho_2 r_1 \pm \rho_1 r_2 = \pm(a_1 - a_2)$ that identify the inner and outer loops of the Cartesian oval.

| | inner loop | outer loop |
|---|---|---|
| $\rho_1 < 1$ | $\rho_2 r_1 + \rho_1 r_2 = a_1 - a_2$ | $\rho_2 r_1 - \rho_1 r_2 = a_1 - a_2$ |
| $\rho_1 > 1$ | $\rho_2 r_1 - \rho_1 r_2 = a_2 - a_1$ | $\rho_2 r_1 - \rho_1 r_2 = a_1 - a_2$ |

Table 2: Identification of Cartesian oval loops (for $\rho_1 < \rho_2$ and $\rho_1, \rho_2 \neq 1$).

An *inversion* (or *reflection*) in a circle $C$ with center $\mathbf{c}$ and radius $\rho$ is a mapping $\mathbf{z} \longrightarrow \tilde{\mathbf{z}}$ of the extended complex plane[3] into itself, defined by

$$\tilde{\mathbf{z}} \;=\; \mathbf{c} \,\pm\, \frac{\rho^2}{|\mathbf{z}-\mathbf{c}|^2}\,(\mathbf{z}-\mathbf{c}) \qquad (12)$$

The + and − sign choices specify [25] "hyperbolic" and "elliptic" inversions. The interior of $C$ is mapped to its exterior, and vice–versa (circumferential points are invariant, and the points $\mathbf{c}$ and $\infty$ are images of each other). Each point $\mathbf{z}$ and its image $\tilde{\mathbf{z}}$ lie on a diametral line through $\mathbf{c}$, with inversely proportional distances from it, i.e., $|\mathbf{z}-\mathbf{c}|\,|\tilde{\mathbf{z}}-\mathbf{c}| = \rho^2$, with $\mathbf{z}$ and $\tilde{\mathbf{z}}$ on the same side of $\mathbf{c}$ for the + sign in (12) and on opposite sides for the − sign. Inversion maps lines/circles into lines/circles (see Figure 12), and preserves

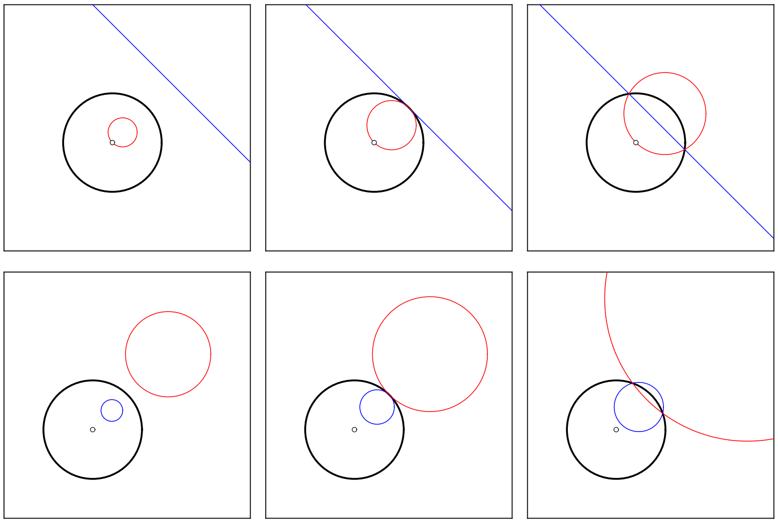

Figure 12: Inversion in a circle — lines that do not pass through the center and circles that do pass through the center are images of each other (upper); circles that do not pass through the center are images of each other (lower).

the magnitudes of angles (but reverses their sense). A complete treatment may be found in standard texts on complex analysis [23, 25].

For a Cartesian oval generated by the Minkowski product of circles $C_1, C_2$ with center 1 and radii $\rho_1, \rho_2$, the poles $a_1 = 1 - \rho_1^2$, $a_2 = 1 - \rho_2^2$ are images of the origin under inversion in $C_1, C_2$. Moreover, the Cartesian oval maps onto itself under an inversion in any of the three circles with centers $\mathbf{c}$ and radii $\rho$ defined (see Figure 13) by

$$\mathbf{c} = 0\,,\ \rho^2 = a_1 a_2\,;\quad \mathbf{c} = a_1\,,\ \rho^2 = a_1(\rho_2^2 - \rho_1^2)\,;\quad \mathbf{c} = a_2\,,\ \rho^2 = a_2(\rho_1^2 - \rho_2^2)\,.$$

Among the three poles $0, a_1, a_2$ the one that is the center of inversion remains fixed, and the other two are swapped. Note that, depending on $\rho_1, \rho_2$ and the chosen circle of inversion, the inner and outer loops may be individually mapped onto themselves, or onto each other.

Curves that map onto themselves by inversion in a circle are known [4] as *anallagmatic curves*. Any circle that cuts the circle of inversion orthogonally is anallagmatic. Furthermore, a family of circles, all of which meet the circle of inversion orthogonally, has an anallagmatic *envelope curve* [4]. This is a

[3] The set of all finite complex values augmented by a single point "∞" at infinity.

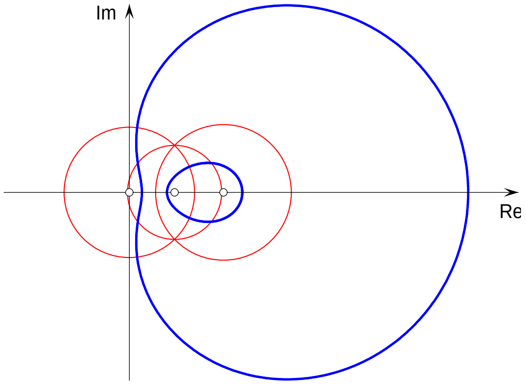

Figure 13: The Cartesian oval defined by $\rho_1 = 0.5$ and $\rho_2 = 0.8$, shown with the three inversion circles (left to right) identified by (1) $\mathbf{c} = 0$, $\rho^2 = |\,a_1 a_2\,|$; (2) $\mathbf{c} = a_2$, $\rho^2 = |\,a_2(\rho_1^2 - \rho_2^2)\,|$; (3) $\mathbf{c} = a_1$, $\rho^2 = |\,a_1(\rho_2^2 - \rho_1^2)\,|$, with respect to which it maps onto itself. For circle (3) the inner and outer loops map on to each other, but this is not true for (1) & (2) since they intersect both loops.

characteristic feature of the circles of inversion identified above — if we regard the Cartesian oval $\partial(C_1 \otimes C_2)$ as the envelope of the family of circles obtained by scaling/rotating $C_1$ by each point of $C_2$, or vice–versa, these scaled/rotated circles are all orthogonal to each of the three circles of inversion.

The anallagmatic nature of the Cartesian oval may be verified by setting $\mathbf{z} = x + \mathrm{i}\,y$ and $\tilde{\mathbf{z}} = \tilde{x} + \mathrm{i}\,\tilde{y}$ in (12), and showing that $(\tilde{x}, \tilde{y})$ satisfies equation (11) if and only if $(x, y)$ satisfies it. Note that the circles of inversion are *real* circles — we use $\sqrt{|\rho^2|}$ for $\rho$ in (12) and choose the + or − sign according to whether $\rho^2 > 0$ or $\rho^2 < 0$. In the exceptional case $\rho^2 = 0$, the inversion (12) degenerates to the identity map (and the Cartesian oval becomes a limaçon).

# 7 Drawing the Cartesian ovals

Drawing a Cartesian oval based on the bipolar equation can be rather tricky. Perhaps the simplest approach is to use "ordinary" polar coordinates about one the three poles $(0, 0)$, $(a_1, 0)$, $(a_2, 0)$. Consider the equation

$$\rho_2 r_1 \,\pm\, \rho_1 r_2 \,=\, \pm(a_1 - a_2) \tag{13}$$

in terms of the poles $(a_1, 0)$ and $(a_2, 0)$. We assume $\rho_1 < 1$ and $\rho_1 < \rho_2$, as in the first row of Table 2, so that $a_1 > a_2$. For polar coordinates $(r, \theta)$ about

$(a_1, 0)$ we then have $r_1 = r$ and

$$r_2^2 \;=\; r^2 + 2\,(a_1 - a_2)\cos\theta\, r + (a_1 - a_2)^2 \tag{14}$$

by the cosine rule. On squaring (13) twice, substituting (14), and simplifying we obtain the quadratic equation

$$(\rho_2^2 - \rho_1^2)\, r^2 \;-\; 2\,(a_1 - a_2)(\rho_1^2 \cos\theta + \rho_2)\, r \;+\; (a_1 - a_2)^2(1 - \rho_1^2) \;=\; 0$$

for the polar distance $r$ in terms of the polar angle $\theta$ with respect to $(a_1, 0)$. Noting that $a_1 - a_2 = \rho_2^2 - \rho_1^2 \neq 0$ for $\rho_1 \neq \rho_2$, this equation simplifies to

$$r^2 \;-\; 2\,(\rho_1^2 \cos\theta + \rho_2)\, r \;+\; (\rho_2^2 - \rho_1^2)(1 - \rho_1^2) \;=\; 0\,,$$

with solutions

$$r_\pm(\theta) \;=\; \rho_1^2 \cos\theta + \rho_2 \;\pm\; \sqrt{(\rho_1^2 \cos\theta + \rho_2)^2 - (\rho_2^2 - \rho_1^2)(1 - \rho_1^2)}\,, \tag{15}$$

where the − and + signs identify points on the inner and outer loops. Note that the expression under the square–root sign in (15) can be re–formulated as $\rho_1^2\,[\,(\rho_2 + \cos\theta)^2 + (1 - \rho_1^2)\sin^2\theta\,]$, which is evidently non–negative if $\rho_1 < 1$.

From (15) we obtain individual parameterizations for the inner and outer Cartesian oval loops of the form

$$\begin{aligned}(x_i(\theta), y_i(\theta)) \;&=\; (a_1 + r_-(\theta)\cos\theta, r_-(\theta)\sin\theta)\,,\\ (x_o(\theta), y_o(\theta)) \;&=\; (a_1 + r_+(\theta)\cos\theta, r_+(\theta)\sin\theta)\,. \end{aligned} \tag{16}$$

## 8 Cartesian ovals in geometrical optics

Descartes was interested in optics as much as in geometry, as already noted in Section 1. His formulation of the eponymous ovals was, in fact, motivated by the problem of discovering the shape of a refracting surface between two different media that will cause a family of rays emanating from a point source in one medium to converge to a point in another medium [17]. This shape is, in fact, a Cartesian oval when the distances of the points from the surface are inversely proportional to the refractive indices of the media [15].

A ray may be thought of as the trajectory of a “particle” of light. Within a single homogeneous medium, it maintains a constant speed and direction, but on encountering an interface with a different medium it is *refracted* and changes speed and direction. Snell's law describes these changes in terms of the *refractive indices* $p$ and $q$ of the different media:

$$\frac{\sin\theta_q}{\sin\theta_p} = \frac{v_q}{v_p} = \frac{p}{q}\,, \tag{17}$$

where $\theta_p, \theta_q$ are the angles of incidence and refraction (relative to the interface normal line), and $v_p, v_q$ are the speeds in the two media. This can be viewed as a consequence of *Fermat's principle* — namely, a light ray follows the path that requires the least time between points in the different media.

The ray theory of light, championed by Newton, is less sophisticated than the wave theory proposed by Huygens in his *Traité de la Lumière* [16] since it cannot accommodate important optical phenomena such as interference and diffraction. Huygens characterized the propagation of a wavefront surface $W$ through a homogeneous medium with speed $c$ by stating that, after a time interval $\Delta t$ has elapsed, the new wavefront $W'$ is the *envelope* of a family of spherical "wavelets" of radius $c\,\Delta t$ with centers on $W$. Hence, $W'$ is *parallel* to — or *offset* from — $W$ at distance $c\,\Delta t$. $W$ and $W'$ have the same family of normal lines, and also parallel tangent planes. See [1] for a more complete treatment of the history of wavefront propagation.

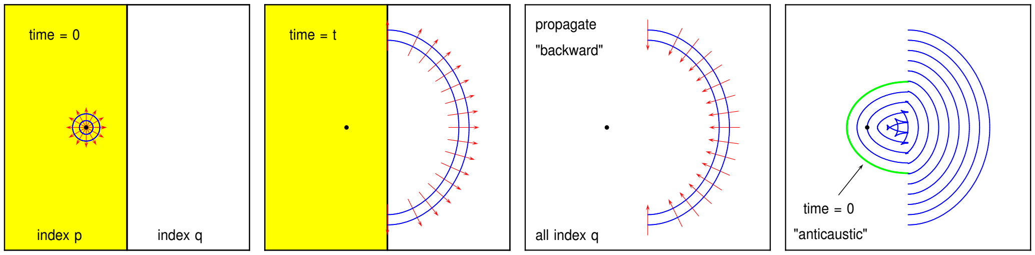

Figure 14: The anticaustic for refraction of a spherical light wave by a planar interface between media with different refractive indices $p$ and $q$ is an ellipse.

If a wavefront propagates through an interface between different media, it is refracted and changes shape. A simple example — a spherical wavefront passing through a planar interface between two media with refractive indices $p$ and $q$ (e.g., air and glass) — can be seen[4] in Figure 14. An ingenious way to describe the shape of refracted wavefronts was discovered by Jakob Bernoulli

[4] For brevity, we consider problems that are rotationally symmetric about a certain axis — the wavefront surfaces are then completely characterized by planar sections.

[2]. Suppose the spherical wavefront is emitted at time $t = 0$ in the medium with index $p$, and at a later time $t = \tau$ has been refracted into the medium with index $q$. We imagine removing the medium of index $p$, and propagating the wavefront backward from time $t = \tau$ to $t = 0$ in a single medium of index $q$. Then, at $t = 0$, it does not collapse back into the point from which the spherical wave emanated, but rather assumes a certain non–trivial "initial" shape, that Bernoulli called the *anticaustic*.

The shape of the refracted wavefront can be obtained from the anticaustic through Huygens' principle in a uniform medium of index $p$. For refraction of a spherical wave by a planar surface, the anticaustic is an *ellipse*, and the refracted wavefronts are therefore parallel or offset curves to an ellipse, which are irreducible algebraic curves with an equation $f(x, y) = 0$ of degree 8 — this equation actually describes both "forward" and "backward" propagating wavefronts (which are not, individually, algebraic curves).

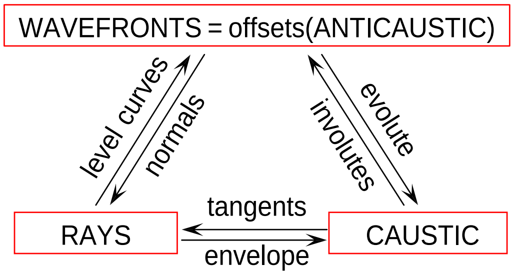

Figure 15: Relationships between rays, wavefronts, caustic, and anticaustic.

To understand the origin of the term "anticaustic" we must review some basic concepts from the differential geometry of plane curves [30]. A system of rays and wavefronts are *dual* to each other, in the sense that the rays are normal lines to the wavefronts, and the wavefronts are "level curves" for the rays (identifying points corresponding to equal travel times along each ray). The *caustic*[5] is the envelope of a system of rays (i.e., the curve along which light tends to concentrate — see Figure 16). The *evolute* of any given curve is the locus of its centers of curvature, which is identical to the envelope of its normal lines. Since wavefronts have common normal lines, they have the

[5] The name is due to Ehrenfried Walther von Tschirnhaus, a contemporary of Leibniz and Newton, who investigated the use of mirrors to focus sunlight.

same evolute — namely, the caustic. There are infinitely many curves, called *involutes*, that possess the same evolute, e.g., the wavefronts corresponding to a given caustic. Among these wavefronts, a distinguished "initial" member exists, that is algebraically simpler than all the other wavefronts — this is what Bernoulli calls the *anticaustic*. A more comprehensive treatment of the geometry of rays, wavefronts, and caustics may be found in [28].

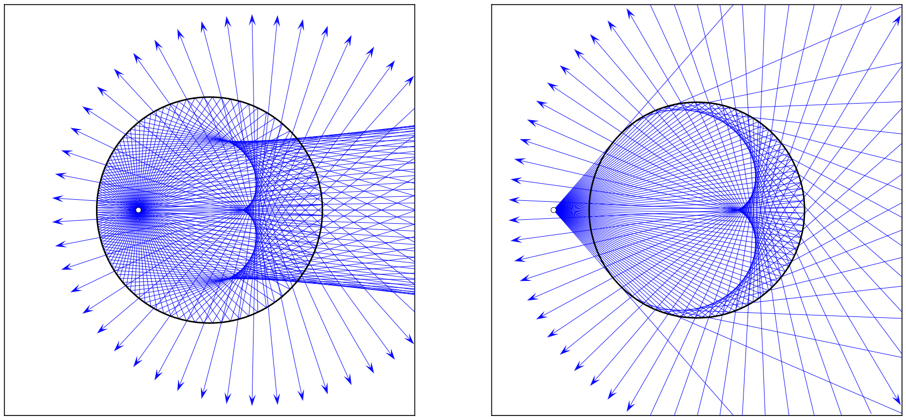

Figure 16: Typical caustic curves for reflection by a spherical surface of rays from a point source that is located inside (left) and outside (right) the sphere.

The example in Figure 14 illustrates the simplest non–trivial instance of an anticaustic. Lenses for cameras, telescopes, microscopes and other optical devices typically involve refraction by spherical air/glass surfaces. When the planar interface in Figure 14 is replaced by a *spherical* interface we obtain a *Cartesian oval* as the anticaustic (for the generic case of a point source not at the center of the sphere). Thus, a spherical wave refracted by a spherical surface is, in general, a parallel/offset curve to the Cartesian oval — which is [6] an irreducible algebraic curve of degree 14, that defines both the "forward" and "backward" propagating wavefronts. Since modern camera lenses often involve dozens of spherical refracting surfaces, the imaging process produces wavefronts of daunting algebraic complexity.

Figure 17 shows examples of waves from a point source being refracted by a spherical interface, for different positions of the source relative to the sphere center (the Cartesian oval anticaustic is shown in green, the blue curve is the refracted wave, and a sampling of rays is indicated in red).

The blue curves are *interior* offsets to the Cartesian oval — the exterior offsets do not contribute to the physical refracted wave. In the first example, the refracted wave is continuous and amounts to the entire interior offset to the inner Cartesian oval loop, which propagates inward and after "collapsing" continues propagating outward. In this case, the interior offset to the outer Cartesian oval loop does not contribute to the physical refracted wave.

The second example illustrates the phenomenon known as *total internal reflection*. For incidence angles $\theta_p > \sin^{-1} q/p$ in Snell's law (17), there is no refracted ray emerging into the medium of index $q$ — it is entirely reflected back into the medium of index $p$. This incurs a *discontinuous* refracted wave, corresponding to a range of angles over which the wave is reflected back into the sphere. In this case, the disjoint portions of the refracted wave are subsets of the interior offsets to both the inner and outer Cartesian oval loops.

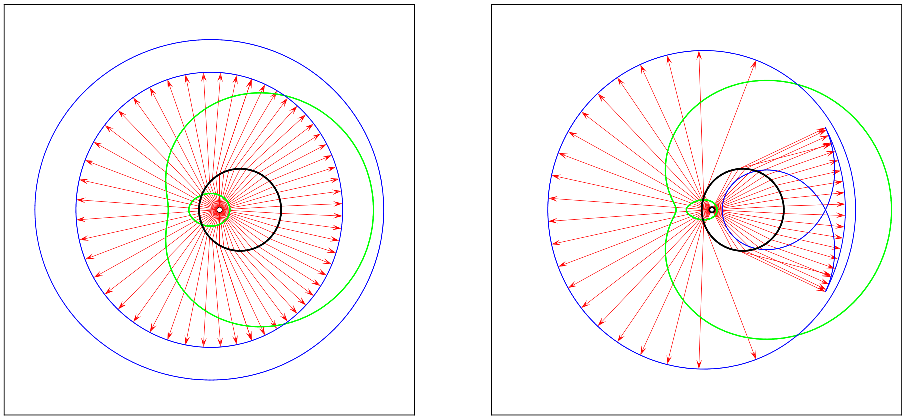

Figure 17: Waves from a point source being refracted by a spherical surface: a continuous refracted wave (left), and a discontinuous refracted wave (right) owing to total internal reflection. The anticaustic is the green Cartesian oval.

These examples illustrate how the remarkable algebraic complexity of an apparently simple problem — the refraction of a spherical wave by a spherical surface — manifests itself in subtle and unexpected geometrical behavior.

## 9 Closure

Notwithstanding its status as a natural generalization of the conics and its many remarkable geometrical properties, the Cartesian oval has (with few exceptions) not attracted much attention in recent decades. By elucidating its algebraic and geometrical properties, and describing its key role in two areas — the Minkowski geometric algebra of complex sets, and the problem of refraction of spherical waves by spherical surfaces — we hope that this brief introduction may serve to spark renewed interest.

There are many other interesting aspects of Cartesian ovals that, in the interests of brevity, we are unable to describe here. We mention, for example, the interpretation of the Cartesian oval as the projection of the intersection curve of two cones with parallel axes onto a plane orthogonal to those axes; or as the boundary between two crystal domains or bacterial colonies that grow from distinct point sources at different speeds [32]. Another fascinating topic is the connection of Cartesian ovals to the theory of elliptic functions — as illustrated, for example, by the fact that under the mapping $\mathbf{w} = u + \mathrm{i}\,v \rightarrow \mathbf{z} = x + \mathrm{i}\,y$ of the complex plane defined by the Weierstrass elliptic function, $\mathbf{z} = \wp(\mathbf{w})$, the images of the coordinate lines $u$ = constant, $v$ = constant in the $\mathbf{w}$–plane are Cartesian ovals in the $\mathbf{z}$–plane [18].

# References

[1] J.–C. A. Chastang and R. T. Farouki (1992), The mathematical evolution of wavefronts, *Optics and Photonics News* **3**, 20–23.

[2] J. Bernoulli (1692), Lineæ cycloidales, evolutæ, ant–evolutæ, causticæ, anti–causticæ, peri–causticæ, *Acta Eruditorum*, May 1692.

[3] C. B. Boyer (2004), *History of Analytic Geometry*, Dover Publications, Mineola, New York (reprint).

[4] J. L. Coolidge (1916), *A Treatise on the Circle and the Sphere*, Clarendon Press, Oxford.

[5] D. E. Smith and M. L. Latham, translators (1954), *The Geometry of René Descartes, with a Facsimile of the First Edition*, Dover Publications, New York.

[6] R. T. Farouki and J.–C. A. Chastang, Exact equations of “simple” wavefronts, *Optik* **91**, 109–121 (1992)

[7] R. T. Farouki, W. Gu, and H. P. Moon (2000), Minkowski roots of complex sets, *Geometric Modeling and Processing 2000*, IEEE Computer Society Press, 287–300.

[8] R. T. Farouki and C. Y. Han (2005), Solution of elementary equations in the Minkowski geometric algebra of complex sets, *Advances in Computational Mathematics* **22**, 301–323.

[9] R. T. Farouki, H. P. Moon, and B. Ravani (2000), Algorithms for Minkowski products and implicitly–defined complex sets, *Advances in Computational Mathematics* **13**, 199–229.

[10] R. T. Farouki, H. P. Moon, and B. Ravani (2001), Minkowski geometric algebra of complex sets, *Geometriae Dedicata* **85**, 283–315.

[11] R. T. Farouki and H. Pottmann (2002), Exact Minkowski products of $N$ complex disks, *Reliable Computing* **8**, 43–66.

[12] F. Gomes Teixeira (1905), *Tratado de las Curvas Especiales Notables*, Gaceta de Madrid, Madrid, Spain.

[13] F. Gomes Teixeira (1908), *Traité des Courbes Spéciales Remarquables Planes et Gauches*, French translation in 3 volumes, Chelsea (reprint), New York.

[14] J. Harkness and F. Morley (1893), *A Treatise on the Theory of Functions*, Macmillan and Co., New York.

[15] E. Hecht and A. Zajac (1974), *Optics*, Addison–Wesley, Reading, Massachusetts.

[16] C. Huygens (1690), *Treatise on Light*, Dover, New York.

[17] M. Klein (1972), *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, Oxford.

[18] D. F. Lawden (1989), *Elliptic Functions and Applications*, Applied Mathematical Sciences, Volume 80, Springer, New York.

[19] J. D. Lawrence (1972), *A Catalog of Special Plane Curves*, Dover, New York.

[20] E. H. Lockwood (1967), *A Book of Curves*, Cambridge Univ. Press.

[21] H. Minkowski (1903), Volumen und Oberfläche, *Mathematische Annalen* **57**, 447–495.

[22] R. E. Moore (1966), *Interval Analysis*, Prentice–Hall, Englewood Cliffs, NJ.

[23] T. Needham (1997), *Visual Complex Analysis*, Clarendon Press, Oxford.

[24] G. Salmon (1960), *A Treatise on the Higher Plane Curves: Intended as a Sequel to A Treatise on Conic Sections*, Chelsea Publishing Co. (reprint), New York.

[25] H. Schwerdtfeger (1979), *Geometry of Complex Numbers*, Dover, New York.

[26] D. H. Von Seggern (1993), *CRC Standard Curves and Surfaces*, CRC Press, Boca Raton.

[27] C. H. Sisam (1907), Review of *Tratado de las Curvas Especiales Notables*, *Bulletin of the American Mathematical Society* **13**, 249–250.

[28] O. N. Stavroudis (1972), *The Optics of Rays, Wavefronts, and Caustics*, Academic Press, New York.

[29] J. Steiner (1846), Geometrische lehrsätze, *Journal für die Reine und Angewandte Mathematik* **32**, 182–184.

[30] D. J. Struik (1961), *Lectures on Classical Differential Geometry*, Dover Publications (reprint), New York.

[31] B. Williamson (1893), *An Elementary Treatise on the Differential Calculus: Containing the Theory of Plane Curves with Numerous Examples*, D. Appleton & Co., New York.

[32] C. Zwikker (1963), *The Advanced Geometry of Plane Curves and Their Applications*, Dover Publications (reprint), New York.

[33] Encyclopedia of Remarkable Mathematical Forms (website) http://mathcurve.com/courbes2d/descartes/descartes.shtml