Taylor & Francis
Taylor & Francis Group

The American Mathematical Monthly

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uamm20

On Cusps of Caustics by Reflection: Billiard Variations on the Four Vertex Theorem and on Jacobi's Last Geometric Statement

Gil Bor & Serge Tabachnikov

To cite this article: Gil Bor & Serge Tabachnikov (2023): On Cusps of Caustics by Reflection: Billiard Variations on the Four Vertex Theorem and on Jacobi's Last Geometric Statement, The American Mathematical Monthly, DOI: 10.1080/00029890.2023.2179842

To link to this article: https://doi.org/10.1080/00029890.2023.2179842

Published online: 17 Mar 2023.

On Cusps of Caustics by Reflection: Billiard Variations on the Four Vertex Theorem and on Jacobi's Last Geometric Statement

Gil Bor and Serge Tabachnikov

Abstract. A point source of light is placed inside an oval. The nth caustic by reflection is the envelope of the light rays emanating from the light source after n reflections off the curve. We show that, for a generic point light source, each of these caustics has at least 4 cusps. This is a billiard variation on Jacobi's Last Geometric Statement concerning the number of cusps of the conjugate locus of a point on a convex surface. We present various proofs, using different ideas, including the curve shortening flow and Legendrian knot theory.

1. INTRODUCTION.

Motivation and background. The *conjugate locus* of a point on a surface is the locus of first conjugate points along geodesics emanating from that point. In his "Lectures on Dynamics" [**11**], published posthumously, Jacobi stated that the conjugate locus of a generic point on an ellipsoid has exactly four cusps. This *Last Geometric Statement of Jacobi* was proved only in this century, see [**18**]. Indeed, as recently as the end of the 20th century, Marcel Berger wrote [**4**]:

> ... this latter assumption depends on the scandalously unproved Jacobi "statement": the conjugate locus of a nonumbilical point of an ellipsoid has exactly four cusps.

A related result is that the conjugate locus of a generic point on a convex surface has at least four cusps, see [**26**] for a recent proof. This theorem was attributed to C. Carathéodory (1912) by W. Blaschke (sect. 103 of [**5**]), who presented a sketch of the proof. This theorem belongs to a long list of results that stem from and are motivated by the celebrated 4-vertex theorem of S. Mukhopadhyaya. See [**2**, **15**] for surveys.

The conjugate locus can be equivalently described as the locus of the first intersections of infinitesimally close geodesics emanating from a point. These geodesics may intersect more than once, and the loci of their intersections are known as second, third, etc., caustics of the point. It is still an open question whether Jacobi's statement generalizes to these higher order caustics and to arbitrary convex surfaces. There is some experimental evidence that if the surface is an ellipsoid then, for a nonumbilic point, each such caustic has exactly four cusps, see [**21**] and Figure 1 from this paper (presented with permission). See also [**22**].

In this article we consider a billiard version of this problem. Let γ be an oval (a smooth strictly convex closed curve in $\mathbb{R}^2$), the boundary of a billiard table or, equivalently, an ideal mirror. Let O be a point inside γ, a source of light. For $n = 1, 2. \dots$, the 1-parameter family of rays that have undergone n optical reflections in γ envelopes a curve Γ_n, the *nth caustic by reflection*. See Figure 2.

These caustics may have singularities, generically, these singularities are semicubical cusps, locally given, in appropriate coordinates, by the equation $y^2 = x^3$. We

doi.org/10.1080/00029890.2023.2179842
MSC: 78A05; 53Axx

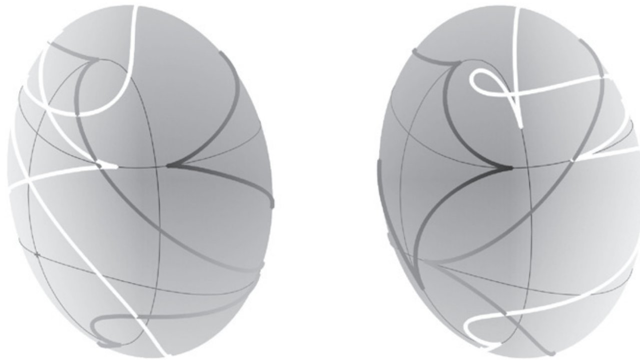

Figure 1. The first three caustics of a nonumbilic point on an ellipsoid.

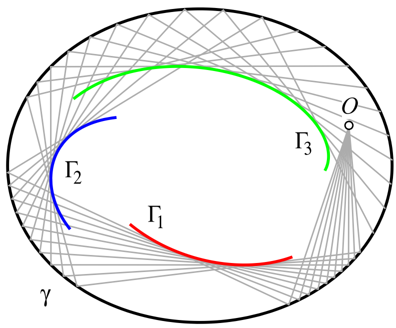

Figure 2. The nth caustic by reflection Γ_n is the envelope of the family of rays emanating from O that have undergone n reflections by γ.

always assume that the caustics Γ_n are in general position in this sense. The singularities of caustics were thoroughly studied by Bruce, Giblin, and Gibson; see [8] and the references therein.

Figure 3a shows that a caustic by reflection may extend beyond the interior of γ, and furthermore, it can be disconnected in the Euclidean plane; however, as the envelope of a 1-parameter family of lines, it is a connected curve in the projective plane $\mathbb{RP}^2$ (possibly, with singularities). Indeed, a 1-parameter family of lines is a curve in the space of lines, and the respective envelope is projectively dual to this curve.

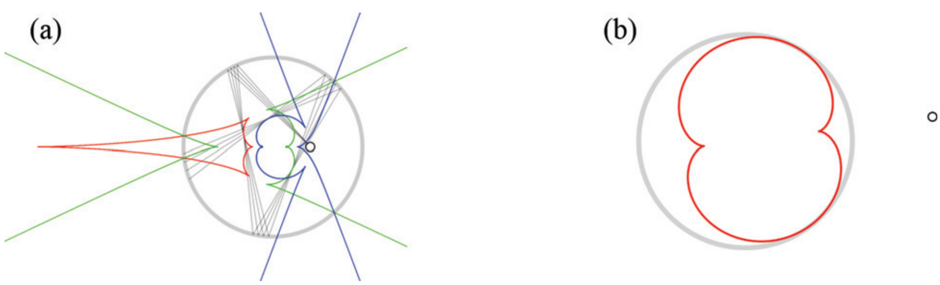

Figure 3. (a) The first three caustics by reflection in a circle, showing 4 cusps on each of them. The small circle is the light source. (b) The first caustic by reflection in a circle with an external source of light (each ray optically reflects at both intersection points with the circle). The caustic in this case has only 2 cusps. In this paper we consider only an internal light source.

The main result and two conjectures. Our main result is as follows.

Theorem 1. *For every oval $\gamma \subset \mathbb{R}^2$, a generic light source inside γ and $n \geq 1$, the nth caustic by reflection $\Gamma_n \subset \mathbb{RP}^2$ has at least four cusps.*

We present three proof sketches.

Let $\mathcal{L}$ be the space of directed lines in $\mathbb{R}^2$. To each caustic Γ_n is associated its dual curve $C_n \subset \mathcal{L}$, corresponding to the tangent lines along Γ_n (the rays of the nth reflected beam). One can identify $\mathcal{L}$ with the complement of the 'north' and 'south' pole of the unit sphere $S^2 \subset \mathbb{R}^3$, so that cusps of Γ_n correspond to inflection points of C_n (points with vanishing spherical geodesic curvature). Using standard properties of convex billiards, we show that C_n is a closed simple smooth curve in S^2, intersecting every great circle. A theorem of B. Segre from 1968 [**20**, **27**] states that such a curve has at least four spherical inflection points, thus completing the proof of Theorem 1.

Another approach starts with a realization of $\mathcal{L}$ as the vertical cylinder circumscribing S^2 and the curve $C_n \subset \mathcal{L}$ representing the tangent lines of Γ_n. Following S. Angenent [**1**], apply the curve shortening flow with respect to the flat metric on the cylinder to the curve C_n to deform it to the graph of a function $F : S^1 \to \mathbb{R}$ with zero mean value. Spherical inflection points of C_n correspond to the zeroes of $F'' + F$, a function with vanishing constant and first order Fourier terms. By the Sturm-Hurwitz theorem, it has at least four zeros.

Yet another approach is to use the relation between the cusps of the caustic Γ_n and the vertices (critical points of the curvature) of its normal front Δ_n, a closed planar curve whose normal lines, parametrized by C_n, are the lines tangent to Γ_n, see Figure 4. The relation between Γ_n and Δ_n is the familiar relation between evolutes and involutes, see, e.g., [**15**].

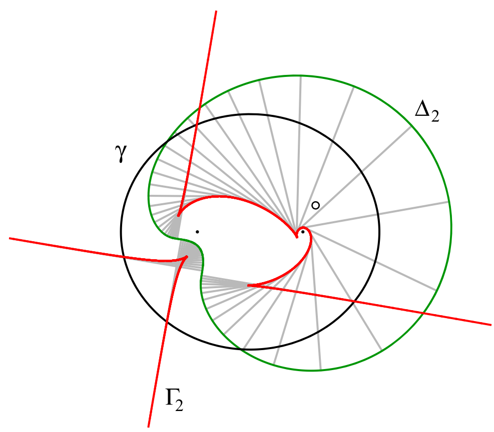

Figure 4. The 2nd caustics Γ_2 with an involute Δ_2 for an elliptical billiard table γ. The rays correspond to C_2, they are normal to the wave front Δ_2 and are tangent to Γ_2.

We show that Δ_n exists as a closed curve, possibly with cusps (in fact, there is a 1-parameter equidistant family of such curves). We show that Δ_n lifts to a Legendrian embedding of a circle in the space of cooriented contact elements in the plane. We next show that this Legendrian lift is 'unknoted', that is, it is Legendrian isotopic to the Legendrian lift of a circle. Finally, we use a theorem of Chekhanov and Pushkar [**9**] stating that the planar projection of a 'Legendrian unknot' has at least 4 vertices.

The rest of the article provides background information and details of these arguments. In the last section we mention some generalizations of Theorem 1 to spherical and hyperbolic geometry, as well as to "projective billiards."

We present two conjectures; the first one is supported by experimental evidence, the second one might be overly optimistic.

Conjecture 2. *If γ is an ellipse, then the caustic by reflection Γ_n for a light source inside γ and different from a focus has exactly four cusps for every $n \geq 1$ (see Figure 5).*

This conjecture is only known to hold in the case of $n = 1$ (the 'catacaustic', see next section).

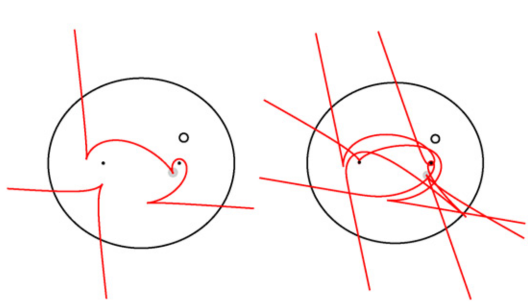

Figure 5. The 2nd, 5th and 8th caustics by reflection in an ellipse, each with 4 cusps (marked by gray disks).

Conjecture 3. *If γ is not an ellipse then, for some choice of light source inside γ and some $n \geq 1$, the caustic by reflection Γ_n has more than four cusps.*

Figure 6 shows caustics with more than 4 cusps for nonelliptical γ.

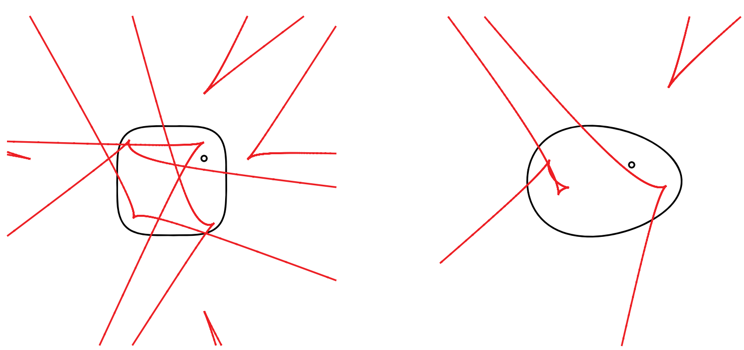

Figure 6. First caustic by reflection with more than four cusps of nonelliptical ovals. Left: $x^4 + y^4 = 1$, $O = (.6, .4)$. Right: $.5x^2 + (1 + .25x)y^2 = 1$, $O = (.5, .3)$.

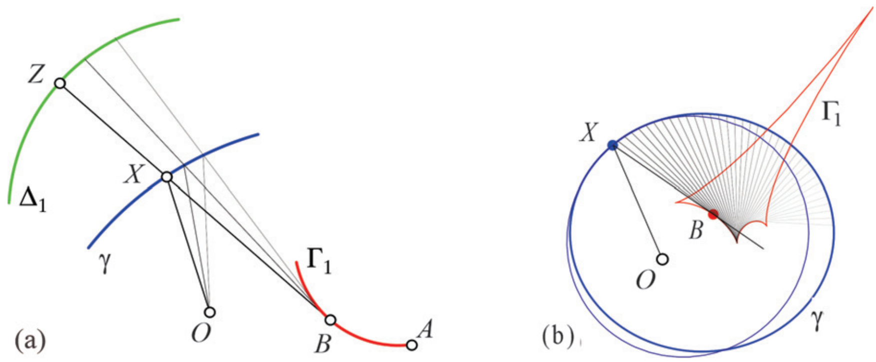

Figure 7. (a) The curve γ is the locus of points X such that $|OX| + |XB| + |\breve{BA}| = const$ (point A is fixed on Γ_1). The point Z is the reflection of O in the tangency line to γ at X. The locus of points Z is the orthotomic curve Δ_1, orthogonal to BZ and whose evolute is Γ_1. (b) The catacaustic Γ_1 is the locus of 2nd foci B of the osculating Kepler conic (thin oval) to the curve γ at X, with 1st foci at O.

Catacaustics. The first caustics by reflections, called *catacaustics*, are well studied. We give a brief summary of what is known about them, referring to [**7**, **8**, **17**] and the literature cited in these articles.

A version of the string construction that recovers a billiard curve from a billiard caustic (see, e.g., [**25**]) makes it possible to reconstruct the curve γ from its first caustic by reflection Γ_1. This construction involves a parameter, the length of the string. See Figure 7 (left).

The orthotomic curve Δ_1 is an involute of the catacaustic Γ_1, see Figure 7 (left), and Γ_1 is the evolute of Δ_1, that is, the envelope of its normals. The cusps of Γ_1 correspond to the vertices of Δ_1. It is known that when γ is an ellipse and O is not one of its foci then Δ_1 has 4 vertices [**8**].

A Kepler conic is a conic with one focus fixed at the origin O. The curve γ has a Kepler conic that has 3-point contact with it at every point, see [**6**]. The locus of the second foci of these osculating Kepler conics is the first caustic Γ_1—this follows from the optical properties of conics (a ray from one focus reflects to another focus). It follows that the cusps of the catacaustics Γ_1 correspond to the points where the Kepler conics hyperosculate the curve γ.

Computer graphics and animations. Most figures in this article were made using the computer program Mathematica. In the web page https://www.cimat.mx/~gil/caustics/ we provide some additional animations.

2. BACKGROUND MATERIAL.

The phase cylinder and the billiard ball map. This section contains some standard material on mathematical billiards, see, e.g., [**25**].

Denote by $\mathcal{L}$ the space of oriented lines in $\mathbb{R}^2$. We use the 'cylinder model' of $\mathcal{L}$, with coordinates (α, p) defined as follows: $\alpha \in S^1 = \mathbb{R}/2\pi\mathbb{Z}$ is the direction of the line and $p \in \mathbb{R}$ is the signed distance from the oriented line to the origin O (which we choose to be the center of the initial beam of light). The sign of p is defined by the right-hand rule, see Figure 8. Thus $\mathcal{L}$ is an infinite cylinder.

The space $\mathcal{L}$ of oriented lines in $\mathbb{R}^2$ admits an area form, unique up to scale, invariant under the Euclidean group action. In coordinates, this area form is $\omega = d\alpha \wedge dp$.

The *phase cylinder* of the billiard system inside an oval $\gamma \subset \mathbb{R}^2$ is the set $M \subset \mathcal{L}$ of oriented lines intersecting γ. It is a bounded cylinder whose two boundary components

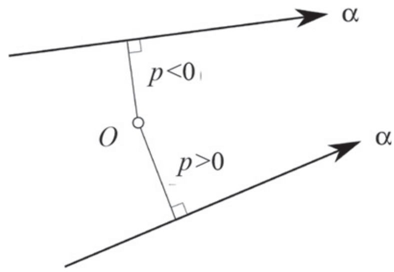

Figure 8. The coordinates (α, p) of an oriented line in $\mathbb{R}^2$.

correspond to the lines tangent to γ, one component for each orientation. The "equator" $p = 0$ corresponds to the lines through O, see Figure 9.

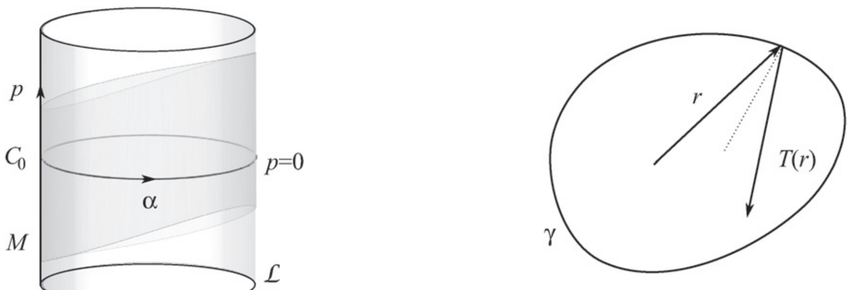

Figure 9. Left: the phase cylinder $M \subset \mathcal{L}$. Right: the billiard ball map $T : M \to M$.

The billiard ball map $T : M \to M$, sending an incoming ray to the reflected one, is an area preserving transformation, that is, $T^*\omega = \omega$. Since $\omega = -d(p d\alpha)$, the differential 1-form $T^*(p d\alpha) - p d\alpha$ is closed. In fact, more is true: as we will now show, it is *exact*, that is, $T^*(p d\alpha) - p d\alpha = dF$ for some function $F : M \to \mathbb{R}$. An example of an area preserving, but nonexact, map is $(\alpha, p) \mapsto (\alpha, p + 1)$.

Proposition 4. *The billiard ball map $T : M \to M$ is exact.*

In order to prove Proposition 4, consider another description of the phase cylinder as the set of unit vectors with a foot point on γ, pointing inwards (the initial position and velocity of the billiard ball). These unit vectors are in one-to-one correspondence with the oriented lines that they generate. Let $\gamma(t)$ be an arc length counterclockwise parameterization and φ be the angle between the tangent $\gamma'(t)$ and the unit vector. See Figure 10a.

Consider the differential 1-form $\cos\varphi\, dt$. Let $L = |\gamma(t_1) - \gamma(t)|$ be the distance between the intersection points of a line with γ. See Figure 10b.

Lemma 5. $T^*(\cos\varphi\ dt) - \cos\varphi\ dt = dL$.

Proof. One has: $T(t, \varphi) = (t_1, \varphi_1)$ and

$$\frac{\partial L(t, t_1)}{\partial t} = -\cos\varphi, \quad \frac{\partial L(t, t_1)}{\partial t_1} = \cos\varphi_1,$$

that is, $dL = \cos\varphi_1\ dt_1 - \cos\varphi\ dt$, as needed. ■

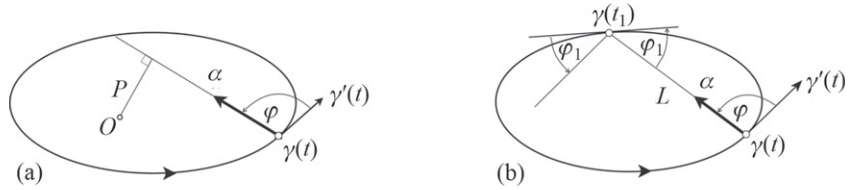

Figure 10. (a) The coordinates (t, φ) on M and their relation to (α, p). (b) The generating function L of the billiard ball map T.

Two differential 1-forms are cohomologous if their difference is the differential of a function.

Lemma 6. *The 1-form $p\,d\alpha$ is cohomologous to $\cos\varphi\,dt$.*

Proof. Using complex notation, see Figure 10a, one has

$$e^{i(\alpha-\varphi)} = \gamma'(t),\ \ p = \det(\gamma(t), e^{i\alpha}).$$

Differentiating these equations, we get

$$d\alpha \equiv d\varphi \ (\mathrm{mod}\ dt),\ \ dp \equiv \sin\varphi\,dt \ (\mathrm{mod}\ d\alpha),$$

so

$$d\alpha \wedge dp = \sin\varphi\,d\varphi \wedge dt = -d(\cos\varphi\,dt),$$

hence $pd\alpha - \cos\varphi\,dt$ is closed.

To show that $p\,d\alpha - \cos\varphi\,dt$ is exact it suffices to show that its integral along a noncontractible closed curve in M vanishes. As such a curve take C, the boundary component of the phase cylinder M given by $\varphi = 0$. Clearly, $\int_C \cos\varphi\,dt$ equals the perimeter of γ.

On the other hand, by the Cauchy-Crofton formula, this perimeter equals π times the average length of the orthogonal projection of γ on a line, that is, $\int_C pd\alpha$. This implies the result. ■

Proof of Proposition 4. Lemma 5 shows that T preserves the integral $\int_C \cos\varphi\,dt$, and Lemma 6 shows that $\int_C \cos\varphi\,dt = \int_C p\,d\alpha$, hence T preserves the integral $\int_C p\,d\alpha$, i.e., it is exact. ■

The curve C_0 representing the initial beam of light is the equator $p = 0$. Define the *signed area* enclosed by an oriented closed curve C in $\mathcal{L}$ as the line integral $\int_C pd\alpha$: this is the algebraic area between C and C_0. Tautologically, C_0 encloses zero area. Since T is an exact map, the same holds for the curves $C_n = T^n(C_0)$. See Figure 11.

Contact elements, Legendrian knots, and wave fronts. A *contact element* in the plane is a pair consisting of a point and a line through it. A coorientation of a contact element is a choice of one of the sides of the line. More conceptually, the space of cooriented contact elements is the spherization of the cotangent bundle $ST^*\mathbb{R}^2$: assign to a covector η its kernel, a tangent line, and define coorientation by choosing the side where η is positive.

The space of contact elements carries a contact structure, a 2-dimensional distribution (a field of tangent planes), defined by the "skating condition": the foot point may

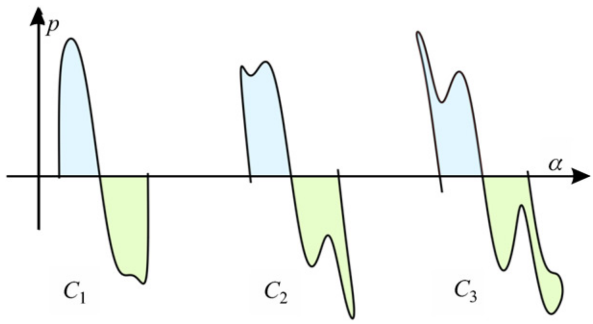

Figure 11. The successive iterates $C_n = T^n(C_0)$, for $n = 1, 2, 3$, drawn on the phase cylinder $\mathcal{L}$ (spread flat). Each is a simple closed smooth curve of 0 signed area. The billiard table is the ellipse $4x^2/5 + y^2 = 1$ and the source is (.6, .2).

move along the line, and the line may rotate about the foot point. Let (x, y) be the standard coordinates in $\mathbb{R}^2$ and θ the angle between the positive x-axis and the direction of the line; then the contact distribution is the kernel of the 1-form $\sin\theta\, dx - \cos\theta\, dy$.

A smooth curve in $ST^*\mathbb{R}^2$ that is tangent to the contact distribution is called *Legendrian*. Its projection to the plane is a *wave front*, a curve that may have singularities, generically semicubical cusps, but that has a tangent line at every point. Conversely, such a curve has a unique lift to the space of contact elements as a Legendrian curve.

We introduce coordinates (α, p, z) in $ST^*\mathbb{R}^2$, where $(\alpha, p) \in T^*S^1$ are the coordinates of the orthogonal line at the foot point A, and z is the (signed) distance of the line to the origin O. See Figure 12.

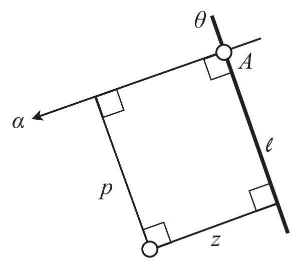

Figure 12. Coordinates (α, p, z) on the space $ST^*\mathbb{R}^2$ of cooriented contact elements in $\mathbb{R}^2$. With a cooriented line ℓ through A we associate the (signed) distance z to O and the coordinates (α, p) of the perpendicular oriented line at A, pointing to the positive side of ℓ.

This defines an identification of $ST^*\mathbb{R}^2$ with J^1S^1, the space of 1-jets of functions $f : S^1 \to \mathbb{R}$, where $z = f(\alpha)$ and $p = f'(\alpha)$. On J^1S^1 there is a standard contact form $dz - pd\alpha$ (the 1-jets of functions $z = f(\alpha)$ are Legendrian curves).

Lemma 7. *The identification $ST^*\mathbb{R}^2 = J^1S^1$ is a contactomorphism.*

Proof. Using the notation of Figure 12,

$$\alpha = \theta + \pi/2, \ \ p = \det(A, e^{i\alpha}), \ \ z = \det(A, e^{i\theta}),$$

so

$$dz - pd\alpha = \det(dA, e^{i\theta}) = \sin\theta dx - \cos\theta dy,$$

as claimed. ∎

We have two projections

$$\mathbb{R}^2 \xleftarrow{\ \pi_1\ } ST^*\mathbb{R}^2 \xrightarrow{\ \pi_2\ } \mathcal{L}$$

where π_1 maps a cooriented contact element (A, ℓ) to A and π_2 maps it to the line through A, orthogonal to ℓ, oriented towards its positive side.

In coordinates, $\pi_2 : (\alpha, p, z) \mapsto (p, \alpha)$ ("forgetting z"). The fibers of π_1 are Legendrian, spanned by the vector field $\partial_\alpha - z\partial_p + p\partial_z$, while the fibers of π_2, spanned by the vector field ∂_z, are transverse to the contact distribution. Hence π_2 projects a smooth Legendrian curve in $ST^*\mathbb{R}^2$ to a smooth curve in $\mathcal{L}$.

Conversely, let C be a closed curve in $\mathcal{L}$. We want to lift it via $\pi_2 : ST^*\mathbb{R}^2 \to \mathcal{L}$ to a Legendrian curve $\widetilde{C} \subset ST^*\mathbb{R}^2$. Since the contact distribution on $ST^*\mathbb{R}^2$ is transverse to the fibers of π_2, once the initial point of the lifting is chosen, the lifting is uniquely determined, but it may fail to close up.

Lemma 8. *The lifted curve $\widetilde{C}$ is closed if and only if $\int_C p\,d\alpha = 0$: the curve C encloses zero signed area.*

Proof. The curve $\widetilde{C}$ is closed if and only if the value of the third coordinate z is the same at the endpoints. Since $dz = p\,d\alpha$ along a Legendrian curve, the values of z at the endpoints are equal if and only if $\int_C p\,d\alpha = 0$. ■

The curve $C \subset \mathcal{L}$ defines a 1-parameter family of oriented lines. The projection of the lifted Legendrian curve $\widetilde{C} \subset ST^*\mathbb{R}^2$ to $\mathbb{R}^2$ is a wave front Δ that is orthogonal to this family of lines and is cooriented by their directions. If a closed front Δ exists, i.e., C encloses zero signed area, then there exists a whole 1-parameter family of fronts that are equidistant from each other. This nonuniqueness corresponds to the choice of the initial point of the lifted curve $\widetilde{C}$.

The situation is the same as in the familiar relation between evolutes and involutes: for an involute of a closed curve to close up it is necessary and sufficient for the curve to have zero signed length (the sign changes after each cusp), and the equidistant family of curves share their normals, and hence their evolutes.

Vertices of wave fronts and Legendrian isotopies. A *vertex* of a plane curve is an extremum of its curvature or, equivalently, a cusp of the evolute, the envelope of its normals. The notion of vertex extends to cooriented wave fronts: the curvature at cusps is infinite, changing from $-\infty$ to ∞ (so cusps are not vertices).

The classical 4-vertex theorem asserts that a simple closed convex curve has at least four vertices. Let Δ be a cooriented wave front whose Legendrian lift to $ST^*\mathbb{R}^2$ is embedded, i.e., is a *Legendrian knot*. V. Arnold conjectured [2, 3] that if this Legendrian knot is homotopic as a Legendrian knot to the Legendrian lift of a circle, then Δ has at least four vertices. This conjecture was proved by Chekanov and Pushkar [9] using Legendrian knot theory.

A generic regular homotopy of a cooriented wave front is a composition of a number of moves, similar to the Reidemeister moves in knot theory, see Figure 13, borrowed from [9]. The first five moves are isotopies of the respective Legendrian knot, but the "dangerous" self-tangency with coinciding coorientations correspond to self-intersection of the Legendrian lifted curve and changes the Legendrian knot type.

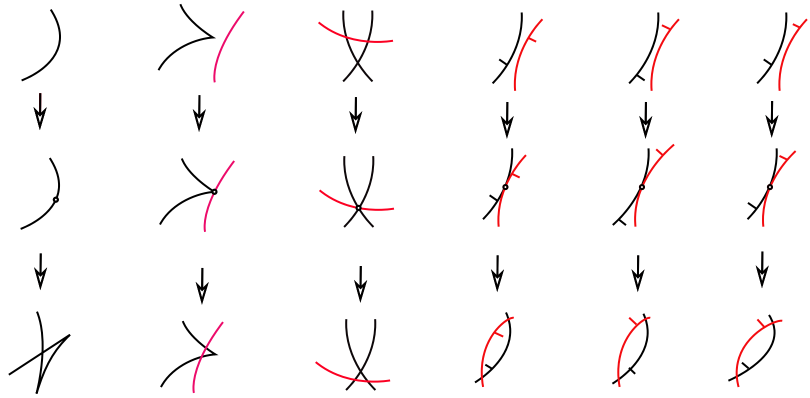

Figure 13. Generic "perestroikas" of cooriented wave fronts.

For example, the curve on the left of Figure 14 has only two vertices, but the curve on the right is Legendrian isotopic to a circle, therefore it has at least four vertices no matter how one draws it. Thus these curves are not Legendrian isotopic. On the other hand, the Whitney winding number of both curves is one, hence they are regularly isotopic.

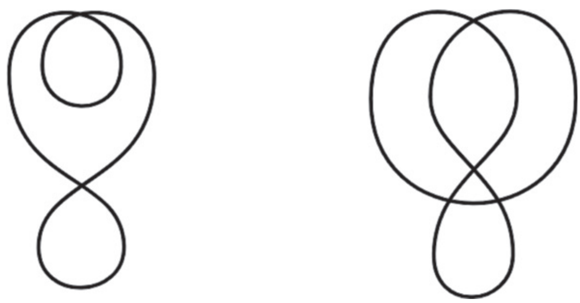

Figure 14. Left: only two vertices; right: at least four vertices.

Summary. With each "beam" of light rays (a 1-parameter family of oriented lines in $\mathbb{R}^2$) we have associated four curves,

$$C \subset \mathcal{L}, \qquad \widetilde{C} \subset ST^*\mathbb{R}^2, \qquad \Delta \subset \mathbb{R}^2, \qquad \Gamma \subset \mathbb{RP}^2.$$

(we recall that the zero enclosed area by the curve C is the necessary and sufficient condition for $\widetilde{C}$ to close up). The correspondences between the cusps, vertices, and inflection points on these curves are depicted in Figure 15.

3. PROOFS OF THEOREM 1.

First proof. Cusps of Γ_n correspond, by projective duality, to inflection points of C_n. These inflection points are 3-point contacts of C_n with the curves in $\mathcal{L}$ corresponding to the 1-parameter families of lines passing through a fixed point. If the point is $(a, b) \in \mathbb{R}^2$, then the respective curve in $\mathcal{L}$ is the graph of the first harmonic

$$p = a \sin \alpha - b \cos \alpha,$$

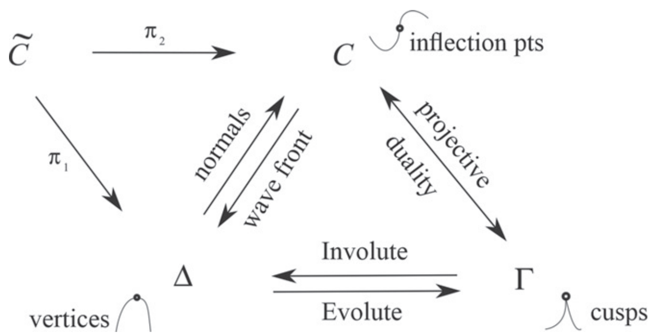

Figure 15.

that is, it is an ellipse obtained as the intersection of the cylinder $\mathcal{L}$ with a plane through the origin. Note that this graph encloses zero signed area.

Consider the central projection of this cylinder to the unit sphere. This projection sends the graphs of the first harmonics to great circles.

Since C_n encloses zero signed area, it intersects every graph of the first harmonics. It follows that $\bar{C}_n$, the image of the curve C_n, is a smooth spherical curve that is not contained in any hemisphere. In particular, the convex hull of $\bar{C}_n$ contains the origin.

The (geodesic) inflections of $\bar{C}_n$ in the standard metric of the sphere are its 3-point contacts with great circles. By the Segre theorem mentioned earlier, $\bar{C}_n$ has at least four inflections. Therefore so does C_n. ∎

Second proof. Following [1], one can use the curve shortening flow to prove that the curve C_n has at least four inflections. Recall that under the curve shortening flow, each point of the curve moves in the normal direction with the speed equal to the curvature; see [14, 16] and the book [10].

Equip $\mathcal{L}$ with the flat Riemannian metric $d\alpha^2 + dp^2$ and apply the curve shortening flow to C_n. Let $C_n(s)$ be arclength parametrization, then the flow is given by the partial differential equation $C_t = C_{ss}$.

A variation of the standard proof shows that the evolution is defined for all $t \geq 0$, deforming C_n through embedded curves, shrinking it to a horizontal curve $p = const$, which is a closed geodesic. A version of the maximum principle implies that the number of inflections does not increase during this evolution, see [1]. This is illustrated in Figure 16.

Next, a version of Lemma 3.1.7 in [14] or Lemma 1.10 of [16] shows that the evolving curves enclose zero signed areas.

Lemma 9. *The curve shortening flow $C_t = C_{ss}$ preserves the signed area $\int_C p d\alpha$.*

Proof. Let $(\alpha(s), p(s))$ be an arc length parameterization of a noncontractible closed curve in $\mathcal{L}$, so that $\alpha_s^2 + p_s^2 = 1$. Then its time evolution under the curve-shortening flow is given by $(\alpha_t, p_t) = kN = k(-p_s, \alpha_s)$, where the subscript denotes the derivative, k is the curvature, and N is the unit normal. It follows that

$$\frac{d \int p d\alpha}{dt} = \int [k\alpha_s^2 - p(kp_s)_s] ds = \int [k\alpha_s^2 + kp_s^2] ds = \int k(s) ds,$$

where the second equality is due to integration by parts.

It remains to note that the total curvature of a closed curve that goes around the cylinder equals zero. ∎

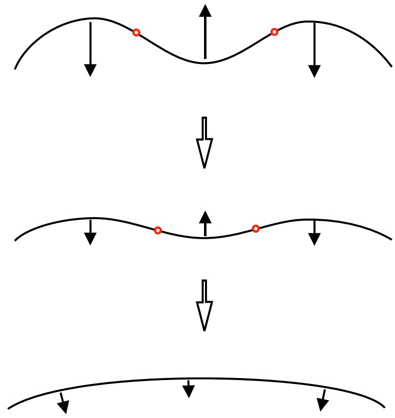

Figure 16. In the curve shortening flow, two nearby inflection points may cancel each other, but they cannot appear on a convex arc.

Recall that C_0 is the equator $p = 0$. As C_n approaches C_0 (with derivatives), it is given by the graph of a function $p = F(\alpha)$. The inflection points of this graph are the points where it is tangent to 2nd order to the graphs of functions of the form $h(\alpha) = a\cos(\alpha) + b\sin(\alpha)$.

For each $\alpha \in S^1$, one can find unique a, b such that $F(\alpha) = h(\alpha)$, $F'(\alpha) = h'(\alpha)$. Since $h'' + h = 0$, the equation $F''(\alpha) = h''(\alpha)$ holds if and only if $F''(\alpha) + F(\alpha) = 0$. So the inflection points of F are the zeros of the function $G := F'' + F$.

To conclude that G has no less than four zeros, apply the Sturm-Hurwitz theorem that states that the number of zeros of a 2π-periodic function is not less than the number of zeros of its first nontrivial harmonic, see, e.g., [**2**, **3**, **13**].

Since the differential operator $d^2 + 1$ preserves the order of Fourier terms and kills the 1st order terms, G has no first harmonics. Since the curve encloses zero signed area, F has zero constant term, and so does G, as needed. ■

Remark. The Sturm-Hurwitz theorem has many proofs, see Section 8.1 of [**19**]. Interestingly, one of them, due to G. Polya, makes use of the heat equation, a close relative of the curve shortening flow.

Third proof. This argument relies on the correspondence between the cusps of Γ_n and the vertices of Δ_n, its normal front.

Since C_n encloses zero signed area, Lemma 8 implies that it admits a Legendrian lift $\widetilde{C}_n \subset ST^*\mathbb{R}^2$, and its projection to $\mathbb{R}^2$ is a closed curve, possibly with cusps, which is normal to the rays of C_n.

A homotopy of the curve C_n to C_0 in the class of smooth closed embedded curves that enclose zero signed area induces a Legendrian isotopy between the Legendrian knots $\widetilde{C}_n$ and $\widetilde{C}_0$. (Such a homotopy is provided by the curve shortening flow but, unlike the second proof, one can use any other homotopy for this purpose).

Now our "black box", the Pushkar-Chekanov theorem [**9**], implies that Δ_n has at least four vertices. ■

4. VARIATIONS. One can extend Theorem 1 to geodesically convex billiards in spherical and hyperbolic geometries (the former lie in one hemisphere); the first proof works with small modifications.

Theorem 1, along with its proofs, also extends to some other initial beams of light. For example, one may consider a 1-dimensional source, an oval that lies inside γ and that emanates rays of light in the outward normal directions.

Projective billiards. For any convex curve $\gamma \subset \mathbb{R}^2$ with a transverse vector field v along it one can define the projective billiard map $T : M \to M$ [12,23,24]. The reflection law is as follows. Consider an incoming ray at a point $x \in \gamma$ in the direction u, decompose $u = u_1 + u_2$, where u_1 is tangent to γ at x and u_2 is a multiple of $v(x)$. Then the outgoing ray passes through x in the direction $u_1 - u_2$. Equivalently, the tangent line, the transverse line, the incoming, and the outgoing ones, form a harmonic quadruple of lines.

If the transverse field consists of the normals, one has the usual law "the angle of incidence equals the angle of reflection".

If γ is an origin-centered ellipse and the transverse field v is given by the gradient of a homogeneous function of two variables, then the projective billiard ball map is again exact area preserving. The area form on the phase space M is the same as the one on the space of oriented geodesics in the hyperbolic plane, considered in the projective, or Cayley-Klein, model of hyperbolic geometry in the interior of the ellipse γ. The total area of M is infinite in this case.

And, as before, Theorem 1 holds: see Figure 17 for an illustration.

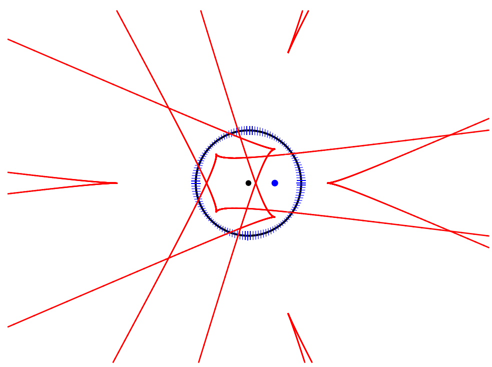

Figure 17. The 1st caustic by reflection, showing 8 cusps, in a projective billiard system with a circular table and the exact transverse field $v = \nabla(x^4 + y^4)$.

ACKNOWLEDGMENTS. GB acknowledges support from the Shapiro visiting program in Penn State and CONACYT Grant A1-S-4588. ST was supported by NSF grant DMS-2005444.

ORCID

Serge Tabachnikov http://orcid.org/0000-0001-5398-7820

REFERENCES

[1] Angenent, S. (1999). Inflection points, extatic points and curve shortening. In: Simó, C., ed. *Hamiltonian Systems with Three or More Degrees of Freedom*. Dordrecht: Kluwer Academic Publishers, pp. 3–10.

[2] Arnold, V. (1996). Topological problems in the theory of wave propagation. *Russ. Math. Surv.* 51(1): 1–47.

[3] Arnold, V. (1994). *Topological Invariants of Plane Curves and Caustics*. University Lecture Series, Vol. 5. Providence, RI: American Mathematical Society.

[4] Berger, M. (2002). *Riemannian Geometry During the Second Half of the Twentieth Century.* University Lecture Series, Vol. 17. Providence, RI: American Mathematical Society.
[5] Blaschke, W. (1930). *Vorlesungen über Differentialgeometrie und geometrische Grundlagen von Einsteins Relativitätstheorie*, Vol. 1, Elementare Differentialgeometrie. Berlin: Springer.
[6] Bor, G., Jackmann, C., Tabachnikov, S. (2021). Variations on the Tait-Kneser theorem. *Math. Intell*. 43(3): 8–14.
[7] Boyle, J. (2015). Using rolling circles to generate caustic envelopes resulting from reflected light. *Am. Math. Monthly*. 122(5): 452–466.
[8] Bruce, J., Giblin, P., Gibson, C. (1984). Caustics through the looking glass. *Math. Intell.* 6(1): 47–58.
[9] Chekanov, Yu., Pushkar, P. (2005). Combinatorics of fronts of Legendrian links, and Arnold's 4-conjectures. *Russian Math. Surv.* 60: 95–149.
[10] Chou, K.-S., Zhu, X.-P. (2001). *The Curve Shortening Problem.* Boca Raton, FL: Chapman & Hall/CRC.
[11] Clebsch, A., ed. (2009). *Jacobi's Lectures on Dynamics.* Delivered at the University of Königsberg in the winter semester 1842–1843, according to the notes prepared by C. W. Brockardt. Springer.
[12] Fierobe, C. (2020). Examples of reflective projective billiards and outer ghost billiards. Preprint: arXiv:2002.09845.
[13] Fuchs, D., Tabachnikov, S. (2007). *Mathematical Omnibus. Thirty Lectures on Classic Mathematics.* Providence, RI: American Mathematical Society.
[14] Gage, M., Hamilton, R. (1986). The heat equation shrinking convex plane curves. *J. Differ. Geom*. 23: 69–96.
[15] Ghys, E., Tabachnikov, S., Timorin, V. (2013). Osculating curves: around the Tait-Kneser theorem. *Math. Intell*. 35(1): 61–66.
[16] Grayson, M. (1987). The heat equation shrinks embedded plane curves to round points. *J. Differ. Geom.* 26: 285–314.
[17] Hungerbühler, N. (2020). The inverse caustic problem. *Am. Math. Monthly.* 127: 387–400.
[18] Itoh, J., Kiyohara, K. (2004). The cut loci and the conjugate loci on ellipsoids. *Manuscr. Math.* 114: 247–264.
[19] Ovsienko, V., Tabachnikov, S. (2005). *Projective Differential Geometry Old and New. From the Schwarzian Derivative to the Cohomology of Diffeomorphism Groups.* Cambridge: Cambridge University Press.
[20] Segre, B. (1968). Alcune proprietè differenziali in grande delle curve chiuse sghembe. *Rend. Mat.* 1: 237–297.
[21] Sinclair, R. (2003). On the last geometric statement of Jacobi. *Exp. Math.* 12: 477–485.
[22] Sinclair, R., Tanaka, M. (2006). Jacobi's last geometric statement extends to a wider class of Liouville surfaces. *Math. Comp.* 75: 1779–1808.
[23] Tabachnikov, S. (1997). Introducing projective billiards. *Ergod. Theory Dyn. Syst*. 17: 957–976.
[24] Tabachnikov, S. (1997). Exact transverse line fields and projective billiards in a ball. *Geom. Funct. Anal.* 7: 594–608.
[25] Tabachnikov, S. (2005). *Geometry and Billiards*. Providence, RI: American Mathematical Society.
[26] Waters, T. (2019). The conjugate locus on convex surfaces. *Geom. Dedicata*. 200: 241–254.
[27] Weiner, J. (1977). Global properties of spherical curves. *J. Diff. Geom.* 12: 425–434.

GIL BOR received his BA and MA degrees in math and physics from the Hebrew University in Jerusalem in 1985, and his PhD in math at UC Berkeley in 1991. After a 3 year postdoc position in Tucson Arizona, he moved to Mexico in 1994 to join the Math Research Center (CIMAT) in Guanajuato. His main research interest is in differential geometry. For the last 25 years, he has been developing in Mexico science workshops for motivated high school students.
Centro de Investigación en Matemáticas (CIMAT), Guanajuato, México
gil@cimat.mx

SERGE TABACHNIKOV received his PhD from Moscow State University in 1987, and since 1990 he has taught in American universities, currently at Penn State. His research interests are in differential geometry, differential topology, and dynamical systems. He served as editor-in-chief of *Experimental Mathematics*, and associate editor of this Monthly (for 20 years), and he is currently a co-editor-in-chief of the *Mathematical Intelligencer*. Serge is substantially involved in popularization of mathematics and in outreach activities.
Department of Mathematics, Penn State, University Park, PA 16802, USA
tabachni@math.psu.edu