

Dual billiards in the hyperbolic plane*

Serge Tabachnikov

Department of Mathematics, Penn State University, University Park, PA 16802, USA

E-mail: tabachni@math.psu.edu

Received 15 June 2001, in final form 31 January 2002

Published 13 May 2002

Online at stacks.iop.org/Non/15/1051

Recommended by L Bunimovich

Abstract

We study an area preserving map of the exterior of a smooth convex curve in the hyperbolic plane, defined by a natural geometrical construction and called the dual billiard map. We consider two problems: stability and the area spectrum. The dual billiard map is called stable if all its orbits are bounded. We show that both stable and unstable behaviours may occur. If the map at infinity has a hyperbolic periodic orbit, then the dual billiard map has orbits escaping to infinity. On the other extreme, if the map at infinity is smoothly conjugated to a Diophantine irrational rotation of the circle, then the dual billiard map is stable.

The area spectrum is the set of extremal areas of n -gons, circumscribed about the dual billiard curve; this is to the dual billiard what the length spectrum is to the usual, inner, one. We show that the area spectrum has an asymptotic expansion in even negative powers of n as $n \rightarrow \infty$. The first coefficient of this expansion is the area of the dual billiard curve, and the next is, up to a constant, the cubed integral of the cube root of its curvature. We describe the curves that are relative extrema of these two functionals and show that they are the trajectories of the pseudospherical pendulum with various gravity directions.

Mathematics Subject Classification: 37E40, 37J50, 70H08

1. Introduction and formulation of results

The dual billiard map is a transformation of the exterior of a closed strictly convex plane curve defined by the following geometric construction. Call the curve γ , and let x be a point in its exterior. There are two tangent lines to γ through x ; choose one of them (say, the right one from x 's view-point) and define $F(x)$ to be the reflection of x in the point of tangency. The map F is called the dual billiard map and the curve γ the dual billiard curve.

* Partially supported by an NSF grant.

The dual billiard map is an outer counterpart of the usual billiard ball map (some authors use the term ‘outer billiard’). Moser [23, 24] put forward the study of dual billiards. The map F is area preserving, and this makes it possible to apply KAM theory in its study. We refer to [1, 3, 11, 12, 15, 27, 29–32, 34] for various aspects of the dual billiard problem, such as existence and non-existence of invariant curves, polygonal dual billiards, multi-dimensional dual billiards, etc.

This paper concerns dual billiards in the hyperbolic plane. The map is defined the same way as in the Euclidean plane and it also enjoys the property to preserve the area associated with the hyperbolic metric. As far as the author knows, dual billiards in the hyperbolic plane have not been studied before. One usually expects hyperbolic geometry to be richer than Euclidean one. In particular, one may expect dual billiards in the hyperbolic plane to have new interesting features, compared to the ones in the affine plane; we will see that it is indeed the case, as far as the stability properties are concerned.

We use the Klein–Beltrami (or projective) model of hyperbolic geometry. The hyperbolic plane is represented by the interior of the unit circle (‘circle at infinity’), straight lines by the chords of this circle, and the distance between points x and y is given by the formula

$$d_H(x, y) = \frac{1}{2} \ln[a, x, y, b], \quad (1)$$

where a and b are the intersection points of the line xy with the circle and

$$[a, x, y, b] = \frac{(y-a)(b-x)}{(x-a)(b-y)}$$

is the cross-ratio.

We consider two aspects of the dual billiard problem in the hyperbolic plane: stability and the area spectrum.

Sections 2 and 3 concern the stability problem. Following Moser, we call the dual billiard map *stable* if every orbit stays bounded. In the Euclidean plane, Moser proved that if the dual billiard curve γ is sufficiently smooth, then the dual billiard map F has an invariant curve outside of every compact domain, and therefore F is stable (on the other hand, if γ is not smooth enough—say, γ is a convex polygon—the stability remains an open problem).

In the hyperbolic plane, the situation is quite different: the stable and unstable behaviours of the dual billiard map F may coexist, depending on the dual billiard curve γ . We assume throughout the paper that γ is infinitely smooth. The dual billiard map F extends to a diffeomorphism of the circle S^1 at infinity and, moreover, to a diffeomorphism of an open Euclidean neighbourhood of S^1 . Denote this smooth circle map by f and the smooth map of the closed disc by \bar{F} , and let $r(f)$ be the rotation number of f . It turns out that, in some cases, the stability of F depends on the properties of the circle map f .

One has the following four possibilities:

- (a) $r(f)$ is rational and f has a hyperbolic attracting periodic point;
- (b) $r(f)$ is rational and all periodic points of f are not hyperbolic attracting;
- (c) $r(f)$ is a Diophantine irrational number;
- (d) $r(f)$ is a Liouville irrational number.

For $r(f) \in \mathcal{Q}$, case (a) is generic and case (b) exceptional. In this paper, we consider only cases (a) and (c); our first result concerns the former.

Theorem 1. *If f has a hyperbolic attracting periodic point, then this point is also hyperbolic attracting as a periodic point of the map \bar{F} , and the dual billiard map F is not stable.*

In C^∞ topology, the diffeomorphism f depends continuously on the dual billiard curve γ . The above case (a) determines an open subset in the space of circle diffeomorphisms. Therefore

instability, described in theorem 1, persists under sufficiently small C^∞ perturbations of the dual billiard curve.

Remark 1.1. In their study of polygonal dual billiards in the hyperbolic plane, Dogru and Tabachnikov [5] discovered the following: if an n -gonal dual billiard table is *large* then *all* orbits of the dual billiard map escape to infinity, namely, they are attracted to an attracting n -periodic orbit of the map f . The term ‘large’ means that $r(f) = 1/n$ and the respective attracting periodic orbit is hyperbolic; although F is discontinuous in the polygonal case, f is still continuous, and the rotation number is defined. A justification for this terminology comes from two facts: that the rotation number depends monotonically on the dual billiard table: if $\gamma_1 \subset \gamma_2$, then $r(f_1) \geq r(f_2)$; and that $1/n$ is the minimal rotation number for an n -gonal dual billiard table—if γ is an n -gon, then $r(f) \geq 1/n$.

As far as the stability problem, an opposite extreme is described in the next theorem.

Theorem 2. *If f is smoothly conjugated to a Diophantine irrational rotation of the circle then F has an invariant curve in every Euclidean neighbourhood of the circle at infinity, and the dual billiard map F is stable.*

Remark 1.2. By a celebrated Herman’s theorem [6], f is smoothly conjugated to a rotation once its rotation number satisfies a certain Diophantine condition that holds for almost all irrational numbers (in the sense of the Lebesgue measure).

Note that the stability problem near the dual billiard curve does not differ in the hyperbolic plane from the Euclidean one: if γ is smooth enough then its every neighbourhood contains an invariant curve of the dual billiard map (see, for example, [1] or [3] for the Euclidean case; the same arguments apply in the hyperbolic plane).

Remark 1.3. An interesting example of the dual billiard map is provided by an elliptic dual billiard curve γ . In this case, the map F is integrable and the invariant curves are ellipses that belong to a pencil of conics generated by γ and the circle at infinity. The rotation number $r(f)$ may be rational or irrational, and cases (b)–(d) may occur. In all cases, the circle map f is smoothly conjugated to a rotation, and this implies the classical Poncelet Porism of projective geometry (see [33] for this approach to the Poncelet theorem).

It is interesting to compare theorems 1 and 2 with the situation in the affine plane. In the latter case, one may also add a ‘circle at infinity’ to the plane; a point of this circle represents a family of parallel orientated lines. The dual billiard map extends to the circle at infinity as a very degenerate map, namely, the central symmetry, and this is a significant difference with the case of the hyperbolic plane. In the affine plane, one obtains more information about the behaviour at infinity by considering F^2 , the second iteration of the dual billiard map. It turns out that F^2 is approximated by a Hamiltonian flow whose Hamiltonian function is determined by the dual billiard curve in a very explicit way (see, for example, [31] for this point of view).

Remark 1.4. Every fixed point free circle diffeomorphism f gives rise to a generalized dual billiard system in the hyperbolic plane. The respective curve γ_f is the envelope of the one-parameter family of lines connecting the points $x \in S^1$ to their images $f(x)$. In general, γ_f is not a smooth curve but rather a wave front, a singular curve with a well-defined tangent line at every point and free from inflections (i.e. the projectively dual curve is smooth). Even for such singular curves, the respective dual billiard map F is well defined in a vicinity of the circle at infinity. The curve γ_f is a smooth convex curve for a set of circle diffeomorphisms open in C^∞ topology. It would be interesting to study the relation between circle diffeomorphisms f and their two-dimensional extensions F in more detail.

Section 4 contains some formulae from hyperbolic geometry, needed for the rest of the paper. Section 5 concerns periodic orbits of the dual billiard map in a vicinity of the dual billiard curve. Connecting consecutive points of such an n -periodic orbit, one obtains an n -gon, circumscribed about γ . Periodic orbits correspond to the area extrema of circumscribed polygons (see [29, 30, 32] for the affine plane; the same argument applies in the hyperbolic setting). The set of extremal areas of circumscribed polygons is called the *area spectrum* of the dual billiard curve γ . We are interested in the asymptotic behaviour of the area spectrum as $n \rightarrow \infty$. In this paper, we restrict attention to simple polygons, i.e. to n -periodic orbits with rotation number $1/n$.

A similar problem for conventional (inner) billiards in the Euclidean plane was studied in [19]: the object of study is the length spectrum, the set of length of periodic billiard trajectories. The technique of interpolating Hamiltonians from [19, 21] (see also [26]) was used in [29, 30] to study the area spectrum of dual billiards in the affine plane, where the following result was proved. Let A_n denote the area of a simple polygon corresponding to an n -periodic orbit of the dual billiard map. Then the asymptotic expansion holds:

$$A_n \sim a_0 + \frac{a_1}{n^2} + \frac{a_2}{n^4} + \dots + \frac{a_i}{n^{2i}} + \dots, \quad (2)$$

where a_0 is the area bounded by γ , and

$$a_1 = \frac{1}{24} \left(\int k^{1/3} ds \right)^3; \quad (3)$$

here k is the curvature and s is the arc length parameter. The integral in (3) is called the affine length of the curve γ ; this quantity is invariant under equiaffine transformations of the plane.

Theorem 3. *If the area, curvature and length are understood in terms of hyperbolic geometry, then the asymptotic expansion (2) and formula (3) hold for dual billiards in the hyperbolic plane.*

An analogue of theorem 3 holds for dual billiards on the unit sphere as well. We would like to emphasize that the existence of the asymptotic expansion (2) follows directly from the work of Melrose [19, 21]; the main work is in identifying the term a_1 as in (3).

Remark 1.5. Let us describe another point of view on the area spectrum. Given a convex curve γ , one wants to approximate it by circumscribed polygons using area as the distance between the curve and an approximating polygon; this problem makes sense in the Euclidean, hyperbolic and spherical geometries. Formula (2) provides an asymptotic expansion for the distance between γ and its best approximating n -gon; in the Euclidean case, the term a_1 was found by Fejes Toth [35, 36] (see [20] for a complete proof, [18] for the value of the term a_2 and [7, 8] for surveys on approximating convex bodies by polytopes). The approach via interpolating Hamiltonians provides a novel view point in the approximation theory of smooth convex curves by polygons.

An important inverse spectral problem in the theory of mathematical billiards is whether the length spectrum determines the billiard table up to isometries (see [10] for a partial positive result). (This problem is closely related to another inverse problem concerning the spectrum of the Laplace operator, known in the famous formulation by Mark Kac: ‘Can one hear the shape of a drum?’ [13].) An analogous problem for dual billiards in the hyperbolic plane is whether the area spectrum determines the curve γ up to isometries. In the affine plane, one can show that the ellipses are characterized by their area spectrum ([29]). This follows from the affine isoperimetric inequality of Blaschke (see, e.g. [28]):

$$(\text{affine length})^3 \leq 8\pi^2 \text{area}, \quad (4)$$

with the equality for the ellipses only. Thus the first two coefficients a_0 and a_1 in (2) characterize the ellipses in the affine plane.

We conjecture that a similar characterization holds for circles in the hyperbolic plane and on the sphere. More precisely, let k denote the geodesic curvature of a curve.

Conjecture 4. *For every closed convex curve bounding area A on the simply connected surface of constant curvature $\sigma = \pm 1$ (unit sphere or pseudosphere), one has*

$$\left(\int k^{1/3} ds \right)^3 \leq A(2\pi - \sigma A)(4\pi - \sigma A), \quad (5)$$

with the equality for circles only.

Being unable to prove this conjecture, we study a related variational problem: to describe immersed curves on the sphere or pseudosphere which are relative extrema of the functionals

$$\int k^{1/3} ds \quad \text{and} \quad \int k ds. \quad (6)$$

A relation with conjecture 4 is provided by the Gauss–Bonnet theorem: the area is equal, up to constants, to the integral curvature. In the affine plane, the solutions to an analogous relative extremum variational problem are curves of constant affine curvature, i.e. conics. We expected a similar result in the hyperbolic plane and on the sphere: that the solutions would be curves of constant geodesic curvature. The actual answer, obtained in section 6, is rather surprising.

Theorem 5. *The extremal curves of the variational problem (6) coincide with the trajectories of the (pseudo)spherical pendulum with various gravity directions.*

The trajectories of the (pseudo)spherical pendulum are described in terms of elliptic functions (see, e.g. [16]). A relevant result, a Puiseux theorem, states that if a trajectory of the (pseudo)spherical pendulum is a simple closed curve, then this curve lies in a plane (see [2]). This provides a certain confirmation to conjecture 4.

2. Instability

In this section we establish some useful facts about the dual billiard map and prove theorem 1. We use the following convention: the hyperbolic distance between points inside the unit disc is denoted by $d_H(x, y)$ and the Euclidean one by $|xy|$.

Let us start with the following question: given a strictly convex invariant curve Γ of the dual billiard map, can one reconstruct the dual billiard curve γ ? For the usual (inner) billiard, the answer is given by a well-known string construction, and for dual billiards in the affine plane, the relevant construction is called the area construction (see, e.g. [32]). The latter extends to the hyperbolic plane as follows.

Lemma 2.1. *The dual billiard curve γ is the envelope of the one-parameter family of lines that cut off a constant hyperbolic area from the curve Γ .*

Proof. Let AB be a segment from the one-parameter family of lines that cut off a constant hyperbolic area from Γ . Let O be the tangency point of AB with the envelope γ . We want to show that $d_H(A, O) = d_H(O, B)$ (see figure 1). Assume that $d_H(A, O) > d_H(O, B)$. Consider another segment $A'B'$ from the same family and let O' be its intersection point with AB . If $A'B'$ is sufficiently close to AB , then $d_H(A, O') > d_H(O', B)$ and $d_H(A', O') > d_H(O', B')$. Therefore, the reflection in O' takes the triangle $BO'B'$ strictly inside the triangle $AO'A'$.

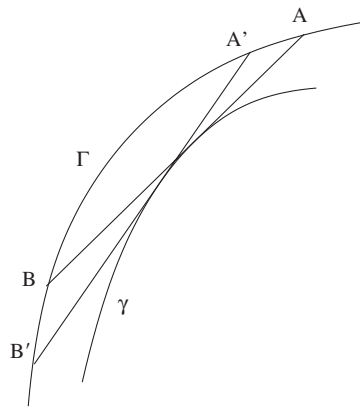


Figure 1. Area construction.

On the other hand, the lines AB and $A'B'$ cut off equal areas from Γ , and thus these two triangles have equal areas. This is a contradiction. \square

Note that the area construction depends on the value of the cut-off area and gives a one-parameter family of dual billiard curves with the same invariant curve.

Now we establish the area-preserving property of the dual billiard map.

Lemma 2.2. *The dual billiard map F preserves the hyperbolic area form.*

Proof. Consider figure 2. It follows from lemma 2.1 that the areas of the two quadrilaterals are equal. In the limit, as Γ' approaches Γ and the angle between the two segments goes to zero, one obtains the area preserving property. \square

Next we address the construction described in remark 1.4. Let α be the angle coordinate on the unit circle S^1 . Consider a diffeomorphism of S^1 given by the formula

$$\alpha \mapsto \alpha_1 = \alpha + f(\alpha),$$

where f is a smooth periodic function with $f' > -1$. Consider the one-parameter family of lines $l(\alpha)$ connecting point $(\cos \alpha, \sin \alpha)$ to $(\cos \alpha_1, \sin \alpha_1)$ and let γ_f be its envelope.

Lemma 2.3. *The curve γ_f is smooth if and only if*

$$\cot\left(\frac{f(\alpha)}{2}\right) \neq \frac{f''(\alpha)}{(1 + f'(\alpha))(2 + f'(\alpha))} \tag{7}$$

for all $\alpha \in [0, 2\pi]$.

Proof. Let $(u(\alpha), v(\alpha))$ be the Cartesian coordinates of the tangency point of the line $l(\alpha)$ with the envelope γ_f . To find these coordinates, let $\phi(\alpha)$ be an affine function in the plane whose zero level curve is $l(\alpha)$. In our case,

$$\phi(\alpha) = (\sin \alpha_1 - \sin \alpha) x - (\cos \alpha_1 - \cos \alpha) y + \sin(\alpha - \alpha_1).$$

Then $u(\alpha)$ and $v(\alpha)$ satisfy the system of linear equations $\phi(\alpha) = \phi'(\alpha) = 0$ where prime denotes the derivative with respect to α . A direct computation yields

$$(u(\alpha), v(\alpha)) = \left(\frac{1 + f'(\alpha)}{2 + f'(\alpha)}\right)(\cos \alpha, \sin \alpha) + \left(\frac{1}{2 + f'(\alpha)}\right)(\cos \alpha_1, \sin \alpha_1).$$

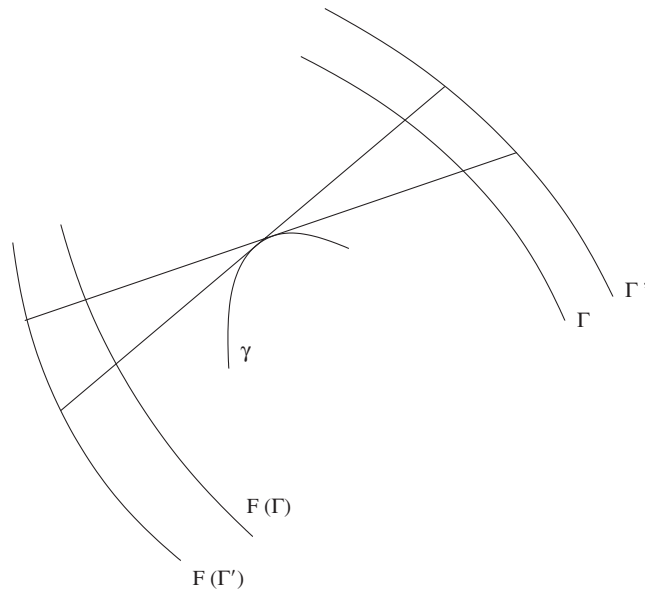


Figure 2. Area preserving property of the dual billiard map.

Differentiating, one obtains

$$(u(\alpha), v(\alpha))' = \left(\frac{1 + f'(\alpha)}{2 + f'(\alpha)} \right) (-\sin \alpha - \sin \alpha_1, \cos \alpha + \cos \alpha_1) - \left(\frac{f''(\alpha)}{2 + f'(\alpha)^2} \right) (\cos \alpha_1 - \cos \alpha, \sin \alpha_1 - \sin \alpha).$$

The right-hand side of this expression equals

$$C \left(-\sin \left(\alpha + \frac{f(\alpha)}{2} \right), \cos \left(\alpha + \frac{f(\alpha)}{2} \right) \right)$$

with

$$\frac{C}{2} = \cos \left(\frac{f(\alpha)}{2} \right) \left(\frac{1 + f'(\alpha)}{2 + f'(\alpha)} \right) - \sin \left(\frac{f(\alpha)}{2} \right) \left(\frac{f''(\alpha)}{2 + f'(\alpha)^2} \right).$$

Thus $(u(\alpha), v(\alpha))' = 0$ if and only if $C = 0$ which is equivalent to (7). □

It follows from lemma 2.3 that the envelope γ_f is smooth for a set of diffeomorphisms open in C^∞ topology, as claimed in remark 1.4.

Our next result is useful for studying the dual billiard map near S^1 , the circle at infinity.

Lemma 2.4. *The dual billiard map F extends to a smooth map of a sufficiently small Euclidean neighbourhood of S^1 .*

Proof. Consider that the dual billiard curve $\gamma(t)$ is some smooth parametrization, and choose a value of the parameter t_0 . Let $C = \gamma(t_0)$, denote by l the tangent line to γ at C , and let A and B be the intersection points of l with S^1 . Set: $|AC| = a$, $|BC| = b$. The restriction of the dual billiard map to l is a projective involution of this line which interchanges A and B and fixes C . Choose the Euclidean coordinate x on l such that C is the origin, A has coordinate $x = -a$ and B the coordinate $x = b$. Then $F|_l$ is given by a fractional linear transformation:

$$x \mapsto \frac{-x}{1 + ux}, \quad u = \frac{1}{a} - \frac{1}{b}. \tag{8}$$

This transformation is originally defined on the open segment $(-a, b)$ but formula (8) smoothly extends it to a neighbourhood of this segment, as long as one avoids zeros of the denominator, namely, as long as $x \neq ab/(a - b)$. Letting t vary in a neighbourhood of the parameter value t_0 , one smoothly extends the dual billiard map F to a Euclidean neighbourhood of the point A . By compactness of γ , one obtains an extension of F to a neighbourhood of the unit circle. \square

We are ready to prove theorem 1. Denote by \bar{F} the extension of F to a neighbourhood of the unit disc, constructed in lemma 2.4. At every point of the unit circle S^1 choose a Euclidean orthonormal frame: the first vector is tangent to the circle and the second points toward the centre. Let $x \in S^1$ and $y = f(x)$. We compute the differential of the dual billiard map \bar{F} at point x , more precisely, its matrix with respect to the introduced frames at points x and y .

Let o be the tangency point of the segment xy with the dual billiard curve. Denote the angle made by the segment xy with the circle by α , and let $\lambda = |yo|/|xo|$.

Lemma 2.5. *The differential of \bar{F} at point x is as follows:*

$$d\bar{F} = \begin{pmatrix} \lambda & -\lambda(\lambda + 1) \cot \alpha \\ 0 & \lambda^2 \end{pmatrix}. \quad (9)$$

Proof. The tangent direction to the circle is invariant under $d\bar{F}$, and we first compute the respective eigenvalue, which also equals $df(x)$. Let x' and $y' = f(x')$ be points on S^1 such that $|xx'| = \varepsilon$ (see figure 3). By plane Euclidean geometry, the triangles xox' and yoy' are similar, therefore $|yy'|/|xx'| = \lambda$. Taking the limit $\varepsilon \rightarrow 0$ yields $df(x) = \lambda$ (this is an argument from theorem XXX, figure 102, in Newton's 'Principia' [25]; Newton studies the gravitational attraction of spherical bodies).

Let x'' and $y'' = F(x'')$ be points on the segment xy such that $|xx''| = \varepsilon$. Let us compute the ratio $|yy''|/|xx''|$. One has $d_H(y'', o) = d_H(x'', o)$, therefore $[y, y'', o, x] = [y, o, x'', x]$. A straightforward computation modulo ε^2 yields: $\lim_{\varepsilon \rightarrow 0} |yy''|/|xx''| = \lambda^2$.

In the frames under consideration, the unit vector at x in the direction to y is $u = (\cos \alpha, \sin \alpha)$ and the unit vector at y in the direction to x is $v = (-\cos \alpha, \sin \alpha)$. We proved that $d\bar{F}(u) = \lambda^2 v$. From this one finds the matrix of $d\bar{F}$ as in (9). \square

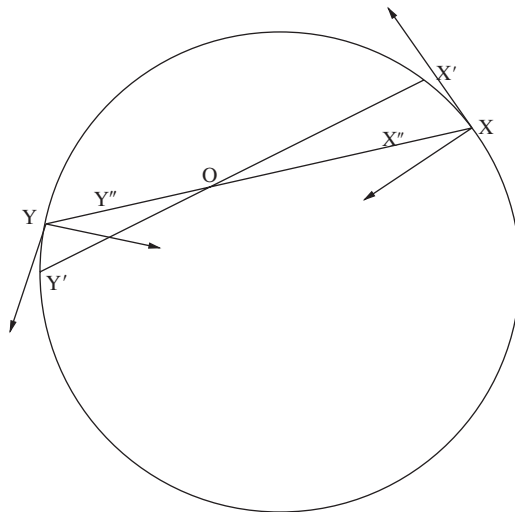


Figure 3. Calculating the differential of the extended dual billiard map.

Consider a periodic orbit x_1, \dots, x_n of the circle map f . Let o_i be the tangency point of the segment $x_i x_{i+1}$ with the dual billiard curve and $\lambda_i = |x_{i+1} o_i|/|x_i o_i|$. Set: $\Lambda_n = \lambda_1 \lambda_2 \cdots \lambda_n$.

Lemma 2.6. *The differential of the n th iteration f^{on} at point x_1 equals Λ_n , and the differential of \bar{F}^{on} has the eigenvalues Λ_n and Λ_n^2 .*

Proof. The product of the matrices (9) is again an upper-triangular matrix with the diagonal entries Λ_n and Λ_n^2 . □

Lemma 2.6 implies theorem 1: if x is a hyperbolic attracting n -periodic point of f then $\Lambda_n < 1$, and the operator $d\bar{F}^{on}$ is contracting at x .

3. Stability

This section contains a proof of theorem 2. We deduce this theorem from the Moser invariant curve theorem [22].

Introduce the following notation. Let $\alpha(t), t \in [0, 1]$, be a smooth parametrization of the circle at infinity such that the map f is conjugated to the shift $t \rightarrow t + r$ where r is the Diophantine irrational rotation number of f . Abusing notation, $\alpha(t)$ will also denote the respective point of the unit circle (whose Cartesian coordinates are $(\cos \alpha(t), \sin \alpha(t))$). Let P be a point inside the unit disc and $Q = F(P)$ its image under the dual billiard map (see figure 4). Let x be the Euclidean distance from P to the circle S^1 along the line QP . We use $(t - r, x)$ as coordinates of point P . Likewise, there is a unique value of parameter t_1 such that the point Q lies on the line l connecting the points $\alpha(t_1)$ and $\alpha(t_1 + r)$. Set $t_1 = t - \varepsilon$, and let y be the distance from Q to S^1 along the line l . Then the dual billiard map acts as follows: $F(t - r, x) = (t - \varepsilon, y)$ where ε and y are smooth functions of t and x .

The plan of the proof of theorem 2 is as follows. We will find new coordinates (ξ, η) (where ξ is a cyclic coordinate) in a Euclidean neighbourhood of the circle S^1 such that the dual billiard map is given by

$$F(\xi, \eta) = (\xi + r - \eta + O(\eta^2), \eta + O(\eta^3)). \tag{10}$$

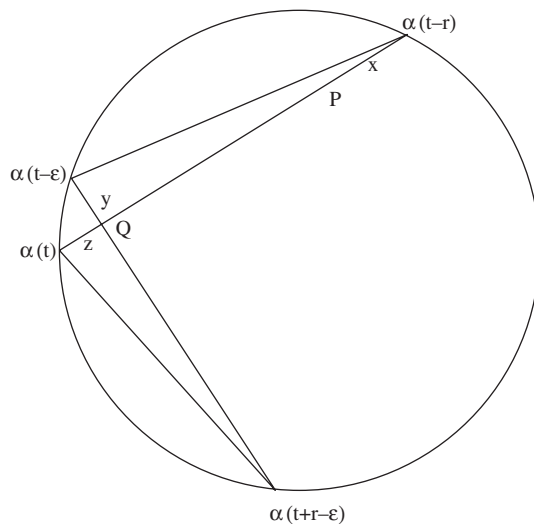


Figure 4. Coordinate description of the dual billiard map.

Since the dual billiard map is area preserving, every simple closed curve, homotopic to S^1 , intersects its image under F . Then, by Moser's small twist theorem [22] (see also [17]), for every positive δ , the map F has an invariant curve inside the annulus $0 < \eta < \delta$, and theorem 2 follows.

Let us return to figure 4. First we compute z as a function of x and t . Denote the derivative with respect to the parameter t by dot, and shorthand $\alpha(t)$ and $\alpha(t-r)$ to α and α_- , respectively. Set $s = \sin((\alpha - \alpha_-)/2)$.

Lemma 3.1. *One has*

$$\left[1 + \left(\left[\frac{\dot{\alpha}}{\dot{\alpha}_-}\right]^2 - 1\right) \frac{x}{2s}\right] \left[1 + \left(\left[\frac{\dot{\alpha}_-}{\dot{\alpha}}\right]^2 - 1\right) \frac{z}{2s}\right] = 1. \quad (11)$$

Proof. The segment of the line PQ inside the disc has length $2s$. The tangency point O with the dual billiard table divides this segment in the ratio $(\dot{\alpha})^2$ to $(\dot{\alpha}_-)^2$ (see lemma 2.5 and its proof). The map F is a projective involution ψ of this segment that interchanges the end points and fixes O . This uniquely determines ψ given by a fractional linear function. A direct computation of this function, along with the fact that $\psi(x) = z$, yields (11). \square

Next we relate the variables y and z .

Lemma 3.2. *One has*

$$\frac{y}{z} = \frac{\sin((\alpha(t-\varepsilon) - \alpha(t-r))/2)}{\sin((\alpha(t+r-\varepsilon) - \alpha(t))/2)}. \quad (12)$$

Proof. It follows from similar triangles in figure 4 that the ratio y/z equals that of the segments $[\alpha(t-r), \alpha(t-\varepsilon)]$ and $[\alpha(t), \alpha(t+r-\varepsilon)]$. This implies (12). \square

As the first step toward formula (10), introduce new coordinates:

$$v = \frac{x}{\dot{\alpha}(t-r)^2} \sin\left(\frac{\alpha(t) - \alpha(t-r)}{2}\right); \quad w = \frac{z}{\dot{\alpha}(t)^2} \sin\left(\frac{\alpha(t) - \alpha(t-r)}{2}\right);$$

$$u = \frac{y}{\dot{\alpha}(t-\varepsilon)^2} \sin\left(\frac{\alpha(t+r-\varepsilon) - \alpha(t-\varepsilon)}{2}\right).$$

The dual billiard map acts as follows: $F(t-r, v) = (t-\varepsilon, u)$, and using (11) and (12), one finds that

$$\varepsilon = f(t)v + O(v^2), \quad u = v + g(t)v^2 + O(v^3), \quad (13)$$

where $f(t) > 0$; the smooth functions $f(t)$ and $g(t)$ depend on the function $\alpha(t)$ and its derivatives. We wish to change coordinates once again to eliminate the term $g(t)$ and to make the term $f(t)$ equal to 1.

Lemma 3.3. *Consider a C^∞ smooth cylinder map $F(t, v) = (T, V)$ with*

$$T = t + r - f(t)v + O(v^2), \quad V = v + g(t)v^2 + O(v^3), \quad (14)$$

$f(t) > 0$ and r Diophantine irrational. Assume that

$$\int_0^1 g(t) dt = 0.$$

Then there exists a diffeomorphism $(t, v) \rightarrow (\xi, \eta)$ such that, in the new coordinates (ξ, η) , the map F is given by (10).

Proof. Let us look for ξ and η in the following form:

$$\xi = t + \phi(t)v + O(v^2), \quad \eta = Cv + \psi(t)v^2 + O(v^3) \quad (15)$$

with C a constant. Then (10) and (15) imply

$$T + \phi(T)V = t + \phi(t) + r - Cv + O(v^2), \quad CV + \psi(T)V^2 = Cv + \psi(t)v^2 + O(v^3). \quad (16)$$

Substitute T and V from (14) to (16) and obtain

$$f(t) = (\Delta\phi)(t) + C, \quad -Cg(t) = (\Delta\psi)(t), \quad (17)$$

where Δ is the difference operator:

$$(\Delta\phi)(t) = \phi(t+r) - \phi(t).$$

Equations (17) are solvable for ϕ and ψ in smooth functions if and only if the functions $f(t) - C$ and $g(t)$ have zero mean values (see [14]). The latter condition holds by assumption, and to satisfy the former one sets

$$C = \int_0^1 f(t) dt > 0.$$

This completes the proof. \square

In view of lemma 3.3, it remains to establish the following fact.

Lemma 3.4. *Let $g(t)$ be as in (13). Then*

$$\int_0^1 g(t) dt = 0.$$

Proof. We use the same notation as before. According to (13),

$$g(t) = \lim_{v \rightarrow 0} \frac{u - v}{v^2}.$$

We will express all the quantities involved in terms of ε ; clearly, it suffices to make all the computations modulo ε^3 .

It follows from (11) that

$$w = v - \frac{v^2(\dot{\alpha}^2 - \dot{\alpha}_-^2)}{2s^2}, \quad (18)$$

and (12) implies that

$$\frac{u}{w} = \left(\frac{\dot{\alpha}(t)}{\dot{\alpha}(t-\varepsilon)} \right)^2 \frac{\sin((\alpha(t-\varepsilon) - \alpha(t-r))/2) \sin((\alpha(t+r-\varepsilon) - \alpha(t-\varepsilon))/2)}{\sin((\alpha(t+r-\varepsilon) - \alpha(t))/2) \sin((\alpha(t) - \alpha(t-r))/2)}. \quad (19)$$

Expanding the right-hand side of (19) in ε yields $1 + \varepsilon P + O(\varepsilon^2)$ with

$$P = 2 \frac{\ddot{\alpha}}{\dot{\alpha}} + \frac{\dot{\alpha}}{2} \left(\cot \left(\frac{\alpha_+ - \alpha}{2} \right) - \cot \left(\frac{\alpha - \alpha_-}{2} \right) \right), \quad (20)$$

where $\alpha_+ = \alpha(t+r)$. Set

$$R = \lim_{\varepsilon \rightarrow 0} \left(\frac{w}{\varepsilon} \right).$$

Combining (18)–(20), one has

$$\lim_{v \rightarrow 0} \frac{u - v}{v^2} = \frac{P}{R} - \frac{(\dot{\alpha}^2 - \dot{\alpha}_-^2)}{2s^2}, \quad (21)$$

and it remains to compute R . The sine rule, applied to the triangle Q , $\alpha(t)$, $\alpha(t+r-\varepsilon)$ (see figure 4 again), implies

$$z = \frac{2 \sin((\alpha(t) - \alpha(t - \varepsilon))/2) \sin((\alpha(t+r-\varepsilon) - \alpha(t))/2)}{\sin((\alpha(t+r-\varepsilon) + \alpha(t - \varepsilon) - \alpha(t) - \alpha(t-r))/2)},$$

and it follows that $w = \varepsilon R + O(\varepsilon^2)$ with

$$R = \frac{\sin((\alpha - \alpha_-)/2) \sin((\alpha_+ - \alpha)/2)}{\dot{\alpha} \sin((\alpha_+ - \alpha_-)/2)}. \quad (22)$$

Thus, by (21), one has

$$g(t) = 2\ddot{\alpha} \left[\cot\left(\frac{\alpha - \alpha_-}{2}\right) - \cot\left(\frac{\alpha_+ - \alpha}{2}\right) \right] + \frac{\dot{\alpha}^2}{2} \left[\cot\left(\frac{\alpha_+ - \alpha}{2}\right) + \cot\left(\frac{\alpha - \alpha_-}{2}\right) \right] \\ \times \left[\cot\left(\frac{\alpha_+ - \alpha}{2}\right) - \cot\left(\frac{\alpha - \alpha_-}{2}\right) \right] - \frac{(\dot{\alpha}^2 - \dot{\alpha}_-^2)}{2s^2}.$$

Using trigonometry, one simplifies the last two terms in the above formula to

$$\frac{\dot{\alpha}^2}{2s_+^2} - \frac{\dot{\alpha}^2}{s^2} + \frac{\dot{\alpha}_-^2}{2s^2}, \quad (23)$$

where $s_+ := \sin((\alpha_+ - \alpha)/2)$. Since

$$\int_0^1 \frac{\dot{\alpha}_-^2}{2s^2} dt = \int_0^1 \frac{\dot{\alpha}^2}{2s_+^2} dt,$$

integrating (23) yields

$$\int_0^1 \dot{\alpha}^2 \left(\frac{1}{s_+^2} - \frac{1}{s^2} \right) dt. \quad (24)$$

Denote the integral (24) by I_2 . Next, consider

$$I_1 = \int_0^1 2\ddot{\alpha} \left(\cot\left(\frac{\alpha - \alpha_-}{2}\right) - \cot\left(\frac{\alpha_+ - \alpha}{2}\right) \right) dt. \quad (25)$$

Integrating (25) by parts, yields

$$I_1 = \int_0^1 \dot{\alpha} \left(\frac{\dot{\alpha} - \dot{\alpha}_-}{s^2} \right) - \dot{\alpha} \left(\frac{\dot{\alpha}_+ - \dot{\alpha}}{s_+^2} \right) dt.$$

Since

$$\int_0^1 \frac{\dot{\alpha}\dot{\alpha}_-}{s^2} dt = \int_0^1 \frac{\dot{\alpha}_+\dot{\alpha}}{s_+^2} dt,$$

one concludes that

$$I_1 = \int_0^1 \dot{\alpha}^2 \left(\frac{1}{s^2} - \frac{1}{s_+^2} \right) dt. \quad (26)$$

Comparing (24) and (26), one finds

$$\int_0^1 g(t) dt = I_1 + I_2 = 0,$$

as claimed. \square

4. Some formulae of hyperbolic geometry

This section contains some formulae of differential geometry of curves in the hyperbolic plane. These formulae will be used in the next two sections.

It will be convenient to use the hyperboloid model of the hyperbolic geometry. Recall that one considers the Lorentz space $\mathbb{R}^2 \times \mathbb{R}$ with coordinates (x, y) (where $x = (x_1, x_2)$) and the Lorentz metric $dx^2 - dy^2$; denote the respective scalar product by $\langle \cdot, \cdot \rangle$. The hyperbolic plane is realized as the upper sheet H of the two-sheeted hyperboloid $x^2 - y^2 = -1$ with the induced Riemannian metric. Straight lines in this model are the intersections of H with planes through the origin. The central projection p onto the tangent plane to H at point $(0, 0, 1)$ provides an isomorphism of this model with the Klein–Beltrami one. If $z = (z_1, z_2)$ is a coordinate inside the unit disc, then the isomorphism p^{-1} is given by the formulae

$$x = \frac{z}{\sqrt{1-z^2}}, \quad y = \frac{1}{\sqrt{1-z^2}}. \quad (27)$$

The position vector of a point on H has Lorentz length 1 and is Lorentz perpendicular to H . The area form ω on H , associated with the hyperbolic metric, is obtained by the convolution of the position vector of a point with the volume form $dx_1 \wedge dx_2 \wedge dy$; using (27), one has, in terms of the coordinate z inside the unit disc,

$$\omega = (1-z^2)^{-3/2} dz_1 \wedge dz_2. \quad (28)$$

A curve inside the unit disc will be considered either in the Euclidean or the hyperbolic arc length parametrization; we will denote the former by t and the latter by τ . Likewise, we will denote Euclidean distances by r, r_1 , etc and the respective hyperbolic ones by ρ, ρ_1 , etc. This convention will be used in the following sections as well. For example, if $z(\tau)$ is a curve, parametrized by the hyperbolic arc length, then

$$\frac{z_\tau^2}{1-z^2} + \frac{(zz_\tau)^2}{(1-z^2)^2} = 1,$$

where the product is the Euclidean scalar multiplication.

Next, we compute the curvature of a curve in the hyperbolic plane. Let $\Gamma(\tau) = (x(\tau), y(\tau))$ be a hyperbolic arc length parametrized curve on H . The geodesic curvature vector of Γ is the Lorentz orthogonal projection on H of the acceleration vector $\Gamma_{\tau\tau}$, and the curvature $\kappa(\tau)$ is the magnitude of the geodesic curvature vector. One has

$$\kappa = |\det(\Gamma, \Gamma_\tau, \Gamma_{\tau\tau})|, \quad (29)$$

where the determinant is associated with the constant volume form $dx_1 \wedge dx_2 \wedge dy$. Let $\gamma = p(\Gamma)$ be the plane projection; then $\gamma(\tau) = x(\tau)/y(\tau)$. Taking (27) into account, one has

$$\kappa = (1-\gamma^2)^{-3/2} |\det(\gamma_\tau, \gamma_{\tau\tau})|, \quad (30)$$

where the determinant is associated with the constant area form $dx_1 \wedge dx_2$. Let $\gamma(t)$ be a Euclidean arc length parametrization of the same curve inside the unit disc. Then $|\det(\gamma_t, \gamma_{tt})| = K(t)$ is the Euclidean curvature of γ . One has

$$\det(\gamma_t, \gamma_{tt}) = \det(\gamma_\tau, \gamma_{\tau\tau}) \left(\frac{d\tau}{dt} \right)^3,$$

and therefore

$$\kappa^{1/3} d\tau = (1-\gamma^2)^{-1/2} K^{1/3} dt. \quad (31)$$

The proof of the next lemma is immediate from the Chain rule; we will use this lemma in section 6.

Lemma 4.1. *Let Γ be a curve in 3-space with a constant volume form, and let \det be the associate determinant. Give the curve an arbitrary parametrization $\Gamma(s)$. Then*

$$(\det(\Gamma, \Gamma', \Gamma''))^{1/3} ds \quad (32)$$

is a well-defined 1-form on Γ which does not depend on the parametrization.

The integral of the above 1-form over the curve is an equiaffine invariant called the (equi)affine length, and the parametrization $\Gamma(s)$ such that

$$\det(\Gamma, \Gamma', \Gamma'') = 1 \quad (33)$$

is called the (equi)affine parametrization. Sometimes we extend the notion of affine parametrization to the case when the determinant (33) is constant not necessarily equal to 1.

Remark 4.2. If Γ lies in an affine plane not through the origin (say, $z = 1$), then the 1-form (32) is also invariant under equiaffine transformations of this plane. This is the affine length element of the plane affine differential geometry (see, e.g. [28]).

Lemma 4.3. *Let $\Gamma(s) \subset H$ be an affine parametrized curve. Then*

$$\Gamma'''(s) = u(s)\Gamma(s) + w(s)\Gamma'(s) \quad (34)$$

with

$$u = 3\langle \Gamma', \Gamma'' \rangle = -\langle \Gamma, \Gamma''' \rangle. \quad (35)$$

Proof. Differentiating (33) yields

$$\det(\Gamma, \Gamma', \Gamma''') = 0.$$

This implies (34). Take inner product of (34) with Γ and use the relation $\langle \Gamma, \Gamma' \rangle = 0$ to get

$$u = -\langle \Gamma, \Gamma''' \rangle,$$

which is half of (35). Differentiate $\langle \Gamma, \Gamma' \rangle = -1$ three times to obtain

$$\langle \Gamma, \Gamma''' \rangle + 3\langle \Gamma', \Gamma'' \rangle = 0,$$

and the other half of (35) follows. \square

5. Area spectrum

This section is devoted to the proof of theorem 3. The proof is based on the theory of interpolating Hamiltonians developed by Melrose in [21] and applied to (inner) plane billiards in [19]. An application to dual billiards in the affine plane is contained in [29, 30].

Let us outline the interpolating Hamiltonians theory in the dual billiard setting. One considers the space C of contact elements (x, l) where x is a point outside of the dual billiard curve γ and l is a line through x tangent to γ . The projection $\pi : (x, l) \rightarrow x$ is a two-folded covering, ramified along the curve γ , which has two sections: 'left' and 'right' tangent lines to γ at point x . The space C has an area form induced by π from the hyperbolic plane. One has two symplectic involutions on C . The first interchanges the two pre-images of a contact element under the projection π ; the second interchanges two points on a fixed tangent line to γ , located at equal distances from the tangency point. The dual billiard map is the composition of these involutions.

For comparison, a similar description for convex inner billiards is as follows. An analogue of C is the space of unit vectors with foot point on the billiard curve; this space has a canonical area form. The first involution interchanges two vectors with the same foot point whose

projections on the tangent line to the billiard curve coincide; it takes a vector with the outward direction to an inward vector. Given a line, intersecting the billiard curve at two points, the second involution moves a unit tangent vector along this line from one intersection point to another. The billiard ball map is the composition of these involutions.

Let us return to the dual billiard in the hyperbolic plane. The theory of interpolating Hamiltonians implies that the dual billiard map F can be approximated in a vicinity of the dual billiard curve by an integrable map. More precisely, there exists a smooth non-negative function h in the exterior of γ , equal to zero on γ , and such that

$$F \equiv \exp(h^{1/2} \operatorname{sgrad} h), \quad (36)$$

where $\operatorname{sgrad} h$ is the Hamiltonian vector field of the function h , and the equality is modulo symplectic maps fixing γ to all orders. In particular, the Taylor series of h along γ is uniquely determined by (36).

The area spectrum appears in this setup as a particular case of the symplectic invariant, called the action spectrum, introduced in [9]. Namely, let M be a symplectic manifold with an exact symplectic structure $\omega = d\lambda$, and F be an exact symplectomorphism with a generating function S , i.e. $F^*(\lambda) - \lambda = dS$. Let x be an n -periodic point. Then the sum

$$\sum_{i=0}^{n-1} S(F^i(x)) \quad (37)$$

is a symplectic invariant of F : it depends neither on the choice of the 1-form λ nor on the choice of the generating function S . Lemma 5.2 below states that (37) identifies with the extremal area of circumscribed n -gons. The length of a periodic billiard trajectory is also a particular case of (37). It is shown in [19] that the action spectrum (37) satisfies the asymptotic expansion (2) in even negative powers of n . Evidently, a_0 is the area of the dual billiard table; our goal is to identify the term a_1 as in (3).

Following [19], consider the next, more easily computable, integral invariants. Let Γ_t be the level curve given by $h = t$. The vector field $\operatorname{sgrad} h$ is tangent to Γ_t ; consider the 1-form α normalized by the condition $\alpha(\operatorname{sgrad} h) = 1$. Then

$$I(t) = \int_{\Gamma_t} \alpha$$

is a smooth function of t near 0, and the coefficients of its Taylor expansion at $t = 0$ are the integral invariants

$$I_{k+1} = \left. \frac{d^k I}{dt^k} \right|_{t=0}.$$

The coefficients a_i algebraically depend on I_k ; in particular, $a_1 = C I_1^3$ where C is a constant—see [19]. We will compute the invariant I_1 and find the constant C .

Example 5.1. Let us illustrate the above discussion by the following example. Consider the cylinder $S^1 \times [0, 1]$ with coordinates (x, y) and the area form $y \, dy \wedge dx$. Let the Hamiltonian be given by

$$h(x, y) = y^2 f(x)^2 + O(y^3), \quad \text{with } f(x) > 0.$$

Then

$$\operatorname{sgrad} h = 2f(x)^2 \partial_x + O(y).$$

Let $F(x, y) = (x_1, y_1)$ be the map $\exp(h^{1/2} \operatorname{sgrad} h)$. Then

$$x_1 = x + 2yf(x)^3 + O(y^2),$$

and one has

$$I_1 = \int \frac{dx}{2f(x)^2}. \tag{38}$$

Let us return to dual billiards in the hyperbolic plane. First, we describe a generating function for the dual billiard map. Consider the dual billiard curve in the hyperbolic arc length parametrization $\gamma(\tau)$. Let z be a point outside of γ . Consider the tangent segments from z to γ ; let $\gamma(\tau)$ and $\gamma(\tau_1)$ be the tangency points, ρ and ρ_1 the hyperbolic lengths of the tangent segments. One can use (τ, ρ) as coordinates in the exterior of γ . Denote the hyperbolic area bounded by the two tangent segments and the dual billiard curve by S ; we consider S as a function of τ and τ_1 .

The following lemma shows that S is a generating function of the dual billiard map.

Lemma 5.2. *There exist smooth positive functions $f(\tau)$ and $g(\rho)$ such that*

$$\frac{\partial S}{\partial \tau} = -f(\tau)g(\rho), \quad \frac{\partial S}{\partial \tau_1} = f(\tau_1)g(\rho_1).$$

The 1-form $\lambda = f(\tau)g(\rho) d\tau$ satisfies $d\lambda = \omega$ where ω is the hyperbolic area 2-form.

Proof. Replace an arc of γ of length $d\tau$ by an arc of its osculating circle at point $\gamma(\tau)$ and denote its hyperbolic radius by $u(\tau)$. The end points of tangent segments of length ρ to the circle of radius u constitute a circle of radius v with

$$\cosh v = \cosh u \cosh \rho; \tag{39}$$

(39) is a hyperbolic trigonometry formula for a right triangle with sides v, u, ρ (see, e.g. [4]). Recall that the area and perimeter length of a circle of radius p in hyperbolic geometry are given, respectively, by

$$A(p) = 2\pi(\cosh p - 1), \quad L(p) = 2\pi \sinh p. \tag{40}$$

To find $\partial S/\partial \tau$, consider the area of the infinitesimal triangle in figure 5. This area equals

$$\frac{A(v) - A(u)}{L(u)} d\tau$$

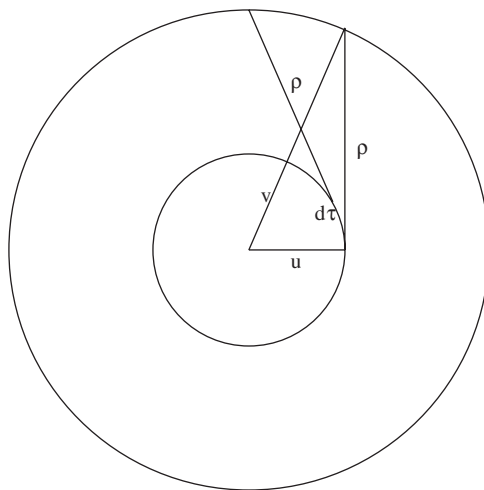


Figure 5. Calculating the hyperbolic area form.

or, in view of (39) and (40),

$$\tanh u(\cosh \rho - 1) d\tau.$$

Therefore, $\partial S/\partial \tau = -f(\tau)g(\rho)$ with $f(\tau) = \tanh u(\tau)$ and $g(\rho) = \cosh \rho - 1$. The formula for $\partial S/\partial \tau_1$ is similar.

By construction,

$$\omega = f(\tau)g'(\rho) d\rho \wedge d\tau,$$

i.e. $\omega = d\lambda$. □

Remark 5.3. An analogue of the above formula in the Euclidean plane is

$$\omega = r dr \wedge d\alpha,$$

where r is a Euclidean analogue of ρ and α is the angle parameter on the curve, i.e. the angle made by its tangent line with a fixed direction. This formula extends the familiar formula for area in polar coordinates.

Next we compute the dual billiard map F near the dual billiard curve. Consider γ in the Euclidean arc length parametrization and characterize a point z outside of γ by the coordinates (t, r) where

$$z = \gamma(t) + r\gamma'(t); \quad (41)$$

here t is the arc length parameter and r is the Euclidean distance from x to $\gamma(t)$ (see figure 6).

Let $F(t, r) = (t_1, r_1)$, and let $K(t)$ be the Euclidean curvature of $\gamma(t)$.

Lemma 5.4. *One has*

$$t_1 = t + 2r - \frac{2}{3} \frac{K'(t)}{K(t)} r^2 + O(r^3), \quad r_1 = r - \left(\frac{2}{3} \frac{K'(t)}{K(t)} + \frac{2\gamma(t)\gamma'(t)}{1 - \gamma(t)^2} \right) r^2 + O(r^3).$$

Proof. One has the vector equality:

$$\gamma(t) + r\gamma'(t) = \gamma(t_1) - R\gamma'(t_1).$$

Write $t_1 = t + \varepsilon$ and expand $\gamma(t_1)$ in powers of ε . Using the equality

$$\gamma''' = -K^2\gamma' + \frac{K'}{K}\gamma''$$

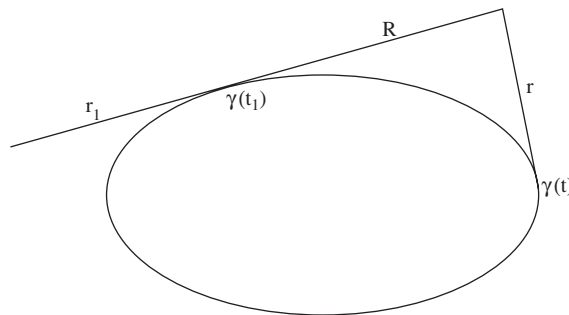


Figure 6. The dual billiard map near the dual billiard curve.

and equating the γ' and γ'' components of the resulting vectors, one finds

$$t_1 = t + 2r - \frac{2}{3} \frac{K'(t)}{K(t)} r^2 + O(r^3), \quad R = r - \frac{2}{3} \frac{K'(t)}{K(t)} r^2 + O(r^3). \quad (42)$$

To find the relation between r_1 and R , parametrize the tangent line at $\gamma(t_1)$ as $\gamma(t_1) + u\gamma'(t_1)$. The two intersection points with the unit circle correspond to the roots $u_{1,2}$ of the quadratic equation $(\gamma(t_1) + u\gamma'(t_1))^2 = 1$, i.e.

$$u^2 + 2u\gamma(t_1)\gamma'(t_1) + \gamma(t_1)^2 - 1 = 0. \quad (43)$$

The dual billiard reflection in point $\gamma(t_1)$ defines a fractional linear involution ψ with 0 a fixed point and interchanging u_1 and u_2 . This involution is given by the formula

$$\psi(x) = \frac{u_1 u_2 x}{(u_1 + u_2)x - u_1 u_2}.$$

It follows from (43) that $u_1 u_2 = \gamma(t_1)^2 - 1$ and $u_1 + u_2 = -2\gamma(t_1)\gamma'(t_1)$. Since $\psi(-R) = r_1$, one has

$$r_1 = \frac{R(1 - \gamma^2)}{2\gamma\gamma'R + 1 - \gamma^2},$$

and it remains to substitute the expression for R from (42). \square

Now we wish to change coordinates from (t, r) to (ξ, η) so that the area has the canonical form

$$\omega = \eta d\eta \wedge d\xi. \quad (44)$$

Lemma 5.5. *There exists a smooth coordinate change in a vicinity of γ given by the formulae*

$$\xi = \phi(t) + O(r), \quad \eta = r + O(r^2) \quad (45)$$

with

$$\phi'(t) = (1 - \gamma(t)^2)^{-3/2} K(t) \quad (46)$$

such that (44) holds. The dual billiard map $F(\xi, \eta) = (\xi_1, \eta_1)$ is given by the formulae

$$\xi_1 = \xi + 2\eta\phi'(t) + O(\eta^2), \quad \eta_1 = \eta + O(\eta^2). \quad (47)$$

Proof. Looking for (ξ, η) as in (45), one has

$$\eta d\eta \wedge d\xi = r\phi'(t) dr \wedge dt + O(r^2).$$

The formula for the area form (28) gives, in view of (41),

$$\omega = (1 - \gamma^2 - 2r\gamma\gamma' - r^2)^{-3/2} r K dr \wedge dt = (1 - \gamma^2)^{-3/2} K r dr \wedge dt + O(r^2),$$

and one obtains (46).

It follows from (45) that

$$\xi_1 = \phi(t_1) + O(r_1), \quad \eta_1 = r_1 + O(r_1^2).$$

Substitute the values of (t_1, r_1) from lemma 5.4 to obtain (47). \square

Now, formulae (44) and (47) take us to the situation of example 5.1 where (x, y) are identified with (ξ, η) and f^3 with ϕ' . It follows that

$$2I_1 = \int \frac{d\xi}{(\phi')^{2/3}}.$$

From (45), one has

$$d\xi = \phi'(t) dt + O(r);$$

therefore

$$2I_1 = \int \phi'(t)^{1/3} dt = \int (1 - \gamma(t)^2)^{-1/2} K(t)^{1/3} dt.$$

Finally, it follows from (31) that the latter integrand is $\kappa(\tau)^{1/3} d\tau$. To summarize, the coefficient a_1 in (2) is given by

$$a_1 = C \left(\int \kappa(\tau)^{1/3} d\tau \right)^3,$$

and it remains to find the constant C .

Lemma 5.6. *One has $C = 1/24$.*

Proof. To determine the constant, consider the case when γ is a circle of radius ρ . The curvature of the circle is related to its area and perimeter by the Gauss–Bonnet theorem:

$$\kappa(\rho)L(\rho) = 2\pi A(\rho).$$

In view of (40), this implies

$$\kappa(\rho) = \coth \rho,$$

and therefore

$$\int \kappa(\tau)^{1/3} d\tau = 2\pi \cosh^{1/3}(\rho) \sinh^{2/3}(\rho). \quad (48)$$

A regular n -gon, circumscribed about the circle, consists of $2n$ congruent right triangles with side ρ and an adjacent angle π/n . Let $\pi/2 - \varepsilon$ be the other acute angle. Then, by hyperbolic trigonometry (see [4]),

$$\sin \varepsilon = \cosh \rho \sin \frac{\pi}{n},$$

and therefore

$$\varepsilon = \frac{\pi}{n} \cosh \rho + \frac{\pi^3 \sinh^2 \rho \cosh \rho}{6n^3} + O\left(\frac{1}{n^5}\right).$$

By the Gauss–Bonnet theorem, the area of the triangle equals

$$\varepsilon - \frac{\pi}{n} = \frac{\pi}{n} (\cosh \rho - 1) + \frac{\pi^3 \sinh^2 \rho \cosh \rho}{6n^3} + O\left(\frac{1}{n^5}\right),$$

and the area of the n -gon is

$$2\pi (\cosh \rho - 1) + \frac{\pi^3 \sinh^2 \rho \cosh \rho}{3n^2} + O\left(\frac{1}{n^4}\right). \quad (49)$$

Equating (49) and

$$A(\rho) + C \left(\int \kappa(\tau)^{1/3} d\tau \right)^3 \frac{1}{n^2} + O\left(\frac{1}{n^4}\right),$$

and taking (48) into account, one finds $C = 1/24$. □

6. (Pseudo)spherical pendulum and relative extrema of the affine length functional

This section contains a proof of theorem 5. We will consider the case of the pseudosphere, the more familiar case of the sphere being completely similar.

First, we describe the pseudospherical pendulum. Let g be a constant vector field in three-space representing gravity. The unit mass-point is constrained to lie on the pseudosphere H . Let $\Gamma(s)$ be the time parametrized trajectory of the point. Newton's second law of motion reads as

$$\Gamma''(s) = g + \nu(s)\Gamma(s), \quad (50)$$

where $\nu\Gamma$ is the force of reaction from H (recall that the position vector Γ is Lorentz orthogonal to H).

Consider the following two functionals on curves in H . The first is the affine length

$$I(\Gamma) = \int \det(\Gamma, \Gamma', \Gamma'')^{1/3} ds,$$

where $\Gamma(s)$ is some parametrization of the curve. The second functional $A(\Gamma)$ is the hyperbolic area bounded by the curve. Both functionals do not depend on the parametrization of the curve. To find the extrema of I relative to A , we consider the curves in the affine parametrization so that (33) holds.

Let v be a variation of Γ , i.e. a vector field along Γ , tangent to H . We compute the variational derivatives of the functionals I and A .

Lemma 6.1. *One has*

$$\frac{\delta I}{\delta v} = \int \left(\det(\Gamma', \Gamma'', v) + \frac{2}{3} \det(\Gamma, \Gamma''', v) \right) ds, \quad \frac{\delta A}{\delta v} = \int \det(\Gamma, \Gamma', v) ds.$$

Proof. By definition of the hyperbolic area on H given in section 4 the area of the parallelogram, spanned by the vectors Γ' and v , equals $\det(\Gamma, \Gamma', v)$. This implies the formula for $\delta A/\delta v$.

Consider $\delta I/\delta v$. One has

$$(\det(\Gamma + v, \Gamma' + v', \Gamma'' + v''))^{1/3} = 1 + \frac{1}{3}(\det(v, \Gamma', \Gamma'') + \det(\Gamma, v', \Gamma'') + \det(\Gamma, \Gamma', v'')).$$

Integrating by parts, one gets

$$\int \det(\Gamma, \Gamma', v'') ds = - \int \det(\Gamma, \Gamma'', v') ds$$

and

$$\int \det(\Gamma, v', \Gamma'') ds = \int \det(\Gamma', \Gamma'', v) ds + \int \det(\Gamma, \Gamma''', v) ds,$$

and the formula for $\delta I/\delta v$ follows. \square

As a consequence, the relative extremum (Euler–Lagrange) equation reads as

$$\int \left(\det(\Gamma', \Gamma'', v) + \frac{2}{3} \det(\Gamma, \Gamma''', v) \right) ds = \lambda \int \det(\Gamma, \Gamma', v) ds \quad (51)$$

for every variation v ; here λ is a constant (Lagrange multiplier).

Notice that (51) holds for every v , tangent to Γ . Therefore, it suffices to consider those v that are collinear with a fixed transverse field of directions along Γ . Choose as such a direction the projection on H of the acceleration vector Γ'' , i.e. set

$$v = f(\Gamma'' + \langle \Gamma, \Gamma'' \rangle \Gamma).$$

Substitute to (51) and take (34) into account to obtain an equation, equivalent to (51):

$$\langle \Gamma, \Gamma'' \rangle + \frac{2}{3}w = \lambda \quad \text{or} \quad w = -\frac{3}{2}\langle \Gamma, \Gamma'' \rangle + \mu, \quad (52)$$

where w is as in (34) and $\mu = 1.5\lambda$ is another constant.

Now we are in a position to prove theorem 5.

First, consider a relative extremal curve $\Gamma(s) \subset H$ in the affine parametrization. Then (52) holds and, according to (34) and (35),

$$\Gamma''' = 3\langle \Gamma', \Gamma' \rangle \Gamma - \frac{3}{2}\langle \Gamma, \Gamma'' \rangle \Gamma' + \mu \Gamma'.$$

Integrating, one has, in view of (35):

$$\Gamma'' = (\mu - \frac{3}{2}\langle \Gamma, \Gamma'' \rangle) \Gamma + g$$

where the vector g is an integration constant. This is an equation of the form (50).

Conversely, let $\Gamma(s)$ be a pendulum trajectory. Differentiate (50) to obtain

$$\Gamma''' = v' \Gamma + v \Gamma'. \quad (53)$$

It follows that

$$\det(\Gamma, \Gamma', \Gamma''') = 0,$$

and therefore $\det(\Gamma, \Gamma', \Gamma'')$ is a constant. Thus $\Gamma(s)$ is an affine parametrization.

Equation (53) is of the form (34) with $u = v'$, $w = v$. We want to establish the relative extremum condition (52) or, equivalently,

$$w' = -\frac{3}{2}\langle \Gamma, \Gamma'' \rangle'. \quad (54)$$

In our case, $w' = u$, and (54) reads as

$$u = -\frac{3}{2}(\langle \Gamma', \Gamma'' \rangle + \langle \Gamma, \Gamma''' \rangle);$$

this equality holds for every affine parametrized curve by (35). This completes the proof of theorem 5.

Acknowledgments

I am grateful to A Veselov who attracted my attention to the Puiseux theorem and provided reference [2] and to M Levi for useful discussions on KAM theory. I am very much indebted to the referees for numerous helpful suggestions. It is a pleasure to acknowledge the hospitality of the Fields Institute in Toronto where the work was completed.

References

- [1] Amiran E 1995 Lazutkin coordinates and invariant curves for outer billiards *J. Math. Phys.* **36** 1232–41
- [2] Appel P 1896 *Traité de mécanique rationnelle* (Paris: Gauthier-Villars)
- [3] Boyland Ph 1996 Dual billiards, twist maps and impact oscillators *Nonlinearity* **9** 1411–38
- [4] Coxeter H S M 1942 *Non-Euclidean Geometry* (Toronto: University of Toronto Press)
- [5] Dogru F and Tabachnikov S 2002 On polygonal dual billiards in the hyperbolic plane *Preprint*
- [6] Herman M 1979 Sur la conjugaison différentiable des difféomorphismes du cercle à des rotations *IHES Publ. Math.* **49** 5–233
- [7] Gruber P 1983 *Approximation of Convex Bodies. Convexity and its Application* (Basle: Birkhauser) pp 131–62
- [8] Gruber P 1993 Aspects of approximation of convex bodies *Handbook of Convex Geometry* (Amsterdam: North Holland) pp 319–46
- [9] Guillemin V and Melrose R 1981 *A Cohomological Invariant of Discrete Dynamical Systems. Christoffel Centennial* vol 672–679 (Basle: Birkhauser)
- [10] Guillemin V and Melrose R 1979 An inverse spectral result for elliptical regions in \mathbf{R}^2 *Adv. Math.* **32** 128–48

- [11] Gutkin E and Katok A 1995 Caustics for inner and outer billiards *Comm. Math. Phys.* **173** 101–34
- [12] Gutkin E and Simanyi N 1991 Dual polygonal billiards and necklace dynamics *Comm. Math. Phys.* **143** 431–50
- [13] Kac M 1966 Can one hear the shape of a drum? *Am. Math. Monthly* **73** 1–23
- [14] Katok A and Hasselblatt B 1995 *Introduction to the Modern Theory of Dynamical Systems* (Cambridge: Cambridge University Press)
- [15] Kolodziej R 1989 The antibilliard outside a polygon *Bull. Pol. Acad. Sci.* **37** 163–8
- [16] Lawden D 1989 *Elliptic Functions and Applications* (Berlin: Springer)
- [17] Lazutkin V 1973 The existence of caustics for a billiard problem in a convex domain *Math. USSR, Izv.* **7** 185–214
- [18] Ludwig M 1994 Asymptotic approximation of convex curves *Arch. Math.* **63** 377–84
- [19] Marvizi S and Melrose R 1982 Spectral invariants of convex planar regions *J. Diff. Geom.* **17** 475–502
- [20] McClure D and Vitale R 1975 Polygonal approximation of plane convex bodies *J. Math. Anal. Appl.* **51** 326–58
- [21] Melrose R 1976 Equivalence of glancing hypersurfaces *Invent. Math.* **37** 165–91
- [22] Moser J 1962 On invariant curves of area preserving mappings of an annulus *Nachr. Akad. Wiss. Gottingen Math-Phys.* II **KL** 1–20
- [23] Moser J 1973 Stable and random motions in dynamical systems *Ann. Math. Stud.* **77**
- [24] Moser J 1978 Is the solar system stable? *Math. Intell.* **1** 65–71
- [25] Newton I 1999 *The Principia: Mathematical Principles of Natural Philosophy* (Berkeley: UC Press)
- [26] Penkov V and Stoyanov L 1992 *Geometry of reflecting rays and inverse spectral problems* (New York: Wiley)
- [27] Shaidenko A and Vivaldi F 1987 Global stability of a class of discontinuous dual billiards *Comm. Math. Phys.* **110** 625–40
- [28] Simon U 2000 Affine differential geometry *Handbook of Differential Geometry* vol 1 (Amsterdam: North-Holland) pp 905–61
- [29] Tabachnikov S 1993 Dual billiards *Russ. Math. Surv.* **48** 75–102
- [30] Tabachnikov S 1995 On the dual billiard problem *Adv. Math.* **115** 221–49
- [31] Tabachnikov S 1996 Asymptotic dynamics of the dual billiard transformation *J. Stat. Phys.* **83** 27–38
- [32] Tabachnikov S 1995 Billiards. SMF ‘Panoramas et Syntheses’, N1
- [33] Tabachnikov S 1993 Poncelet’s theorem and dual billiards *L’Enseign. Math.* **39** 189–94
- [34] Tabachnikov S 1994 Commuting dual billiards *Geom. Dedicata* **53** 57–68
- [35] Fejes Toth L 1948 Approximation by polygons and polyhedra *Bull. Am. Math. Soc.* **4** 431–8
- [36] Fejes Toth L 1953 *Lagerungen in der Ebene, auf der Kugel und im Raum* (Berlin: Springer)