Received: 6 July 2024 | Revised: 12 October 2024 | Accepted: 17 October 2024

DOI: 10.1112/jlms.70033

RESEARCH ARTICLE

Journal of the London Mathematical Society

Cusps of caustics by reflection in ellipses

Gil Bor[1] | **Mark Spivakovsky**[2,3] | **Serge Tabachnikov**[4]

[1]CIMAT, Guanajuato, México

[2]CNRS UMR, Institut de Mathmatiques de Toulouse, Toulouse, France

[3]Instituto de Matemáticas (Cuernavaca) LaSol, UMI CNRS, UNAM, Av. Universidad s/n. Col. Lomas de Chamilpa, Cuernavaca, Morelos, México

[4]Department of Mathematics, Penn State University, University Park, Pennsylvania, USA

Correspondence
Gil Bor, CIMAT, A.P. 402, Guanajuato, Gto. 36000, México.
Email: gil@cimat.mx

Funding information
CONAHCYT, Grant/Award Number: A1-S-45886; NSF, Grant/Award Numbers: DMS-2005444, DMS-2404535

Abstract

This paper is concerned with the billiard version of Jacobi's last geometric statement and its generalizations. Given a non-focal point O inside an elliptic billiard table, one considers the family of rays emanating from O and the caustic Γ_n of the reflected family after n reflections off the ellipse, for each positive integer n. It is known that Γ_n has at least four cusps and it has been conjectured that it has exactly four (ordinary) cusps. The present paper presents a proof of this conjecture in the special case when the ellipse is a circle. In the case of an arbitrary ellipse, we give an explicit description of the location of four of the cusps of Γ_n, though we do not prove that these are the only cusps.

MSC 2020
53A04, 78A05, 37C83 (primary)

Contents

1 | INTRODUCTION AND STATEMENT OF RESULTS

The motivation for this work goes back to Jacobi's 1842-3 'Lectures on Dynamics' [13]. Recall that the conjugate locus of a point on a surface is the locus of the first conjugate points on the geodesics that start at this point. Jacobi considered the conjugate locus of a non-umbilic point on the surface of a triaxial ellipsoid in 3-space. What is known as the *last geometric statement of Jacobi* is the claim that this conjugate locus has exactly four cusps; see Figure 1. We refer to [17] for a detailed historical discussion.

The last geometric statement of Jacobi was proved only recently [10]. In contrast, it was known for a long time that the conjugate locus of a generic point on a convex surface has at least four cusps; see [2] where this theorem is attributed to Carathéodory and [22] for a recent proof.

The conjugate locus of a point is also called the first caustic. One considers the loci of the second, third, etc., conjugate points on the geodesics emanating from a point; these are the second, third, etc., caustics. These curves are also the components of the envelope of the 1-parameter family of geodesics that start at this point. Figure 1 (right) depicts the first and second such caustics.

This article concerns the billiard versions of these problems. Birkhoff [1] suggested to consider billiard trajectories in a convex plane domain as the geodesics on a 'pancake', the surface obtained from the domain by infinitesimally 'thickening' it. This leads to the following set-up.

Consider an oval C, a smooth strictly convex closed curve in the plane, the boundary of a billiard table. Let O be a point inside C and consider the billiard trajectories that start at O. After n reflections off C, we obtain a 1-parameter family of lines whose envelope is a closed connected curve

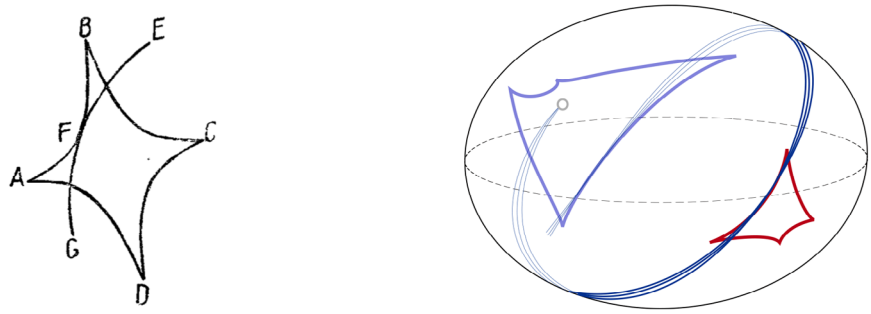

FIGURE 1 Left: A sketch of the conjugate locus from [13]. Right: The first (red) and second (blue) conjugate locus of a point on a triaxial ellipsoid.

in the real projective plane $\mathbb{RP}^2$, possibly with some cusps and self-intersections, called the nth *caustic by reflection* from O. The term caustic, meaning 'capable of burning', comes from optics, where C is an ideal mirror and O is a light source.

We refer to [4, 5] and the literature cited therein for the study of the first caustics by reflection, also known as *catacaustics*. In particular, Cayley studied the first caustics by reflection and refraction in a circle in his memoire [6] where he considered the cases when the source of light was inside the circle, on the circle, and outside the circle, including at infinity.

We proved in [3] that, for every $n \geqslant 1$, if O is a generic point inside an oval, then the nth caustic by reflection from O has at least 4 cusps. This is one of many variations on the classic 4-vertex theorem. Here are refined versions of two conjectures made in [3].

Conjecture 1. *If C is an ellipse and O is an interior point which is not a focus of C then, for all $n \geqslant 1$, the nth caustic by reflection from O has exactly four cusps, and all four are ordinary ones.*

See the Section 2.2 for a precise definition of 'ordinary cusp'.

Remark 1. The $n = 1$ case of Conjecture 1 (without the 'ordinary' part) can be thought of as a 'limiting case' of the Jacobi's last geometric statement, as one of the axes of the ellipsoid tends to 0.

Conjecture 2. *If an oval C is not an ellipse then there exists an $n \geqslant 1$ and an open set U inside C such that for every $O \in U$ the number of cusps of the nth caustic by reflection from O is greater than four.*

An analogue of Conjecture 1 for the caustics of geodesics emanating from a point on a tri-axial ellipsoid was experimentally studied in [17]. That paper contains numerous computer generated images of first, second, third and fourth caustics, each having exactly four cusps.

This article is a step toward proving Conjecture 1. To state our first result, we recall a well-known property of billiards in an ellipse.

An ellipse C defines two 1-parameter families of *confocal conics*, those conics which share their foci with C. One family consists of ellipses, the other of hyperbolas (including the major and minor axes of C). They form, in the complement of the foci of C, a double foliation so that through each point pass one confocal ellipse and one confocal hyperbola, intersecting orthogonally at the point. A ray (directed line), incident to the interior of C, is tangent to exactly one of these confocal conics (or incident to one of the foci), and after reflection off C it is tangent to the same conic; see Figure 2 (left).

Theorem 1. *Let O be a non-focal point inside an ellipse C, and let E and H be the ellipse and hyperbola (respectively), passing through O and confocal to C. Consider the four rays emanating from O and tangent to E and H (two each). Then after n reflections, the four rays are tangent to E and H at four points which are cusps of the nth caustic by reflection from O; see Figure 2 (right).*

Remark 2.

(a) If O lies on one of the axes of C, then the role of H in the above theorem is played by this axis. The location of the corresponding cusps along this axis is then determined by the 'mirror equation' of geometric optics; see Section 5.2.

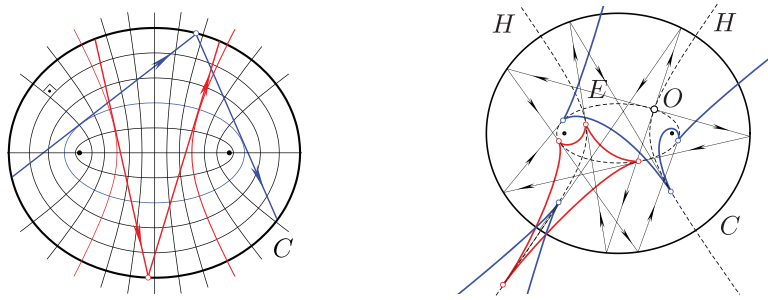

FIGURE 2 Left: A light ray reflected off an elliptical table C stays tangent to the same confocal conic, either an ellipse (blue) or a hyperbola (red). Right: The first (red) and the second (blue) caustics by reflection from O each have cusps at the four tangency points with the confocal conics through O of the four reflected rays emanating from O tangent to these conics.

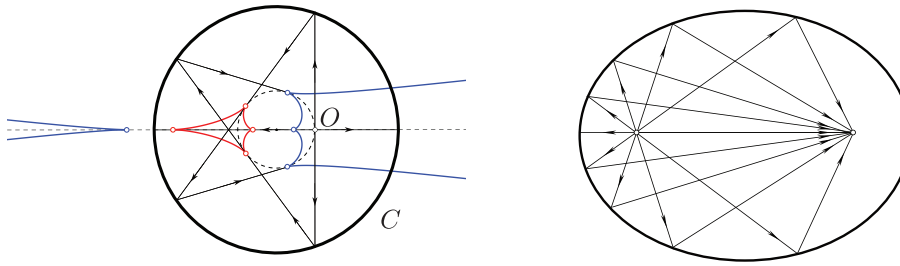

FIGURE 3 Left: Theorem 1 for a circle. Right: The nth caustic from a focus of an ellipse is the other focus for odd n, the same focus for even n.

(b) The limiting case when C is a circle is not excluded: in this case, the role of the two confocal conics through O are played by the concentric circle through O and the line through O and the center; see Figure 3 (left).
(c) If the point O is a focus of the ellipse, then the nth caustic by reflection degenerates to one of the two foci, depending on the parity of n; see Figure 3 (right).
(d) The stated location of the 4 cusps in Theorem 1 can be deduced from the conjectures made in [17] about the location of the cusps of caustics of envelopes of geodesics from a point on an ellipsoid.
(e) It is straightforward to extend Theorem 1 to an arbitrary non-degenerate conic section C (parabola and hyperbola). The complement of the closure of C in $\mathbb{RP}^2$ consists of two components, diffeomorphic to a disc and to a Möbius band, respectively. The former can serve as a billiard table, and our proof of Theorem 1 applies, mutatis mutandis, to it as well; see Figure 4.
(f) In Section 5.1, Theorem 1 is further extended to 'Liouville billiards', where the billiard table is formed by a coordinate line on a Liouville surface.

Thus, after Theorem 1, proving Conjecture 1 amounts to showing that the four cusps described by Theorem 1 are the *only* cusps of the nth caustic by reflection from O and that all four cusps are ordinary. We were able to show this only in the case when C is a circle, which is our next result.

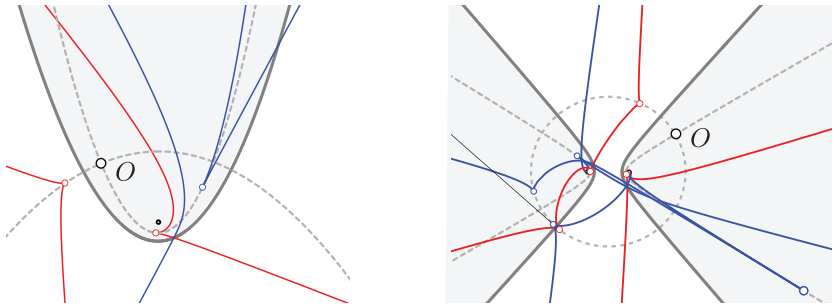

FIGURE 4 Theorem 1 holds for any convex billiard table in the projective plane, bounded by a conic; shown are a parabola (left) and a hyperbola (right). Some cusps are out of sight.

Theorem 2. *Conjecture 1 holds if C is a circle. Namely, if O is an interior point of a circle C, different from its center, then for every $n \geqslant 1$ there are exactly four cusps on the nth caustic by reflection from O; two of these cusps lie on the line passing through O and the center of the circle, the other two on the circle through O concentric with C. Furthermore, these four cusps are ordinary.*

The content of this article is as follows. In Section 2 we recall relevant facts about billiards in ellipses, envelopes of families of lines and their cusps. In Section 3 we prove Theorem 1 and in Section 4 we prove Theorem 2. Section 5 contains various additional results and suggested problems.

2 | PRELIMINARIES

2.1 | Billiards in ellipses

Let us recall relevant facts concerning billiards in ellipses, in particular, their complete integrability; see, for example, [9, 12, 20].

Consider a billiard table C bounded by an ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \quad \text{where } 0 < b \leqslant a.$$

Associated with the billiard table C is a dynamical system whose phase space $\mathcal{L}$ (topologically a cylinder) is the space of rays (oriented lines) that intersect the interior of C. The billiard transformation T is the transformation of the phase space that sends an incoming ray to the outgoing one upon reflection off C; see Figure 6 (left).

The phase cylinder $\mathcal{L}$ admits a T-invariant area form. If a ray is characterized by its direction α and the signed distance from the origin p (see Figure 5), then the area form is $dp \wedge d\alpha$. This fact is not specific to ellipses: this area form is invariant under the billiard transformation in a billiard table of any shape.

The ellipse C is included in a confocal family of conics

$$C_\lambda : \frac{x^2}{a^2 - \lambda} + \frac{y^2}{b^2 - \lambda} = 1, \quad \lambda \in (-\infty, b^2) \cup (b^2, a^2).$$

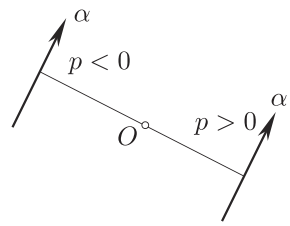

FIGURE 5 The coordinates (α, p) on the space of oriented lines.

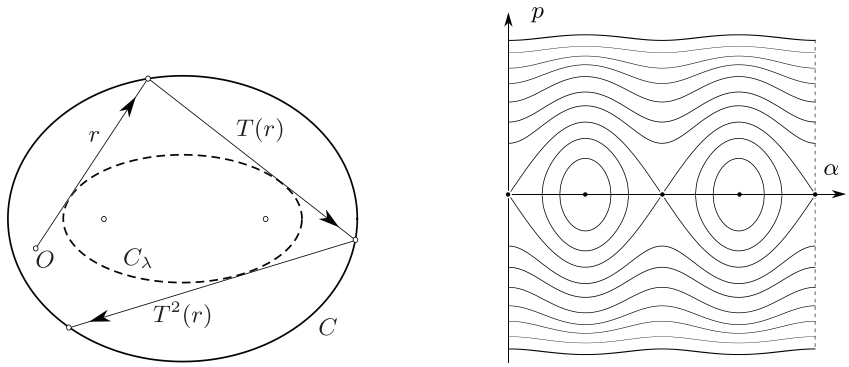

FIGURE 6 Left: A billiards trajectory in an ellipse and the associated confocal conic. Right: The phase space $\mathcal{L}$ of the billiard transformation T in an elliptical table (topologically a cylinder), and its T-invariant foliation, which is regular away from the four marked points on the α axis, corresponding to rays aligned with the major and minor axis of the table. Reversing the orientation of a ray corresponds to the 'glide-reflection' $(\alpha, p) \mapsto (\alpha + \pi, -p)$. The ∞-shaped curve corresponds to rays incident to the foci of the ellipse, phase curves inside it correspond to rays tangent to confocal hyperbolas (including their asymptotes), phase curves outside it to rays tangent to confocal ellipses.

This is an ellipse for $\lambda < b^2$ and a hyperbola for $b^2 < \lambda < a^2$. For $0 < \lambda < b^2$ the confocal ellipse C_λ is contained in the interior of C, for $\lambda < 0$ it is contained in the exterior of C. For $\lambda = 0$ one has $C_0 = C$.

As λ tends to b^2 on the left, the confocal ellipse C_λ tends to the line segment on the x-axis connecting the two foci of C; the right limit is the closure of the complement of this segment in the x-axis. As λ tends to a^2 on the left, C_λ tends to the y-axis.

A ray $r \in \mathcal{L}$, not incident to one of the foci of C, is tangent to a unique conic C_λ from this confocal family, so λ can be considered as a function on $\mathcal{L}$. As is easy to show, it is given by

$$\lambda = (a \sin \alpha)^2 + (b \cos \alpha)^2 - p^2.$$

This formula shows that λ extends smoothly to all of $\mathcal{L}$, including rays incident to the foci.

After reflection, the ray $T(r)$ is tangent to the same conic [20, Theorem 4.4]. Thus the level curves of λ define a (singular) T-invariant foliation of the phase space $\mathcal{L}$, whose leaves consist of the rays tangent to a fixed conic; see Figure 6 (right).

Note that the resulting foliation is non-singular away from the four marked points on the α-axis (the critical points of λ), corresponding to the rays aligned with the major and minor axes of C. Note also that each level curve of a regular value $\lambda \in (0, b^2) \cup (b^2, a^2)$ has *two* connected

components. For $\lambda \in (0, b^2)$ (rays tangent to a fixed confocal ellipse), each of the two components is T-invariant. For $\lambda \in (b^2, a^2)$ (rays tangent to a fixed confocal hyperbola, including its asymptots), the two components are interchanged by T. The figure ∞ (the level curve $\lambda = b^2$) corresponds to rays passing through the foci. Orientation reversing acts on $\mathcal{L}$ by $R : (\alpha, p) \mapsto (\alpha + \pi, -p)$, satisfying $R^2 = (RT)^2 = id$. The two reflections about the major and minor axes of C induce maps of $\mathcal{L}$ commuting with T.

The following proposition is a special case of the Arnold–Liouville theorem on completely integrable Hamiltonian systems [21]. We will give a self-contained proof in our case, following [20, Chapter 4].

Proposition 1. *On each leaf γ of the T-invariant foliation of $\mathcal{L}$ there is a T-invariant non-vanishing 1-form, well defined up to multiplicative constant. Consequently, there is a local coordinate t on γ in which T is given by $T(t) = t + c$ for some constant c.*

Proof. Choose a smooth function f without critical points in a neighborhood of γ, which is constant on each leaf of the T-invariant foliation (for example, $f = \lambda$). Then $f \circ T = f$ implies $T^* df = df$. Let X_f be the Hamiltonian vector field associated to f, that is, $\omega(X_f, \cdot) = df$, where $\omega = dp \wedge d\alpha$ is the T-invariant area form on $\mathcal{L}$. Since both df and ω are T-invariant, the same holds for X_f. Since X_f is non-vanishing and tangent to γ, there is a unique 1-form α on γ such that $\alpha(X_f) = 1$. Since X_f is T-invariant, so is α. In neighborhoods of a point $r \in \gamma$ and its image $T(r)$, one can find coordinates t and t_1, respectively, such that $\alpha = dt$ near r and $\alpha = dt_1$ near $T(r)$. It follows that $0 = T^*\alpha - \alpha = d(t_1 \circ T - t)$, thus $t_1 \circ T - t = c$ for some constant c; that is, T is given near r by $t_1 = t + c$.

If one replaces f by another function, say $g = \phi(f)$, then the corresponding vector field changes to $X_g = (\phi' \circ f) X_f$, that is, a non-zero constant multiple of X_f along γ, so α is also changed by a constant multiple. □

Once a choice of coordinate t is made on each level curve of λ, one can use (t, λ) as coordinates on $\mathcal{L}$ (away from singular leaves); the ray $r(t, \lambda)$ is tangent to the confocal conic corresponding to the parameter λ, such that $T(r(t, \lambda)) = r(t + c(\lambda), \lambda)$.

Remark 3. It is important to note that the choice of the t coordinate in the last proposition depends only on the T-invariant foliation of $\mathcal{L}$, which in turn depends on the family of conics confocal to the billiard table C, and not on a particular choice of conic within this family as a billiard table. That is, if one chooses, as a billiard table, any conic confocal to C, say C_1, then the associated billiard map T_1 with respect to C_1 admits the same invariant foliation of $\mathcal{L}$ as T, and is thus given in the t coordinate on an invariant leaf by the same kind of formula as the formula for T in Proposition 1, $T_1(t) = t + c_1$ (though the constant shift c_1 may differ from c).

For example, consider an ellipse E from a confocal family. One may think of t as a coordinate on E. Then the locus of the intersection points of the tangent to E, whose t-coordinates differ by a constant, is a confocal ellipse, and if the half-sum of the two t-coordinates is constant, then this locus is a confocal hyperbola. We refer to [9, 12] and to detailed discussions in [18, 19].

Remark 4. Note that the T-invariant leaf γ in Proposition 1 need not be connected for the proposition to hold. Indeed, each level curve of λ has two components (each topologically a circle); in the elliptic case (level curves above and below the ∞ shape in Figure 6 (right)), each of these

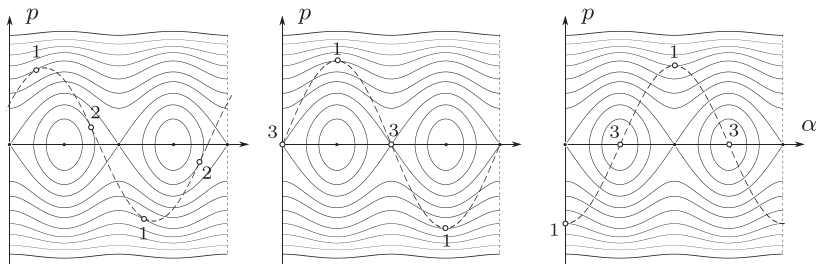

FIGURE 7 The three types of inflection points of pencils. In each figure, the pencil is the dotted curve, with four points and their type marked on it. Left: O does not lie on an axis of C (the generic case). Middle: O lies on the major axis, between a vertex and the nearby focus. Right: O lies on the minor axis.

components is T-invariant, while in the hyperbolic case (level curves inside the ∞ shape in Figure 6 (right)), the two components are interchanged by T. By Proposition 1, even in this hyperbolic case, one can put a coordinate on each of the two components, say t on one component and t_1 on the other, such that T is given by $T(t,\lambda) = (t_1 + c, \lambda)$, $T(t_1,\lambda) = (t + c, \lambda)$, for some constant c (depending on λ).

2.2 | Families of rays, envelopes, cusps

The conjectures and theorems of the Introduction concern families of rays, their envelopes (or caustics) and cusps. We briefly review here the pertinent definitions; see, for example, Section 8.4 of [8].

Define the 'line' dual to a point in $\mathbb{R}^2$ as the curve in $\mathcal{L}$ corresponding to the set of rays incident to the point (a 'pencil' of rays); see the dotted curves in Figure 7. We also include 'lines' dual to 'points at infinity', corresponding to pencils of parallel rays sharing a common direction (vertical lines in the (α, p) coordinates on $\mathcal{L}$). This defines a 2-parameter family of curves in $\mathcal{L}$, a unique curve through each given point in a given tangent direction at this point.

Definition 1. Given a 1-parameter family of rays, that is, a smooth curve $\gamma \subset \mathcal{L}$, an *inflection* point of γ of order $m \geqslant 2$ is a point where the tangent 'line' to γ at this point has contact of order m with γ (the tangent 'line' to a curve has typically contact of order 1).

Definition 2. The *envelope* (or *caustic*) of γ is an oriented plane curve Γ whose set of tangent lines is γ.

Note that Γ is a curve in the projective plane $\mathbb{RP}^2$, possibly singular, or even reduced to a single point, if γ consists of all lines though this point (the 'line' in $\mathcal{L}$ dual to the point).

Definition 3. An m-cusp of a plane curve Γ, $m \geqslant 2$, is a point for which there is a C^1-diffeomorphism taking a neighborhood of the point to a neighborhood of the origin in the (x, y)-plane, taking the point to (0,0) and Γ to the curve $y^m = x^{m+1}$. An *ordinary cusp* is a 2-cusp (or a semi-cubical cusp).

A useful basic characterization of m-cusps is the following. Let γ be a smooth 1-parameter family of rays with envelope Γ. Then an m-cusp of Γ corresponds to an inflection point of γ of order m; see [8, Example 8.2].

3 | PROOF OF THEOREM 1

We restate here Theorem 1 from the Introduction.

Theorem 1. *Let O be a non-focal point inside an ellipse C, and let E and H be the confocal ellipse and hyperbola (respectively) passing through O. Consider the four rays emanating from O and tangent to E and H (two each). Then after n reflections, the 4 rays are tangent to E and H at 4 points which are cusps of the nth caustic by reflection from O.*

To prove Theorem 1, we first reformulate it as a statement about the inflection points of a curve in the phase cylinder $\mathcal{L}$, as explained in Section 2.2.

Consider the pencil of rays incident to O and let γ be the corresponding curve in $\mathcal{L}$ (a 'line'). There are four points on γ, corresponding to the four rays tangent to the confocal conics E and H at O. The dual statement to Theorem 1 is then that *T^n maps these four points to inflection points of $T^n(\gamma)$*.

We proceed as follows. Let $r_0 \in \gamma$ be one of these four rays. We separate the proof into three cases (see Figure 7).

1. The ray r_0 is one of the two rays tangent to the confocal ellipse E through O. In this case, O may not lie on the line segment connecting to two foci.
2. The ray r_0 is one of the two rays tangent to the confocal hyperbola H through O. In this case, O may not lie on the minor axis of C, nor on the major axis, on the complement of the line segment connecting the two foci.
3. The point O lies on one of the axes of C and r_0 is one of the two rays aligned with this axis. In this case O may be the center of C.

3.1 | Case 1

Let $E = C_{\lambda_0}$, $b^2 < \lambda_0 < a^2$, be the confocal ellipse passing through O and $r_0 \in \gamma$ one of the 2 rays tangent to C_{λ_0} at O. The T-invariant curve in $\mathcal{L}$ passing through r_0 is given by $\lambda = \lambda_0$ in the (t, λ) coordinates.

Note. We use r_0 to denote both a point in $\mathcal{L}$ and the corresponding ray in $\mathbb{R}^2$.

Lemma 1. *r_0 is a tangency point of γ with the T-invariant phase curve $\lambda = \lambda_0$.*

Proof. The rays of the pencil close to r_0 are tangent to confocal ellipses with a greater value of the parameter λ; see Figure 8 (left). It follows that γ, near r_0, drawn in the (t, λ) plane, lies above the horizontal line $\lambda = \lambda_0$ and is therefore tangent to it at r_0; see Figure 8 (right). □

Lemma 2. *$T(r_0)$ is an inflection point of $T(\gamma)$.*

Proof. Let $r_0 = (t_0, \lambda_0)$, $r_1 = (t_1, \lambda_0) = T(r_0)$, the reflection of r_0 by C, where $t_1 = t_0 + c(\lambda_0)$. Then r_1 is tangent to C_{λ_0} at some point, O_1. Let γ_1 be the 'line' dual to O_1, corresponding to the pencil of rays through O_1. To show that r_1 is an inflection point of $T(\gamma)$ it is then enough to show that the 2-jets at r_1 of $T(\gamma)$ and γ_1 coincide; see Figure 9.

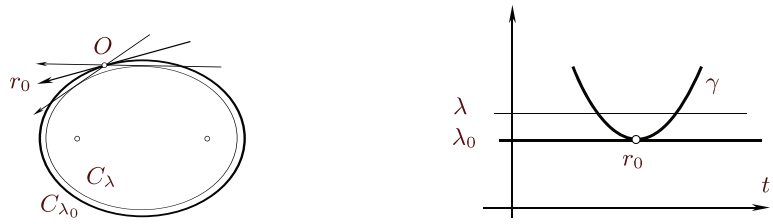

FIGURE 8 Lemma 1.

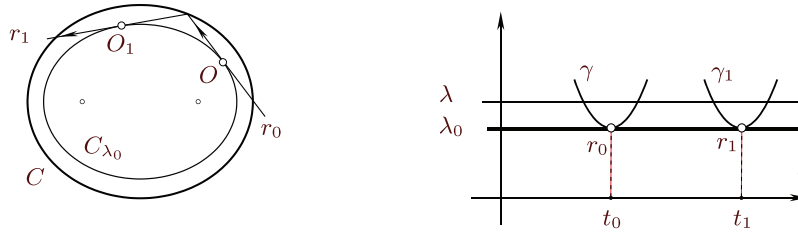

FIGURE 9 Proof of Lemma 2.

First, $r_1 \in \gamma_1$, so γ_1 and $T(\gamma)$ intersect at r_1 (their 0-jets coincide). Second, γ is tangent to the T-invariant horizontal line $\lambda = \lambda_0$ at r_0 (Lemma 1) hence $T(\gamma)$ is tangent to $\lambda = \lambda_0$ at $r_1 = T(r_0)$. The same holds for γ_1, by Lemma 1, hence γ_1 and $T(\gamma)$ are tangent at r_1 (their 1-jets coincide).

Next, the curve γ intersects the horizontal line at a level $\lambda > \lambda_0$ at two points, corresponding to the rays shown in Figure 8 (left). The billiard reflection in the ellipse with parameter λ_0 (the outer ellipse in Figure 8 (left)) takes one of these rays to the other one. The difference of the t-coordinates of these two intersection points depends only on λ and λ_0, but not on t_0 (see Remark 3 of Section 2.1). It follows that the 2-jets of γ and γ_1, at r_0 and r_1 (respectively), are parametrized by

$$\gamma : \varepsilon \mapsto (t_0 + \varepsilon, \lambda_0 + a\varepsilon^2), \quad \gamma_1 : \delta \mapsto (t_1 + \delta, \lambda_0 + a\delta^2), \tag{1}$$

where $a = a(\lambda_0)$, $t_1 = t_0 + c(\lambda_0)$.

Note. All calculations for the rest of the proof of this lemma are mod ε^3 and δ^3.

Now $T(t, \lambda) = (t + c(\lambda), \lambda)$, hence the 2-jet of $T(\gamma)$ at r_1 is parametrized by

$$\begin{aligned} T(\gamma) : \varepsilon \mapsto & (t_0 + \varepsilon + c(\lambda_0 + a\varepsilon^2), \lambda_0 + a\varepsilon^2) \\ & = (t_0 + \varepsilon + c(\lambda_0) + ac'(\lambda_0)\varepsilon^2, \lambda_0 + a\varepsilon^2) \\ & = (t_1 + \varepsilon + ac'(\lambda_0)\varepsilon^2, \lambda_0 + a\varepsilon^2). \end{aligned}$$

Next we reparametrize this 2-jet by setting

$$\delta = \varepsilon + ac'(\lambda_0)\varepsilon^2,$$

with inverse (mod δ^3),

$$\varepsilon = \delta - ac'(\lambda_0)\delta^2.$$

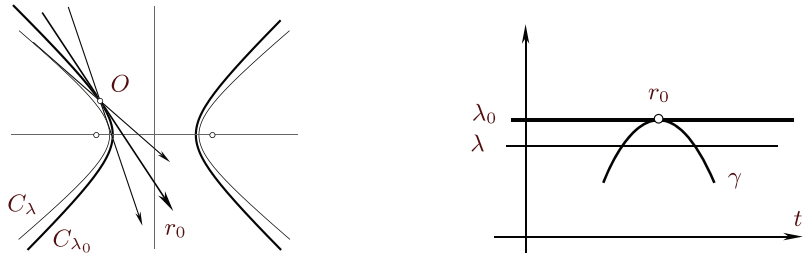

FIGURE 10 Case 2 of Lemma 2. Compare to Figure 8.

It follows that the 2-jet of $T(\gamma)$ at r_1 can be reparametrized as

$$T(\gamma) : \delta \mapsto (t_1 + \delta, \lambda_0 + a\delta^2),$$

coinciding with the expression (1) for the 2-jet of γ_1 at r_1, as needed. □

Note that the last two lemmas are statements about the 2-jet of γ at r_0. That is, they remain valid if one replaces γ with a curve whose 2-jet at r_0 coincides with that of γ. We thus conclude: if r_0 is an inflection point of a curve $\gamma \subset \mathcal{L}$, which is also a point of tangency of γ with the leaf of the T-invariant foliation of $\mathcal{L}$ dual to an ellipse E confocal to C, then the same holds for $T(r_0) \in T(\gamma)$. It follows by induction on n that the same holds for $T^n(r_0) \in T^n(\gamma)$. This proves Case 1 of Theorem 1.

3.2 | Case 2

This case is very similar to the previous one, so we omit the details. We only note that in this case, like in Case 1, the T-invariant leaf $\lambda = \lambda_0$ consists of two components, but unlike Case 1, T^n, for n odd, interchanges the two components; the argument however is unaffected; see Remark 4 and Figure 10.

3.3 | Case 3

This case is simpler than the previous two. First, a lemma.

Lemma 3. *Let ρ denote the involution of $\mathcal{L}$ induced by the reflection about one of the axes of C, major or minor. Let r_0 be one of the two fixed points of ρ (a ray aligned with the axis of reflection) and $\gamma \subset \mathcal{L}$ a ρ-invariant curve containing r_0. Then r_0 is an inflection point of γ.*

Proof. Assume that ρ is given by reflection about the major axis of C (the x-axis) and r_0 is the ray along this axis, oriented eastwards. We use the coordinates (α, p) on $\mathcal{L}$, see Figure 5. Then $\rho(\alpha, p) = (-\alpha, -p)$ and $r_0 = (0, 0)$. Assume the tangent to γ at r_0 is not vertical. Then the 2-jet of γ at r_0 can be parametrized by

$$\varepsilon \mapsto (\varepsilon, a\varepsilon + b\varepsilon^2), \tag{2}$$

for some $a, b \in \mathbb{R}$. This is mapped by ρ to

$$\varepsilon \mapsto (-\varepsilon, -a\varepsilon - b\varepsilon^2).$$

Renaming $-\varepsilon$ by ε, this 2-jet of $\rho(\gamma)$ at r_0 can be reparametrized as

$$\varepsilon \mapsto (\varepsilon, a\varepsilon - b\varepsilon^2). \tag{3}$$

Since $\rho(\gamma) = \gamma$ and r_0 is a fixed point of ρ, the 2-jets (2) and (3) must coincide. It follows that $b = 0$, hence the 2-jet of γ at r_0 is parametrized by

$$\varepsilon \mapsto (\varepsilon, a\varepsilon). \tag{4}$$

On the other hand, the tangent 'line' to γ at r_0 is the graph of $p = a \sin\alpha$ (see Lemma 4 below). Its 2-jet at r_0 is given by formula (4). This shows that r_0 is an inflection point of γ.

If the tangent to γ at r_0 is vertical then the tangent 'line' at r_0 is $\alpha = 0$ and we can parametrize the 2-jet of γ at r_0 by $\varepsilon \mapsto (a\varepsilon^2, \varepsilon)$. As before, ρ-invariance of γ implies that $a = 0$, hence the 2-jets of γ and the $\alpha = 0$ at r_0 coincide. Thus in this case r_0 is an inflection point of γ as well.

The other three cases, where $r_0 = (\pi, 0)$ and ρ is the reflection about the x-axis, or ρ is the reflection about the y-axis and $r_0 = (\pm\pi/2, 0)$, are treated similarly and their proof is omitted. □

We can now complete the proof of Case 3 of Theorem 1. Let O be a point on one of the axes of C (major or minor, or both, when O is the center of C, if C is not a circle). Let $\gamma \subset \mathcal{L}$ be the dual 'line' (the curve corresponding to the pencil of rays through O). Let $r_0 \in \gamma$ be one of the two rays aligned with the axis through O. Then γ is ρ-invariant and r_0 is a fixed point of ρ. Clearly, ρ and T commute, hence $T^n(\gamma)$ is ρ-invariant and $T^n(r_0) \in T^n(\gamma)$ is a fixed point of ρ. Lemma 3 implies that $T^n(r_0)$ is an inflection point of $T^n(\gamma)$, as needed.

4 | PROOF OF THEOREM 2

4.1 | Two lemmas

The billiard table C here is the unit circle $x^2 + y^2 = 1$. We use the same coordinates (α, p) in the space of oriented lines in $\mathbb{R}^2$ that were introduced in Section 2.1, Figure 5.

Lemma 4. *The pencil of rays through a point $(a, b) \in \mathbb{R}^2$, the 'line' dual to (a, b), is given by the equation*

$$p(\alpha) = a \sin\alpha - b \cos\alpha. \tag{5}$$

See Figure 11.

Proof. Let $(a, b) = r(\cos\theta, \sin\theta)$ and $\alpha' = \pi/2 - \alpha$. Then

$$\begin{aligned} p &= r\cos(\theta + \alpha') = r(\cos\theta\cos\alpha' - \sin\theta\sin\alpha') \\ &= r(\cos\theta\sin\alpha - \sin\theta\cos\alpha) = a\sin\alpha - b\cos\alpha. \end{aligned}$$

□

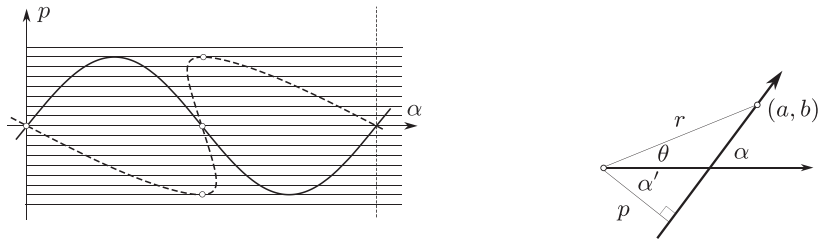

FIGURE 11 Left: The solid curve represents the pencil of rays through a point inside a circular table C. The dotted curve is its image under the billiard map T. The 4 marked points on it are its inflection points. The horizontal lines are the leaves of the T-invariant foliation of the phase cylinder $\mathcal{L}$. Right: The proof of Lemma 4.

Let γ be a curve in the phase space $\mathcal{L}$. Using the same terminology as in Section 3, an inflection point of γ is a second-order tangency with the 'line' tangent to γ at the point. If γ is the graph of a function $p(\alpha)$, the tangent 'line' is a graph of a function given by (5), that is, a solution to the ODE $f'' + f = 0$, hence the inflection points of γ are given by the zeros of the function $p''(\alpha) + p(\alpha)$.

If a line tangent to γ is vertical, that is, γ is tangent at $r_0 = (\alpha_0, p_0)$ to the vertical line $\alpha = \alpha_0$, then γ is the graph of a function $\alpha(p)$ near r_0, and r_0 is an inflection point if and only if $\alpha(p) = \alpha_0 + O(|p - p_0|^3)$, degenerate if $\alpha(p) = \alpha_0 + O(|p - p_0|^4)$.

Next consider a map $T : \mathcal{L} \to \mathcal{L}$ given by

$$T(\alpha, p) = (\tilde{\alpha}, p),\ \tilde{\alpha} = \alpha + \phi(p) \pmod{2\pi},$$

where $\phi(p)$ is some function. Let (α_0, p_0) be the coordinates of a point r_0 on a curve $\gamma \subset \mathcal{L}$, the graph of a function $p(\alpha)$. We ask: What is the condition on the second-order jets of $p(\alpha)$ and $\phi(p)$ at α_0 and p_0 (respectively) so that $T(\gamma)$ has an inflection point at $T(r_0)$? The answer is given by the following lemma.

Lemma 5. *Let γ be the graph of $p(\alpha)$, $r_0 = (\alpha_0, p_0) \in \gamma$, with*

$$p(\alpha_0 + \varepsilon) = p_0 + p_1\varepsilon + \frac{p_2}{2}\varepsilon^2 + O(\varepsilon^3),$$

$$\phi(p_0 + \delta) = \phi_0 + \phi_1\delta + \frac{\phi_2}{2}\delta^2 + O(\delta^3).$$

Then $T(r_0)$ is an inflection point of $T(\gamma)$ if and only if

$$p_2 + p_0(1 + p_1\phi_1)^3 = p_1^3\phi_2. \tag{6}$$

Proof. Calculating mod ε^3, δ^3 throughout, set

$$\delta = p_1\varepsilon + \frac{p_2}{2}\varepsilon^2,$$

then

$$\phi(p(\alpha_0 + \varepsilon)) = \phi(p_0 + \delta) = \phi_0 + \phi_1\delta + \frac{\phi_2}{2}\delta^2$$

$$= \phi_0 + \phi_1 p_1\varepsilon + \frac{p_1^2\phi_2 + p_2\phi_1}{2}\varepsilon^2.$$

The 2-jet of γ at $r_0 = (\alpha_0, p_0)$ is parametrized by

$$\varepsilon \mapsto \left(\alpha_0 + \varepsilon,\ p_0 + p_1\varepsilon + \frac{p_2}{2}\varepsilon^2\right),$$

hence the 2-jet of $T(\gamma)$ at $T(r_0) = (\alpha_0 + \phi_0, p_0)$ is parametrized by

$$\varepsilon \mapsto \left(\alpha_0 + \phi_0 + (1 + p_1\phi_1)\varepsilon + \frac{p_1^2\phi_2 + p_2\phi_1}{2}\varepsilon^2,\ p_0 + p_1\varepsilon + \frac{p_2}{2}\varepsilon^2\right).$$

Let

$$\tilde{\varepsilon} := (1 + p_1\phi_1)\varepsilon + \frac{p_1^2\phi_2 + p_2\phi_1}{2}\varepsilon^2,$$

then, assuming $1 + p_1\phi_1 \neq 0$, one can invert this (mod $\tilde{\varepsilon}^3$),

$$\varepsilon = \frac{\tilde{\varepsilon}}{1 + p_1\phi_1} - \frac{p_1^2\phi_2 + p_2\phi_1}{2(1 + p_1\phi_1)^3}\tilde{\varepsilon}^2.$$

Thus the 2-jet of $T(\gamma)$ at $T(r_0)$ is parametrized by

$$\tilde{\varepsilon} \mapsto \left(\alpha_0 + \phi_0 + \tilde{\varepsilon},\ p_0 + \tilde{p}_1\tilde{\varepsilon} + \frac{\tilde{p}_2}{2}\tilde{\varepsilon}^2\right),$$

where

$$\tilde{p}_1 = \frac{p_1}{1 + p_1\phi_1},\ \tilde{p}_2 = \frac{p_2 - p_1^3\phi_2}{(1 + p_1\phi_1)^3}.$$

The inflection condition at r_1 is then $\tilde{p}_2 + p_0 = 0$, which reduces to the stated formula (6).

If $1 + p_1\phi_1 = 0$ then $p_1 = p'(\alpha_0) \neq 0$ so one can invert $p(\alpha)$ near α_0,

$$\alpha(p_0 + \delta) = \alpha_0 + \alpha_1\delta + \frac{\alpha_2}{2}\delta^2,$$

where

$$\alpha_1 = \frac{1}{p_1},\ \alpha_2 = -\frac{p_2}{p_1^3} \tag{7}$$

and

$$p_1 = \frac{1}{\alpha_1},\ p_2 = -\frac{\alpha_2}{\alpha_1^3}. \tag{8}$$

The inflection condition for $p(\alpha)$ at α_0 is $p_2 + p_0 = 0$. Substituting for p_2 from Equation (8), this is

$$\alpha_2 = p_0(\alpha_1)^3. \tag{9}$$

Now $T(\gamma)$ is the graph of $\alpha(p)+\phi(p)$, hence the inflection condition at $T(r_0)$ is

$$\alpha_2+\phi_2=p_0(\alpha_1+\phi_1)^3.$$

Substituting for α_1,α_2 from Equation (7), one obtains Equation (6). □

4.2 | Cusps by reflection in a circle

The billiard ball map inside the unit circle C is given by $T(\alpha,p)=(\alpha+2\arccos p,p)$. Fix a point $O=(a,b)$ inside C and let γ be the dual 'line' (5). This takes us to the setting of Lemma 2 with

$$p(\alpha)=a\sin\alpha-b\cos\alpha,\ \phi(p)=2n\arccos(p),\ -1<p<1.$$

We are looking for points $r_0=(\alpha_0,p_0)\in\gamma$ such that $T^n(r_0)$ is an inflection point of $T(\gamma)$. Using circular symmetry, we may assume, without loss of generality, that $\alpha_0=0$, $0\leqslant b<1$ and $0<a^2+b^2<1$. We substitute in formula (6)

$$p_1=a,\ p_2=-p_0=b,\ \phi_1=\frac{-2n}{\sqrt{1-b^2}},\ \phi_2=\frac{2bn}{(1-b^2)^{3/2}},$$

obtaining the inflection condition at $T^n(r_0)$:

$$b-b\left[1-\frac{2an}{\sqrt{1-b^2}}\right]^3=\frac{2a^3bn}{(1-b^2)^{3/2}}. \tag{10}$$

This is satisfied if $a=0$ or $b=0$, corresponding to four inflection points of the curve $T^n(\gamma)$, as described by Theorem 2.

We claim that there are no other solutions to Equation (10) with $n\geqslant 1$ and $0<a^2+b^2<1$. Set $x=a/\sqrt{1-b^2}$. Assuming $a,b\neq 0$, Equation (10) becomes

$$(4n^2-1)x^2-6nx+3=0. \tag{11}$$

The discriminant of this quadratic equation in x is a positive multiple of $1-n^2$. Thus Equation (11) has a solution with $n\geqslant 1$ only for $n=1$. But in this case the solution is $x=1$, that is, $a^2+b^2=1$, which is out of range.

4.3 | The four cusps are ordinary

Dually, this amounts to proving the non-degeneracy of the four inflection points of $T^n(\gamma)$. Suppose, without loss of generality, that $O=(a,0)$, $a>0$, hence γ is given by $p=a\cos\alpha$, and the inflection points of $T^n(\gamma)$ are $T^n(r_0)$, where $r_0=(0,0),(\pi,0)$ or $\pm(\pi/2,a)$.

Begin with $r_0=(0,0)$. The 3-jet of γ at this point is parametrized by

$$\varepsilon\mapsto\left(\varepsilon,a\varepsilon-\frac{a}{6}\varepsilon^3\right).$$

Then $r_n := T^n(r_0) = (n\pi, 0)$. We calculate mod ε^4:

$$\arccos\left(a\varepsilon - \frac{a}{6}\varepsilon^3\right) = \frac{\pi}{2} - a\varepsilon + \frac{a(1-a^2)}{6}\varepsilon^3,$$

hence the 3-jet of $T^n(\gamma)$ at r_n is parametrized by

$$\varepsilon \mapsto \left(n\pi + (1-2na)\varepsilon + \frac{na\left(1-a^2\right)}{3}\varepsilon^3, a\varepsilon - \frac{a}{6}\varepsilon^3\right). \tag{12}$$

Let

$$\tilde{\varepsilon} := (1-2na)\varepsilon + \frac{na\left(1-a^2\right)}{3}\varepsilon^3.$$

If $1-2na \neq 0$ this can be inverted,

$$\varepsilon = \frac{\tilde{\varepsilon}}{1-2na} - \frac{na\left(1-a^2\right)\tilde{\varepsilon}^3}{3(1-2na)^4},$$

so that

$$a\varepsilon - \frac{a}{6}\varepsilon^3 = \frac{a}{1-2na}\tilde{\varepsilon} - \frac{a\left(1-2na^3\right)}{6(1-2na)^4}\tilde{\varepsilon}^3.$$

The 3-jet of $T^n(\gamma)$ at r_n can thus be reparametrized as

$$\tilde{\varepsilon} \mapsto \left(n\pi + \tilde{\varepsilon}, \frac{a}{1-2na}\tilde{\varepsilon} - \frac{a\left(1-2na^3\right)}{6(1-2na)^4}\tilde{\varepsilon}^3\right).$$

The tangent 'line' at $r_n = (n\pi, 0)$ is the graph of

$$p(\alpha) = \frac{a}{1-2na}\sin(\alpha - n\pi),$$

with 3-jet at r_n parametrized by

$$\varepsilon \mapsto \left(n\pi + \varepsilon, \frac{a}{1-2na}\varepsilon - \frac{a}{6(1-2na)}\varepsilon^3\right).$$

This coincides with the 3-jet of $T^n(\gamma)$ at r_n if and only if

$$\frac{a\left(1-2na^3\right)}{(1-2na)^4} = \frac{a}{1-2na},$$

which simplifies to

$$(4n^2-1)a^2 - 6na + 3 = 0. \tag{13}$$

The only solution is $a = n = 1$, which is excluded.

Remark 5. We notice a mysterious coincidence between Equations (11) and (13). We could not find an explanation.

If $1 - 2na = 0$ then the parametrized 3-jet (12) becomes

$$\varepsilon \mapsto \left(n\pi + \frac{1-a^2}{6}\varepsilon^3, a\varepsilon - \frac{a}{6}\varepsilon^3 \right). \tag{14}$$

Let

$$\tilde{\varepsilon} := a\varepsilon - \frac{a}{6}\varepsilon^3,$$

with inverse

$$\varepsilon = \frac{\tilde{\varepsilon}}{a} + \frac{\tilde{\varepsilon}^3}{6a^3}.$$

Then (14) can be reparametrized as

$$\tilde{\varepsilon} \mapsto \left(n\pi + \frac{(1-a^2)}{6a^3}\tilde{\varepsilon}^3, \tilde{\varepsilon} \right).$$

This is vertical at $r_n = (n\pi, 0)$, so the tangent 'line' at r_n is the vertical line $\alpha = n\pi$. It coincides with the 2-jet of the above, but not with the 3-jet, as claimed. The argument for $r_0 = (\pi, 0)$ is similar and is omitted.

For $r_0 = (\pi/2, a)$ we proceed in a similar way. The 'line' γ is the graph of $p = a \sin \alpha$, whose 3-jet at r_0 is parametrized by

$$\varepsilon \mapsto \left(\frac{\pi}{2} + \varepsilon, a - \frac{a}{2}\varepsilon^2 \right).$$

The image of this 3-jet under T^n is the 3-jet at $T^n(r_0)$ parametrized by

$$\varepsilon \mapsto \left(\alpha_n + \varepsilon + \frac{an}{\sqrt{1-a^2}}\varepsilon^2, a - \frac{a}{2}\varepsilon^2 \right), \quad \alpha_n = \frac{\pi}{2} + 2n \arccos a.$$

Let

$$\tilde{\varepsilon} := \varepsilon + \frac{an}{\sqrt{1-a^2}}\varepsilon^2,$$

with inverse

$$\varepsilon = \tilde{\varepsilon} - \frac{an}{\sqrt{1-a^2}}\tilde{\varepsilon}^2 + \frac{2a^2n^2}{1-a^2}\tilde{\varepsilon}^3.$$

We get the parametrization of the 3-jet of $T^n(\gamma)$ at $T^n(r_0)$,

$$\tilde{\varepsilon} \mapsto \left(\alpha_n + \tilde{\varepsilon}, a - \frac{a}{2}\tilde{\varepsilon}^2 + \frac{a^2 n}{\sqrt{1-a^2}}\tilde{\varepsilon}^3 \right). \tag{15}$$

The 'line' tangent to $T^n(\gamma)$ at $T^n(r_0)$ is given by $p = a\cos(\alpha - \alpha_n)$, with 3-jet at $T^n(r_0)$ parametrized by

$$\varepsilon \mapsto \left(\alpha_n + \varepsilon, a - \frac{a}{2}\varepsilon^2 \right).$$

This coincides with the 2-jet of (15), but not the third, as claimed. The case $r_0 = (-\pi/2, -a)$ is similar and is omitted.

5 | MISCELLANEA

We present here briefly some results and conjectures, inspired by the previous sections.

5.1 | Liouville billiards

Recall that a Riemannian metric in a two-dimensional domain is called a *Liouville metric* if there exist coordinates (x, y) in which it is given by the formula

$$(f(x) + g(y))(dx^2 + dy^2),$$

where f and g are smooth functions of one variable, such that $f(x) + g(y) > 0$ for all x, y. The coordinate lines form a *Liouville net*, consisting of two families of mutually orthogonal curves.

The Euclidean metric in the plane admits a Liouville net consisting of confocal conics, corresponding to the respective elliptic coordinates. The degenerations of this net include the net of confocal parabolas and the net consisting of concentric circles and the radial lines (as well as the trivial net consisting of the horizontal and vertical lines).

The elliptic coordinates in 3-space, restricted to a triaxial ellipsoid which is a level surface of one of the coordinates, define a Liouville metric whose Liouville net consists of the lines of curvature; see Figure 12.

One considers a billiard system in a geodesically convex domain with a smooth closed boundary on a Riemannian surface: the trajectories are made of geodesic segments, and the law of reflection is the same as in the Euclidean case (the angle of incidence equals the angle of reflection). Similar to the case of billiards in an ellipse in the plane, the billiard system on a Liouville surface whose billiard table is bounded by a coordinate line from the Liouville net is integrable: a generic trajectory has all its segments tangent to a fixed curve of the Liouville net; see [9, 12, 14–16] for details.

The main ingredient in the proof of Theorem 1 was the complete integrability of the billiard ball map in ellipses and its consequences, such as a version of the Arnold–Liouville theorem (Proposition 1). For this reason, Theorem 1 and its proof extend, with appropriate adjustments, to Liouville billiards as well.

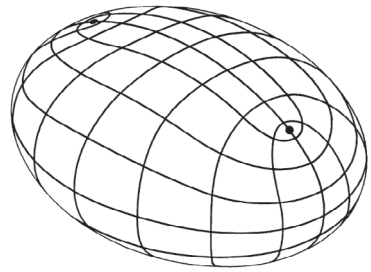

FIGURE 12 The lines of curvature on an ellipsoid form a Liouville net, associated with the elliptic coordinates in $\mathbb{R}^3$. The billiard system on the ellipsoid, whose table is bounded by one of these curves, is completely integrable.

We note that this set-up includes billiards bounded by conics in the hyperbolic and spherical geometries, the closest 'relatives' of the Euclidean billiard inside an ellipse. Concerning spherical and hyperbolic conics, see, for example, [11].

5.2 | Cusps on axes

As noted in Remark 2(a), when a light source O is placed on one of the axes of an ellipse, two of the cusps on the nth caustic by reflection from O will be located on this axis, but Theorem 1 does not give their location. Here we fill this gap, using the classical 'mirror equation' of geometric optics [20, Equation (5.9)].

Proposition 2. *Let $O = (x_0, 0)$, $|x_0| < a$, and let O_n (resp., O'_n) be the cusp of the nth caustic by reflection from O along the trajectory leaving O in the positive (resp., negative) direction of the x-axis. Then*

$$O_n = (-1)^n f^n(O), \quad O'_n = (-1)^{n+1} f^n(-O),$$

where f is a hyperbolic Möbius transformation of the x-axis with fixed points at the foci $\pm F = (\pm c, 0)$, $c = \sqrt{a^2 - b^2}$. Furthermore, F is an unstable fixed point of f and $-F$ is stable. Thus, as $n \to \infty$,

$$O_{2n} \to -F, \quad O_{2n+1} \to F, \quad O'_{2n} \to F, \quad O'_{2n+1} \to -F.$$

Explicitly,

$$f(x) = \frac{(a^2 + c^2)x - 2ac^2}{-2ax + a^2 + c^2}. \tag{16}$$

Exception: If C is a circle then f is parabolic, with a single fixed point at (0,0)*. Thus,* $\lim O_n = \lim O'_n = (0, 0)$*, as $n \to \infty$.*

Proof. Let $(R(x), 0)$ be the image of $(x, 0)$ after reflection off C at $(a, 0)$ and $(L(x), 0)$ the image after reflection at $(-a, 0)$. The x-coordinate of the successive images of $(x_0, 0)$, starting with a reflection at $(a, 0)$, are then

$$R(x_0), LR(x_0), RLR(x_0) \ldots$$

Note that $L(x) = -R(-x)$, hence the nth term in the above sequence is

$$x_n = (-1)^n(-R)^n(x_0).$$

It remains to find an explicit formula for $f(x) := -R(x)$.

The 'mirror equation' states that if an object is placed on the line normal to a convex mirror, where the curvature of the mirror is k, at a distance d from the mirror, then a reflected image of the object will form at a distance d' from the mirror, given by

$$\frac{1}{d} + \frac{1}{d'} = 2k. \tag{17}$$

The curvature of C at $(a, 0)$ is a/b^2, so setting $d = a - x, d' = a + f(x)$ in formula (17), we obtain

$$\frac{1}{a-x} + \frac{1}{a+f(x)} = \frac{2a}{b^2}. \tag{18}$$

Formula (16) for $f(x)$ follows. From formula (16) follows that

$$f'(c) = \frac{(a+c)^2}{(a-c)^2}, \quad f'(-c) = \frac{(a-c)^2}{(a+c)^2}.$$

Thus $f'(c) > 1$ and $0 < f'(-c) < 1$. It follows that c is an unstable fixed point of f and $-c$ is stable.

The formula for O'_n is obtained is a similar manner by considering the sequence $L(x_0), RL(x_0), LRL(x_0), \ldots$ □

Next, we study the case when O is on the minor axis.

Proposition 3. *Let $O = (0, y_0)$, $|y_0| < b$, and let O_n (resp. O'_n) be the cusp of the nth caustic by reflection from O along the trajectory leaving O in the positive (resp. negative) direction of the y-axis. Then*

$$O_n = (-1)^n g^n(O), \quad O'_n = (-1)^{n+1} g^n(-O),$$

where g is an elliptic Möbius transformation of the y-axis, conjugate to a rotation by 4θ, where $c + ib = ae^{i\theta}$ (that is, θ is the angle between the x-axis and line through $(0, b)$ and $-F = (-c, 0)$).

Explicitly,

$$g(y) = \frac{y\left(c^2 - b^2\right) - 2bc^2}{2by + c^2 - b^2}. \tag{19}$$

Proof. The proof of formula (19) is very similar to the above proof of formula (16) and is omitted. Using formula (19), one finds that $g(y)$ has no fixed points, hence it is elliptic, that is, conjugate to a rotation. The angle of rotation is given by the derivative at the complex fixed points. The complex fixed points of (19) are $\pm ic$, with

$$g'(ic) = \left(\frac{c-ib}{c+ib}\right)^2, \; g'(-ic) = \left(\frac{c+ib}{c-ib}\right)^2.$$

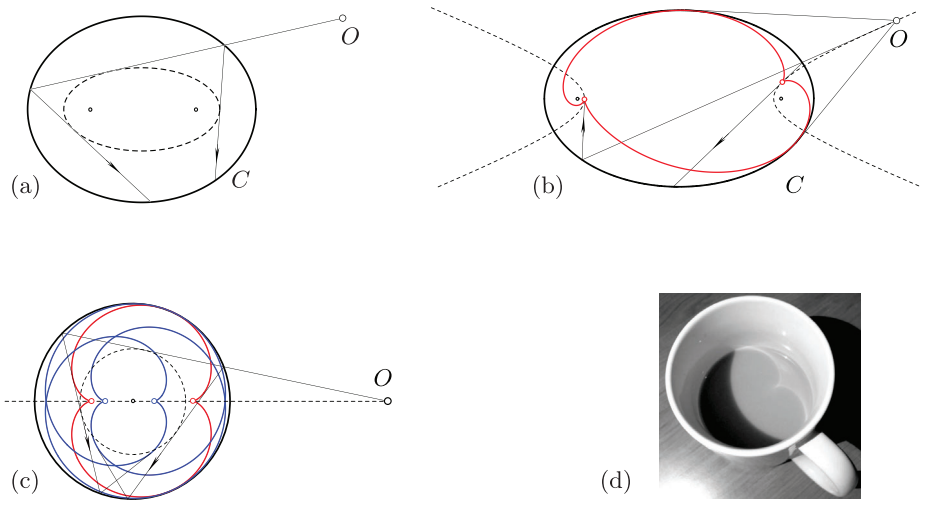

FIGURE 13 Caustics by reflection from an external light source O: (a) each line through O, incident to the interior of C, produces 2 billiards trajectories, tangent to the same conic confocal to C; (b) the 1st caustic by reflection off an ellipse, showing two cusps, lying on the confocal hyperbola through O. (c) The first two caustics by reflection off a circle, from an exterior light source O, showing two cusps for each caustic, lying on the line through O and the center of the circle. (d) A coffee cup 'half-caustic', showing a single cusp.

Let $c + ib = ae^{i\theta}$. Then $g'(ic) = e^{-4i\theta}$, $g'(-ic) = e^{4i\theta}$, from which follows the statement about the angle of rotation. □

5.3 | A light source outside an ellipse

Let us place a light source O outside an ellipse C. For each line through O intersecting the interior of C we consider the two billiard trajectories in the interior of C, whose initial rays are aligned with the line. One then finds analogues of the two conjectures and two theorems of this article, with '4' replaced by '2' throughout: the nth caustic by reflection of these rays is tangent to C at the contacts points with C of the two tangents to C through O, and has 2 cusps, located on the hyperbola confocal with C and passing through O; see Figure 13.

5.4 | The complexity of caustics by reflection

Figure 14 illustrates the observation that the complexity of the nth caustic by reflection in an ellipse increases with n. There are many ways to measure 'complexity'; for example, one may consider the number of times that the caustic goes to infinity (these points correspond to the vertical tangents of the curve $T^n(\gamma) \subset \mathcal{L}$). It would be interesting to make conjectures in this direction.

5.5 | Pseudo-integrable billiards

One may consider billiard tables bounded by arcs of confocal conics; such billiards were introduced in [7]. Since confocal conics intersect at right angles, these billiard tables have angles that

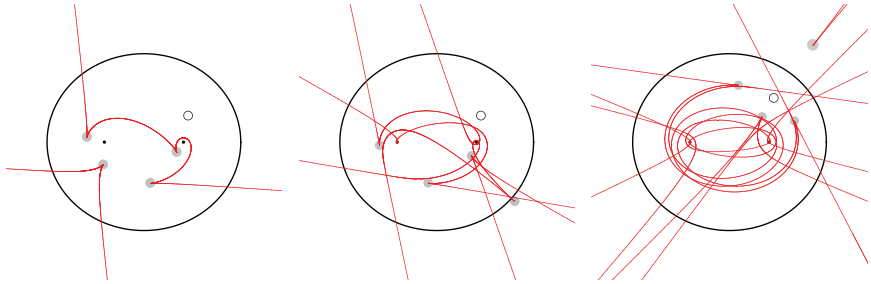

FIGURE 14 The second, fifth, and eighth caustics by reflection in an ellipse. The cusps are marked by gray circles.

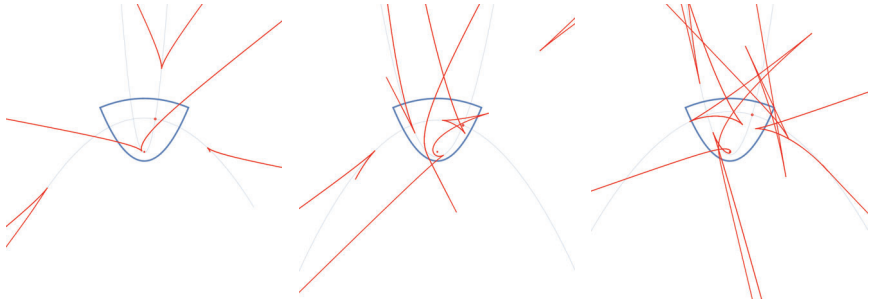

FIGURE 15 The first three caustics by reflection in a table bounded by two confocal parabolas.

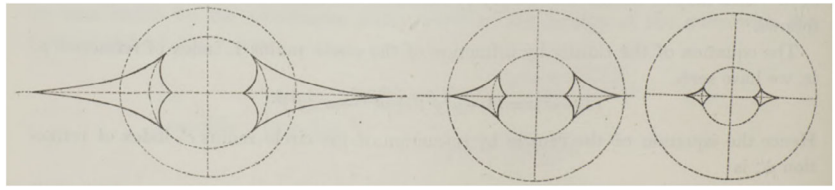

FIGURE 16 Caustics by refraction of a parallel beam in a circle. The figures show three distinct values of the index of refraction μ (from left to right): $1 < \mu < 2$, $\mu = 2$, $\mu > 2$. The cusps occur on the line through the center of the circle and parallel to the rays and on a concentric circle of radius $1/\mu$.

are multiples of $\pi/2$. Figure 15 shows caustics by reflection in a table bounded by two confocal parabolas. Although four cusps still lie on the confocal parabolas that pass through the source of light, there are additional cusps, and their number increases with n.

5.6 | Caustics by refraction

One could extend the experimental study and make conjectures about caustics by refraction in ellipses. Cayley considered the first such caustic in the case of a circle in [6]; see Figure 16 taken from p. 286 of Cayley's text.

ACKNOWLEDGMENTS

G. Bor acknowledges hospitality of the Toulouse Mathematics Institute during visits in 2023-4 and a CONAHCYT Grant A1-S-45886. ST participated in the special program 'Mathematical Billiards: at the Crossroads of Dynamics, Geometry, Analysis, and Mathematical Physics' at Simons Center for Geometry and Physics, where this work started; he is grateful to the Center for its hospitality and the inspiring atmosphere. ST was supported by NSF grants DMS-2005444 and DMS-2404535.

JOURNAL INFORMATION

The *Journal of the London Mathematical Society* is wholly owned and managed by the London Mathematical Society, a not-for-profit Charity registered with the UK Charity Commission. All surplus income from its publishing programme is used to support mathematicians and mathematics research in the form of research grants, conference grants, prizes, initiatives for early career researchers and the promotion of mathematics.

REFERENCES

1. G. Birkhoff, *Dynamical systems. With an addendum by Jurgen Moser*, American Mathematical Society Colloquium Publications, vol. IX, American Mathematical Society, Providence, RI, 1966.
2. W. Blaschke, *Vorlesungen über Differentialgeometrie und geometrische Grundlagen von Einsteins Relativitätstheorie*, vol. 1, Elementare Differentialgeometrie, Springer, Berlin, 1930.
3. G. Bor and S. Tabachnikov, *Cusps of caustics by reflection: a billiard variation on Jacobi's last geometric statement*, Amer. Math. Monthly **130** (2023), 454–467.
4. J. Boyle, *Using rolling circles to generate caustic envelopes resulting from reflected light*, Amer. Math. Monthly **122** (2015), 452–466.
5. J. Bruce, P. Giblin, and C. Gibson, *Caustics through the looking glass*, Math. Intelligencer **6** (1984), no. 1, 47–58.
6. A. Cayley, *A memoir upon caustics*, Philos. Trans. Roy. Soc. Lond. **147** (1857), 273–312.
7. V. Dragović and M. Radnović, *Pseudo-integrable billiards and arithmetic dynamics*, J. Mod. Dyn. **8** (2014), 109–132.
8. D. Fuchs and S. Tabachnikov, *Mathematical omnibus: thirty lectures on classic mathematics*, American Mathematical Society, Providence, RI, 2007.
9. A. Glutsyuk, I. Izmestiev, and S. Tabachnikov, *Four equivalent properties of integrable billiards*, Israel J. Math. **241** (2021), 693–719.
10. J. Itoh and K. Kiyohara, *The cut loci and the conjugate loci on ellipsoids*, Manuscripta Math. **114** (2004), 247–264.
11. I. Izmestiev, *Spherical and hyperbolic conics*, Eighteen essays in non-Euclidean geometry, IRMA Lectures in Mathematics and Theoretical Physics, vol. 29, European Mathematical Society, Zürich, 2019, pp. 263–320.
12. I. Izmestiev and S. Tabachnikov, *Ivory's theorem revisited*, J. Integrable Systems **2** (2017), 1–36.
13. C. G. J. Jacobi, *Vorlesungen über Dynamik*, Druck and Verlag von G. Reimer, Berlin, 1884.
14. G. Popov and P. Topalov, *Liouville billiard tables and an inverse spectral result*, Ergodic Theory Dynam. Systems **23** (2003), 225–248.
15. G. Popov and P. Topalov, *Discrete analog of the projective equivalence and integrable billiard tables*, Ergodic Theory Dynam. Systems **28** (2008), 1657–1684.
16. G. Popov and P. Topalov, *On the integral geometry of Liouville billiard tables*, Comm. Math. Phys. **303** (2011), 721–759.
17. R. Sinclair, *On the last geometric statement of Jacobi*, Experiment. Math. **12** (2003), 477–485.
18. H. Stachel, *The geometry of billiards in ellipses and their Poncelet grids*, J. Geom. **112** (2021), 29.
19. H. Stachel, *On the motion of billiards in ellipses*, Eur. J. Math. **8** (2022), 1602–1622.
20. S. Tabachnikov, *Geometry and billiards*, American Mathematical Society, Providence, RI, 2005.
21. A. Veselov, *Integrable mappings*, Russian Math. Surveys **46** (1991), no. 5, 1–51.
22. T. Waters, *The conjugate locus on convex surfaces*, Geom. Dedicata **200** (2019), 241–254.