Billiards

Serge Tabachnikov

Introduction

Mathematical billiards is a rich and beautiful subject. It is very extensive as well. The choice of material for this survey reflects the taste of the author, who has attempted to make the exposition as geometrical as possible.

The survey consists of five chapters: the first provides some general background; the second concerns convex smooth billiards (elliptic case); the third deals with billiards in polygons and polyhedra (parabolic case); the fourth discusses the lesser-known topic of dual billiards, which are of particular interest to the author; and the fifth is a very brief treatment of chaotic billiards (hyperbolic case). Each chapter has a brief introduction of its own and is subdivided into sections; it goes without saying that "Lemma 1.2.3" means "Lemma 3 from Section 2, Chapter 1".

The author is grateful to the mathematicians he had the opportunity to discuss billiards with and learn from: V. Arnold, M. Audin, M. Berger, M. Bialy, Ph. Boyland, N. Chernov, D. Fuchs, G. Galperin, E. Ghys, A. Givental, M. Gromov, E. Gutkin, P. Iglesias, A. Katok, R. de Llave, J. Moser, Ya. Pesin, L. Polterovich, Ya. Sinai, J. Smillie, S. Troubetzkoy, A. Veselov, M. Wojtkowski. They taught him far more than these notes show.

Special gratitude goes to I. Monroe for his help with the numerical study of dual billiards, to J. Duncan for patiently reading the text and giving stylistic advice and to E. Gutkin, G. Galperin and the referees whose suggestions helped to improved the exposition. Last but not least, it is a pleasure to acknowledge the partial support of an ASTA grant 94-B-25.

Contents

1. General Theory and Mathematical Background

This chapter concerns the definition of the billiard transformation and the billiard flow. It also provides a necessary mathematical background for the rest of the survey. Section 2 introduces the area form invariant under the billiard transformation in the two-dimensional case, and the variational approach to the billiard problem. Section 3, in the spirit of the geometrical optics, deals with the billiard transformation of the space of rays in the plane. Section 5 is a very brief introduction to symplectic geometry, which we use in the next section to understand the results of Sections 2 and 3 from a more general viewpoint and to generalize them to the higher-dimensional case. Section 7 concerns discontinuities of the billiard transformation; we show that the transformation and the billiard flow for all times are defined almost everywhere in the sense of measure. Section 10 introduces the hierarchy of stochastic properties such as ergodicity, minimality, topological transitivity, mixing, etc. The contents of other section is self-explanatory.

1.1 Mathematical Billiards. Phase and Configuration Spaces

A billiard table is a Riemannian manifold M with a piecewise smooth boundary. The billiard dynamical system in M is generated by the free motion of a mass-point (called a billiard ball) subject to the elastic reflection in the boundary. This means that the point moves along a geodesic line in M with a constant (say, unit) speed until it hits the boundary. At a smooth boundary point the billiard ball reflects so that the tangential component of its velocity remains the same, while the normal component changes its sign. In dimension two this collision is described by a well known law of geometrical optics: the angle of incidence equals the angle of reflection. Thus the theory of billiards and the theory of geometrical optics have many features in common. If the billiard ball hits a corner, its further motion is not defined (there are some exceptions to this to be discussed later).

fig. 1

The time-t billiard transformation acts on unit tangent vectors to M, more precisely, on those pairs (x, v) with $x \in M, v \in T_xM$ whose trajectories undergo finitely many reflections in the boundary and avoid corners on the time interval $[0, t]$. The unit tangent bundle to M is the phase space of the billiard, and the manifold M is its configuration space.

Billiards are the geodesic flows on Riemannian manifolds with boundaries. They can also be treated as a limit case of the geodesic flows on boundaryless manifolds, at least heuristically. Let M be a smooth plane billiard table. Consider its "thickening", i.e. an infinitely thin three dimensional body whose boundary N is obtained by pasting two copies of M along their boundaries and smoothening the edge. Then a billiard trajectory in M can be viewed as a geodesic line on the boundary of N, that goes from one copy of M to another each time the billiard ball bounces off the boundary. This construction is due to G. Birkhoff ([Bi 1]).

fig. 2

1.2 Invariant Measure and Generating Function for Plane Billiards. Variational Formulation

So far billiards were defined as a continuous time dynamical system. One can reduce the dimension by one and replace continuous time by discrete time, i.e. replace a flow by a mapping.

Let M be a bounded plane billiard table. Consider the manifold V of unit tangent vectors (x, v) with the inward direction v and the footpoint x on the boundary ∂M. If the boundary consists of one component then V is an annulus $S^1 \times I$. We now define the billiard transformation T of V. A vector (x, v) moves along the straight line through x in the direction of v to the next point of its intersection x_1 with ∂M, and then v reflects in ∂M to a new vector v_1: $T(x, v) = (x_1, v_1)$.

fig. 3

A very remarkable property of the billiard transformation T is the existence of an invariant area form. Parametrize ∂M by the length parameter t and let α be the angle between v and the direction of the boundary at the point $x(t)$. Use (t, α) as the coordinates in V, $\alpha \in [0, \pi]$.

Lemma. *The area form* $\sin\alpha \; d\alpha \wedge dt$ *is* T*-invariant.*

Proof. Let $T(t, \alpha) = (t_1, \alpha_1)$. One wants to prove that $\sin\alpha_1 \; d\alpha_1 \wedge dt_1 = \sin\alpha \; d\alpha \wedge dt$. Let $H(t, t_1)$ be the distance between the points $x(t)$ and $x(t_1)$. It readily follows from elementary geometry that

$$\frac{\partial H(t,t_1)}{\partial t} = -\cos\alpha \quad \text{and} \quad \frac{\partial H(t,t_1)}{\partial t_1} = \cos\alpha_1.$$

Hence

$$\cos\alpha_1 \; dt_1 - \cos\alpha \; dt = \frac{\partial H(t,t_1)}{\partial t_1} dt_1 + \frac{\partial H(t,t_1)}{\partial t} dt = dH,$$

and taking differentials,

$$\sin\alpha_1 \; d\alpha_1 \wedge dt_1 - \sin\alpha \; d\alpha \wedge dt = 0.$$

Q.E.D.

Consider three consecutive points: $(t_1, \alpha_1) = T(t, \alpha), (t_2, \alpha_2) = T(t_1, \alpha_1)$. It follows from the previous proof that

$$\frac{\partial H(t,t_1)}{\partial t_1} = \cos\alpha_1, \;\; \frac{\partial H(t_1,t_2)}{\partial t_1} = -\cos\alpha_1.$$

Hence

$$\frac{\partial H(t,t_1)}{\partial t_1} + \frac{\partial H(t_1,t_2)}{\partial t_1} = 0.$$

This formula has the following interpretation. Suppose that one wants to start the billiard ball at the point x so that after one reflection in the boundary at some point x_1 it arrives to the given point x_2. How does one find the unknown point x_1? Answer: this point is a critical point of the functional $dist(x, x_1) + dist(x_1, x_2)$. This variational principle plays an important role in the study of billiards.

1.3 Billiard Transformation of the Space of Rays in the Plane

It would be more in the spirit of geometrical optics to deal with oriented lines (or rays). Such an approach to billiards is possible and indeed fruitful.

Suppose that a plane billiard table M is convex (this assumption is not really necessary; what follows will hold true locally for a generic M). Let U be the set of oriented lines in the plane that intersect M. To parametrize the set of rays choose the origin O inside M. Given a ray l, drop the perpendicular OP onto it. Fix a direction in the plane and let l make the angle of ϕ with it. Let $p = \pm|OP|$ depending on the orientation of the frame $(\vec{l}, \vec{OP})$. Then (p, ϕ) are the coordinates in the set of oriented lines in the plane. U is given by the inequality $|p| \leq f(\phi)$, where

the function $f(\phi)$ depends on the shape of M (it is called the support function). It follows that U is diffeomorphic to an annulus.

fig. 4

Define the billiard transformation T' of U: the ray, that contains a segment of the trajectory of the billiard ball, oriented by the direction of its motion, is sent to the ray, that contains the next segment of this trajectory after the reflection in the boundary.

The set of lines in the plane is an object of study in integral geometry. It is known that there exists a unique, up to a constant factor, measure on the set of lines invariant under the motions of the plane. In our notation this is given by the 2-form $dp \wedge d\phi$ (see [Sa], and also [Ig] on the relation between symplectic and integral geometry).

Identify the manifold V, introduced in the previous section, with U: the ray through x in the direction of v corresponds to a point $(x, v) \in V$. Thus one identifies the transformations T and T'. Compare the two 2-forms $\sin\alpha \ d\alpha \wedge dt$ and $dp \wedge d\phi$.

Lemma. *These forms are equal.*

Proof. An equation of the line, whose coordinates in U are (p, ϕ), is

$$y\cos\phi - x\sin\phi = p.$$

Differentiate:

$$\cos\phi \ dy - \sin\phi \ dx - (y\sin\phi + x\cos\phi) \ d\phi = dp.$$

Hence

$$\cos\phi \ dy \wedge d\phi - \sin\phi \ dx \wedge d\phi = dp \wedge d\phi.$$

The angle made by the direction of ∂M at the point $x(t), y(t)$ and the fixed direction is $\alpha + \phi$. Therefore

$$dy = \sin(\alpha + \phi) \ dt, \quad dx = \cos(\alpha + \phi) \ dt.$$

fig. 5

Hence

$$(\cos\phi\sin(\alpha + \phi) - \sin\phi\cos(\alpha + \phi)) \ dt \wedge d\phi = \sin\alpha \ dt \wedge d\phi = dp \wedge d\phi.$$

Since $\frac{d(\alpha+\phi)}{dt} = K(t)$ –the curvature of the curve ∂M, one has: $d\phi = -d\alpha + K \ dt$. Therefore $dt \wedge d\phi = d\alpha \wedge dt$ and, finally,

$$\sin\alpha \ d\alpha \wedge dt = dp \wedge d\phi$$

Q.E.D.

It follows that the billiard transformation T' preserves the natural area form of the space of rays in the plane.

1.4 Language of Symplectic Geometry

To put the observations, made so far about plane billiards, into a proper context, and to generalize them to higher dimensions, one needs some basic concepts of symplectic geometry (see the excellent surveys [A-G, Ar 1, Gro]).

Definition. A symplectic manifold (M, ω) is a smooth manifold M with a closed nondegenerate differential 2-form ω, called the symplectic structure.

Since the 2-form is nondegenerate, the dimension of a symplectic manifold is even. A symplectic manifold has a canonical volume form ω^n, where $2n = dim\ M$.

Example. A $2n$–dimensional linear space with coordinates $x_1, ..., x_n, y_1, ..., y_n$ has a linear symplectic structure $dx \wedge dy = dx_1 \wedge dy_1 + ... + dx_n \wedge dy_n$.

Unlike Riemannian manifolds symplectic manifolds are all locally equivalent (Darboux's theorem): there exists a local diffeomorphism that carries one symplectic form to another. Hence the previous example provides a local normal form of a symplectic manifold. The coordinates x, y in which $\omega = dx \wedge dy$ are called Darboux coordinates.

The following example is of fundamental importance to classical mechanics, and in particular, to the theory of billiards.

Example. The cotangent bundle T^*M of a smooth manifold M has a symplectic structure. Let λ be the tautological differential 1-form on T^*M (called the Liouville form), i.e., the form whose value on a vector ξ, tangent to T^*M at a point (q, p) with $q \in M,\ p \in T^*M$, is equal to the value of the covector p on the projection of ξ to the tangent space T_qM. The natural symplectic structure on T^*M is the 2-form $\omega = d\lambda$.

Identify a linear $2n$-dimensional space with the cotangent bundle of an n-dimensional space, and choose the "position" coordinates $q_1, ..., q_n$ and the dual "momentum" coordinates $p_1, ..., p_n$. In these coordinates $\lambda = p\ dq = p_1\ dq_1 + ... + p_n\ dq_n$ and $\omega = dp \wedge dq = dp_1 \wedge dq_1 + ... + dp_n \wedge dq_n$.

Definition. An n-dimensional submanifold L of a symplectic manifold (M^{2n}, ω) is called Lagrangian if the restriction of ω to L vanishes.

Since ω is a nondegenerate 2-form, n is the greatest possible dimension of a submanifold on which the symplectic form vanishes.

Examples. A smooth curve in the plane (with any area form) is a Lagrangian manifold. Fibers of a cotangent bundle are Lagrangian manifolds. Given a smooth function on a manifold M the graph of its differential, considered as a section of the cotangent bundle, is a Lagrangian submanifold in T^*M.

Definition. A diffeomorphism of symplectic manifolds that carries one symplectic structure to another is called a symplectomorphism.

Symplectomorphisms carry Lagrangian manifolds to Lagrangian manifolds. If $f\colon (M_1, \omega_1) \to$

(M_2, ω_2) is a symplectomorphism then its graph is a Lagrangian submanifold of the product manifold $M_1 \times M_2$ with the symplectic structure $\omega_1 \ominus \omega_2$. The converse is also true.

Definition. A symplectic manifold is called exact if its symplectic structure is the differential of a 1-form: $\omega = d\lambda$.

Cotangent bundles are exact symplectic manifolds. An exact symplectic manifold cannot be compact since its symplectic volume form is exact. Let T be a symplectomorphism of an exact symplectic manifold. Then $T^*\omega = \omega$, or $d\,(T^*\lambda - \lambda) = 0$.

Definition. A symplectomorphism is called exact if the closed 1-form $T^*\lambda - \lambda$ is exact: $T^*\lambda - \lambda = dH$ for a function H. The function H is called a generating function of the exact symplectomorphism.

To a function on a symplectic manifold a vector field naturally corresponds (the function is sometimes called a Hamiltonian function and the field – the Hamiltonian vector field). A symplectic structure, being a nondegenerate 2-form, defines an isomorphism between the tangent and the cotangent bundles. The Hamiltonian vector field (a section of the tangent bundle) corresponds under this isomorphism to the differential of the Hamiltonian function (a section of the cotangent bundle). In Darboux coordinates the Hamiltonian field of a function f is $-\frac{\partial f}{\partial y}\frac{\partial}{\partial x} + \frac{\partial f}{\partial x}\frac{\partial}{\partial y}$. The flow of a Hamiltonian vector field preserves the symplectic structure. In particular, it preserves the symplectic volume (Liouville's theorem). The Hamiltonian vector field of a function f is also called its symplectic gradient and denoted by $sgrad\ f$.

Hamiltonian vector fields form a Lie algebra under the usual commutator of vector fields. Hamiltonian functions also form a Lie algebra whose operation is called Poisson bracket. The Poisson bracket of two functions f and g is the derivative of one of them along the Hamiltonian vector field of the other. In Darboux coordinates $\{f, g\} = \frac{\partial f}{\partial x}\frac{\partial g}{\partial y} - \frac{\partial f}{\partial y}\frac{\partial g}{\partial x}$. The map $f \to sgrad\ f$ is a homomorphism of Lie algebras.

Given a hypersurface in a symplectic manifold, the restriction of the symplectic structure to it is not nondegenerate any more: it has a one-dimensional kernel in each tangent hyperplane.

Definition. This kernel is called the characteristic direction. Integral lines of the field of characteristic directions are called characteristic lines, or simply characteristics. Characteristic lines constitute the characteristic foliation of a hypersurface.

If a smooth function is constant on a hypersurface then its symplectic gradient is tangent to the hypersurface along its characteristic lines.

Suppose that the set of characteristics of a hypersurface in a symplectic manifold is itself a manifold (locally it is always the case). Then this new manifold of characteristics carries a symplectic structure: its value at a pair of tangent vectors $\bar{\xi}$ and $\bar{\eta}$ equals the value of the original symplectic structure at vectors ξ and η, tangent to the hypersurface at some point, that project to $\bar{\xi}$ and $\bar{\eta}$ (the result does not depend on the choice involved).

Let M be a Riemannian manifold. Consider the hypersurface in T^*M that consists of unit covectors. The characteristics of this hypersurface are identified with nonparametrized oriented geodesic lines in M. Hence if the set of oriented geodesics of M is a manifold, it is a symplectic

manifold.

Example. The set of oriented lines in a Euclidean space is a symplectic manifold. Up to the sign of the symplectic structure it is symplectomorphic to the cotangent bundle of the unit sphere in the space. The diffeomorphism of the space of rays with T^*S^{n-1} is shown in the figure (exercise for the reader: formulate it explicitly).

fig. 6

1.5 Symplectic Properties of Billiards

Now we are in a position to reconsider the results of Sections 1.2–1.3 from a broader point of view. To fix ideas, consider a bounded strictly convex domain M with a smooth boundary in an n-dimensional Euclidean space. This will be the billiard table (in the general case the results of this section hold locally, that is, in a neighbourhood of a generic point of the phase space). Identify the tangent and cotangent bundles using the Euclidean structure, and consider the natural symplectic structure of T^*M. Two hypersurfaces in T^*M are of importance: the one that consists of unit (co)vectors, and the one that consists of (co)vectors whose footpoints lie on the boundary ∂M. Call these hypersurfaces Y and Z, correspondingly. Their intersection W consists of unit tangent vectors with the footpoints on ∂M. This intersection is transversal.

Consider the characteristic foliations of the hypersurfaces Y and Z. We already know that the characteristics of Y are oriented lines in the space, that intersect M. Let U be the manifold of these lines; it has the induced symplectic structure, introduced in the previous section. U is diffeomorphic to the unit disc subbundle of the cotangent bundle of the unit sphere in the space.

To describe the characteristics of Z, consider the map $Z \to T(\partial M)$ that projects a tangent vector in the ambient space onto the tangent hyperplane to ∂M.

Lemma 1. *The characteristics of Z are the fibers of this projection.*

Proof. Introduce the usual coordinates $q_1, ..., q_n, p_1, ..., p_n$ in the cotangent bundle of the space; $\omega = dp \wedge dq$. Let $f(q) = 0$ be a local equation of ∂M. Then the fibers of the projection $Z \to T(\partial M)$ are generated by the normal vectors to ∂M, i.e., by the vectors $\frac{\partial f(q)}{\partial q}\frac{\partial}{\partial p}$. The inner product of this vector with ω equals df – a 1-form that vanishes on tangent hyperplanes to ∂M. *Q.E.D.*

Thus the space of characteristics of Z is the (co)tangent bundle of the boundary of the billiard table. Call it V. The identification $V = T^*(\partial M)$ carries the symplectic structure of V to that of $T^*(\partial M)$. The image of the composite map $W \to Z \to V$ is the unit disc subbundle of $T(\partial M)$. This image can be identified with the set of inward unit tangent vectors to M whose footpoints lie on ∂M. In the two-dimensional case the symplectic structure of V is given by the formula $\sin\alpha \; d\alpha \wedge dt$ from Section 1.3.

Let $\Sigma \subset W$ be the set of points where the restriction of the symplectic structure of T^*M to W degenerates. Denote T^*M by X and collect all the spaces and maps in the commutative diagram:

$$\begin{array}{ccc} & X & \\ Y & & Z \\ & W & \\ U & & V \\ & \Sigma & \end{array}$$

Here (X, ω) is a symplectic manifold; Y and Z are two hypersurfaces, whose transversal intersection is W; U and V are the spaces of characteristics of Y and Z; and Σ is the singular set of the restriction of ω onto W. This hexagonal diagram was introduced by R. Melrose ([Me 1,2; Ar 2,3]) in his study of diffraction singularities near gliding rays. This diagram translates the differential geometry of submanifolds in a Riemannian manifold to the symplectic geometry of pairs of hypersurfaces in a symplectic manifold (not necessarily in the context of billiards).

Lemma 2. *Σ equals the set of critical points of the projection $W \to U$ as well as of the projection $W \to V$.*

Proof. A point of W is a critical point of the projection $W \to U$ if and only if the characteristic of Y trough this point is tangent to W. In this case the restriction of ω to W has a kernel – this characteristic direction. Conversely, let ξ be a vector from the kernel of the restriction of ω to W, and let η be a characteristic vector of Y at the same point. If ξ and η are not collinear, then the tangent space to Y is generated by η and the tangent space to W. Since $\omega(\xi, \eta) = 0$, ξ also belongs to $ker\ \omega|Y$. Hence ξ also has the characteristic direction and is collinear to η. Thus η is tangent to W. *Q.E.D.*

In the case of billiards Σ consists of unit (co)tangent vectors to the boundary ∂M. Convexity of the billiard table implies that the only singularity of the projection $W \to U$ is a fold along Σ (and similarly for $W \to V$). Two involutions on W arise: the ones that interchange the inverse images of a point under the projections $W \to U$ and $W \to V$. Call them σ and τ, correspondingly. The two involutions have the common set of fixed points Σ. The billiard transformation of W is the composition $\tau\sigma$.

fig. 7

The results of Sections 1.2-1.3 in our more general situation will follow from an almost tautological remark.

Lemma 3. *The projections $W \to U$ and $W \to V$ preserve the symplectic structures off the singular set Σ. The involutions σ and τ preserve the restriction of ω to W.*

Proof. Given two tangent vectors to U, one computes the value of the symplectic structure of U on these vectors as follows: lift the vectors to W and evaluate ω on them. This implies the first statement.

Given two tangent vectors to W, one computes the value of $\omega|W$ on them as follows: lift the vectors to Y and evaluate $\omega|Y$ on them. Since the result does not depend on the lift, one could

equally well use the images of these tangent vectors to W under the involution σ. This implies the second statement. *Q.E.D.*

Finally one defines the billiard transformations $T' : U \to U$ and $T : V \to V$. Define T'. Given a ray in the space, that intersects M (i.e., a point in U), consider its unit tangent vector at the second point of its intersection with ∂M (lift the point of U to W). Apply τ (reflect in the boundary) and project back to U. Likewise one defines T. Given a tangent vector to ∂M, whose length is not greater than 1 (a point in V), consider the inward unit vector in the ambient space at the same point of ∂M whose projection to $T(\partial M)$ is the given vector (lift to W). Apply σ (move the billiard ball until it hits the boundary) and project to V.

It follows from the previous lemma that the transformations T and T' preserve the symplectic structures of V and U. Therefore they also preserve the symplectic volumes therein. We also mention that T and T' are conjugate by the natural diffeomorphism of U and V.

To complete this discussion describe the generating function of the billiard transformation $T : V \to V$. As before, use the position coordinates q in the linear space and the corresponding momentum coordinates p in the tangent space. The symplectic structure is $d\lambda$, where $\lambda = p\,dq$. Let (q, p) and (q_1, p_1) be two points of W (so $q, q_1 \in \partial M$ and $|p| = |p_1| = 1$) such that $\sigma(q, p) = (q_1, p_1)$. Let $H(q, q_1)$ be the distance between q and q_1. Then as in Section 1.2, $\frac{\partial H}{\partial q} = -p$, $\frac{\partial H}{\partial q_1} = p_1$. Hence

$$\sigma^*\lambda - \lambda = p_1\ dq_1 - p\ dq = dH.$$

fig. 8

Thus H is the generating function of the billiard transformation, and in particular, we again see that this transformation is symplectic. The variational interpretation from Section 1.2 holds as well.

1.6 Poincaré's Recurrence Theorem

We have seen that the billiard transformation is symplectic, and therefore, volume-preserving. A general fundamental property of volume-preserving transformation was discovered by H.Poincaré.

Theorem 1. *Let T be a volume-preserving transformation of a manifold with a finite volume. Then for any neighbourhood U of any given point there exists a point $x \in U$ which returns to this neighbourhood: $T^n x \in U$ for some positive n. The set of points in U that never return to U has zero volume.*

Proof. Consider the images of $U : U, TU, T^2U, ...$ They have equal positive volumes. Since the volume of the manifold is finite, some images intersect. Hence for $k > l \geq 0$ one has: $T^kU \cap T^lU \neq \emptyset$. Therefore $T^{k-l}U \cap U \neq \emptyset$. Let $T^{k-l}x = y$ for $x, y \in U$. Then x is the desired point with $n = k - l$.

Let $V \subset U$ be the set of points that never return to U, that is $V = U - \cup_{n\geq 1} T^nU$. Then V is measurable. For any $n > 0, T^nV \cap V = \emptyset$ – otherwise a point of V will return to V, and therefore

to U. Hence the sets $V, TV, T^2V, ...$ do not intersect each other and, as before, one concludes that the volume of V equals zero. *Q.E.D.*

One can strengthen the theorem by proving that almost all points of U return to it infinitely many times. One can also change the formulation of the theorem, replacing a manifold by a set with a finite measure, U by a measurable subset of nonzero measure and T – by a measure-preserving transformation.

Poincaré's theorem has numerous applications to classical mechanics and to billiards, in particular. As an example, consider a point (x, v) in the phase space of a compact billiard M, that is, a position x of the billiard ball on the boundary ∂M and its inward unit velocity v. Then for any positive ϵ there exists an ϵ-close position $y \in \partial M$ and an ϵ-close unit velocity u such that the ball (y, u) will eventually reflect in the boundary at a point ϵ-close to x in the direction ϵ-close to v.

In the next section we will use the following corollary of the Poincaré's theorem in its strengthened form.

Corollary 2. *Let T be a measure-preserving transformation of a space M with a finite measure. Given a positive measurable function f on M, for almost all points $x \in M$ (in the sense of measure)*

$$\sum_{k=1}^{\infty} f(T^k x) = \infty.$$

Proof. Let M_n be the set of points where $f(x) \geq 1/n$. Then M_n is measurable and, according to the Poincaré's theorem, almost all points of M_n return to it infinitely many times. For such points the sum $\sum_{k=1}^{\infty} f(T^k x)$ is clearly infinite. The result follows from the fact that M is the union of its subsets M_n. *Q.E.D.*

1.7 Billiard Transformation Revisited: Measure-Theoretic View-Point

The aim of this section is to show that the billiard transformation is well defined off a set of zero measure in the phase space. Let M be a compact billiard table with a piecewise smooth boundary; let $N_1, ..., N_k$ be smooth components of the boundary, and assume that they have transversal pairwise intersections. Let V be the set of unit tangent vectors (x, v) with the footpoints x at the smooth part of the boundary and the inward directions v. The measure μ in V is the one induced by the symplectic structure, which is equivalent to the product of the Riemannian measure on the boundary and the Lebesgue measure on the unit sphere of vectors v in the tangent space $T_x M$.

The billiard transformation and the billiard flow for all times are defined at (x, v) if neither of the two "bad" things happen: the footpoint of $T^k(x, v)$ belongs to one of the intersections $N_i \cap N_j$, or the billiard ball (x, v) makes an infinite number of reflections in the boundary on a finite time interval. Since each intersection $N_i \cap N_j$ has a positive codimension in the boundary ∂M, its measure equals zero. It follows that the set of points (x, v) for which the first possibility occurs is of zero measure.

Consider the second possibility; let Q be the set of such points in V.

Lemma. $\mu(Q) = 0$.

Proof. Define a positive function f on V. Given a point (x, v), let $f(x, v)$ be the length of the geodesic segment through x in the direction of v until its first intersection with the boundary. Apply Corollary 1.6.2 to conclude that for almost all points (x, v)

$$\sum_{k=1}^{\infty} f(T^k x) = \infty.$$

If a point (x, v) belongs to Q then the sums $\sum_{k=1}^{\infty} f(T^k x)$ stay bounded as $n \to \infty$. Hence $\mu(Q) = 0$. *Q.E.D.*

Thus the billiard flow is defined on a subset of full measure in the phase space.

The second of the above mentioned possibilities, namely a trajectory that makes an infinite number of reflections on a finite time interval, can occur even for a plane strictly convex billiard whose boundary is three times differentiable (but this third derivative is unbounded). Such an example was constructed by B.Halpern. His construction involves rather meticulous estimates; its idea is, however, quite transparent.

Consider a sequence of points p_n of the unit circle that monotonically converges to a point of the circle (in the example the angular coordinate of p_n is equal to $n^{-1/2}$). The points p_n are the consecutive points of reflection of a billiard trajectory in the billiard curve γ to be constructed. Join the points p_n to obtain a polygonal trajectory. The law of reflection determines the direction of γ at the points p_n. One constructs a small portion γ_n of γ through each point p_n and then connects γ_n in a smooth way to obtain γ. We omit the details; the interested reader is referred to [Ha]. It is worth mentioning that, in a sense, the Halpern's example is the best possible. Namely, Halpern proved the following theorem: if a billiard curve has a bounded third derivative and nowhere vanishing curvature, then the billiard flow is defined for all times.

1.8 Complete Integrability and the Arnold–Liouville Theorem

The billiard dynamics may vary from very regular to extremely chaotic. Here we discuss the most regular type of dynamical behavior: complete integrability, or integrability in the sense of Liouville. The considerations of this section are usually applied to continuous time dynamical systems; we will stick to a discrete time case.

Definition. A symplectomorphism T of a symplectic manifold (M^{2n}, ω) is called completely integrable if there exist T-invariant smooth functions $f_1, ..., f_n$ (integrals) whose pair-wise Poisson brackets vanish and that are functionally independent almost everywhere on M (it means that their differentials are linearly independent in the complement of a set of zero measure).

Consider a nondegenerate level set P of the functions $f_1, ..., f_n$. It is an n-dimensional manifold, whose tangent space is generated by the symplectic gradients of the integrals $f_1, ..., f_n$. Since $\{f_i, f_j\} = \omega(sgrad\ f_i, sgrad\ f_j) = 0$ it follows that P is a Lagrangian manifold. Such manifolds constitute a Lagrangian foliation (that is, a foliation whose leaves are Lagrangian manifolds) off

the set $d\ f_1 \wedge ... \wedge d\ f_n = 0$. The symplectomorphism T preserves the foliation leaf-wise. Sometimes the existence of such a foliation is taken as the definition of complete integrability.

Definition. An affine structure on a manifold is an atlas whose transition functions are affine transformations of the coordinate space (i.e., compositions of linear transformations and parallel translations).

Lemma 1. *A leaf of a Lagrangian foliation carries a canonical affine structure.*

Proof. Locally a foliation is a fibration; let $\pi : M \to Q$ be its local projection. Then smooth functions on the base Q lift to M and their symplectic gradients commute on M. If the differential of a function at a point $x \in Q$ vanishes then its symplectic gradient vanishes in the fiber $\pi^{-1}(x) = P$. Therefore a neighbourhood of each point of P is the local orbit of a locally free action of the additive group of the space $T^*_x Q$. Whence the affine structure. *Q.E.D.*

Said otherwise, the vector fields *sgrad* f_i define a locally free action of $\mathbf{R}^n$ on P.

The following result is a discrete version of the Arnold–Liouville theorem ([Ve 1, Ar 2]).

Theorem 2. *Let T be a completely integrable symplectomorphism. Then a compact connected component of a nondegenerate level set P of its integrals is diffeomorphic to an n-dimensional torus. In appropriate coordinates the restriction of T to this torus is a translation: $x \to x + a$.*

Proof. Since P is connected the action of the group $\mathbf{R}^n$, defined in the previous proof, is transitive. Hence P is a quotient space of $\mathbf{R}^n$ by a discrete cocompact subgroup, that is, a torus. Since T preserves the Lagrangian foliation leaf-wise, its restriction to a leaf preserves the affine structure therein. Since T commutes with the flows of the fields *sgrad* f_i, it is a translation. *Q.E.D.*

Thus the phase space of a completely integrable system is foliated by its invariant tori, and the transformation on each torus is simply a translation.

The following corollary will be useful in the sequel.

Corollary 3. *If two completely integrable symplectomorphisms T and T' have the same foliation on invariant tori, then they commute: $T\ T' = T'\ T$.*

Proof. Restricted to an invariant torus both maps are translations in the same affine structure, whose definition depends on the Lagrangian foliation only. Translations commute; hence the result. *Q.E.D.*

Let $\phi = (\phi_1, ..., \phi_n)$ be coordinates on an invariant torus P in which T is a translation. A neighbourhood of P in M is foliated by invariant tori, parametrized by the values of the integrals $f = (f_1, ..., f_n)$. Extend the coordinates ϕ to these tori in a continuous way so that T is a translation in these coordinates on each torus. The coordinates (f, ϕ) are not necessarily Darboux coordinates. It is possible, however, to choose a new set of variables $I = I_1, ..., I_n$, depending on f, so that $\omega = dI \wedge d\phi$. These coordinates are called the action-angle coordinates.

1.9 On KAM Theory

Integrable systems are very exceptional: only a handful of examples is known. However, many important systems are small perturbations of integrable ones. A classical example is the solar system. The total mass of the planets is about 0.1% of the mass of the sun; therefore, in the first approximation, one can neglect the gravitational forces between the planets and consider only their attraction to the sun. The result is an integrable system, in which the orbit of each planet is an ellipse with a focus at the sun.

The methods of classical perturbation theory, developed in celestial mechanics, are not quite satisfactory, because they lead to divergent series. The reason is that in computations one has to divide by "small denominators" – integer linear combinations of the frequencies of the unperturbed motion. If such a combination is small, the corresponding term in the perturbation series becomes too big.

A breakthrough in this problem was made by A. Kolmogorov in 1954. He announced a theorem, that described the behaviour of Hamiltonian systems close to integrable, the theorem, later proved and generalized by V. Arnold and J. Moser. The whole theory is known as the Kolmogorov-Arnold-Moser (KAM) theory. We refer to [Ar 2, A-K-N, Mo 3,4, Mo-S, Pos, Bo] for a detailed discussion of the present state of the art in this rich area.

We know from the previous section, that the motion in an integrable (discrete time) system reduces to a rotation on each its invariant torus. Let $\alpha(I) = (\alpha_1, ..., \alpha_n)$ be the translation vector as a function of the action variable I. Generically, the function $\alpha(I)$ is nondegenerate, that is, $det(\frac{\partial \alpha}{\partial I}) \neq 0$. In the nondegenerate case the frequencies $\alpha_1, ..., \alpha_n$ can be chosen as parameters, enumerating the invariant tori. These frequencies can be either linearly independent or linearly dependent over rational numbers. The former case is generic, in a sense: the corresponding tori constitute a set of full measure in the phase space. However, the latter case is unavoidable too: the corresponding tori are dense in the phase space. The situation here is analogous to the coexistence of rational and irrational numbers.

For the sake of an illustration, consider a two-dimensional phase space, foliated by invariant circles. The transformation of each circle is a rotation. Consider a circle rotated through a π-rational angle. A certain iteration of the transformation leaves each point of this circle fixed. As Poincaré already noticed it is highly nontypical for an area-preserving map to have a whole circle of fixed points. Thus one expects such circles to be destroyed by a small perturbation of an integrable map. Likewise, the tori with linearly dependent frequencies most frequently vanish under a small perturbation of the transformation.

However, if the frequencies are independent, and, moreover, their linear combinations are badly approximated by rational numbers in a certain precise sense, then the corresponding tori survive a small perturbation of a map in the class of symplectomorphisms (of course, the tori get perturbed slightly). More precisely, consider an analytic symplectomorphism of an $2n$-dimensional symplectic "annulus" $\mathbf{T}^n \times \mathbf{D}^n$, exact with respect to the 1-form $Id\phi$ and close to an integrable

one:

$$(\phi, I) \to (\phi + \alpha(I) + \epsilon\ \beta(\phi, I, \epsilon),\ I + \epsilon\ \gamma(\phi, I, \epsilon)); \quad \phi \in \mathbf{T}^n, I \in \mathbf{D}^n.$$

Let the unperturbed map be nondegenerate: $det(\frac{\partial \alpha}{\partial I}) \neq 0$.

Theorem. *If the functions β and γ are C^r-smooth with $r > 2n + 1$, and ϵ is sufficiently small, then there exist invariant tori in the "annulus", close to the unperturbed tori $I = const$. The measure of the complement of the union of these invariant tori goes to zero as $\epsilon \to 0$. The motion on an invariant torus is conjugate to a translation.*

This result found in [Do 1,2] is an example of a theorem of KAM type. Many other versions can be found in the literature, in particular, concerning Hamiltonian vector fields and Hamiltonian systems in a neighbourhood of a fixed point.

In the two-dimensional case invariant circles separate the phase space. Thus each orbit is bound to stay between two such circles and cannot escape to infinity. If the dimension is greater, then invariant tori do not separate the space anymore, and orbits can escape to infinity (a specific example of such a process has been found by V. Arnold; it is known now as Arnold's diffusion). This diffusion is very slow – its speed is of order $exp(-1/\epsilon^{const})$, where ϵ is the parameter of the perturbation. This result is due to N. Nekhoroshev; it is one of the finest results of classical perturbation theory.

1.10 Stochastic Properties of Dynamical Systems

Completely integrable systems exhibit an extremely regular behavior. In this section we consider an opposite extreme: the maps that are, to a certain extent, chaotic. Our point of view is that of the ergodic theory, that studies maps of spaces with a measure. References are the books [C-F-S] and [Si 1].

Let M be a set with a σ-algebra of subsets and μ – a measure on this algebra. Assume that the measure is normed: $\mu(M) = 1$, and complete, that is, all subsets of measure zero sets are measurable. A dynamical system on M is a measure-preserving transformation T of M; this means that for any measurable set A one has: $\mu(A) = \mu(T^{-1}A)$. An example to keep in mind is, naturally, the billiard transformation.

The following definition is fundamental in ergodic theory.

Definition. A dynamical system (M, T) is called ergodic if the measure of any T-invariant set equals either 0 or 1.

In an ergodic system any T-invariant measurable function f on M is a constant off a set of zero measure. Indeed, the set $\{x :\ f(x) < a\}$ is measurable and T-invariant, hence its measure is equal to 0 or 1. Mention that a completely integrable system is not ergodic: it has an abundance of invariant sets consisting of invariant tori.

The following result is a consequence of the Birkhoff-Khinchin ergodic theorem for an ergodic dynamical system (M, T), that strengthens Corollary 1.6.2.

Theorem. *For almost all points $x \in M$ the time average of a measurable function f equals its space average:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) = \int_M f \, d\mu.$$

If f is the indicator of a measurable set (it equals 1 in the set and 0 outside of it), then the left-hand side of the above formula is the average number of times the orbit of the point x visits this set. Thus, in an ergodic system, the trajectory of almost every point spends the time in a measurable set asymptotically proportional to its measure. Moreover, any two measurable sets are statistically independent in the sense that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu(T^k A \cap B) = \mu(A)\mu(B).$$

Another stochastic property of dynamical systems, stronger than ergodicity, is mixing.

Definition. A system (M, T) is called mixing if, for any two functions $f, g \in L^2(M, \mu)$ the following equality holds:

$$\lim_{n\to\infty} \int_M f(T^n x) \; g(x) \; d\mu = \int_M f(x) \; d\mu \; \int_M g(x) \; d\mu.$$

Again if f and g are the indicators of two sets A and B then

$$\lim_{n\to\infty} \mu(T^{-n} A \cap B) = \mu(A) \; \mu(B).$$

It means that the iterations of the map uniformly mix a measurable set in the space: after a number of iteration it becomes virtually impossible to distinguish the points that originally belonged to a given set.

Assume that the set M is a compact topological space, the algebra of measurable sets consists of Borel sets (the minimal σ-algebra that contains open and closed sets) and the transformation T is a homeomorphism.

Definitions. A homeomorphism T is called uniquely ergodic if there exists a unique normed Borel T-invariant measure on M. A homeomorphism T is called minimal if the trajectory $\{T^n x, -\infty < n < \infty\}$ of every point x is dense in M. A homeomorphism is called topologically transitive if there exists a point whose trajectory is dense.

A uniquely ergodic homeomorphism is ergodic with respect to its unique invariant measure μ. Indeed if A is an invariant set with $\mu(A) \neq 0, 1$ then one can define a new invariant normed measure: $\mu'(B) = \mu(A \cap B)/\mu(A)$.

Clearly a minimal homeomorphism is topologically transitive. The properties of being minimal and uniquely ergodic are independent, although they often appear simultaneously in examples of dynamical systems. Minimal systems are, in a sense, similar to ergodic ones: they do not have nontrivial invariant closed subsets. If a dynamical system in a compact metric space is uniquely

ergodic, then the convergence of the time average to the space average in the ergodic theorem holds uniformly for any continuous function and for all (instead of only for almost all) points of the space.

Finally we mention a statistical property stronger than mixing.

Definition A measure preserving transformation T has K-property if for any measurable sets $A_0, A_1, ..., A_k$

$$\lim_{n\to\infty} \; sup \; |\mu(A_0 \cap B) \; - \; \mu(A_0)\mu(B)| = 0,$$

where supremum is taken over sets B from the σ-algebra generated by the sets $T^j(A_i)$ with $j \geq n$ and $i = 1, ..., k$.

Roughly speaking, this property means that the "present" is independent of the "past". "K" in "K-property" stands for Kolmogorov.

1.11 Entropy

The notion of entropy was introduced into physics by Clausius in the middle of the 19-th century, and about a century later it became a mathematical concept, first in information theory, due to Shannon, and then in ergodic theory, due to Kolmogorov. We discuss some very basic facts concerning entropy, referring to [Wa; Si 1,; Pet; D-G-S] for a detailed account.

For a physicist the entropy of a system is a quantity proportional to the logarithm of the relative probability of its state. To motivate the definition, consider a simple model. Suppose a certain system consists of N identical particles, each of which can be at one of k different states. The number of ways to achieve the state in which exactly N_i particles are in the i-th state is equal to $\frac{N!}{N_1!...N_k!}$. This is proportional to the probability of this state. Using Stirling's approximation $\ln n! \sim n \ln n - n$ this quantity approximately equals $N \ln N - \sum N_i \ln N_i$. Introducing the probability $p_i = N_i/N$ of a particle to be in the i-th state, one gets for the probability per particle the following quantity:

$$\ln N - \sum p_i \ln p_i - \ln N \sum p_i = - \sum p_i \ln p_i.$$

Given a set M with a probabilistic measure μ and its finite or countable partition on measurable sets $\xi = \{A_1, A_2, ...\}$, define the entropy of the partition

$$H(\xi) = - \sum \mu(A_i) \ln \mu(A_i).$$

For a measure-preserving transformation T of M, define the partition $T^{-n}\xi$ as $\{T^{-n}A_1, T^{-n}A_2, ...\}$. Entropy of the transformation T with respect to the partition ξ is

$$h(\xi, T) = \lim_{n\to\infty} \frac{1}{n} H(\xi \vee T^{-1}\xi \vee ... \vee T^{-n+1}\xi),$$

where $\xi \vee \eta \vee ...$ denotes the partition into the pairwise intersections of the elements of the partitions $\xi, \eta, ...$ (this definition incorporates the lemma saying that the limit exists). The number $h(\xi, T)$ is a measure of the average uncertainty per unit time about which element of the partition ξ the point

x will enter next, given its preceding history. A bad choice of ξ might prove not very informative, though. One defines metric entropy of the dynamical system (M, T) as $h(T) = \sup h(\xi, T)$, taken over all finite or countable partitions ξ with finite entropy.

Metric entropy is an invariant of dynamical systems, that had made it possible to distinguish between the systems undistinguishable by the previously known invariants (such as the spectrum). However the above definition can be rarely used for computations. The following theorem by A. Kolmogorov and Ya. Sinai makes the computations possible. A finite or countable partition ξ is called a generator if $\vee_{-\infty}^{\infty} T^i(\xi)$ generates the σ-algebra of measurable sets.

Theorem 1. *If ξ is a generator then $h(T) = h(\xi, T)$.*

A useful corollary is that if there exists a one-sided generator ξ, i.e., $\vee_{i=0}^{\infty} T^{-i}(\xi)$ generates the σ-algebra of measurable sets, then the metric entropy of the transformation T vanishes. This agrees with the one's intuitution: the existence of such a partition means that the present and future of the system is completely determined by its past. We mention two properties of entropy. First, $h(T^n) = |n|h(T)$ for an integer n; secondly, entropy of the product of two dynamical systems $(M_1 \times M_2, T_1 \times T_2)$ is equal to $h(T_1) + h(T_2)$.

The next concept of the topological entropy of a transformation is independent of an invariant measure, unlike the metric entropy. Given a compact Hausdorff topological space M and its open cover $\mathcal{U}$ define $N(\mathcal{U})$ to be the minimal cardinality of the subcovers of $\mathcal{U}$. For two covers $\mathcal{U}$ and $\mathcal{V}$ let $\mathcal{U} \vee \mathcal{V}$ be the cover, consisting of the pairwise intersections of their elements. Given a homeomorphism T of M define

$$h_{top}(\mathcal{U}, T) = \lim_{n\to\infty} \frac{1}{n} \ln N(\mathcal{U} \vee T^{-1}\mathcal{U} \vee ... \vee T^{-n+1}\mathcal{U}),$$

and $h_{top}(T) = \sup h_{top}(\mathcal{U}, \mathcal{T})$ over all open covers of M.

The following result is known as the variational principle for topological entropy.

Theorem 2. *Let T be a homeomorphism of a compact metric space M. Then $h_{top}(T) = \sup h(T)$, where the supremum of metric entropy is taken over all T-invariant Borel probability measures on M.*

In particular, topological entropy is an upper bound for the metric one.

For the later use we mention shift mappings. Consider a finite set $\{1, ..., k\}$ and let X be the set of sequences (x_i), $-\infty < i < \infty$ with $x_i \in \{1, ..., k\}$. Put the discrete topology on the finite set and the product topology on X. Then X is a compact metrizable space. Let S be the shift transformation: $S(x_n) = y_n$ where $y_n = x_{n+1}$ for all n. The shift is a homeomorphism of X whose topological entropy equals $\ln k$. If $Y \subset X$ is a closed shift invariant subspace then the dynamical system (Y, S) is called a subshift. Its topological entropy is equal to $\lim_{n\to\infty} \frac{1}{n} \ln N_n$, where N_n is the number of distinct words of length n in Y.

A Bernoulli shift is the shift of the space M of sequences $\{(x_n)\}$, $n \in \mathbf{Z}$, $x_n \in A$ where A is a probabilistic space and M has the product-measure. A Bernoulli shift is measure-preserving;

the above introtuced full shift corresponds to $A = \{1, ..., k\}$ and the uniform measure on A. Any Bernoulli shift is ergodic and mixing.

2. Convex Billiards

The main object of study in this chapter are smooth convex billiards in the plane. We start with an elementary study of the billiard in ellipses; we use nothing but high-school mathematics here. These results are applied in the next section to some geometrical problems: "the most elementary theorem of Euclidean geometry", which we prove in a non-elementary way, the problem of constructing a trap for a beam of light, and the illumination problem. Section 3 concerns the classical results by Jacobi on integrability of the geodesic flow in ellipsoids with applications to billiards in ellipsoids bounded by confocal quadrics. Section 4 deals with Birkhoff's conjecfture that the only integrable plane billiards are the ones in ellipses; "the mirror equation" appears there to play an important role in what follows. Section 5 gives a variational proof of the existence of Birkhoff's periodic trajectories of the billiard ball, the topic revisited in the next section from the point of view of symplectic geometry, and again in Section 10 in the framework of the theory of monotone twist maps. Section 7 concerns the relation between the set of length of periodic trajectories of the billiard ball and the spectrum of the Laplace operator in the billiard table with the Dirichlet boundary condition. Sections 8 and 9 discuss the existence and nonexistence of invariant circles of the billiard transformation: they exist if the billiard table is sufficiently smooth and strictly convex, and do not exist if it has a flat point. The methods used are those of the KAM theory and the theory of twist maps. The last section contains some isolated but nice results on smooth convex billiards not directly related to other topics of the chapter.

2.1 An Excursion to Elementary Geometry. Billiards in Conics

The simplest billiard table is a circular one. There is not much to say about it: each trajectory makes a constant angle with the boundary and remains tangent to a concentric circle. The induced transformation on this tangent circle is a rotation through a fixed angle, that is, a translation. Translations of a circle deserve some attention; we delay their study until Section 3.1.

fig. 9

Before we proceed any further we introduce a new concept.

Definition. A caustic of a plane billiard is a curve such that if a trajectory is tangent to it, then it again becomes tangent to it after every reflection.

Thus the billiard in a circle has a family of caustics, consisting of concentric circles.

The next case to consider is that of conics. Recall that an ellipse consists of points whose sum of distances to two given points is fixed; these two points are called the foci of an ellipse. An ellipse can be constructed using a string, whose ends are fixed at the foci – the method carpenters and gardeners actually use. A hyperbola is defined similarly with the sum of distances replaced by the absolute value of their difference; and a parabola is the set of points at equal distances from a given point (the focus) and a given line (the directrice). Ellipses, hyperbolas and parabolas all have second order equations in Cartesian coordinates.

fig. 10

The first result is the following optical property of ellipses.

Lemma 1. *A ray of light, emanating from one focus, comes to another focus after a reflection in the ellipse. Said otherwise, the segments, that join a point of an ellipse with its foci, make equal angles with the ellipse.*

Proof. Consider an extremal problem: given a line l and two points F_1 and F_2 on one side of it, find a point X on l such that the distance $F_1X + XF_2$ is minimal. Solution: reflect F_1 in the line and join with F_2 by a straight segment. The point of intersection with l is X. It follows that the angles made by F_1X and F_2X with l are equal.

fig. 11

On the other hand, X can be obtained as follows. Consider the family of ellipses with the fixed foci F_1 and F_2. Then X is the point where an ellipse from this family touches l for the first time. Hence X is the point of tangency of an ellipse with the foci F_1 and F_2 and the line l. *Q.E.D.*

To this we add that the billiard trajectory through the foci converges to the major axis of an ellipse. Indeed, consider consecutive segments of the trajectory. The point A_2 is closer to

the major axis than A_1, etc. Hence there exists the limit A_∞ of the sequence $A_1, A_2, ...$; and, likewise, the limit B_∞ of the sequence $B_1, B_2, ...$ The segment $A_\infty B_\infty$ is itself a "to and fro" billiard trajectory: the billiard ball goes from A_∞ to B_∞ and back ad infinitum. Thus the segment $A_\infty B_\infty$ is perpendicular to the ellipse at both ends, and therefore is the diameter MN.

fig. 12

Likewise one proves the optical properties of a hyperbola and a parabola, shown in figure 10. These properties are extensively used in construction of various optical instruments. For example, if one puts a source of light in the focus of a parabolic mirror, then the reflected rays form a parallel beam – the property used in headlights' design.

Ellipses and hyperbolas with the same foci are called confocal. In the appropriate Cartesian coordinates (x, y) they are given by the equation:

$$\frac{x^2}{a^2+\lambda} + \frac{y^2}{b^2+\lambda} = 1; \quad 0 < a < b.$$

Here λ is the variable parameter; for $-b^2 < \lambda < -a^2$ the curve is a hyperbola, and for $-a^2 < \lambda$ it is an ellipse.

The main result on elliptical billiards says that they are completely integrable.

Theorem 2. *An elliptic billiard table has a family of caustics, that consists of the confocal ellipses and hyperbolas. More precisely, if a segment of a billiard trajectory does not intersect the segment, joining foci F_1 and F_2, then all the segments of this trajectory do not intersect F_1F_2 and are all tangent to the same ellipse with foci F_1 and F_2; if a segment of a trajectory intersects F_1F_2, then all the segments of this trajectory intersect F_1F_2 and are all tangent to the same hyperbola with foci F_1 and F_2.*

Proof. Let A_0A_1 and A_1A_2 be consecutive segments of a trajectory. Assume that A_0A_1 does not intersect the segment F_1F_2 (the other case is dealt with similarly). It follows from the optical property that the angles $A_0A_1F_1$ and $A_2A_1F_2$ are equal.

fig. 13

Reflect F_1 in A_0A_1 to F_1', and F_2 in A_1A_2 to F_2', and set: $B = F_1'F_2 \cap A_0A_1, C = F_2'F_1 \cap A_1A_2$. Consider the ellipse with foci F_1 and F_2, that is tangent to A_0A_1. Since the angles F_2BA_1 and F_1BA_0 are equal, this ellipse touches A_0A_1 at the point B. Likewise an ellipse with foci F_1 and F_2 touches A_1A_2 at the point C. One wants to show that these two ellipses coincide, or, equivalently, that $F_1B + BF_2 = F_1C + CF_2$, which boils down to $F_1'F_2 = F_1F_2'$.

To this end one observes that the triangles $F_1'A_1F_2$ and $F_1A_1F_2'$ are congruent: $F_1'A_1 = F_1A_1, F_2A_1 = F_2'A_1$ by symmetry, and the angles $F_1'A_1F_2$ and $F_1A_1F_2'$ are equal. Hence $F_1'F_2 = F_1F_2'$, and the result follows. *Q.E.D.*

It might be instructive to compare this elementary proof with a more conceptual and general one we will give in the next section, concerning billiards in ellipsoids.

As an addition to the theorem we remark that billiards bounded by different confocal conics are integrable as well, because their caustics are still confocal conics.

fig. 14

To finish this section, consider the problem of reconstructing a billiard table Γ from its convex caustic γ. The following "well known" string construction produces a one-parameter family of billiard curves (see [Por] and [Tu]).

Lemma 3. *Wrap a closed inelastic string around γ, pull it tight at a point and move the point around γ to construct the billiard curve Γ.*

fig. 15

Proof. Recall that the involute of a curve γ is a curve, traced by the end of an inelastic (nonclosed) string, wrapped around γ. Fix a point of reference y on γ, and, given a point x outside of γ, draw the tangent segment xz to γ. Define a function $f(x)$ as the length of the arc yz plus the length of the segment zx. Then the involute is a level curve of $f(x)$. Since the string, that traces the involute, is inelastic, the tangential component of the velocity of the point x along zx vanishes. Hence the involute is perpendicular to the segment zx at each point x (see [B-G]). Therefore the gradient of $f(x)$ is the unit vector in the direction of xz.

Consider another involute, constructed by wrapping a string around γ in the opposite direction, and let $g(x)$ be the corresponding function in the exterior of γ. Then the curve Γ, described in the lemma, is a level curve of the function $f(x) + g(x)$. Its gradient at x is the sum of two unit vectors in the directions of the tangent lines to γ. Hence the gradient makes equal angles with these tangent lines, and therefore, so does Γ. This is the desired billiard property. *Q.E.D.*

In particular, wrapping a closed string around an ellipse produces a confocal ellipse – this is Graves theorem ([Be 1]).

2.2 Three Geometrical Applications

The first is "the most elementary theorem of Euclidean geometry", formulated by M. Urquhart (see [El]): if $AB + BF = AD + DF$, then $AC + CF = AE + EF$.

fig. 16

This theorem has synthetic proofs (the author knows some); however, we will deduce it from integrability of elliptic billiards. The theorem states that if B and D belong to an ellipse with foci

A and F, then C and E belong to a confocal ellipse.

Consider two confocal ellipses. The billiards therein have the same collection of caustics that consists of confocal ellipses. The family of rays, tangent to a caustic, is an invariant curve of the billiard transformation of the set of rays. Corollary 1.8.3 implies that the two billiard transformations commute.

fig. 17

Back to Urquhart's theorem. Construct the ellipses Γ_1 and Γ_2 with foci A and F, that contain the points B and C, respectively. The former ellipse contains D; one wants to show that the latter contains E. The billiards' reflection in Γ_2, and then in Γ_1 sends the ray AB to DA. Hence the reflection, first in Γ_1, and then in Γ_2 sends AB again to DA. This means that the points of intersection $BF \cap \Gamma_2$ and $AD \cap \Gamma_2$ coincide. Hence $E = BF \cap AD$ lies on Γ_2.

Urquhart's theorem has more or less straightforward generalizations (for instance, consider confocal hyperbolas instead of ellipses). We will not dwell on it.

The next application is a construction of a trap for a beam of light, that is, a reflecting curve such that a collection of parallel rays of light, shone into it, gets permanently trapped. The question, asked in [Con], was answered by several authors (see [Guy] and [Fr]). We describe the trap, constructed by R. Peirone ([Pei]).

fig. 18

The curve γ is a part of an ellipse with foci F_1 and F_2; the curve Γ is a parabola with focus F_2. These curves are joined in a smooth way to produce a trap: it follows from the optical properties of conics, that a vertical ray, entering the curve through a "window", will tend to the major axis of the ellipse and therefore never escapes.

It is tempting to change the problem slightly and to try to trap the set of rays sufficiently close to a given one in the space of rays (in other words, to allow the rays to make a small angle with a given one). However, in this case the trap does not exist, as readily follows from Poincaré's recurrence theorem. Indeed, let U be the neighbourhood in the space of rays, one wants to trap. Close the window of the trap and consider the billiard inside this closed curve. There exists a ray in U, that returns back to U after a number of reflections. The only way to return to U is to get reflected in the part of the curve that constitutes the window. Hence this ray is not trapped in the first place.

The third application concerns the illumination problem. Consider a room (a plane domain) with mirror walls; is it possible to illuminate it with a point source of light, that emits rays in all directions?

fig. 19

An example of a room, that cannot be illuminated from any of its points, is shown in the figure (see [C-F-G]). The upper and the lower curves are half-ellipses with foci F_1, F_2 and G_1, G_2. Since a ray, passing between the foci, reflects back again between the foci, no ray can enter the shaded area from the one between the lines F_1F_2 and G_1G_2, and vice versa. A modification of this construction yields a region that, for any positive n, requires at least n sources for illumination. Likewise one constructs a bounded region, whose boundary is smooth at all but one point, that cannot be illuminated by any finite number of sources. However, for an everywhere smooth bounded region a finite number of sources will always do ([Ra]).

2.3 Geodesic Flow and Billiards in Ellipsoids

The results of this section – integrability of the geodesic flow on an ellipsoid and of the billiard inside it – go back to the works of Chasles and Jacobi. In our exposition we follow [Mo 1,2, Ar 2]. For an approach via factorization of matrix polynomials see [Mo-V].

A quadric Q in a Euclidean space V with the scalar product $(\ ,\ )$ is determined by a self-adjoint operator $A = A^* : V \to V$ as follows:

$$Q = \{x \in V : \ \frac{1}{2}(Ax, x) = 1\}.$$

A quadric can be included in the following one-parameter families of quadrics.

Definition. An Euclidean pencil of quadrics is a family of the form:

$$\{x \in V : \ \frac{1}{2}(\ (A - \lambda E)\ x, x) = 1\};$$

a confocal family of quadrics is a family:

$$Q_\lambda = \{x \in V : \ \frac{1}{2}(\ (A - \lambda E)^{-1} x, x) = 1\},$$

where A is a self-adjoint operator and E is the unit operator.

The next figure shows an Euclidean pencil and a confocal family of conics.

fig. 20

Thus a confocal family consists of quadrics dual to the ones in an Euclidean pencil. One can easily verify that, in the plane, confocal quadrics have common foci. Unlike the plane case, in space there is no transparent geometrical description of confocal quadrics (however, see [H-CV] for a string construction due to O. Staude, the construction Hilbert considered one of the most remarkable results of 19-th century mathematics!).

Given a quadric, one introduces coordinates in space.

Definition. The elliptic coordinates of a point are the values of the parameter λ, for which quadrics, confocal with the given one, pass through this point.

The geometry of confocal quadrics is described by the following two theorems, of which the first one justifies the term "elliptic coordinates". Fix an ellipsoid

$$Q = \{x : \ \frac{1}{2}(A^{-1}\ x, x) = 1\}$$

in an Euclidean n-dimensional space, whose axes have pairwise distinct lengths.

Theorem 1 (Jacobi). *A generic point of space is contained in exactly n quadrics, confocal with the given ellipsoid. Two confocal quadrics are perpendicular at their intersection points.*

Proof. A non-zero vector in space determines an affine hyperplane in the dual space that consists of the covectors whose values at the given vector equal 1. The statement of the theorem, formulated in terms of the dual space, reads: *each affine hyperplane not through the origin is tangent to exactly n quadrics from a given Euclidean pencil; and the position vectors of the points of tangency are pairwise orthogonal.*

Let y be the original point in space; then the dual hyperplane is $\{x : \ (x, y) = 1\}$. Let the Euclidean pencil be $\frac{1}{2}(\ (A - \lambda E)\ x, x) = 1$. Introduce a new quadratic form $B(x) = (Ax, x) - 2(x, y)^2$, and consider a one-dimensional eigenspace L of this form with eigenvalue λ. Then the quadratic form $B - \lambda E$ vanishes on L together with its differential. Consider the point of intersection z of L with the hyperplane $(x, y) = 1$. Since $(B - \lambda E)\ z = 0$, the point z lies on the quadric $\frac{1}{2}(\ (A - \lambda E)\ x, x) = 1$; and since the differential vanishes as well, this quadric is tangent to the hyperplane at z.

Hence the points of tangency of the given hyperplane with quadrics of the Euclidean pencil $\frac{1}{2}(\ (A - \lambda E)\ x, x) = 1$ are eigenvectors of the quadratic form $B(x)$. There are exactly n such vectors and they are pairwise orthogonal. *Q.E.D.*

Theorem 2 (Chasles). *A generic line in n-dimensional Euclidean space is tangent to $(n-1)$ distinct quadrics from a given confocal family. The tangent hyperplanes to these quadrics at the points of tangency with the line are pairwise orthogonal.*

fig. 21

Proof. Project the space along the given line onto its $(n-1)$-dimensional orthogonal complement. A quadric determines a hypersurface in this $(n-1)$-space, the set of critical values of its projection (the apparent contour). If one knows that these hypersurfaces also constitute a confocal family of quadrics, the statement will follow from the previous theorem.

To prove that it is indeed the case, one applies duality. Duality interchanges projections and sections. The apparent contour of a quadric is transformed by duality to the section of the dual quadric by the hyperplane through the origin, orthogonal to the direction of the projection. A confocal family is dual to a Euclidean pencil. The sections of quadrics from an Euclidean pencil by a hyperplane constitute a Euclidean pencil in it. Applying duality again, it follows that the apparent contours of quadrics from a confocal family constitute a confocal family. *Q.E.D.*

The next theorem says that the geodesic flow on an ellipsoid is completely integrable.

Theorem 3 (Jacobi and Chasles). *The tangent lines to a fixed geodesic on a quadric in n-dimensional Euclidean space are tangent to $(n-2)$ other fixed quadrics, confocal with the given one. The set of oriented lines, tangent to $(n-1)$ fixed confocal quadrics, is a Lagrangian submanifold in the space of rays in space.*

Proof. The proof consists of several steps.

Step 1. Let l be an oriented line tangent to the given quadric Q_0 at the point x. By Theorem 2 it is tangent to $(n-2)$ confocal quadrics $Q_1, ..., Q_{n-2}$. Consider an infinitesimal rotation of the tangent line l along the geodesic on Q_0 through x in the direction of l. Modulo infinitesimals of the second order, this line rotates in the 2-plane generated by l and the normal vector n to Q_0 at x. By Theorem 2 the tangent hyperplane to the quadric Q_i, $i = 1, ..., n-2$, at the point of its tangency with l contains the vector n. Hence, modulo infinitesimals of the second order, the line l remains tangent to Q_i, and thus remains tangent to each one of them.

Step 2. This step consists of the following lemma. Let Q be a hypersurface in space. Consider the hypersurface of oriented lines, tangent to Q, in the symplectic manifold of oriented lines in space. *Then the characteristic lines of this hypersurface consist of rays, tangent to a fixed geodesic on Q.*

To prove this statement we use the Melrose diagram from Section 1.5:

$$\begin{array}{ccc} & X & \\ Y & & Z \\ & W & \\ U & & V \\ & \Sigma & \end{array}$$

As before, X is the (co)tangent bundle of the space; Y consists of unit vectors and Z – of vectors with the footpoints on Q; U is the space of oriented lines and V is the (co)tangent bundle of Q; and Σ consists of unit tangent vectors to Q.

First, consider Σ as a hypersurface in V, that consists of unit tangent vectors to Q. The characteristics of this hypersurface are identified with geodesics on Q. The same characteristic directions are obtained by restricting the symplectic structure of X to its codimension-3 submanifold Σ. And still the same characteristics Σ has as a hypersurface in U. This hypersurface consists of oriented lines in space, tangent to Q. The statement follows.

Step 3. Use the notation of Step 1. At a neighbourhood of the point of tangency of l with Q_i fix a smooth function f_i, whose level sets are the quadrics, confocal with Q_i. Any line l' close to l is tangent to a close confocal quadric Q_i'. Define a function F_i on oriented lines, whose value at l' is the value of f_i on Q_i'. To show that a common level set of these functions is a Lagrangian manifold, one needs to show that the functions F_i pairwise Poisson commute.

Fix a ray l, tangent to the quadrics Q_i, and compute the derivative of F_i along the symplectic gradient of F_j at the point l of the space of rays. The field $sgrad\ F_j$ is tangent to the characteristics

of the hypersurface $F_j = const$, that is, of the hypersurface, that consists of the lines, tangent to Q_j. These characteristics are known to consist of the rays, tangent to a fixed geodesic line on Q_j (Step 2). These rays are all tangent to Q_i (Step 1); hence the function F_i does not change along the flow of $sgrad\ F_j$. Therefore $\{F_i,\ F_j\} = 0$. *Q.E.D.*

Corollary 4. *A trajectory of the billiard inside an ellipsoid in n-dimensional space is tangent to $(n-1)$ confocal quadrics.*

Proof. As explained in Section 1.1, the billiard flow inside an ellipsoid in n-dimensional space is the limit case of the geodesic flow on an ellipsoid in $(n+1)$-dimensional space, whose minor axis goes to zero. The result follows from the previous theorem. *Q.E.D.*

Explicit formulas for the integrals of the billiard in an ellipsoid can be found in [Mo 2]. Let the ellipsoid be given by the equation $(A^{-1}\ x, x) = 1$, where A is a diagonal matrix with the diagonal elements A_i, and let (x, v) be a unit inward tangent vector, whose footpoint x lies on the ellipsoid. The following functions are invariant under the billiard transformation:

$$F_i\ (x, v) = v_i^2 + \sum_{j \neq i} \frac{(v_i x_j - v_j x_i)^2}{A_j - A_i};\quad i = 1, ..., n.$$

These functions Poisson commute.

Another corollary, one deduces from Theorem 3, concerns integrability of a billiard on a quadric in n-dimensional space, bounded by its intersections with confocal quadrics.

Corollary 5. *The tangent lines to a trajectory of such a billiard are all tangent to $(n-2)$ fixed confocal quadrics.*

Proof. Let γ and γ_1 be two consecutive segments of a billiard trajectory on a quadric Q, and let Q' be the confocal quadric, whose intersection with Q is the boundary of the billiard. Denote by l and l_1 the tangent lines to γ and γ_1 at the reflection point.

fig. 22

One knows that the tangent lines to the geodesic γ are tangent to $(n-2)$ quadrics, confocal with Q. In particular, so is the line l. Consider the billiard inside Q'. Since Q' is orthogonal to Q this billiard transformation sends l to l_1. By Corollary 4 the line l_1 is tangent to the same $(n-2)$ quadrics. Hence the tangent lines to the geodesic γ_1 are tangent to the same quadrics as well. *Q.E.D.*

We refer to [Ve 2, D-G-K-R, C-S] for a discussion on billiards on a Euclidean sphere. The paper [Ve 2] by Veselov also concerns billiards in hyperbolic spaces.

Let us mention that the lines of intersection of an ellipsoid in 3-space with confocal quadrics are its lines of curvature, i.e. the lines that have the eigendirections of the second quadratic form of the surface of the ellipsoid. Applying the string construction from Section 2.1, one concludes that these lines are analogous to confocal ellipses and hyperbolas in the the following sense: the

sum (or difference) of distances from all points of a line of curvature to two umbilic points of the ellipsoid is constant (umbilic points are the points where the principal curvatures are equal ; they are the singularities of the foliation by the lines of curvature) – see [H-CV].

fig. 23

In conclusion we remark that the geodesic flow on an ellipsoid was first integrated by Jacobi using elliptic coordinates to separate the variables in the equation of motion, which is now called the Hamilton-Jacobi equation.

2.4 Birkhoff's Conjecture

We know that the billiard inside an ellipse is integrable: its caustics are confocal ellipses and hyperbolas. The partition of the phase cylinder into invariant curves of the billiard transformation is shown in the figure (the ∞ -shaped curve corresponds to the rays through the foci). Notice the topological difference between the cases of an ellipse and a circle.

fig. 24

G. Birkhoff conjectured that ellipses are characterized by integrability of the billiard transformation.

Conjecture. *If a neighbourhood of a strictly convex smooth billiard curve is foliated by caustics, then the curve is an ellipse.*

Several attempts had been made to prove this conjecture, but, so far, it remains open. A partial result was obtained recently by M. Bialy [Bia]. In our exposition we combine the ideas of Bialy and M. Wojtkowski [Wo 1].

We start with two observations of independent interest. Let H be the function on the phase cylinder V, whose value at the point (x, v) is the length of the trajectory of the "ball" (x, v) until it hits the billiard curve. Recall that V has an invariant measure μ, introduced in Section 1.2.

fig. 25

The following result is well known in integral geometry.

Lemma 1. *Let A be the area of a convex billiard table. Then*

$$\int_V H \; d\mu = 2\pi A.$$

Proof. Recall (Section 1.3) that the set of oriented lines in the plane has a natural measure, associated with the form $d\, p \wedge d\, \phi$, where p is ($\pm$) the distance from the origin to the line and ϕ is

its angle with a fixed direction. This measure is equal to μ under the identification of the cylinder V with the set of lines U, that intersect the billiard curve. So

$$\int_V H\ d\mu = \int_U H\ dp\ d\phi = A\int_0^{2\pi} d\phi = 2\pi A;$$

the second equality is due to the fact, that for a fixed direction ϕ the integral $\int H\ dp$ is the area. *Q.E.D.*

The next result is the classical "mirror equation" of geometric optics (see, e.g., [Por]). Consider a caustic of a billiard and introduce the variables a, b, θ, as shown in the figure. Let K be the curvature of the billiard curve at the point x.

fig. 26

Lemma 2.

$$\frac{1}{a} + \frac{1}{b} = \frac{2K}{\sin\theta}$$

.

Proof. Replace the billiard curve by its circle of curvature at the point x, which has the second order tangency with the curve. Consider a nearby point x', seen from the center of the circle at the angle of ϵ. Draw the tangent segments to the caustic from x', and denote the angles xyx' and xzx' by α and β.

fig. 27

It follows from figure 27 that the beam of light, emanating from the point y, focuses at the point z after reflection in the curve.

The Sin theorem for the infinitesimal triangles xyx' and xzx' yields:

$$\frac{\sin\theta}{a} = \frac{\alpha K}{\epsilon}\quad,\quad \frac{\sin\theta}{b} = \frac{\beta K}{\epsilon}.$$

Hence

$$\frac{1}{a} + \frac{1}{b} = \frac{(\alpha+\beta)\ K}{\epsilon\ \sin\theta}.$$

It remains to show that $\alpha + \beta = 2\epsilon$. Consider the closed path $yxx'zxx'y$; being homotopic to a twice traversed circle, its total rotation equals 4π. This rotation consists of the exterior angles at the vertices and twice the turn of the arc xz', that is, 2ϵ. Thus

$$\theta + \psi + (\pi - \beta) + (\pi - \theta) + (\pi - \psi) + (\pi - \alpha) + 2\epsilon = 4\pi,$$

or: $\alpha + \beta = 2\epsilon$. *Q.E.D.*

Now we are in a position to prove a particular case of Birkhoff's conjecture.

Theorem 3. *If a billiard table is foliated by smooth closed convex caustics so that almost every trajectory is tangent to a caustic, then the table is a disc.*

Proof. Fix a caustic. Given a point x on the billiard curve, let v be the unit tangent vector in the direction of xz. Denote the billiard transformation by T. Lemma 2 reads:

$$\frac{2\sin\theta}{K(x)} = \frac{4b(x,v)\ (H(x,v) - b(T(x,v))}{b(x,v) + (H(x,v) - b(T(x,v))}.$$

In view of the inequality between the harmonic and the arithmetic mean the right-hand side is not greater than $H(x,v) + b(x,v) - b(T(x,v))$. Integrate both sides over the phase space. Since the measure is T-invariant,

$$\int_V (H(x,v) + b(x,v) - b(T(x,v))\ d\mu = \int_V H(x,v)\ d\mu = 2\pi\ Area.$$

Since $\mu = \sin\theta\ d\theta\ dl$, where l is the length parameter (Section 1.2), the integral of the other side equals

$$\int_0^L \int_0^\pi \frac{2\sin^2\theta}{K(l)}\ dl\ d\theta = \pi \int_0^L \frac{1}{K(l)}\ dl,$$

where L is the length of the billiard curve. By the Cauchy-Schwartz inequality

$$\int_0^L \frac{1}{K(l)}\ dl\ \int_0^L K(l)\ dl \geq L^2.$$

Since the second of these integrals equals 2π, one concludes:

$$\frac{L^2}{2} \leq 2\pi\ Area.$$

This inequality is opposite to the isoperimetric one, hence it is actually an equality, and the curve is a circle. *Q.E.D.*

Theorem 3 is slightly weaker than Bialy's result. His theorem assumes that the phase cylinder is foliated by continuous closed noncontractible invariant curves, with the same conclusion that the billiard table is a disc. Some additional work is needed to handle this continuous case; we will not dwell on it.

A result somewhat related to Birkhoff's conjecture was obtained by E. Amiran ([Am]). Consider strictly convex smooth integrable billiards with the following evolution property: if any caustic is taken as a new billiard curve, then this new billiard is also integrable and shares its caustics with the original one. This evolution property holds for elliptic billiards; it follows from Amiran's work that the converse is also true.

A different version of integrability is the analytic one: a piecewise smooth billiard is analytically integrable if there exists an invariant function, continuous in the position variable and analytic in the velocity variable. S. Bolotin [Bol] proved that such a billiard curve consists of pieces of straight lines and conics. The methods of the proof are those of algebraic geometry.

Notice that, although Birkhoff's conjecture is not proved yet, integrable billiards certainly constitute a "small" subset in the space of convex billiards. More precisely, it is a set of the first Baire category (a countable union of nowhere dense sets) in an appropriate topology in the set of analytic – see [K-T], and smooth convex curves – see [La 3].

2.5 Periodic Trajectories

Given a C^1 smooth strictly convex billiard curve, does the billiard map have periodic trajectories, that is, points in the phase space that return back after a number of iterations? These periodic trajectories are inscribed polygons, whose consecutive sides make equal angles with the billiard curve. We distinguish star-shaped n-gons with rotation r (i.e., the sum of the exterior angles at the vertices, divided by 2π). These polygons are obtained by choosing n points $x_1, ..., x_n$ on the curve in the clockwise order and connecting x_1 to x_{r+1} to x_{2r+1}, etc.

fig. 28

The answer to the above question is given by the following theorem by G. Birkhoff ([Bi 1]).

Theorem 1. *For any $n \geq 2$ and $r \leq n/2$, coprime to n, there exist two geometrically distinct n-periodic trajectories with the rotation r.*

Proof. Consider the set M of inscribed n-gons $x_1, ..., x_n$, where x_i is a point on the curve. M is an n-dimensional torus. Let H be the perimeter length function $|x_1x_{r+1}| + |x_{r+1}x_{2r+1}| + ...$ This function is smooth off the singular set $M_0 = \cup\{x_i = x_{i+1}\}$ (we adopt the convention that $n + 1 = 1$), and at a generic point of M_0 the function has the absolute value type singularity. We learned in Section 1.2 that critical points of the function H are billiard trajectories (this is Maupertuis' Principle from classical mechanics –see [Ar 2] – as applied to billiards).

First we prove the easy part: for any n there exists an n-periodic trajectory. M being compact, H attains its maximum on it. This maximum is not attained on M_0: the perimeter of a k-gon with $k < n$ can be increased by introducing a new vertex. Thus this maximum corresponds to a faithful n-gonal trajectory.

fig. 29

To prove the result in full, consider the set of inscribed star-shaped n-gons with rotation r. Abusing the notation, call it M. This set is the product of a circle and an $(n-1)$-dimensional disc, and its boundary consists of degenerate polygons from M_0. The function H has at least n maxima in M, obtained by a cyclic permutation of the vertices of geometrically the same polygon of the maximal perimeter in M.

Use the min-max (or the Buridan's Ass) argument to show, that there are other critical points of H inside M. Connect two maxima by a curve inside M and consider the minimum of H on it.

Take the maximum of these minima over all such curves. This is also a critical point of H, other than the maxima. A subtle point is to establish its existence. It follows from the fact, that one does not need to come close to the boundary M_0, since the function H increases as one moves from the boundary – see somewhat technical details in [C-S] or [K-T].

fig. 30

The reason to consider r coprime to n, is that, otherwise, the above argument might yield a trajectory which is a polygon with fewer vertices, traversed by the billiard ball several times. *Q.E.D.*

It follows, that there are at least $\phi(n)$ distinct n-periodic orbits, where $\phi(n)$ is the number of integers, less than n and coprime to n.

The following particular case may shed some light on the proof. Consider 2-periodic "to and fro" trajectories, that is, the chords of the billiard curve, perpendicular to it at both ends. One such curve is easily found: it is the diameter, i.e., the greatest chord of the curve. Since the billiard curve is strictly convex, there is a unique chord of maximal length in each family of parallel chords. Thus one has a one-parameter family of these longest chords, parametrized by their directions, whose initial element is the diameter, and whose terminal element is the same diameter with the end-points interchanged. In this family there exists a shortest chord, and this chord is perpendicular to the curve at both ends. This is the desired second 2-periodic billiard trajectory.

fig. 31

One naturally wonders whether there are any higher-dimensional generalizations. The situation with 2-periodic orbits is particularly clear: a convex smooth body in n-dimensional space has at least n chords, perpendicular to the boundary at both end-points (due to N. Kuiper. Proof: assign to a line in space the length of the orthogonal projection of the body on this line. This is a function on projective space whose critical points correspond to the desired chords). The case of trajectories with a greater number of links is much harder, and not much is known about it. Recently I. Babenko proved that a three-dimensional smooth strictly convex billiard has at least n distinct periodic trajectories for any prime number n ([Bab 1]). His approach is similar to the one in dimension two, and it involves the Morse theory of the length functional on the space of inscribed polygons. An obvious difficulty is the lack of a rotation number in this setting.

After the existence of periodic trajectories is established, one would like to learn about their stability. The billiard transformation T is a mapping of the phase annulus; its n-periodic orbits are fixed points of the map T^n.

Definition. A fixed point x of a map F is stable (in the sense of Lyapunov), if for any neighbourhood U of x there exists a neighbourhood V such that the orbit $\cup F^i\ (V),\ -\infty < i < \infty$,

is contained in U.

Given an n-periodic trajectory of the billiard ball, consider the corresponding fixed point of the mapping T^n. The derivative of T^n at a fixed point is an area-preserving linear transformation of the plane; hence its determinant equals 1. Generically two cases are possible: the eigenvalues are reciprocal distinct reals, or they are distinct conjugate complex numbers with the absolute values of 1. In the former case (called hyperbolic) the linear map is a hyperbolic rotation; the fixed point is not stable (even in the linear approximation). In the latter case (called elliptic) the linear map is a rotation, and the fixed point is stable in the linear approximation. It follows from KAM theory, that in the general position, there exist invariant curves of the map T^n near the fixed point. Thus an elliptic periodic trajectory is generically stable.

fig. 32

Without going into details we mention, that the maximal length Birkhoff's periodic trajectory is always unstable. The stability condition for a 2-periodic orbit reads as follows. Let l be the length of the segment of such a trajectory, and r_1, r_2 be the radii of curvature of the billiard curve at its end-points, which we allow to be negative as well.

Lemma 2. *A 2-periodic orbit is elliptic if and only if*

$$\frac{r_1 + r_2 - l}{r_1 r_2} > 0 \quad \textit{and} \quad \frac{(l - r_1)(l - r_2)}{r_1 r_2} > 0.$$

Accordingly, the major axis of an ellipse is an unstable, and the minor one is a stable 2-periodic trajectory of the billiard ball. We refer to [Wo 2] for the proof of the lemma.

Since a periodic trajectory is a critical point of the length functional, another approach to the stability problem is via the study of the Hessian matrix at this critical point. The eigenvalues of the linearization of the appropriate iteration of the billiard transfromation at its fixed point can be expressed in terms of the Hessian – see [Ma-M] and [K-T] (and [Kl] for a similar question in the study of closed geodesics in Riemannian manifolds). An application to 2-periodic trajectories is given in [Bar].

In conclusion of this section we mention the works by Ph. Levallois and M. Tabanov ([L-T, Lev]) and V. Donnay ([Don 1]) on non-integrability of billiards obtained by small perturbations of an ellipse. The general mechanism responsible for this non-integrability had been discovered by H. Poincaré in his study of the three body problem; it is called the splitting of separatrices (see, e.g., [Ar 2]).

In an elliptic billiard the major axis is a hyperbolic 2-periodic trajectory. The set of rays through one or another focus constitute a curve, invariant under the billiard transformation, that connects the two points of this 2-periodic trajectory (this curve is called a separatrix) – see figure 24. Any point of the separatrix goes to one of the two points (B in figure 33) under positive iterations of the billiard map, and to another one (point A) under negative ones.

fig. 33

Make a small perturbation of the billiard curve. The 2-periodic hyperbolic trajectory will persist (it will still be the diameter of the curve), however the above constructed curves for the new points A and B will not necessarily coinside. More precisely, let S_A be the curve consisting of points x with $T^{-n}x \to A$ as $n \to \infty$, and $S_B = \{x \mid T^n x \to B, \quad n \to \infty\}$. If the curves S_A and S_B intersect once, they will intersect infinitely many times because both are invariant under the billiard map. The two curves constitute a complicated web, whose existence does not agree with integrability.

Levallois and Tabanov analyzed the splitting of separatrices in the case when the perturbed billiard table is the following curve of degree four:

$$x = a\cos\alpha, \quad y = b\sin\alpha(1 + \epsilon\cos^2\alpha),$$

ϵ being a small parameter. They proved that for all sufficiently small ϵ the separatrices intersect transversally, and estimated the angles made by them at intersection points. V. Donnay obtained a similar result for a small perturbation of an ellipse in a neighbourhood of one point; this perturbation changes the curvature, but preserves the point and the tangent direction at it.

2.6 Periodic Trajectories, Poincaré's Last Geometric Theorem and Symplectic Topology

Another approach to the existence of periodic trajectories consists in using a theorem on fixed points of area-preserving maps of an annulus. In this section we describe this approach and put it into the broader perspective of symplectic topology.

Recall that the phase space of the billiard transformation T inside a smooth strictly convex region is an annulus with the coordinates (t, α), where t is the length parameter along the billiard curve and α is the angle made by the velocity vector with the boundary. Assume for convenience that the length of the billiard curve is 2π. The two boundary circles of the phase annulus are $\alpha = 0$ and $\alpha = \pi$; these curves are pointwise preserved by the transformation T.

Fix a point x on the billiard curve and consider the unit tangent vectors, whose footpoint is x. This set is the vertical segment, shown in figure 34. The image of this segment under T is a curve, that makes one complete turn around the annulus. One can lift T to a map $\tilde{T}$ of the universal covering of the annulus, which is an infinite strip. The choice of $\tilde{T}$ is made by requesting that it pointwise fixes the lower boundary $\alpha = 0$. Then the upper boundary is translated by $\tilde{T}$ through 2π. In down-to-earth terms, one says that T fixes one boundary and rotates another one in the positive direction through 2π.

fig. 34

Let R be the rotation of the annulus in the negative direction through 2π (that is, a negative

translation of the covering). Then n-periodic orbits of T with rotation number r are fixed points of the map $T^n R^{-r}$. Notice that this map translates the boundary circles in the opposite directions. The existence of two fixed points follows from Poincaré's Last Geometrical Theorem (H. Poincaré announced it shortly before death; the proof was given by G. Birkhoff in 1917).

Theorem. *An area-preserving transformation of an annulus, that moves the boundary circles in the opposite directions, has at least two distinct fixed points.*

Modern proofs of this theorem use the methods of symplectic topology we will discuss later in this section. We chose to give here the elegant topological proof by Birkhoff ([Bi 1]), which, from the modern point of view, might seem somewhat of an anachronism. More precisely, we show that the transformation in question has fixed points (this is the most nontrivial part of the argument).

Proof. To fix ideas, let the universal covering of the annulus be the strip $0 \leq y \leq 1$, and let $\tilde{T}$ move the lower boundary to the right and the upper one – to the left. Assume that $\tilde{T}$ has no fixed points, and let ϵ be smaller than the distance between any point and its image under $\tilde{T}$ (it exists due to compactness of the annulus). Denote by S_ϵ the vertical translation of the plane, in which the strip lies, through ϵ, and set: $\tilde{T}\ S_\epsilon = \tilde{T}_\epsilon$.

Consider the strip $0 \leq y \leq \epsilon$. Its images have equal areas, and their interiors are disjoint. Hence there is an image of this narrow strip that intersects the upper boundary $y = 1$; let k be the least number of iterations needed. Let P_k be the point of the upper boundary of the k-th image of the ϵ-strip with the greatest y-coordinate, and let $P_0 = \tilde{T}_\epsilon^{-k}\ (P_k)$ be the corresponding point of the line $y = 0$. Set $\tilde{T}_\epsilon^i(P_0) = P_i$. Join P_0 and P_1 by a segment and consider its consecutive images under $\tilde{T}_\epsilon$. They constitute a simple arc from P_0 to P_k.

fig. 35

For a point Q of the strip consider the vector $v(Q) = (Q, \tilde{T}_\epsilon\ Q)$. If Q is on the lower boundary, its direction is close to that of the positive x-ray, and for Q on the upper one its direction is almost opposite. Let Q move along the above constructed arc from P_0 to P_{k-1}. Then, for a sufficiently small ϵ, the rotation of $v(Q)$ is the same as that of the tangent vector to the arc (one adds the external angles at the corners), that is, almost equals π. Therefore the rotation of $v(Q)$ almost equals π for any continuous motion of Q from the lower to the upper boundary. This conclusion "survives" the limit $\epsilon \to 0$, in which case one gets rid of the word "almost".

Consider now the inverse map $\tilde{T}^{-1}$, that moves the lower boundary to the left and the upper – to the right. By entirely analogous argument, the rotation of the vector $u(Q) = (Q, \tilde{T}_\epsilon^{-1}\ Q)$ equals $(-\pi)$ as Q moves from the line $y = 0$ to $y = 1$. But $u(Q) = -v(Q)$, so the rotations of these vectors are equal. The contradiction means that $\tilde{T}$ has a fixed point. *Q.E.D.*

The theorem of symplectic topology, that generalizes (and implies) Poincaré's theorem, reads as follows. Consider a transformation T of a two-dimensional torus: $T(x, y) = (x + f(x, y), y + g(x, y))$, that preserves the area form $dx \wedge dy$, and assume that it also preserves the center of gravity:

the average values of f and g are equal to zero. Then T has at least 4 fixed points, counted with the multiplicities, and at least 3 of them are geometrically distinct. Given an annulus transformation, as in Poincaré's theorem, one constructs a torus transformation, pasting together two copies of the annulus along two narrow strips (see [Ar 3]).

In the case of an $2n$-dimensional standard symplectic torus the numbers of fixed points of a symplectomorphism, that preserves the center of gravity, are equal to 4^n (counted with the multiplicities) and $2n+1$ (geometrically distinct). This theorem by Conley and Zehnder ([C-Z]), published in 1983 and conjectured, among other things, by V. Arnold in the 60-s, was a breakthrough result in symplectic topology. A more general conjecture states that a symplectomorphism of a compact symplectic manifold, symplectically homological to the identity (that is, the time-1 transformation of a Hamiltonian vector field with a time-dependent Hamiltonian function), has at least as many fixed points as the least number of critical points of a smooth function on this manifold. This conjecture is proved in many cases, in particular, for surfaces and toric varieties.

Symplectic topology is a very exciting subject, but a more detailed discussion would lead one too far away from billiards. The interested reader may consult the surveys [Ar 1,4; Gro; Sik; Ben].

2.7 Length Spectrum and Laplace Operator

The existence of periodic trajectories being established, what can be said about their lengths? This is a counterpart to the question one asks in differential geometry concerning the lengths of closed geodesic lines in a Riemannian manifold; and indeed, there is a strong similarity between the results in both areas. We will outline some results on the set of lengths of the closed trajectories in a convex billiard, the set that is called the length spectrum of a billiard.

To start with, let us explain the symplectic meaning of the length spectrum, following V. Guillemin and R. Melrose ([G-M]). Let T be an exact symplectomorphism of an exact symplectic manifold (M, ω). This means that $\omega = d\alpha$ for a 1-form α, and that the closed 1-form $T^*\alpha - \alpha$ is exact, i.e., equal to dF for a generating function F. Let x be an n-periodic point: $T^n x = x$. Assign to this point the number

$$L(x) = \sum_{i=1}^{n} F(T^i x).$$

This number depends on the choice of the generating function, determined up to an additive constant. One normalizes F in a natural way: say, by requiring that its average vanish, or that F vanish on a boundary component (in the case of billiards). Notice that $L(x)$ depends on the periodic orbit, but not on a particular choice of the point: $L(x) = L(Tx)$. Finally $L(x)$ is independent of a change of α by a coboundary: if $\alpha' = \alpha + d\phi$, then $F' = F + \phi \circ T - \phi$ and

$$L'(x) = \sum_i F(T^i x) + \phi(T^{i+1}x) - \phi(T^i x) = \sum F(T^i x) = L(x).$$

In particular, $L(x)$ is a symplectic invariant of a periodic orbit for a simply connected symplectic manifold.

In the billiard setting the symplectic phase manifold is the unit disc subbundle of the (co)tangent bundle of the boundary of the billiard table, and the generating function is (negative) the length of the trajectory segment between two consecutive reflections. This function vanishes on the boundary of the phase space, that consists of tangent vectors to the billiard hypersurface. Thus the symplectic invariant of a periodic orbit specializes to the length of a closed billiard trajectory.

S. Marvizi and R. Melrose ([M-Me]) studied the asymptotics of the lengths of n-periodic billiard trajectories in a smooth strictly convex plane domain as $n \to \infty$. Let L_n be the supremum and l_n – the infimum of the perimeters of simple billiard n-gons.

Theorem 1.

$$\lim_{n\to\infty} n^k \ (L_n - l_n) = 0$$

for any positive k. Moreover, L_n has an asymptotic expansion as $n \to \infty$:

$$L_n \sim l + \sum_{i=1}^{\infty} \frac{c_i}{n^{2i}},$$

where l is the length of the billiard table and c_i are constants, depending on the curvature of the table.

Y. Collin de Verdière ([Col]) similarly studied periodic trajectories near a stable 2-periodic billiard orbit.

The idea of the proof is to deal with the billiard transformation T as if it were integrable. More specifically, there exists a smooth function f in a vicinity of the boundary of the phase annulus, such that $T^{-1} \circ exp\ sgrad\ (f^{3/2})$ is a smooth map, that fixes this boundary to infinite order. In particular, the Taylor expansion of f is uniquely determined along the boundary. The appearance of the exponent $3/2$ is due to the fact, that T has a square root type singularity at the boundary. For the integrable map $exp\ sgrad\ (f^{3/2})$ one expresses the length spectrum in terms of the Taylor coefficients of the function f, which depends on the geometry of the billiard curve, and the result follows.

In the case, considered by Collin de Verdière, one deals with a vicinity of an elliptic fixed point of an area-preserving map. If the derivative of the map is a rotation through a π-irrational angle, then the map is approximated to infinite order by an integrable map: a rotation through a variable angle, depending on the radius (Birkhoff's normal form – see [Ar 2, Mo-S]). After this the proof proceeds along similar lines.

A remarkable relation exists between the length spectrum of a convex billiard and the spectrum of the Laplace operator in it with Dirichlet boundary condition:

$$\Delta f = \lambda f \ \text{ in } M, \quad f|_{\partial M} = 0.$$

From the physical point of view, the eigenvalues λ are the eigenfrequencies of the membrane M with a fixed boundary.

Roughly speaking, one can recover the length spectrum from that of the Laplacian. More precisely, the following theorem of K. Anderson and R. Melrose ([A-M]) holds.

Theorem 2. *The sum*

$$\sum_{\lambda_i \in spec\ \Delta} \cos\left(t\sqrt{-\lambda_i}\ \right)$$

is a well-defined generalized function (distribution) of t, smooth away from the length spectrum. That is, if $l > 0$ belongs to the singular support of this distribution, then there exists either a closed billiard trajectory of length l, or a closed geodesic of length l in the boundary of the billiard table.

A relation between the Laplacian and the length spectrum is, of course, not a coincidence. Geometric optics is not a very accurate description of light. In wave optics light is considered as electromagnetic waves, and geometric optics gives a realistic approximation only when the wave length is small. This small-wave appproximation is based on the assumption that the waves are locally almost harmonic, while their amplitudes change slowly from point to point. The substitution of such a function into the corresponding PDE's gives, in the first approximation, the equations of wave fronts, that is, of geometric optics (see [G-S 1,2]). Precise results in this direction are not easy to prove, and even the formulations are rather technical. See, for instance, the works of Lazutkin ([La 2,3]), in which approximations of the Laplacian's eigenvalues are constructed from caustics of the corresponding billiard.

2.8 Existence of Caustics

An important property, that makes it possible to apply the KAM theory to the billiard transformation, is that it is close to an integrable map near the boundary of the phase cylinder. Let $T(t,\alpha) = (t_1, \alpha_1)$. One can express the values of t_1 and α_1 in terms of t, α, and the curvature of the billiard curve. We omit these straightforward, but rather cumbersome, computations, and state the answer ([La 1]):

$$t_1 = t + 2\rho(t)\alpha + \frac{4}{3}\rho(t)\rho'(t)\alpha^2 + f(\alpha,t)\alpha^3,$$

$$\alpha_1 = \alpha - \frac{2}{3}\rho'(t)\alpha^2 + g(\alpha,t)\alpha^3.$$

Here $\rho(t)$ is the radius of curvature of the strictly convex billiard curve, f and g are some smooth functions. Using ideas from the geometrical optics Lazutkin found a smooth change of variables $x(t,\alpha), y(t,\alpha)$ with $y(t,0) = 0$ that reduces the billiard map to a simple form:

$$x_1 = x + y + f(x,y)y^3, \quad y_1 = y + g(x,y)y^4.$$

It follows that near the boundary curve $y = 0$ the map is a small perturbation of the integrable shear map $(x,y) \to (x+y, y)$. It is also area-preserving, hence the KAM theory is applicable. One arrives to the following result by Lazutkin ([La 1]).

Theorem 1. *If the billiard curve is sufficiently smooth and its curvature never vanishes, then there exists a collection of smooth caustics in a vicinity of the billiard curve, whose union has a positive area.*

Originally this theorem asked for 553 continuous derivatives of the billiard curve; later R. Douady using results of Russman and Herman reduced the number to 6 ([Do 1,2]).

Notice that a smooth closed caustic in a vicinity of the billiard curve corresponds to an invariant circle of the billiard transformation in the phase cylinder, homotopic to its boundary cirle. This invariant circle separates the cylinder into two invariant subsets of positive measure. Therefore strictly convex smooth billiards are not ergodic.

Let us mention that caustics do not have to be convex curves. For example, consider a curve of constant width. It has a chord in any direction, perpendicular to the curve at both ends. These chords are 2-periodic billiard trajectories, and their envelope is a caustic. This caustic is the involute of the billiard curve (see [B-G]); it necessarily has cusps, that correspond to the critical points of the curvature of the billiard curve. This example is considered in [Gu-K 1]. Another example from the same paper is the caustic, given by the equation $x^{2/3} + y^{2/3} = 1$, called the astroid. Using a modification of the string construction, one recovers a one-parameter family of convex billiard tables from this caustic.

fig. 36

The situation with caustics of higher-dimensional billiards is very different. Given a smooth strictly convex billiard table in space, its caustic is a hypersurface that enjoys the same property: a ray tangent to it, remains tangent after the reflection in the billiard hypersurface. An ellipsoid has a collection of caustics, the confocal ellipsoids. M. Berger proved the converse. Suppose that a billiard hypersurface has a caustic. Then the collection of rays through a point of the hypersurface, tangent to the caustic, is a symmetric cone, whose axis is perpendicular to the billiard hypersurface. M. Berger's result reads (see [Be 2,3]):

Theorem 2. *If M is an open C^2- smooth hypersurface and N is an open hypesurface such that for each point $x \in M$ there exists a hypercone with the vertex at x, tangent to N, symmetric with respect to its axis, which is perpendicular to M, then M is a part of a hyperquadric and N is a part of a confocal quadric.*

Notice that this result is local: the existence of a piece of a caustic already proves to be a very rigid condition, in sharp contrast with the plane case. Also notice that this theorem does not imply the higher-dimensional version of Birkhoff's conjecture: only billiards in ellipsoids are integrable. Indeed, if the billiard transformation of the space of rays has an invariant hypersurface, this hypersurface does not necessarily consist of rays tangent to some hypersurface in the configuration space. The proof of Berger's theorem is computational and makes use of the classical differential geometry. Its key ingredient is the mirror equation from Lemma 2.4.2.

To this we add that, according to [Gru], most convex billiards (in the sense of Baire's category; the topology in the set of convex sets being determined by the Hausdorff metric) have no convex caustics.

2.9 Twist Maps, Birkhoff's Theorem and Nonexistence of Caustics

Consider an area-preserving diffeomorphism T of a cylinder, that preserves the boundary components.

Definition. T is called a twist map if each vertical tangent vector is turned in the same sense by its differential (that is, either all vectors are turned to the right, or all vectors are turned to the left).

In other words, the image of every vertical line has everywhere positive (negative) deviation from the vertical. Let $(\overline{x}, y)$ be coordinates in the cylinder $S^1 \times \mathbf{R}^1$, and lift the map to the universal covering of the cylinder, that is to the plane with coordinates (x, y). Denote by p the projection on the the first factor, and let the map be $f(x, y)$. Then the twist condition reads as follows: for each x the maps $y \to p \circ f(x, y)$ and $y \to p \circ f^{-1}(x, y)$ are diffeomorhisms of $\mathbf{R}^1$. It is also convenient to include into the definition the existence of a uniform estimate for the deviation from the vertical direction, that is for the derivatives $\partial(p \circ f^{\pm 1}(x, y))/\partial y$.

The billiard transformation for a smooth convex billiard acts in the cylinder with the coordinates (t, α); t is the length parameter along the billiard curve as measured from some fixed point, and $\alpha \in [0, \pi]$ is the angle, made by the billiard trajectory with the positive direction of the billiard curve. The following remark is of fundamental importance.

Lemma 1. *The billiard transformation is a twist map.*

Proof. Fix a vertical line $t = const$ in the phase cylinder. This line corresponds to a fixed position of the billiard ball on the billiard curve with all possible directions α allowed. Clearly, the greater this angle α, the greater is the arc from the fixed point to the point, at which the ball hits the boundary. Thus the image of the vertical line has a positive deviation from the vertical. *Q.E.D.*

fig. 37

We now make a digression to discuss Birkhoff's theorem on invariant curves of twist maps. There exist several versions of this theorem, that can be found in [Bi 2,3; He 1; Ma 1,2]. Some higher-dimensional generalizations are available too – [B-P]. In our exposition we follow [He 1] and [Ma 2].

Theorem 2. *Let T be an orientation and area-preserving diffeomorphism of an open cylinder $S^1 \times (-1, 1)$, that maps each topological end to itself. Let U be an open T-invariant subset, which contains the lower end $S_1 \times (-1, -1 + \epsilon)$ and whose closure does not intersect the upper*

end $S^1 \times (1-\epsilon, 1)$, $\epsilon > 0$ *being sufficiently small. Suppose also that the circle* $S^1 \times \{-1+\epsilon\}$ *is a deformation retract of* U. *Then the boundary of* U *is the graph of a Lipschitz continuous function* $S^1 \to (-1, 1)$.

We will outline the idea of a proof, leaving many details aside.

Proof. To fix ideas, assume that T is tilted to the right, that is, T turns vertical directions to the ones with a positive first component. Fix a circle $S = S^1 \times \{-1+\epsilon\}$, contained in U. Let U_+ be the set of positively accessible points, i.e., points $x \in U$ such that there exists an imbedded curve in U, that starts at a point of S, ends at x and has a positive deviation from the vertical. Similarly one defines the set of negatively accessible points U_-. Let $V = U_+ \cap U_-$. Then V is the set of points, vertically accessible from S.

fig. 38

Notice that $U_\pm$ are open. Because T is a positive twist map, it preserves the property of an imbedded curve to have a positive deviation from the vertical. Hence $T(\bar{U}_+ \cap U) \subset U_+$, and, likewise, $T^{-1}(\bar{U}_- \cap U) \subset U_-$. Since T is area-preserving, there exists T- invariant subsets of full measure in U_+ and U_- (take $\cap_i\, T^i(U_+)$). The same property is enjoyed by V.

Suppose that the invariant curve (the boundary of U) is not a graph. Then V has vertical segments on its boundary. Consider such a segment and suppose that V locally lies to the left of it. The image of this segment is a curve with a positive deviation from the vertical, and the image of V lies above it. Pick a point x of the T-invariant subset of V sufficiently close to the vertical segment; then $T(x)$ belongs to V, and therefore is vertically accessible from S. This is a contradiction. If V lies to the right of the vertical segment, the same argument works with T being replaced by T^{-1}.

fig. 39

Let us indicate an alternative approach: one shows that $U_+ = \cup_{i \geq 0} T^i V$ and $U_- = \cup_{i \leq 0}\, T^i V$; since T is area-preserving, it follows that $U_- - \bar{V}$ and $U_+ - \bar{V}$ are empty (see [LeC]).

To see that the invariant curve is the graph of a Lipschitz function, use the following trick. Given a C^1-differentiable function f on $(-1, 1)$, consider the diffeomorphism of the cylinder:

$$F_f(x, y) = (x + f(y),\ y); \quad x \in S^1,\ y \in (-1, 1).$$

Let $T_f = F_f \circ T \circ F_f^{-1}$. Then $F_f U$ is invariant under T_f. If f is sufficiently close to zero in the Whitney C^1 topology, then T_f is also a twist map. Hence the boundary of $F_f U$ is also the graph of a continuous function. This holds for all functions f in a small neighbourhood of zero; hence the boundary of U satisfies the Lipschitz condition. *Q.E.D.*

As an application of Birkhoff's theorem consider results by J. Mather on noexistence of invariant curves of a billiard transformation ([Ma 1]). These results are in sharp contrast with those

of Lazutkin.

First, we formulate the twist property of the billiard map analytically. As before, $H(t,t')$ denotes the Euclidean distance between two points on the billiard curve whose length parameters are t and t'. The following statement simply means that the curve is convex.

Lemma 3.

$$\frac{\partial^2 H(t,t')}{\partial t\ \partial t'} > 0$$

Proof. Consider the segment tt' of a billiard trajectory, and let α and α' be the corresponding angle variables. Then $\frac{\partial H(t,t')}{\partial t} = -\cos\alpha$ (Section 1.2). Differentiate with respect to t':

$$\frac{\partial^2 H(t,t')}{\partial t\ \partial t'} = \sin\alpha\ \frac{\partial \alpha}{\partial t'}.$$

Since $\sin\alpha > 0$, one only needs to show that $\frac{\partial\alpha}{\partial t'} > 0$.

Consider the twist property. The billiard transformation $(t,\alpha) \to (t',\alpha')$ turns vertical vectors to the right: $\frac{\partial t'}{\partial \alpha} > 0$. Hence $\frac{\partial\alpha}{\partial t'} > 0$. *Q.E.D.*

fig. 40

We now proceed to Mather's results.

Theorem 4. *If the curvature of a C^2 smooth convex billiard curve vanishes at some point, then the billiard transformation has no invariant circles.*

Proof. Let t_{-1}, t_0, t_1 be the first coordinates of three consecutive points of a billiard trajectory. One of them can be viewed as a function of the other two.

Let $\bar{t}_0$ be the coordinate of a point with zero curvature, and consider a two-link trajectory that reflect at this point. A straightforward computation, which we omit, shows that

$$\frac{\partial^2 H(\bar{t}_{-1},\bar{t}_0)}{\partial \bar{t}_0{}^2} + \frac{\partial^2 H(\bar{t}_0,\bar{t}_1)}{\partial \bar{t}_0{}^2} > 0$$

(the reason is that the curvature vanishes at $\bar{t}_0$: for the sake of the computation, one replaces the curve by its tangent line at $\bar{t}_0$, which can be done because the curve has second order contact with the tangent at this point).

One knows from Section 1.2 that three consecutive points are related by the equation:

$$\frac{\partial H(t_{-1},t_0)}{\partial t_0} + \frac{\partial H(t_0,t_1)}{\partial t_0} = 0.$$

Considering t_0 as a function of t_{-1} and t_1, implicit differentiation yields:

$$\frac{\partial t_0}{\partial t_{-1}} = -\frac{\partial^2 H(t_{-1},t_0)}{\partial t_{-1}\partial t_0} \Bigg/ \left(\frac{\partial^2 H(t_{-1},t_0)}{\partial t_0{}^2} + \frac{\partial^2 H(t_0,t_1)}{\partial t_0{}^2} \right),$$

$$\frac{\partial t_0}{\partial t_1} = -\frac{\partial^2 H(t_0,t_1)}{\partial t_0 \partial t_1} \Big/ \left(\frac{\partial^2 H(t_{-1},t_0)}{\partial {t_0}^2} + \frac{\partial^2 H(t_0,t_1)}{\partial {t_0}^2} \right)$$

(since the invariant circle is the graph of a Lipshitz function the derivatives exist almost everywhere).

In view of the above inequality, the denominators are positive whenever (t_{-1}, t_1) is sufficiently close to $(\bar{t}_{-1}, \bar{t}_1)$. By Lemma 3 the numerators are positive as well. Thus

$$\frac{\partial t_0}{\partial t_{-1}} < 0, \quad \frac{\partial t_0}{\partial t_1} < 0$$

for (t_{-1}, t_1) sufficiently close to $(\bar{t}_{-1}, \bar{t}_1)$.

Assume now that an invariant curve exists. By Birkhoff's theorem it is a graph of a continuous function. If the three consecutive points whose first coordinates are t_{-1}, t_0, t_1, are taken on this curve, it follows that t_0 increases as a function of t_{-1}, as well as of t_1. If, in addition, t_0 is sufficiently close to $\bar{t}_0$, then the above inequalities for the partial derivatives hold, and they imply that t_0 decreases as a function of t_{-1}, as well as of t_1. This is a contradiction. *Q.E.D.*

It follows that a convex billiard with vanishing curvature at some point does not have caustics inside it. This statement can be proved more easily: the mirror formula from Section 2.4, applied to a point with zero curvature, implies that the sum of tangent segments from this point to a caustic is zero, which is absurd. This observation is due to M. Wojtkowski ([Wo 2]).

Mather's theorem has a somewhat paradoxical consequence. Call a billiard trajectory ϵ-glancing if, for some bounce, the angle of reflection is smaller than ϵ. One distinguishes positive and negative ϵ-glancing, according to whether the angle with the positive or negative direction of the billiard curve is taken into account.

Corollary 5. *If the curvature of a smooth convex billiard curve vanishes at some point then, for every $\epsilon > 0$, there exists a trajectory that is both positively and negatively ϵ-glancing.*

fig. 41

Proof. *Let V be the phase cylinder $S^1 \times [-\pi, \pi]$ of the billiard transformation T. Given a positive ϵ, consider the ϵ-strips around the boundary circles:*

$$V_- = S^1 \times (-\pi, -\pi + \epsilon), \quad V_+ = S^1 \times (\pi - \epsilon, \pi),$$

and set:

$$W = \cup_{-\infty}^{\infty} T^n(V_-) \cup (S^1 \times \{-\pi\}).$$

Assume that the statement is false. Then for some ϵ the intersection of W and V_+ is void. Consider the connected component of $V - W$, that contains the upper boundary circle of the cylinder, and let U be the complement of this component. Then U is an invariant domain of

the billiard transformation, satisfying the assumptions of Birkhoff's theorem. Its boundary is an invariant circle, which contradicts Theorem 4. *Q.E.D.*

In fact, even a stronger result holds: under the same assumptions there exists a billiard trajectory, such the angle of reflection in the oriented billiard curve tends to π, as "time" $n \to +\infty$, and tends to 0, as $n \to -\infty$. We refer to [Ma 3] for strong results on chaotic dynamics in the so-called instability zones of twist maps.

The above results on nonexistence of caustics were strengthened in a recent preprint by E. Gutkin and A. Katok [Gu-K 1]. For example, the following estimate holds. Denote by A, L, D the area, the perimeter length and the diameter of a convex billiard table, and let k_{min} and k_{max} be the minimal and the maximal values of its curvature.

Theorem 6. *If $\sqrt{2}D^2 k_{min} k_{max} \leq 1$, then the billiard table contains a convex region, free of convex caustics, whose area is not less than $A - \sqrt{2}k_{min} L D^2$.*

Let us also mention the results by A. Hubacher ([Hu]): if the curvature of a convex piecewise C^2-smooth billiard curve is discontinuous at some point, then the billiard does not have caustics in a neighbourhood of the boundary.

2.10 Aubry-Mather Theory

Numerous results, concerning smooth convex billiards (periodic orbits, existence and nonexistence of invariant curves), hold in a more general situation, namely, for area-preserving twist maps. In this section we discuss some of the developments in this field, known as the Aubry-Mather theory. We refer to [LeC, Mo 5, Che, Ban, Ka 1] for surveys of this topic.

Let T be an area-preserving positive twist C^1 diffeomorphism of an annulus $S^1 \times I$ with the area form $dx \wedge dy$. Let $T(x_1, y_1) = (x_2, y_2)$. Then $y_2 dx_2 - y_1 dx_1 = dH$, the function H being a C^2 smooth generating function of the map T. Fix x_1 and vary y_1. The twist property implies that y_1 is a function of x_2. Hence the generating function can be considered as a function of x_1 and x_2, and

$$y_1 = -\frac{\partial H}{\partial x_1}, \; y_2 = \frac{\partial H}{\partial x_2}.$$

The twist condition reads:

$$\frac{\partial x_2}{\partial y_1} > 0, \quad \text{or} \quad \frac{\partial^2 H(x_1, x_2)}{\partial x_1 \partial x_2} < 0.$$

For convex billiards, x is the length coordinate and H is (negative) the distance between the corresponding points.

To proceed further, recall some facts about orientation preserving homeomorphisms of a circle, that go back to Poincaré and Denjoy (see, e.g., [Ni, He 2]). Let $f : S^1 \to S^1$ be such a homeomorphism ($S^1 = \mathbf{R}/\mathbf{Z}$), and let $\tilde{f} : \mathbf{R}^1 \to \mathbf{R}^1$ be a lifting. Then there exists a real number ρ, well defined modulo integers, such that

$$-1 < \tilde{f}^n(x) - x - n\rho < 1$$

for all real x and integer n. The class of ρ modulo integers is independent of the choice of the lifting and this ρ is called the rotation number of the homeomorphism f.

If $\rho = p/q$ is rational, p and $q > 0$ being coprime, there exists $x \in \mathbf{R}$ such that $\tilde{f}^q(x) = x + p$. The projection of this x to the circle has a q-periodic orbit under f, referred to as a (p, q)-type orbit (once the lifting is chosen). If ρ is irrational and f is twice continuously differentiable, then f is conjugate to a rotation by a homeomorphism (the Denjoy theorem), and every orbit is dense in the circle. However, there exists a C^1 smooth map f none of whose orbits are dense; the orbits have a common limit set which is an invariant Cantor set in S^1 (Denjoy minimal set).

Now we can formulate a result on periodic orbits of area-preserving twist maps. Let $\alpha < \beta$ be the rotation numbers of the restriction of the map T to the lower and upper boundary circles of the cylinder.

Theorem 1. *For any rational number $p/q \in [\alpha, \beta]$ (p and q coprime, $q > 0$) there exist two (p, q)-type periodic orbits of T, such that the first coordinates of the points of these orbits (i.e., points in S^1) are pairwise distinct. Moreover, the orbits have the following monotonicity property: if z_1 and z_2 are liftings of two points from the union of these two orbits to the universal covering $\mathbf{R} \times I$, and z_2 lies to the right of z_1, then $\tilde{T}(z_2)$ lies to the right of $\tilde{T}(z_1)$ ($\tilde{T}$ being a lifting of T).*

Specializing to billiards, one recovers two star-shaped q-periodic trajectories with rotation number p. The proof makes use of the variational principle: the orbits in question are critical points of the "length" functional

$$H(x_0, x_1) + H(x_1, x_2) + ... + H(x_{q-1}, x_q),$$

defined on sequences of real numbers (x_i), $i \in \mathbf{Z}$ with $x_{i+q} = x_i + p$.

Assume that a twist map is close to an integrable map $(x, y) \to (x + \phi(y), y)$ near the boundary circles of the annulus. By KAM theory the map has invariant curves near the boundary. These invariant curves are graphs of continuous functions. Parametrize an invariant curve by two continuous functions: $x = u(t)$, $y = v(t)$, where u is increasing, and $u(t) - t$ and $v(t)$ are 1-periodic. The dynamics on an invariant circle, provided by the KAM theory, is conjugate to the rotation through an irrational number badly approximated by the rationals. Let ρ be such a number for the invariant curve under consideration. Then $T\ (u(t), v(t)) = (u(t + \rho), v(t + \rho))$. One can exclude the second function v, using the partial derivatives of the generating function:

$$H_1(u(t), u(t + \rho)) + H_2(u(t - \rho), u(t)) = 0,$$

where the indices denote the partial derivatives with respect to the first and the second variables. Solutions to this equation, that is increasing continuous functions $u(t)$, with $u(t) - t$ being 1-periodic, determine invariant curves with rotation number ρ.

Dropping the continuity requirement on $u(t)$, one arrives to the notion of the Mather set with the rotation number ρ, given parametrically by $x = u(t), y = -H_1(u(t), u(t + \rho))$ where $u(t)$ is a monotone function, satisfying the above equation, and such that $u(t + 1) = u(t) + 1$.

Theorem 2. *For any real number $\rho \in [\alpha, \beta]$, there exists an invariant Mather set with the rotation number ρ. This set lies on the graph of a Lipschitz continuous function.*

Mather's original proof is based on the study of critical points of the functional $\int_0^1 H(u(t), u(t+\rho))\, dt$, whose Euler equation defines the Mather set.

Mather sets generalize both the periodic trajectories from the previous theorem ($\rho = p/q$; u is piecewise constant) and invariant curves (u is continuous). If ρ is irrational, u may be discontinuous. In this case the Mather set is a Cantor set lying on a graph of a Lipschitz continuous function, whose gaps correspond to the discontinuities of u. The graph itself is not uniquely determined and, by no means an invariant curve. These Cantor sets are in a sense the remnants of the invariant curves.

The Aubry-Mather theory explains how the invariant curves provided by KAM theory disintegrate as the parameter ϵ of a perturbation of an integrable map increases. Fixing a rotation number badly appriximated by the rationals, one considers the corresponding Mather invariant sets. For small values of ϵ they are invariant curves, but for greater ϵ they may become Cantor sets. Let ϵ^* be the greatest value of ϵ, such that for all $\epsilon \leq \epsilon^*$ the invariant sets are curves. Then to the value ϵ^* of the perturbation parameter corresponds an invariant curve which disintegrates under an arbitrary small increase of the parameter. This curve is necessarily not too smooth, because, otherwise, by KAM theory it would "survive" an increase in ϵ. Thus invariant curves break up through loss of smoothness.

We mention in conclusion, that besides an interest in the area-preserving twist maps per se, a motivation for the peresent theory came from the solid state physics, namely from the study by Aubry of the Frenkel-Kontorova model. Moreover, similar results in differential geometry go back to G. Hedlund , who studied minimal geodesics in two-dimensional tori, that is, the geodesics, that globally minimize the distance between any pair of their points.

2.11 Miscellanea

We collect some more results on smooth convex billiards in this section.

First, following E. Gutkin ([Gu 1]), consider a billiard map, which admits a horizontal circle $\alpha = const$ as an invariant curve (in the usual coordinates in the phase cylinder). The corresponding trajectories of the billiard ball make a constant angle with the billiard curve. An obvious example of such a situation is provided by a disc. A more interesting example is a billiard curve of constant width. For any direction such a curve has a chord in this direction, that is perpendicular to the curve at both end-points. This chord is a 2-periodic trajectory, and their totality is the invariant curve $\alpha = \pi/2$. What are other possible values of the constant angle α?

fig. 42

Theorem 1. *There exists a non-circular billiard table with an invariant curve $\alpha = const$,* $0 <$

$\alpha < \pi/2$, if and only if this constant angle α satisfies the equation $\tan n\alpha = n \tan \alpha$ *for some integer $n > 1$.*

We outline the proof, leaving aside some computations.

Proof. Parametrize the billiard curve by the angle ϕ it makes with a fixed direction. Let $r(\phi)$ be the radius of curvature of the curve. If $(x(\phi), y(\phi))$ are the Cartesian coordinates of a point on the curve, then

$$x'(\phi) = r(\phi)\cos\phi, \quad y'(\phi) = r(\phi)\sin\phi.$$

fig. 43

Consider a segment of the trajectory, corresponding to an orbit of the billiard map on its horizontal invariant circle. It makes the angle α with the billiard curve. Hence

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

Differentiate twice and make the necessary substitutions to arrive at:

$$\cos\alpha\ (r(\phi+\alpha) - r(\phi-\alpha)) = \sin\alpha\ (r'(\phi+\alpha) + r'(\phi-\alpha)).$$

This equation is equivalent to the existence of the horizontal invariant circle at height α. Let

$$r(\phi) = r_0 + \sum_1^\infty\ a_n \cos n\phi + b_n \sin n\phi$$

be the Fourier decomposition of the radius of curvature. The previous equation implies:

$$a_n\ (\sin n\alpha\ \cos\alpha - n\cos n\alpha\ \sin\alpha) = b_n\ (\sin n\alpha\ \cos\alpha - n\cos n\alpha\ \sin\alpha) = 0$$

for all n.

If the billiard curve is not a circle, then, for some n, $a_n \neq 0$ or $b_n \neq 0$. Hence

$$\sin n\alpha\ \cos\alpha = n\ \cos n\alpha\ \sin\alpha.$$

This is the desired equation. *Q.E.D.*

The second result we want to discuss here concerns the following problem: how big can the set of n-periodic points of a billiard transformation be? In particular, can a billiard transformation have an (non-void) open set of periodic points? The motivation for this question comes from the theory of the Laplace operator (Section 2.7).

It is easy to answer this question in the real analytic category: the measure of the set of periodic orbits is zero. In the smooth case, the answer is readily found for $n = 2$: a 2-periodic trajectory is perpendicular to the billiard curve, and therefore lies on the horizontal circle $\alpha = \pi/2$. However, the problem becomes quite hard for greater values of n. The case of $n = 3$ was settled by M. Rychlik ([Ry]); his proof involved symbolic computations. Later the proof was drastically simplified by L. Stojanov ([Sto 1]) and M. Wojtkowski ([Wo 1]). We follow the later paper.

Theorem 2. *The set of 3-periodic points of the billiard transformation of a smooth convex plane billiard is nowhere dense.*

Proof. Assume that there is an open subset in the phase space that consists of 3-periodic trajectories. Then a 3-periodic trajectory can be included into a 1-parameter family of 3-periodic trajectories with one reflection point being fixed.

fig. 44

Denote the curvature at X_i by K_i, and set

$$C_i = \frac{\sin \alpha_i}{2K_i},\ i = 1, 2, 3.$$

Let H_i be the length of the corresponding trajectory segment. Consider two infinitely close rays X_1X_2 and $X_1'X_2'$. Let a_1 and a_2 be the (signed) distances from their intersection point Y to the points X_1 and X_2 (so, if Y lies to the left of X_2 on the line X_1X_2, then $a_2 < 0$, etc). Then $a_1 + a_2 = H$. Applying the mirror equation (Lemma 2.4.2) twice, one gets:

$$\frac{1}{H_1} + \frac{1}{a_2} = \frac{1}{C_2},\quad \frac{1}{H_2} + \frac{1}{a_1} = \frac{1}{C_1}.$$

It follows that

$$\frac{H_1}{H_1C_2 - 1} + \frac{H_2}{H_2C_1 - 1} = H_3.$$

Rewrite this equation as

$$(H_2 + H_3 - H_2H_3/C_1)\ (H_1 + H_3 - H_1H_3/C_2) = H_1H_2.$$

Two similar equations are obtained by a cyclic permutation of the indices. The system of three equations is easily solved: multiply all three, take the square root, and divide by one of them. One gets:

$$H_1 + H_2 - H_3 = H_1H_2/C_3 = 2H_1H_2K_3/\sin\alpha_3.$$

Now apply the Cosine Theorem to the triangle $X_1X_2X_3$. It boils down to

$$H_1 + H_2 - H_3 = \frac{4H_1H_2\sin^2\alpha_3}{H_1 + H_2 + H_3}.$$

Notice that a 3-periodic trajectory is an extremum of the perimeter length functional on inscribed triangles. This functional is constant on its critical set: $H_1 + H_2 + H_3 = L$ for some constant L and all nearby 3-periodic trajectories. Compare the above two formulas to conclude:

$$\frac{K_3}{\sin\alpha_3} = \frac{2\sin^2\alpha_3}{L}.$$

But this is clearly absurd, because one can change the angle α_3, keeping the pont X_3, and therefore, the curvature K_3, fixed. *Q.E.D.*

With a little more effort one shows that the set of 3-periodic orbits has the zero measure. We refer to the papers of Rychlik or Wojtkowski for this refinement.

Almost nothing has been said about higher-dimensional billiards. To make up with this shortcoming, consider a smooth convex billiard in 3-space that has the following property: each billiard trajectory belongs to a 2-plane. An example is provided by the billiard inside a ball; are there other billiards with this property? The answer was found by R. Sine ([Sin]).

Theorem 3. *Balls are the only billiard tables with the above property.*

Proof. Let M be a billiard surface. Given a point $x \in M$, denote by $l(x)$ the normal line to M at x. Pick a point x on M and consider a plane P through $l(x)$. Given a point y of the curve $P \cap M$, the billiard ball, shot from x to y, will stay in the plane P. It follows that $l(y) \subset P$. Let x' be the "antipodal" point of intersection of $l(x)$ with M. As y goes to x', one concludes that $l(x') \subset P$. This holds for any plane P through $l(x)$; thus $l(x') = l(x)$.

fig. 45

Consider the point $z \in M$ such that the line $l(z)$ is perpendicular to P. Let Q be the plane through $l(z)$ and y. Arguing as above, one concludes that $l(y) \subset Q$. Hence $l(y)$ passes through the point of intersection of $l(z)$ and P – the "center" of the section of the billiard table by the plane P. This conclusion holds for all points $y \in P \cap M$; hence this section is a circle and xx' is its diameter.

Finally, revolve the plane P about $l(x)$. Each section of M by this plane is a circle with the same diameter xx'. Thus the billiard surface is a sphere. *Q.E.D.*

The result holds in dimensions greater than 3 as well – see [S-K].

Let us mention in conclusion a recent work by Gutkin and Knill [Gu-Kn]. Given a convex curve δ consider the one-parameter family of billiards for which δ is a caustic; these billiards are obtained by the string construction. Let l be the variable string length and γ_l the corresponding billiard curve. The billiard transformation associated to γ_l induces a homeomorphism of δ; let $\rho(l)$ be its rotation number. Then $\rho(l)$ is a continuous nondecreasing function. One of the results of [Gu-Kn] reads: for a generic δ the function $\rho(l)$ is a "devil-staircase"; this means that for any rational q the interval $\rho^{-1}(q)$ has a non-empty interior.

3. Billiards in Polygons

This chapter concerns billiards in polygons and polyhedra, mostly convex. Section 1 starts with the method of unfolding billiard trajectories to be used throughout the chapter. The rest of the section concerns dynamical properties of a torus translation, to which the billiard transformation in a cube reduces. Section 2 deals with encoding billiard trajectories with irrational slopes in a cube according to the faces they reflect in. Such encoding sequences are quasi-periodic; their complexity is found in the two and three dimensional cases. Section 3 starts with a criterion for stability of periodic billiard trajectories in polygons; this means that a trajectory "survives" a perturbation of the billiard table. Next we describe the set of trajectories that reflect in the same sides of the billiard polygon; the results imply that the entropy of polygonal billiards vanishes. Section 4 concerns rational billiard polygons: we partition the phase space into invariant surfaces and study the billiard flow on them. Section 5 is a brief discussion of recent strong results on rational billiards via the theory of quadratic differentials; we only outline the main ideas and concepts. The last section concerns the relation between the dynamics of point masses and billiards, the billiard in a polyhedral angle and periodic billiard trajectories in triangles.

3.1 Square Billiards

We start the study of billiards in polygons wih the simplest case of a square. It is surprising how much one can actually say about it!

First, let us describe the procedure of unfolding a billiard trajectory. Instead of reflecting the trajectory in a side of a polygon reflect the polygon in this side. In this way the trajectory, straightened to a line, piercing a number of isometric copies of the billiard polygon.

fig. 46

Thus a correspondence is established between the billiard trajectories in a square and straight lines in the plane with a square grid, which may be assumed unit (we disregard the fact that the billiard trajectories stay away from the vertices; the unfolding argument shows that for a square the trajectories can be extended through the vertices). Two lines in the plane correspond to the same billiard trajectory if they differ by a translation through a vector from the lattice $2\mathbf{Z} + 2\mathbf{Z}$.

Notice that two neighbouring squares have the opposite orientations – they are symmetric with respect to their common side. Consider a bigger square that consists of four unit squares with a common vertex, and identify its opposite sides to obtain a torus. A billiard trajectory becomes a geodesic line on the flat torus: closed if the trajectory is periodic, and infinite dense geodesic otherwise.

fig. 47

Consider the trajectories in a fixed direction α. The billiard flow gives rise to the map f of the circle $S^1 = \mathbf{R^1}/\mathbf{Z}$, shown in the figure. This map is a rotation:

$$f(x) = x + \cot\alpha \ \ (mod 1).$$

The dynamical properties of the billiard flow in a fixed direction correspond to those of this rotation; say, if the rotation is ergodic, then so is the flow etc. Likewise the billiard dynamical system in an n-dimensional cube decomposes according to the directions of the trajectories; each component has a section which is a $(n-1)$-dimensional torus rotation.

Thus one is led to the study of a rotation of the torus $\mathbf{T}^n$:

$$T_a : (x_1, ..., x_n) \to (x_1 + a_1 \ (mod 1), ..., x_n + a_n \ (mod 1)),$$

equipped with the Lebesgue measure $\mu = dx_1...dx_n$.

Notice that if the trajectory of some point is dense, then any other trajectory is dense as well: it follows from the fact that the torus is a homogeneous space of $\mathbf{R}^n$, acting by translations. The main properties of the torus rotations are contained in the following two theorems, due to Weyl and to Kronecker-Weyl, respectively (see, e.g., [Si 1]).

Theorem 1. *The rotation T_a is ergodic if and only if the numbers $a_1, ..., a_n$ are independent over the integers (that is, if $r_1 a_1 + ... + r_n a_n \in \mathbf{Z}$ for $r_1, ..., r_n \in \mathbf{Z}$ then $r_1 = ... = r_n = 0$).*

Proof. Let $a_1, ..., a_n$ be independent over $\mathbf{Z}$, and let f be a T_a-invariant *mod* 0 measurable function (which means that $f(T_a x) = f(x)$ almost everywhere). One wants to show that f is a constant off a set of zero measure.

First, one may assume that f is bounded. Otherwise, consider the set $S = \{x : |f| \leq c\}$, and replace f by $\chi(S)f$, where χ is the indicator of the set (1 inside, 0 outside). This new function is still T_a-invariant and bounded, so the result for f will follow from that for $\chi(S)f$ as $c \to \infty$.

Let

$$f(x) \sim \sum_r c_r e^{2\pi i < r,x >}$$

be the Fourier expansion of f, where $r = (r_1, ..., r_n)$ is an integer vector, and the summation is over all such vectors. Since f is T_a-invariant,

$$f(T_a x) \sim \sum c_r \ e^{2\pi i < r,a >} e^{2\pi i < r,x >} = \sum c_r \ e^{2\pi i < r,x >} \sim f(x) \ (mod\ 0).$$

The Fourier coefficients are unique, hence

$$c_r \ e^{2\pi i < r,a >} = c_r$$

for all r. If $c_r \neq 0$ then $< r, a >$ is an integer, which is impossible by the assumption.

Conversely, if $< r, a > \in \mathbf{Z}$ then the function $e^{2\pi i < r,x >}$ is T_a-invariant. *Q.E.D.*

As a matter of fact irrational rotations are uniquely ergodic.

In an ergodic system the trajectory of almost every point spends the time in a measurable set asymptotically proportional to its measure; it follows that the trajectory of almost every point is dense. In view of the above remark, all orbits of T_a are dense, provided that $a_1, ..., a_n$ are independent over the integers (this fact has an elementary proof; the reader is encouraged to find it). Apply this to the orbit of the "origin" (0, ... , 0).

Theorem 2. *Let the irrational numbers $a_1, ..., a_n$ be linearly independent over the rationals. Then for every positive ϵ there exist the integers k and $m_1, ..., m_n$ such that*

$$|k \ a_i - m_i| < \epsilon, \quad i = 1, ..., n.$$

Proof. There exists a number k such that the k-th image of the origin under T_a is ϵ- close to itself: $k \ a_i \approx_\epsilon 0 \ (mod\ 1)$. These are the desired inequalities. *Q.E.D.*

As far as the entropy is concerned, the following result holds.

Theorem 3. *The metric entropy (with respect to the Lebesgue measure) of a torus rotation equals zero.*

Proof. This follows from two properties of entropy, mentioned in Section 1.11. A rotation of a torus is the product of rotations of circles, so it is enough to consider this 1-dimensional case. If

the angle of a rotation T is π-rational, then $T^n = id$ for some n. Since $h(T^n) = nh(T)$, the entropy of T vanishes. Otherwise the partition of the circle in two semicircles is a one-sided generator, hence $h(T) = 0$. *Q.E.D.*

Consider an irrational rotation T of the circle $\mathbf{R}^1/\mathbf{Z}$. The orbit of each point spends the average time in an interval, asymptotically equal to its length (Weyl's theorem):

$$\lim_{n\to\infty} \{k : 0 \le k < n,\ T^k(x) \in I\} = \mu(I).$$

One says that the orbit is equidistributed in the circle.

One cannot help mentioning an application to a problem from number theory (see [Ar 5, Ki]). Consider the powers of 2 : 1, 2, 4, 8, ... ; do they start more often with 7 than with 8?

2^k starts with 7 if and only if $7 \cdot 10^m \le 2^k < 8 \cdot 10^m$ for some integer m. Take the logarithm with the base 10:

$$\lg 7 + m \le k \lg 2 < \lg 8 + m.$$

Consider the rotation T of the circle $\mathbf{R}^1/\mathbf{Z}$ through $\lg 2$. The average number of times 2^k starts with 7 is the number of times $T^k(0) \in [\lg 7, \lg 8)$. This number equals $\lg(8/7)$. Likewise one deals with the first digit of 8. Thus the powers of 2 start with 7 more often than with 8, namely,

$$\frac{\lg(8/7)}{\lg(9/8)}$$

times as often.

Return back to the billiard inside a square. Its phase space is partitioned into disjoint components, corresponding to different directions α of the trajectories. If $\tan\alpha$ is rational, all trajectories are periodic; otherwise the billiard flow is ergodic in this component and each trajectory is dense in the configuration space (that is, in the square). This dichotomy – a trajectory is either periodic or dense – does not hold for a general polygon, as we will see later.

To conclude this section we analyse periodic trajectories of the billiard in the square. The unfolding of such a trajectory is a segment in the plane whose end-points differ by a translation through a vector from the lattice $2\mathbf{Z} + 2\mathbf{Z}$. A nearby parallel trajectory is also periodic with the same number of links and the same length. Thus periodic trajectories come in parallel beams.

fig. 48

To describe the length spectrum, assume that an unfolded trajectory goes from the origin to the lattice point $(2p, 2q)$. A trajectory in the south-east direction will go to the north-east after a reflection, so, without loss of generality, we assume that p and q are nonnegative. The length of the trajectory equals $2\sqrt{p^2+q^2}$, and to a choice of p and q two orientations of the trajectory correspond. Hence the number of periodic trajectories of length less than L is the number of nonnegative integers, satisfying the inequality $p^2 + q^2 < L^2/2$.

In the first approximation, this number is the number of integer points inside the quarter of the circle of radius $L/\sqrt{2}$. Modulo terms of lower order, it equals the area, that is, $\pi L^2/8$. Hence the number of periodic families of length less than L satisfies $N(L) \sim \pi L^2/8$. The error $N(L) - \pi L^2/8$ is of the order of the circumference length, i.e.$N(L) = \pi L^2/8 + O(L)$ (see [Be 4]).

We mention in conclusion that billiards in rectangles, equilateral triangles and the right triangles with an acute angle $\pi/4$ or $\pi/6$ are investigated in a similar way. Reflecting any one of these polygons in their sides, one generates a plane lattice; thus the billiard flow reduces to a linear flow on a torus.

3.2 Symbolic Description of Billiards in Squares and Cubes

Unfolding a nonperiodic billiard trajectory in a square yields a line L in the plane with an irrational slope λ. The plane is equipped with the unit grid, and, travelling along the line, one meets the vertical and the horizontal segments of the grid (we assume the line avoids the vertices; in any case it can happen at most once). Encode the former event by 1 and the latter by 0. Thus to the line L an infinite sequence ξ of 0's and 1's corresponds, which we call the cutting sequence of the line. This sequence determines in which sides of the square – vertical or horizontal – occur consecutive reflections of the billiard ball.

fig. 49

Assume that $\lambda > 1$ and let $[n_0, n_1, n_2, ...]$ be the continuous fraction expansion of λ:

$$\lambda = n_0 + \cfrac{1}{n_1 + \cfrac{1}{n_2 + \cfrac{1}{n_3 + ...}}}.$$

In particular, $n_0 = [\lambda]$. The line L never intersects two consecutive vertical segments, so the cutting sequence consists of blocks of 0's, separated by isolated 1's. How many 0's can there be in a block? By comparizon of the slope of L with those of the two broken lines in the figure one concludes that the number of 0's in a block equals either n_0 or $n_0 + 1$. Following [Se], call the minimum of the two the *value* of the sequence . Likewise, the number N of 0's in the segment 10...010...01...10...01 (k blocks) satisfies the inequalities $k\lambda - 1 < N < k\lambda + 1$, and, therefore, may assume only two different values. If the slope $\lambda < 1$, the roles of 0's and 1's are reversed, and λ gets replaced by $1/\lambda$ in the previous formulas.

fig. 50

The cutting sequence ξ can be encoded in the new symbols: $0' = 0,\ 1' = 0...01$ (n_0 zeroes). For example, if the value of the sequence equals 1, then

$$...0101001001... = ...1'1'0'1'0'1'...$$

Call this new sequence ξ' the predecessor of ξ. It might happen so that ξ' also has a predecessor etc. The following observation was made more than a century ago by Christoffel and H.J.S.Smith.

Theorem 1. *The cutting sequence ξ has predecessors of all generations, and their values are $n_1, n_2, ...$*

Proof. Consider another grid in the plane – the image of the original one under the linear transformation

$$A = \begin{pmatrix} 1 & 0 \\ n_0 & 1 \end{pmatrix}.$$

Encode the sides of the parallelograms of the new grid by the symbols $0'$ and $1'$, according to whether their preimages are horizontal or vertical. Then the pedecessor ξ' is the cutting sequence of the line L with respect to the new grid. This is the same as the cutting sequence of the line $A^{-1}(L)$ with respect to the original square grid. The slope of $A^{-1}(L)$ equals $\lambda - n_0 < 1$. Thus the value of the new cutting sequence is $\left[\frac{1}{\lambda - n_0}\right]$. *Q.E.D.*

fig. 51

As an instructive example, consider the line L through the origin whose slope is the golden ratio $\frac{1+\sqrt{5}}{2}$. Its cutting sequence ξ enjoys a remarkable property.

Theorem 2. *ξ is invariant under the substitution $0 \to 01, \quad 1 \to 0$ (so one creates the sequence starting with 0 and iterating the substitution).*

Proof. Consider the linear transformation

$$\begin{pmatrix} -1 & 1 \\ 1 & 0 \end{pmatrix}.$$

It preserves the line L and sends the integer lattice to itself. Hence the cutting sequence η with respect to the new grid is the same as the old one ξ. The figure depicts that 0 in ξ corresponds to 01 in η, and 1 in ξ – to 0 in η. *Q.E.D.*

fig. 52

Leaving aside other substitution rules, corresponding to the lines whose slopes are quadratic irrationalities, and the relation to the theory of quasicrystals (see, e.g., [Br, S-T]), we proceed to quasiperiodicity of the cutting sequences of lines with irrational slopes.

Theorem 3. *The cutting sequence is not periodic; however, each finite segment appears in it infinitely many times.*

Proof. If the positions of two billiard balls in a square are close enough and they go in the same direction, then they make sufficiently many consecutive reflections in the same sides of the square. One knows from the previous section that a billiard trajectory with an irrational

slope will come infinitely many times to a neighbourhood of any point in the same direction. The quasiperiodicity follows. Since the slope of the line equals the average number of 0's per block of the cutting sequence, periodicity of the sequence implies rationality of the slope. *Q.E.D.*

The next property of the cutting sequences was discovered by M. Morse and G. Hedlund in their study of symbolic dynamics ([M-H]).

Definition. The complexity function $p(n)$ of an infinite sequence in some symbols is the number of its distinct n-element segments.

The following result says that the complexity of the cutting sequences of the lines with irrational slopes is the minimal complexity of nonperiodic sequences in two symbols (such sequences are called Sturmian sequences).

Theorem 4. $p(n) = n + 1$.

Proof. Since a billiard trajectory with an irrational slope comes arbitrarily close to any point of the square, the sets of finite segments of the cutting sequences of any two parallel trajectories coincide. Thus one can find complexity by computing the number of different initial segments of length n in the cutting sequences of the lines with a given slope . Partition the square grid into "ladders', as shown in the figure. The k-th symbol in the cutting sequence is 0 or 1, according to whether the line L meets a horizontal or a vertical segment of the k-th ladder.

fig. 53

Let e_1 and e_2 be the unit orthogonal frame in the plane. Project the plane onto the diagonal $x + y = 0$ along L, and factorize the diagonal by the translation through $e_1 - e_2$. Identify the quotient space with the unit circle S; the projections of the vertices of the first ladder partition S into two irrational arcs. Let T be the rotation of the circle through the length of an arc, that is, through the projection of e_1.

The number of different initial n-segments of the cutting segments of the lines with the given slope equals the number of segments into which the projections of the vertices of the first n ladders partition the circle S. Each ladder is obtained from the first one by translation through e_1. Hence the complexity equals the number of points in the orbit $T^i(0),\ i = 0, ..., n$. Since T is an irrational rotation, all points of the orbit are distinct; thus $p(n) = n + 1$. *Q.E.D.*

The theorem was recently generalized to the billiard in a 3-dimensional cube – see [A-M-S-T]. Consider a line in 3-space in a totally irrational direction (this means that the components of the direction vector are rationally independent). In space the "grid" of the planes is given: $x_i = n;\ i = 1, 2, 3;\ n \in \mathbf{Z}$. One defines the cutting sequence in three symbols, according to the direction of the consecutive planes the line meets.

Theorem 5. *The complexity of the cutting sequence equals* $n^2 + n + 1$.

fig. 54

The proof proceeds along the lines of the previous one. Instead of the circle S one gets a 2-torus, represented by a hexagon (the projection of a unit cube along the line onto the diagonal plane $x+y+z=0$), whose opposite parallel sides are pasted together. The projection of the first ladder (which is now a polyhedral surface) partitions the torus into three parallelograms. Also an irrational rotation T of the torus is given (through the projection of a unit basic vector onto the diagonal plane); and the problem consists in computing the number of pieces in the partition of the torus, obtained by the intersection of the first n iterations of the initial partition under T. The numbers of vertices and edges of this intersection are found to be n^2+2 and $2n^2+n+3$, respectively. Since the Euler characteristics vanishes, one finds the number of faces to be n^2+n+1.

See Theorem 3.4.4 for another generalization of Theorem 4.

3.3 Unfolding Trajectories in General Polygons

A billiard trajectory in a general polygon P can be unfolded exactly in the same way as we did in Section 3.1. Each successive copy of the billiard polygon is obtained from the previous one by reflection in the side, met by the straightened trajectory.

fig. 55

Consider the group of motions of the plane $G(P)$, generated by the reflections in the sides of P. Denote the reflection in side s by σ_s. Notice that $\sigma_{\sigma_s(t)}=\sigma_s\sigma_t\sigma_s^{-1}$ for any two sides s and t. It follows that every copy of P, involved in the unfolding, is the image of P under a motion from the group $G(P)$. Notice that the product of an even number of generators of the group preserves the orientation of the plane, while an odd number reverses it.

To study the directions of the trajectories one projects the group of motions of the plane onto the orthogonal group $\mathcal{O}(2,\mathbf{R})$. Denote the image of $G(P)$ in the orthogonal group by $\mathcal{O}(\mathrm{P})$. This group is generated by the reflections in the lines through the origin, parallel to the sides of the polygon. The product of two reflections is the rotation through the angle twice the angle between the corresponding lines.

Definition. A polygon P is called rational if $\mathcal{O}(\mathrm{P})$ is a finite group and irrational otherwise.

Thus the angles of a rational polygon are rational multiples of π; if a polygon is simply connected this condition is also sufficient for being rational. A finite subgroup of $\mathcal{O}(2)$, generated by reflections, is the group of symmetries of a regular polygon, called the dihedral group. Since the direction of a billiard trajectory changes by an element of the group $\mathcal{O}(\mathrm{P})$, when it hits the boundary, the number of directions of a given trajectory in a rational polygon is finite.

Consider an n-periodic billiard trajectory and unfold it to a line l. The motion g from the group $G(P)$, that sends P to its n-th copy along l, preserves this line. One of two things hold: either g is orientation reversing – then n is odd and g is a glide reflection in l, or g is orientation preserving – then n is even and g is a parallel translation in the direction of l (we refer to [Be 1]

for these facts from plane geometry). In the latter case the trajectory is included in a strip, that consists of a one-parameter family of periodic trajectories with the same periods and lengths. In the former the trajectory is also included in a one-parameter family of periodic trajectories, but now their periods and lengths are twice as big.

fig. 56

Definition. A periodic billiard trajectory in a polygon is called stable if an arbitrary small perturbation of the vertices leads to a perturbation of the trajectory, but not to its destruction.

For example, a 2-periodic trajectory in a rectangle is not stable.

fig. 57

Enumerate the sides of the polygon by $1, 2, ..., k$ clockwise. Consider an even-linked periodic trajectory and let $i_1, ..., i_{2n}$ be the numbers of the sides, visited successively. The following criterion of stability was found in [G-S-V].

Lemma 1. *The trajectory is stable if and only if the numbers in the list $i_1, ..., i_{2n}$ can be partitioned in pairs of equal numbers, so that the number from each pair appears once at an even position, and once at an odd one.*

Proof. Let α_i be the angle made by the i-th side with a fixed direction. The $2n$-th copy of the polygon along the unfolded trajectory is parallel to it. The composition of two reflections is a rotation, and the angle, through which the consecutive reflections in the sides $i_1, ..., i_{2n}$ rotate the plane, equals

$$\alpha_{i_1} - \alpha_{i_2} + \alpha_{i_3} - ... + \alpha_{i_{2n-1}} - \alpha_{i_{2n}}.$$

Hence this angle is a multiple of 2π, and, for the trajectory to be stable, the variation of the angle under a perturbation of the polygon should satisfy:

$$\delta\alpha_{i_1} - \delta\alpha_{i_2} + ... + \delta\alpha_{i_{2n-1}} - \delta\alpha_{i_{2n}} = 0.$$

However, an arbitrary variation of the vertices induces an arbitrary variation of the directions of its sides. Hence the previous relation holds if and only if the terms in it cancel pairwise. *Q.E.D.*

Corollary 2. *If the angles of a polygon are independent over the rational numbers then every periodic billiard trajectory in it is stable.*

Proof. In the previous notation,

$$\alpha_{i_1} - \alpha_{i_2} + ... + \alpha_{i_{2n-1}} - \alpha_{i_{2n}} = 0 \ (mod 2\pi).$$

If the indices $i_1, ..., i_{2n}$ do not satisfy the previous criterion, one obtains a nontrivial rational linear relation on the angles $\alpha_1, ..., \alpha_k, \pi$, which implies a rational dependence of the interior angles of the polygon. *Q.E.D.*

For instance, a generic convex $2n$-gon does not admit a $2n$-link nonintersecting billiard trajectory. A necessary condition for such a trajectory to exist is that the alternating sum of the interior angles vanishes (see [Be 1]). We refer to [G-S-V] for other results and examples concerning stability of billiard trajectories in polygons.

We proceed to some results from [G-K-T] that strengthen the ones from [B-K-M]. assume that the billiard polygon is simply connected. As before, the sides are assigned the numbers $1, ..., k$, and an orbit gets encoded by the sequence of the numbers of the sides it successively hits. Consider the part of the phase space of the billiard transformation T, that consists of the positions of the billiard ball on the boundary and its inward unit velocities such that the forward orbit of the ball never hits a vertex. Abusing the notation, call it V. For $x \in V$ denote by $w(x)$ the sequence that encodes the forward orbit of x. Given a sequence w denote by $X(w)$ the set of phase points whose forward orbits are encoded by w (in particular, their foot points belong to the same side of the polygon). Call a set $S \subset X(w)$ a strip if all $x \in S$ are parallel vectors whose foot points constitute an interval. If the foot points form an open interval, call it an open strip. The image of a strip under the billiard transformation is again a strip of the same width.

fig. 58

We start with two remarks. First, if the directions of two vectors x and y from V are not parallel, then $w(x) \neq w(y)$. Indeed, their unfolded trajectories linearly diverge, so some vertex of the copies of P will fall into the angle between them. The first time it happens the reflections of the two trajectories occur in different sides.

fig. 59

Secondly, given a word w and two parallel vectors x and y in $X(w)$, any parallel vector, whose foot point lies between those of x and y, also belongs to $X(w)$. Hence $X(w)$ is the maximal strip, corresponding to the word w. The trajectories of its boundary points come arbitrarily close to some vertices of the polygon. We mention also, that the maximal strip, that contains a vector x, is sent to the maximal strip for $T(x)$ by the billiard transformation T.

Theorem 3. *Let w be an n-periodic word. Then each vector from $X(w)$ has a periodic trajectory (of period n if n is even; if n is odd then one element of $X(w)$ has an n-periodic trajectory and the rest – $2n$-periodic ones). The set $X(w)$ is an open strip.*

Proof. Assume n is even (otherwise, double it). Pick $x \in X(w)$ and unfold its trajectory. We claim that the n-th copy of P along the trajectory is parallel to P. Indeed, otherwise x and $T^n(x)$ have the foot points on the same side of the polygon (because the word w is n-periodic), but their directions are not parallel. Hence $w(x) \neq w(T^n(x))$, which contradicts the periodicity of w. Thus the n-th copy of P is obtained from P by a parallel translation in the direction of x. This

means that x is a fixed point of T^n, so its trajectory is periodic. Since a periodic trajectory stays a bounded distance away from the vertices, the strip $X(w)$ is open. *Q.E.D.*

Unlike periodic trajectories, non-periodic ones are never included in strips.

Theorem 4. *For a non-periodic sequence w the set $X(w)$ consists of at most one point.*

Proof. If there are two points in $X(w)$ then $X(w)$ is a nonzero width strip. Let $x \in X(w)$ have its foot point on the strip's median. Consider the forward limit set Y of x under the billiard transformation T. Since the width of $X(w)$ is positive, the trajectories of the points from Y stay a bounded distance away from the vertices. Hence $Y \subset V$, and the restriction of T to Y is continuous. Being bounded and closed, Y is compact.

Use the strengthened version of the recurrence theorem from Section 1.6 due to H. Furstenberg ([Fu]) to conclude that there exists a uniformly recurrent point $y \in Y$ (it means: for any neighbourhood W of y there exists a constant $C > 0$ such that the return times $n_i > 0$, defined by $T^{n_i}(y) \in W$, satisfy: $n_{i+1} - n_i < C$). Pick a positive ϵ and consider the unfolded maximal strip S for y together with its ϵ-neighbourhood S_ϵ.

fig. 60

According to the remark, preceding Theorem 3, some vertices of the copies of P, unfolded along the strip S, fall into S_ϵ. We claim that they fall with uniformly bounded gaps into each of the two components of $S_\epsilon - S$. Consider the leftmost point y' of S. Find m such that the foot point of $T^m(y')$ is within $\epsilon/2$ of a vertex of the polygon, and assume it to be the first such vertex along the strip. Consider the neighbourhood W of y such that for any $z \in W$ the m-th iterate $T^m(z)$ is $\epsilon/2$-close to $T^m(y)$. Let z' be the vector, parallel to z, whose foot point concides with that of y'. Then $T^m(z')$ is $\epsilon/2$-close to $T^m(y')$. Let n_i be the recurrence times for y to W. Then $m + n_i$ are the recurrence times for y' to the ϵ-neighbourhood of the vertex under consideration, and this sequence has uniformly bounded gaps.

Recall that y is a forward limit of x: there is a sequence $n_i \to \infty$ such that $x_i = T^{n_i}(x) \to y$. No two vectors x_i and x_j may be parallel. Indeed, they do not coincide because the trajectory of x is not periodic. Also both distances from the foot point of x to the two boundaries of its unfolded maximal strip can only increase under T. Hence if x_i is parallel to x_j, $i < j$, then the maximal strip for x_j contains at least one boundary of the maximal strip for x_i in its interior. This would imply that there are some vertices inside the unfolded strip for x_j, which is impossible.

fig. 61

Finally, consider the intersections of the unfolded strips for x_i with $S_\epsilon - S$. As $i \to \infty$, the diameter of this intersection goes to infinity. Because the gaps between the vertices that fall into $S_\epsilon - S$ are uniformly bounded, a vertex will eventually appear in the interior of the unfolded strip

for x_i. This is a contradiction. *Q.E.D.*

It follows that the closure of a non-periodic billiard trajectory in a polygon contains at least one vertex. Previously this was proved in [B-K-M] for every position of the billiard ball and almost every direction at it. Similar results hold for the geodesic flow on a polyhedral surface and for 3-dimensional polyhedral billiards ([G-K-T]).

An important consequence of the two previous theorems concerns the topological entropy of the billiard transformation T for a polygon. Topological entropy was defined for continuous maps, while the billiard transformation has discontinuities. The authors of [G-K-T] refer to [P-P] for an appropriate definition of h_{top} in the discontinuous case.

Corollary 5. $h_{top}(T) = 0$.

The above statement was first appeared in the paper by A. Katok [Ka 2]. It implies that the metric entropy vanishes as well (proved in [B-K-M] and [Si]).

We only outline the proof. Consider the set Y of sequences in $1, 2, ..., k$ that encode the billiard trajectories for time from $-\infty$ to ∞. Let S be the shift transformation. Then the encoding map $E : V \to Y$ conjugates the billiard transformation and the shift : $ET = SE$. By the previous results, S has a one-sided generator – the partition into k parts by the value of the zero symbol in the sequence (i.e., the side to which the foot point of a vector belongs). Hence for a shift-invariant measure on Y its metric entropy vanishes.

To apply the variational principle for topological entropy, consider the closure of Y in the space of sequences. A. Katok proved in [Ka 2] that every ergodic non-atomic shift-invariant measure on $\bar{Y}$ is supported on Y. Thus topological entropy of S on $\bar{Y}$ vanishes. To conclude that topological entropy of T vanishes too, one uses a refinement of a theorem by R. Bowen (see [G-K-T] for details).

It follows that certain quantities, associated with the billiard in a polygon, grow slower than exponentially (subexponentially in the terminology of [Ka 2]): the number of different words of a fixed length in Y; the number of families of periodic trajectories with the period not greater than a given number; the number of generalized diagonals, that is, billiard trajectories from one vertex to another, etc. It was conjectured that these numbers grow at most polynomially, but it is not proved yet (see [Gu 2] and [Ka 2] for a discussion). The property of having zero topological entropy is also proved in [Gu-H] for a broader class of transformations – the generalized polygon exchanges, that include polygonal billiards.

We mention in conclusion a result by M. Boshernitzan ([Bos]). Recall that the phase space of the billiard is $\partial P \times S^1$. A direction $v \in S^1$ is called exceptional if, for some $x \in \partial P$, the trajectory of (x, v) is a generalized diagonal. A direction $v = \exp(2\pi\phi) \in S^1$ is called rational if ϕ is a rational number. Boshernitzan's theorem states that there exist only finitely many directions that are exceptional and rational at the same time.

3.4 Rational Polygons

Recall that rational polygons are the polygons P such that the group $\mathcal{O}(\mathrm{P})$, introduced in the

previous section, is finite; for a simply connected polygon this is equivalent to the condition that all the angles are π-rational. In a rational polygon a given billiard trajectory may have only finitely many different directions. This makes it possible to decompose the phase space into a collection of smaller invariant subspaces. This decomposition first appeared in [F-K], and then had been rediscovered in [K-Z, Ke] and [R-B] (see [Gu 2] for a reference).

To fix ideas let P be a simply connected rational polygon. The group $\mathcal{O}$(P) is the dihedral group D_N generated by the reflections in lines through the origin that meet at angles π/N, where N is a positive integer. This group has $2N$ elements, and the orbit of a generic point $\theta \neq k\pi/N$ of the unit circle consists of $2N$ points. If the angles of the polygon are $\pi m_i/n_i$, where m_i and n_i are coprime integers, then N is the least common multiple of n_i's. Consider the phase space of the billiard flow $P \times S^1$, and let R_θ be its subset of points whose second coordinate belongs to the orbit of θ under D_N. Since a trajectory changes its direction by an element of D_N under each reflection, R_θ is an invariant surface of the billiard flow in P.

To construct R_θ (for $\theta \neq k\pi/N$) consider $2N$ disjoint parallel copies of P in the plane. Call them $P_1, ..., P_{2N}$, and orient the even ones clockwise and the odd ones counterclockwise. We will paste their sides together pairwise, according to the action of the group D_N. Let $0 < \theta = \theta_1 < \pi/N$ be some angle, and let θ_i be its i-th image under the action of D_N (see figure 62). Consider P_i and reflect the direction θ_i in one of its sides. The reflected direction is θ_j for some j. Paste the chosen side of P_i to the identical side of P_j. After these pastings are done for all the sides of all the polygons one obtains an oriented closed surface R. The result does not depend on the choice of the original angle θ, hence we suppress it from the notation. Figure 62 illustrates our construction for a right triangle whose acute angle equals $\pi/8$. The surface R has genus 2 in this example.

fig. 62

Consider the general case. Let the angles of the billiard k-gon be $\pi m_i/n_i$, and N be the least common multiple of n_i's. In the process of pasting one consecutively attaches $2n_i$ copies of the i-th angle of the polygon around its i-th vertex to obtain a multiple of 2π (in the above example 16 copies of the angle $3\pi/8$ are needed; they sum up to 6π). The number of copies of the i-th vertex before pasting was $2N$, hence, after pasting, there remain N/n_i copies of this vertex. Thus the total number of vertices in R is $N\sum \frac{1}{n_i}$. The total number of edges is $2kN/2 = kN$, and the number of faces is $2N$. Therefore the Euler characteristic of R equals

$$N\sum \frac{1}{n_i} - kN + 2N = -N\sum \frac{m_i - 1}{n_i};$$

the last equality follows from the formula for the sum of angles of a k-gon:

$$\sum \frac{\pi m_i}{n_i} = \pi(k-2).$$

Consequently the genus equals

$$1 + \frac{N}{2}\sum \frac{m_i - 1}{n_i}.$$

The billiard flow on the surface R is obtained from the constant flows in the directions θ_i in the polygons P_i. The result is a vector field on R with singularities at the vertices. The i-th vertex of R is the result of pasting $2n_i$ copies of the angle $\pi m_i/n_i$, which sums up to $2\pi m_i$. However the angle at the corresponding vertex of R is 2π. Thus one geometrically realizes the pasting by scaling down the angles at the i-th vertex by the factor m_i. The result is a multisaddle singularity, shown in figure 63 for $m_i = 3$.

fig.63

We remark that the flow on R is free from singularities if and only if $m_i = 1$ for all i. Since the sum of all angles equals $\pi(k-2)$ one gets the following equation in this case:

$$\frac{1}{n_1} + ... + \frac{1}{n_k} = k - 2.$$

The only integer solutions $n_i \geq 2$ are, up to a permutation:

$$(1/3, 1/3, 1/3), \quad (1/2, 1/4, 1/4), \quad (1/2, 1/3, 1/6), \quad (1/2, 1/2, 1/2, 1/2)$$

(exercise for the reader). The corresponding polygons, sometimes called integrable, were already mentioned in Section 3.1: an equilateral triangle, a right isosceles triangle, a right triangle with the acute angles $\pi/6$ and $\pi/3$, and a rectangle. In each case the billiard flow reduces to a constant flow on a torus.

After the surface R is constructed one can cut it along some of its edges to get a connected polygon Q in the plane (if its pieces happen to overlap we consider them as belonging to different copies of the plane). This is similar to cutting a paper cube (or a more general polyhedron) along its edges to flatten it. The resulting polygon Q is by no means unique. Its sides are partitioned into pairs of parallel equal sides that are to be pasted together pairwise to recover the surface R. The billiard flow in a fixed direction is realized as a constant flow, with the understanding that once the billiard ball hits a side it instantaneously reappears at the corresponding point of the parallel side to proceed in the same direction.

fig. 64

What we have achieved so far is a decomposition of the billiard flow in a rational polygon into a one-parameter family of flows F_θ on the surface R. The flows in different directions are obtained one from another by a rotation. Recall that a flow is called minimal if any of its orbits is dense (see section 1.10).

Theorem 1. *For all but countably many directions θ the flow F_θ is minimal.*

Proof. Given a rational billiard polygon P consider the unfolding of the billiard trajectories in it. Each copy of P one obtains in this way is an image of P under the group $G(P)$. This

group is countable, hence there are countably many different copies. Call a direction in the plane exceptional if it is the direction from a vertex of P to a vertex of one of its images under $G(P)$. There are countably many exceptional directions. Let θ be a non-exceptional direction.

We claim that the flow F_θ is minimal on R. First, a trajectory in this direction cannot be periodic – otherwise its unfolded maximal strip will contain vertices on the boundary, which implies that θ is exceptional. Suppose that a trajectory t of F_θ is not dense. Then there exists a maximal strip of non-zero width in the direction θ in the above described polygon Q, which is never visited by the trajectory t. Since θ is not exceptional, the boundary of this strip does not contain a segment through two vertices of Q. Thus each segment of this boundary is either a segment of t or the limit of segments of t. Since the width of the strip is positive and the area of R is finite, the strip must be periodic. Then t is periodic as well, which was already found to be impossible. *Q.E.D.*

Projecting back to the billiard polygon P we see that for all but countably many directions every billiard trajectory is dense in P. Following [K-Z] we apply the previous considerations to arbitrary billiard polygons. Recall (Section 1.10) that a flow (or a map) is called topologically transitive if it has a dense orbit. Consider the space X of simply connected k-gons with its obvious topology (the product of k copies of the plane). Recall that a G_δ set is a countable intersection of open sets.

Theorem 2. *There is a dense G_δ subset in X consisting of polygons with topologically transitive billiard flows.*

Proof. Identify the phase space of the billiard flow in each polygon from X with $D^2 \times S^1$; assume that this identification continuously depends on the polygon. Choose a countable basis B_i for the topology in $D^2 \times S^1$. Denote by X_n the set of k-gons P which have a billiard trajectory that visits all the images of the sets $B_1, ..., B_n$ in the phase space of the billiard flow in P. Each X_n is open, and $\cap X_n$ is a G_δ set.

To show that it is dense consider the set Y_m of rational k-gons such that the least common denominator of their angles is not less than m. For a polygon $P \in Y_m$ any invariant surface R_θ is $1/m$-dense in the phase space. Hence for any n there exists m such that for any $P \in Y_m$ the surface R_θ intersects all the images of $B_1, ..., B_n$ in the phase space of the billiard in P. Since R_θ has a dense trajectory for almost all θ, we conclude that $Y_m \subset X_n$. Finally, Y_m is dense in X for every m, so X_n is dense for every n, and so is $\cap X_n$ in view of Baire's theorem. *Q.E.D.*

We refer to [Gu-K 2] for similar results concerning weak mixing.

In Part 1 we reduced the dimension of the phase space of the billiard flow from 3 to 2 replacing the flow with a map. Likewise, in the case of rational polygons, the billiard flow on an invariant surface can be reduced to a one-dimensional transformation.

Definition. Consider a partition $(I_1, ..., I_n)$ of the interval $[0, 1)$ into nonintersecting semi-closed intervals, enumerated from left to right, and let $\sigma = (\sigma_1, ..., \sigma_n)$ be a permutation of n elements. An interval exchange transformation $T : [0, 1) \to [0, 1)$, associated to the partition and

the permutation, is a transformation whose restriction to each I_i is a parallel translation and such that the intervals $T(I_1), ..., T(I_n)$ follow from left to right in the order $\sigma_1, ..., \sigma_n$.

For example, the exchange of two intervals $[0, a)$ and $[a, 1)$ is identified with the rotation through $1 - a$ of the circle $\mathbf{R^1}/\mathbf{Z}$.

The reduction goes as follows. Choose a side s_1 and an initial angle θ_1, and consider the collection of all pairs of sides and angles visited, starting from s_1 in the direction θ_1. This collection consists of sides $s_1, ..., s_k$ together with angles $\theta_{ij},\ j = 1, ..., l_i,\ i = 1, ..., k$, belonging to the sides s_i. Draw l_i copies of s_i side by side, and contract the j-th copy of s_i by the factor $\sin \theta_{ij}$. Then, by the results from Part 1 on the invariant measure of the billiard transformation, this transformation induces a piecewise isometry of the collection of intervals under consideration. If the intervals are correctly oriented this transformation is order preserving, and therefore, is an interval exchange transformation (see [C-F-S, Gu 2] or [B-K-M]).

An equivalent way to construct this reduction is by way of unfolding the invariant surface in the plane. If Q is the corresponding polygon then its sides sre partitioned into n pairs (s_i, s_i') of equal parallel segments. A constant flow has an invariant transversal measure, "the width of a beam". Let $I = s_1 \cup ... \cup s_n$, and define a measure in I as the length of the orthogonal projection of s_i along the constant flow. Then the flow induces an interval exchange in I.

The reduction of the billiard flow to an interval exchange transformation provides an alternative approach to the study of rational billiards. For instance, Theorem 1 can be proved along these lines. An important result by A. Katok ([Ka 3]), obtained in this way, states that the billiard flow on an invariant surface is not mixing.

Next we show that the billiard in a rational polygon always has a periodic trajectory. This simple argument was found by A. Stepin (see [G-C, G-Z]) and was also mentioned in [Bos]. Shoot the billiard ball in the direction perpendicular to a side of the polygon. By Poincaré's recurrence theorem, the ball will return back arbitrarily close to its position and will meet the side at an angle arbitrarily close to $\pi/2$. Since the number of directions of a trajectory is finite the ball will return to the original side in the direction perpendicular to it. After the ball bounces off it will repeat its motion along the same trajectory backwards, so this trajectory is periodic. A similar argument applies to a rational polytope in any dimension, that is a polytop with the property that the group, generated by the reflections in hyperspaces, parallel to its faces, is finite.

fig. 65

We also mention almost integrable billiards studied by E. Gutkin ([Gu 2,3]). These billiards occupy an intermediate position between the integrable ones and the ones in rational polygons. A polygon P is called almost integrable if the group $G(P)$ is a discrete subgroup of the group of motions of the plane. There are four such subgroups, generated by reflections in the sides of the four integrable polygons. An almost integrable polygon is a polygon that can be drawn on the corresponding lattice.

fig. 66

Choose a basis e_1, e_2 for this lattice; a direction $a_1e_1 + a_2e_2$ is called rational if a_1/a_2 ia a rational number (this does not depend on the choice of the basis). E. Gutkin proved that the following conditions are equivalent for an almost integrable polygon: a direction is irrational; the billiard flow on the invariant surface in this direction is minimal; this flow is ergodic; this flow is aperiodic (i.e., for no t its time t map is equal to identity). This result is specific for almost integrable polygons: it does not hold for an arbitrary rational polygon.

In conclusion we formulate a generalization of Theorem 3.2.4 due to P. Hubert ([Hub]) to rational polygons. Let the angles of a rational billiard k-gon be $\pi m_i/r$ where $m_1, ..., m_k, r$ are coprime. Encode a billiard trajectory by the sides it reflect in.

Theorem 4. *If a trajectory is dense on the corresponding invariant surface of the billiard flow then the complexity of its encoding sequence*

$$p(n) = (k-2)rn + 2r$$

for all sufficiently large n.

In the case of a square one gets $p(n) = 4n + 4$ which appears to be different from the result of Theorem 3.2.4. The reason for this discrepancy is that the encoding in that theorem was different from the present one: parallel sides of a square corresponded to the same symbol in the cutting sequence.

3.5 Rational Billiards and Quadratic Differentials

A recent breakthrough in the study of rational billiard polygons is due to applications of Teichmuller theory. A wealth of results obtained by S. Kerckhoff, H. Masur, J. Smillie and W. Veech relies heavily on the rather technical theory of quadratic differentials (see, e.g. [Str]). We are not in a position to discuss these results in detail; we will only outline some of them. Let us remark that this theory is specific for the two dimensional case – no analogs are known so far for higher-dimensional billiards.

We saw in the previous section that the billiard flow in a rational polygon has invariant surfaces R. The surface R comes equipped with a special geometric structure: an atlas of charts mapping open subsets of R to $\mathbf{R}^2$ with the change-of-coordinate functions of the form $v \to v+c$. The surface R has a finite number of singular points at which the charts are n-fold branched coverings of $\mathbf{R}^2$. This structure on R is a particular case of a *quadratic differential*, that is defined in a similar way with the only difference that the transition functions are $v \to \pm v + c$ and that at the singular points one has branched coverings of $\mathbf{R}^2/(\pm 1)$ (analytically, a quadratic differential on a Riemann surface is $f(z)dz^2$ with the singularities $z^k dz^2$).

A quadratic differential determines several other structures. Identifying $\mathbf{R}^2$ with $\mathbf{C}$ one obtains a complex structure on R off singular points. This complex structure extends to all of R. A

quadratic differential defines a metric on R, flat off singularities; at singular points this metric has a cone type singularity with the cone angle a multiple of π. The metric determines the area of R. A quadratic differential also defines a pair of foliations. The horizontal foliation is induced by the foliation of $\mathbf{R}^2$ by horizontal lines, the vertical foliation is induced by the foliation of $\mathbf{R}^2$ by vertical lines. These foliations have transverse measures determined by the quadratic differential. The measures are invariant: given two transverse intervals to the same leaf, the intersections with the leaves of the foliation define a mapping from one interval to another; this mapping is measure-preserving. A geodesic line on R is given by a line segment in each nonsingular coordinate chart; such a segment has vertical and horizontal components. Two isotopic geodesic lines have the same vertical and horizontal components.

Given an element $g \in \mathbf{PSL}(2,\mathbf{R})$ and an atlas $\{\phi_i\}$ the products $\{g\phi_i\}$ form a new atlas. Thus $\mathbf{PSL}(2,\mathbf{R})$ acts on quadratic differentials. One distinguishes the following one parameter subgroups:

$$g_t = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}, \quad r_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \quad h_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix},$$

referred to as the geodesic, circular and horocyclic flow, respectively. Notice that

$$\lim_{t\to\infty} g_t \; r_{s\exp(-t)} g_{-t} = h_s$$

(a direct computation with matrices).

In Teichmuller theory one considers the Teichmuller space of complex structures on a surface modulo its diffeomorphisms isotopic to the identity. Similarly one considers the space Q of quadratic differentials with the area one modulo diffeomorphisms isotopic to the identity (this space is naturally identified with the unit cotangent bundle of the Teichmuller space). The mapping class group M of diffeomorphisms modulo diffeomorphisms isotopic to the identity acts on this space. The quotient space $QD = Q/M$ is the main object of study; the group $\mathbf{PSL}(2,\mathbf{R})$ acts continuously on QD.

Many notions of ergodic theory extend to foliations with transverse invariant measure. Such a foliation is called ergodic if, given a saturated set (a union of the leaves), its intersection with any transverse disc has either full or zero measure. A foliation is called uniquely ergodic if there is a unique invariant transverse measure (which automatically is ergodic).

Now we can formulate the results from [K-M-S] (see also [Arn]).

Theorem 1. *Given a quadratic differential q on a closed surface of genus not less than 2, for almost all θ (in the sense of Lebesgue measure) the vertical foliation of $r_\theta(q)$ is uniquely ergodic.*

H. Masur ([M 2]) proved also that the Hausdorff dimension of the set of angles θ for which the vertical foliation of $r_\theta(q)$ is not ergodic does not exceed $1/2$.

We already know that the billiard flows in different directions on the surface R are obtained one from another by the rotation r_θ. Applying the previous theorem to rational billiards one gets the following result.

Corollary 2. *The billiard flow on the invariant surface is uniquely ergodic for almost all directions.*

An approximation of arbitrary polygons by rational ones, similar to the one used in Theorem 3.4.2, leads to the next result.

Theorem 3. *There is a dense G_δ subset in the space of polygons consisting of polygons for which the billiard flow is ergodic.*

We now very briefly outline the proof of Theorem 1. Given a quadratic differential q consider its geodesic orbit $g_t(q)$. If it eventually leaves all compact sets in QD as t increases to infinity, q is called divergent; otherwise it is called recurrent. One proves that the set of angles θ for which $r_\theta(q)$ is divergent is of measure zero. Then one considers the set S of θ for which $r_\theta(q)$ is recurrent and its vertical foliation is not uniquely ergodic. There exists a closed set B in QD that contains all accumulation points of the orbits $g_t(q)$ for $q \in S$. Each quadratic differential in B possesses a closed leaf of its vertical foliation (may be through singular points).

If α is a closed geodesic of a quadratic differential q with the horizontal and vertical components h and v then the horizontal component of α with respect to $h_s(q)$ is $h + sv$. Hence α is vertical for at most one value of s. If β is isotopic to α it has the same horizontal and vertical components. Since there are only countably many isotopic classes of geodesics the intersection of the horocyclic orbit $h_s(q)$ with B is at most countable. Thus $h_s(q)$, $-1 \le s \le 1$ spends most of the time away from B. On the other hand, there is a sequence $t_i \to \infty$ such that $g_{t_i}(q) \to B$. Since $h_s g_{t_i}(q)$ is approximated by $g_{t_i} r_{s\exp(-t_i)}(q)$ (see above), $g_{t_i}\ r_\theta(q)$ is close to B for the majority of $\theta \in [-e^{-t_i}, e^{-t_i}]$. This shows that 0 is not a point of density for the set S. By changing coordinates one shows that S has no points of density, and therefore has zero measure. A subtlety in the actual proof is that all convergences involved are to be uniform, and the considerations should deal with compact sets.

We would like to make a comment on this proof. Note that a result, concerning a given quadratic differential, is obtained by way of studying the space of classes of quadratic differentials and certain flows on it. The situation somewaht resembles that in knot theory after the works by V. Vassiliev: to construct invariants of an individual knot one studies the space of all knots.

Another collection of results concern periodic orbits. We already saw that any rational polygon has a periodic billiard trajectory. The following theorem is due to H. Masur ([M 1]).

Theorem 4. *Given a quadratic differential q on a closed surface of genus not less than 2, there exists a dense set of angles θ such that the vertical foliation of $r_\theta(q)$ has a closed nonsingular leaf.*

Corollary 5. *For any rational billiard polygon there is a dense set of directions each with a periodic orbit.*

The following result from [B-G-K-T] is a strengthening of the previous one.

Theorem 6. *Periodic points are dense in the phase space of the billiard flow in a rational polygon.*

We reiterate that it remains unknown whether an arbitrary polygon has a closed billiard trajectory.

The next result by Masur ([M 3]) concerns the growth function of periodic billiard trajectories in rational polygons. Let $N(t)$ be the number of strips of periodic trajectories of length not greater than t.

Theorem 7. *For any rational polygon there exist constants c and C such that for sufficiently large t the inequalities hold: $ct^2 < N(t) < Ct^2$.*

W. Veech proved that the growth function is asymptotically quadratic

$$N(t) \sim \frac{C(P)\ t^2}{Area P}$$

for the following polygons P: regular polygons and isosceles triangles with equal angles π/n ([V 1,2]). Veech found explicit formulas for the coefficients $C(P)$ in each case. A similar result: the growth function is asymptotically quadratic, was proved for almost integrable polygons by E. Gutkin ([Gu 2]).

Another remarkable result by W. Veech ([V 1,2]) strengthens Theorem 3.4.1 for regular polygons.

Theorem 8. *Any billiard trajectory in a non-exceptional direction is infinite or semi-infinite and uniformly distributed in the regular billiard polygon. Any billiard trajectory in an exceptional direction, that does not hit a vertex, is periodic.*

3.6 Miscellanea

First, we discuss a reduction of the dynamics of point-masses in the line with elastic reflection to that of billiards. Consider two points with masses m_1 and m_2 in the half-line $x \geq 0$ (so if a point hits the "wall" $x = 0$ its velocity changes sign). The configuration space of this system is given by the inequalities $0 \leq x_1 \leq x_2$, where x_1 and x_2 are the coordinates of the two points. Rescale the variables: $\bar{x}_i = \sqrt{m_i} x_i;\ i = 1, 2$. The configuration space is now the angle $\arctan \sqrt{m_1/m_2}$.

fig. 67

Consider a collision of the two points. Let v_1, v_2 be the velocities before, and u_1, u_2 – after the collision. Then the conservation of momentum and energy reads:

$$m_1 u_1 + m_2 u_2 = m_1 v_1 + m_2 v_2,$$

$$\frac{m_1 u_1^2}{2} + \frac{m_2 u_2^2}{2} = \frac{m_1 v_1^2}{2} + \frac{m_2 v_2^2}{2}.$$

In the rescaled coordinates the velocities are rescaled by the same factor, so:

$$\sqrt{m_1}\ \bar{u}_1 + \sqrt{m_2}\ \bar{u}_2 = \sqrt{m_1}\ \bar{v}_1 + \sqrt{m_2}\ \bar{v}_2,$$

$$\bar{u}_1^2 + \bar{u}_2^2 = \bar{v}_1^2 + \bar{v}_2^2.$$

The second equation says that the length of the velocity vector $(\bar{v}_1, \bar{v}_2)$ does not change after the collision. The first one says that the dot product of the velocity vector with the vector $(\sqrt{m_1}, \sqrt{m_2})$ does not change as well. This vector is orthogonal to the boundary line of the configuration space: $\bar{x}_1/\sqrt{m_1} = \bar{x}_2/\sqrt{m_2}$. Hence the configuration trajectory reflects in this line according to the billiard law. Likewise one considers a reflection in the wall $x = 0$ to conclude that the system is isomorphic to the billiard in the configuration angle. Unfolding the trajectory yields the following conclusion: the number of collisions is always finite, and does not exceed the number

$$\left\lceil \frac{\pi}{\arctan \sqrt{m_1/m_2}} \right\rceil.$$

fig. 68

Similar considerations reduce the dynamics of several particles in the line, half-line or a segment to that of the billiard in a higher dimensional polyhedral angle or a polyhedron.

The billard in a convex polyhedral angle was considered by Ya. Sinai in [Si 2]. The result is similar to the one in the plane.

Theorem 1. *The number of reflections of any trajectory is finite; it is bounded by a constant that depends on the angle only.*

Proof. We follow [G-Z]. Consider the unit sphere, centered at the vertex of the polyhedral angle, and project the angle onto the sphere from its center. The projection is a convex spherical polyhedron P, and a billiard trajectory in the angle projects to a billiard trajectory in P (in the spherical metric, of course). Unfolding a trajectory in the angle one gets a line in space, whose projection on the sphere is half a great circle. Hence the total length of the billiard trajectory in P equals π.

The proof is by induction on the dimension. Consider the union of the ϵ-neighbourhoods of the faces of P whose codimension is 2 or greater. Inside each neighbourhood the billiard is reduced to the one in a lower dimensional angle. Therefore the number of reflections of the billiard ball in each neighbourhood, until the ball leaves it, is uniformly bounded. The union of the codimension 1 faces with the ϵ-neighbourhoods removed is disconnected, and consists of the union of pieces of the faces. The distance between two such pieces is bounded away from zero by a constant c, depending on the polyhedron and ϵ. Hence the billiard ball will travel for the length of at least c between two different ϵ-neighbourhoods, and since the length of the trajectory is π, the number of such "trips" does not exceed π/c. Thus the total number of reflections is uniformly bounded. *Q.E.D.*

fig. 69

The figure illustrates the situation in the 2-dimensional sphere, corresponding to a polyhedral angle in 3-space. Another proof of this result can be found in [Sev].

The next example, due to G. Galperin ([Ga]), provides a non-periodic billiard trajectory that is not dense in a convex polygonal billiard table.

fig. 70

Consider a centrally symmetric hexagon such that the perpendicular to the side AF bisects the angle BAC, and the perpendicular to BC bisects the angle ABF. Then a vertical ray becomes vertical again after two reflections. Projecting the vertical rays onto the segment XY, the second iteration of the billiard transformation, acting on the rays, induces the exchange of the intervals XZ and ZY in it (one needs the following inequalities to hold: $\cot 2\alpha > \tan\beta,\ \cot 2\beta > \tan\alpha$). An exchange of two intervals is the same as a rotation of a circle; for generic α and β this rotation is irrational. Its orbits are infinite and dense, so the vertical billiard trajectories are dense in the hexagon. Extend the non-vertical sides of the hexagon to form a parallelogram whose triangular "ears" are never visited by any vertical trajectory.

A question of interest is whether every triangle has a periodic billiard trajectory. As far as an acute triangle is concerned, the following "Fagnano" trajectory is, probably, known to the reader from his/her high-school years: it is an inscribed triangle whose vertices are the feet of the triangle's altitudes.

fig. 71

It is unknown whether a periodic trajectory exists in every obtuse triangle. Following [G-S-V], we mention several explicit constructions of periodic billiard trajectories.

The first construction concerns the triangles whose acute angles α and β satisfy the relation $k\alpha = n\beta < \pi/2$ for some positive integers k and n. After $(k-1)$ reflections about one vertex and n reflections about another vertex two sides become parallel. Their common perpendicular, being folded back, is a periodic trajectory, perpendicular to the sides of the triangle.

fig. 72

Another kind of periodic trajectory is shown in the next figure. The acute angles of this triangle satisfy $\alpha + n\beta = \pi/2,\ 2\alpha + \beta > \pi/2$.

fig. 73

The above trajectories are not stable in the sense discussed in Section 3 (i.e., they are destroyed by a small perturbation of the triangle). The last construction is free from this shortcoming. Let the acute angles satisfy the inequalities

$$\frac{\pi-\beta}{2} < k\alpha < \frac{\pi}{2} \le (k+1)\alpha,\ \frac{\pi-\alpha}{2} < n\beta < \frac{\pi}{2} \le (n+1)\beta.$$

Make $(k-1)$ reflections about one vertex and $(n-1)$ about another one. One obtains an acute triangle, whose Fagnano billiard trajectory folds back to a periodic trajectory in the original triangle.

fig. 74

Finally, consider right triangles. If it is a rational triangle, we know from Section 3.4 that almost every trajectory, perpendicular to a side, returns to this side in the same direction, and therefore is periodic. Remarkably, the same happens in the irrational case – see [C-H-K].

Theorem 2. *Given a right triangle with π-irrational acute angles, almost every (in the sense of measure) billiard trajectory, that starts at a side of the right angle in the perpendicular direction, returns to this side in the same direction.*

fig. 75

Proof. Let α be an acute angle of the triangle. Reflect the triangle in the sides of the right angle to obtain a rhombus R. The study of the billiard in the triangle reduces to that in the rhombus.

fig. 76

Construct the two dimensional invariant subspace of the phase space of the billiard flow in the rhombus, corresponding to the beam of horizontal trajectories which start at the upper half of the vertical diagonal. This is done as described in Section 3.4: consider the disjoint union of rhombi obtained from R by the action of the group $\mathcal{O}(\mathrm{R})$, and identify their sides pairwise in an appropriate manner. One obtains a noncompact surface (partially) foliated by trajectories from the beam. This foliation has an invariant transversal measure ("width of a beam").

Each rhombus involved is obtained from R by a clockwise rotation through the angle $2n\alpha,\ n \in \mathbf{Z}$; such a rhombus will be refered to as the n-th rhombus R_n (thus $R = R_0$). A trajectory from the beam may leave the n-th rhombus through one of its two sides; call a side positive if the trajectory enters the $(n+1)$-st, and negative if it enters the $(n-1)$-st rhombus.

fig. 77

One wants to show that almost all trajectories return to R_0 (where they stop at the vertical diagonal). Since α is π-irrational, for every $\epsilon > 0$ there exists $n > 0$ such that the vertical projection of the positive side of R_n is smaller than ϵ – see Section 3.1 on irrational circle rotations. Hence the set of trajectories that make it to R_{n+1} has measure less than ϵ.

The rest of trajectories is bound to stay in $R_0, ..., R_n$; call the set of these trajectories S. The union of the rhombi 0 through n is compact, and the Poincaré recurrence argument applies as in Section 3.4. Namely, almost every trajectory in S is recurrent; since it makes only finitely many different angles with the vertical diagonal of the rhombus R, it will become perpendicular to it, that is return to R_0.

Said differently, the total width of the beams that constitute S is positive, and the total area of the rhombi 0 through n is finite. Since beams cannot overlap, almost all trajectories from S come back to R_0.

Since the above ϵ is arbitrary small, the result follows. *Q.E.D.*

It seems that this argument can be extended with appropriate changes to other polygons (perhaps all?!)

Another result on periodic trajectories in right triangles is due to G. Galperin and A. Zvonkin ([Ga-Zv]).

Theorem 3. *There exists a periodic billiard trajectory through every point of a right triangle. Moreover these trajectories can be chosen so that their lengths are bounded by a constant depending on the triangle only.*

Concerning the numerical study of the billiard in a triangle we also refer to [Ru], [C-H-K] and [Bos].

4. Dual Billiards

The topic of this chapter is the lesser known dual billiard transformation. We define it in Section 1 and discuss its basic features such as the area-preserving property. We also comment on the relation between billiards and dual billiards via projective duality. Section 2 gives an approximation of the dual billiard dynamics far away from a dual billiard; if the billiard is smooth enough and strictly convex the map has invariant curves, and all orbits are bounded and separated from the table. We introduce the area spectrum of a dual billiard and obtain some results similar to those from Section 2.7. Section 3 contains several dynamical proofs of the classical Poncelet's theorem from projective geometry; the first one is by way of dual billiards in the hyperbolic plane. Section 4 concerns polygonal dual billiards; it provides a sufficient condition for all orbits to be bounded. We also describe the dual billiard dynamics for an affine-regular pentagon. The last section contains the definition of higher dimensional dual billiards; we prove that this map has periodic trajectories of all odd prime periods.

4.1 Definition, Area-Preserving Property and Generating Function

Dual billiards (also known as outer billiards) are in many ways similar to billiards. The game of dual billiard is weird in a sense: the "balls" move outside of the dual billiard "table", and the motion is defined for discrete time only. Let γ be an oriented strictly convex closed curve in the plane (a dual billiard table). Given a point x outside of γ draw the segment xy, $y \in \gamma$ of the supporting line to γ, whose orientation from x to y agrees with that of γ, and extend it through y to the point $T(x)$ such that $dist\ (T(x), y) = dist\ (y, x)$. The map T of the exterior of γ to itself is the *dual billiard transformation*.

fig. 78

If γ is convex but has straight segments the dual billiard transformation and its inverse are defined off the extensions of these straight segments. This is analogous to billiards: one does not define an extension of a trajectory through a corner of a billiard curve. If γ is not convex T is defined as a multiple-valued map.

We start with a fundamental property of dual billiard maps.

Lemma 1. *T is area-preserving.*

fig. 79

Proof. Let δ be a convex curve; fix $c > 0$ and consider the one-parameter family of straight lines that cut off segments of area c from δ. Let γ be the envelope of this family. Claim: the segment of any line from this family determined by its intersection with δ is bisected by the point of tangency with γ. Indeed, suppose $OA > OB$ in figure 79. Let $A'B'$ be a nearby line from the family and let $O' = AB \cap A'B'$. Then $O'A > O'B$. The central symmetry in O' sends the "triangle" $BO'B'$ inside $AO'A'$. On the other hand, $Area(BO'B') = Area(AO'A')$, which is a contradiction (this elementary argument is contained in [F-T]).

Consider now two nearby curves δ and δ_1 – see figure 79. It follows from the previous argument that the shaded areas are equal. Hence T is area preserving. *Q.E.D.*

Corollary 2 (of the proof). *Given a convex curve δ there exists a one-parameter family of curves γ such that δ is an invariant curve for the dual billiard transformation with respect to γ. The curves γ are the envelopes of the segments of equal areas.*

The reader will recognize in this construction the dual billiard analog of the string construction from Section 2.1. Notice that the "duality" interchanges length and area. We will comment on this later. Another consequence of the proof concerns periodic trajectories of the dual billiard map. Connecting consecutive points of such an orbit one obtains a circumscribed n-gon whose sides are

bisected by their tangency points to γ. Such a polygon may make r turns about γ, $1 \leq r < n/2$; the number of turns is called its rotation.

fig. 80

It follows from the proof of Lemma 1 that n-periodic orbits correspond to extrema of areas of n-gons circumscribed about γ. For a smooth strictly convex curve this area functional attains its minimum on the set of circumscribed n-gons with the rotation r.

Corollary 3. *If a dual billiard curve is smooth and strictly convex then for any coprime numbers $1 \leq r < n/2$ there exists an n-periodic orbit of the dual billiard map with the rotation r.*

This corollary also follows from the general results concerning area-preserving twist maps (Sections 2.6 and 2.10). Introduce the coordinates $x(\alpha, r)$ in the exterior of γ: α is the angle made by the positive tangent line to γ in the direction of x with a fixed direction in the plane, and r is the distance along this line from γ to x.

fig. 81

First, we observe that T is a twist map with respect to the "vertical" foliation $\alpha = const$ – see figure 81. Secondly, let $T(\alpha, r) = (\beta, R)$, and denote the area of the shaded curvilinear triangle in figure 81 by $S(\alpha, \beta)$. Let ω be the standard area form in the plane. The next result follows from a direct computation.

Lemma 4.

$$\omega = rdr \wedge d\alpha$$

and

$$\frac{r^2}{2} = -\frac{\partial S(\alpha, \beta)}{\partial \alpha}, \qquad \frac{R^2}{2} = \frac{\partial S(\alpha, \beta)}{\partial \beta}.$$

It follows that $S(\alpha, \beta)$ is a generating function of the map T: if $\lambda = r^2 d\alpha/2$ then $d\lambda = \omega$ and $T^*\lambda - \lambda = dS$. We again see that T is area preserving, and that the area of the circumscribed polygon corresponding to a periodic orbit is a particular case of the symplectic invariant, introduced at the beginning of Section 2.7. We also notice for the second time that area plays the role in the dual billiard problem analogous to that of length in the billiard one.

To compare billiards and dual billiards consider both on the sphere. We use the projective duality, interchanging points and lines (that is, great circles): to a pole its oriented equator corresponds. A point on a curve γ is sent to a tangent line to the dual curve γ^*, which, by definition, consists of the tangent lines to γ. Notice that the angle between two lines equals the distance between the dual points.

Lemma 5. *Billiards and dual billiards are projective dual in the sphere.*

fig. 82

Proof. The billiard map acts on oriented lines: to a ray the reflected ray corresponds. Consider a moment of reflection in γ. The dual configuration consists of a tangent line to γ^* and two points on it. Since $Angle(a,b) = Dist(A,B)$ the billiard rule of angles translates to the dual billiard rule: $dist\ (A,L) = dist\ (L,B)$. *Q.E.D.*

Closed billiard trajectories are extrema of the length functional on inscribed polygons. Therefore closed dual billiard trajectories are extrema of the "sum of angles" functional on circumscribed polygons. By the Gauss-Bonnet theorem this sum equals, up to constants, the area of the polygon. Whence the length – area duality. However in the flat case (which is the limit of the spherical one: the radius goes to infinity) the two problems become different: dual billiards are equivariant with respect to the group of affine transformations of the plane, while billiards are equivariant with respect to the smaller group of similarities. Another important difference is that there is no configuration space for the dual billiard transformation.

We mention the dual billiard analog of the mirror equation from Section 2.4 found by E. Gutkin and A. Katok ([Gu-K 1]). Let δ be an invariant curve of the dual billiard transformation T with respect to a curve γ, and let $\rho(y)$ be the curvature radius of γ at point y

fig. 83

Lemma 6.

$$\cot\alpha + \cot\beta = \frac{2\rho(y)}{r}.$$

This lemma is proved by a direct computation which we omit.

The results and constructions of this section had been discovered and rediscovered by several authors. Corollary 3 is found in [Da]; the construction of dual billiards and their area-preserving property were considered by J. Moser in the framework of KAM theory ([Mo 4]). Ph. Boyland lectured on dual billiards in the late 80-s; his paper on the subject is in preparation. We also refer to [Ta 1,2] for a general discussion of dual billiards.

4.2 Invariant Curves, Integrability, Area Spectrum

A natural question to ask about dual billiards is whether orbits can escape to infinity or "fall" on the dual billiard curve. If the curve is sufficiently smooth (C^7 is enough) and has nonvanishing curvature the negative answer is provided by KAM theory. In a vicinity of γ the dual billiard map T is a small perturbation of the shear map $(\alpha, r) \to (\alpha + 2r, r)$, therefore T possesses an abundance of invariant curves near the dual billiard curve. These curves separate orbits from γ.

Consider the situation near infinity. We use the variables α and $\rho = 1/r$ as the local coordinates. In these coordinates the second iteration T^2 is again a small perturbation of a shear map near the curve $\rho = 0$. Hence there exist invariant curves outside of any disc containing γ, and

these curves prevent orbits from escaping to infinity. J. Moser ([Mo 6]) considered the dual billiard transformation as a crude model for planetary motion; thus for a sufficiently smooth dual billiard curve this "planetary motion" is stable.

One can be more specific concerning the dual billiard dynamics near infinity. Let γ be a strictly convex smooth closed oriented curve. Denote by $y(\alpha)$ the point of γ at which the tangent direction to γ makes the angle α with a fixed direction in the plane. Set $v(\alpha) = y(\alpha + \pi) - y(\alpha)$. We extend the definition of $v(\alpha)$ to strictly convex piecewise smooth curves, replacing tangent lines by the supporting ones. $v(\alpha)$ is a continuous odd vector function of α. We further extend the definition to convex piecewise smooth curves, in particular, to polygons. Let l be a straight segment of γ in the direction α. Then $v(\alpha - \epsilon)$ and $v(\alpha + \epsilon)$ are defined for small ϵ, and $v(\alpha)$ has a discontinuity of the first kind at α.

fig. 84

Consider in another copy of the plane with polar coordinates (α, r) a homogeneous vector field $V(\alpha, r) = v(\alpha)$. The following result is contained in [M-T].

Lemma 1. *The integral curves of the field V are closed homothetic curves, centrally symmetric with respect to the origin. If γ^* is one of these curves then corners of γ^* correspond to straight segments of γ, and straight segments of γ^* – to corners of γ. Up to a dilation, $\gamma^{***} = \gamma^*$, and if γ is centrally- symmetric then $\gamma^{**} = \gamma$. The flow defined by the vector field V satisfies the Kepler law: area, swept by the position vector of a point, is linear in time.*

Proof. Choose an origin inside γ and let $\rho(\alpha)$ be the support function, that is the distance from the origin to the supporting line of γ in the direction α. The point of tangency of this supporting line with γ has the coordinates:

$$\rho(\alpha)\sin\alpha + \rho'(\alpha)\cos\alpha, \quad -\rho(\alpha)\cos\alpha + \rho'(\alpha)\sin\alpha$$

(see, e.g. [Sa] or make a direct computation). Let $d(\alpha) = \rho(\alpha + \pi) + \rho(\alpha)$ be the width of γ in the direction α. It follows that the vector $v(\alpha)$ has the components

$$-(d(\alpha)\sin\alpha + d'(\alpha)\cos\alpha), \quad d(\alpha)\cos\alpha - d'(\alpha)\sin\alpha.$$

fig. 85

Let a trajectory γ^* of the vector field V be given by a function $r(\alpha)$ in polar coordinates. Since $x = r\cos\alpha,\ y = r\sin\alpha$ one concludes that

$$\frac{r'(\alpha)\sin\alpha + r(\alpha)\cos\alpha}{r'(\alpha)\cos\alpha - r(\alpha)\sin\alpha} = \frac{d'(\alpha)\sin\alpha - d(\alpha)\cos\alpha}{d'(\alpha)\cos\alpha + d(\alpha)\sin\alpha}.$$

Hence $r(\alpha) = const/d(\alpha)$ is the polar equation of γ^*.

Geometrically it means that γ^* is obtained from γ in two steps: symmetrize, i.e., replace $\rho(\alpha)$ by $\rho(\alpha) + \rho(\alpha + \pi)$; polar dualize, i.e., replace $\rho(\alpha)$ by $1/\rho(\alpha)$. In invariant terms, polar duality assigns to a point other than the origin the affine line in the dual space that consists of the linear functionals whose value at this point equals 1. From this description all but the last statement follow.

To prove the Kepler law we compute the cross-product of the velocity vector with the position vector and use the equation $r(\alpha) = const/d(\alpha)$:

$$\begin{vmatrix} d'(\alpha)\cos\alpha + d(\alpha)\sin\alpha, & const\ \cos\alpha/d(\alpha) \\ d'(\alpha)\sin\alpha - d(\alpha)\cos\alpha, & const\ \sin\alpha/d(\alpha) \end{vmatrix} = const.$$

Q.E.D.

Note that the position of the origin in the previous consideration is irrelevant.

The following figure shows several curves γ and the corresponding curves $\gamma^*,\ \gamma^{**}$. If γ is a semi-circle, γ^* consists of two arcs of equal parabolas intersecting at right angle.

fig. 86

The significance of the above described continuous motion is that is provides an approximation of the second iteration of the dual billiard map T near infinity. Choose an origin O inside γ and let x be a point outside of γ at distance d from the origin. Consider the orbit of x under T^2, and let $x_0 = x, x_1, ..., x_n$ be its segment that makes one turn about the origin:

$$Angle(x_0\ O\ x_1) + ... + Angle(x_{n-1}\ O\ x_n) \leq 2\pi,$$

$$Angle(x_0\ O\ x_1) + ... + Angle(x_{n-1}\ O\ x_n) + Angle(x_n\ O\ x_{n+1}) > 2\pi.$$

Let Γ be the dilation with the center O and the coefficient $1/d$. Consider the motion along the vector field V, and let $\{y_t\}$ be the closed trajectory of the point $\Gamma(x)$. Normalize time so that $y_1 = y_0$.

Theorem 2. *For all* $1 \leq i \leq n$,

$$Dist\ (y_{i/n},\ \Gamma(x_i)) \to 0 \quad as \quad d \to \infty.$$

The proof is a straightforward estimation, and we omit it (see [M-T]). The following figure shows that for large d the map T^2 is almost a translation through the vector $2v(\alpha)$.

fig. 87

If the smoothness of a dual billiard curve is less than required by KAM theory one cannot guarantee the existence of invariant curves "at infinity". In such a case almost nothing is known

concerning whether orbits may escape to infinity (however, a sufficient condition for stability in the case when the dual billiard curve is a polygon is known; we will discuss it in the next section). There is a strong computer evidence and an heuristic argument to the effect that if the dual billiard curve is a semi-circle then some orbits go to infinity, and the set of such orbits has a nonvoid interior (see [M-T]). The dual billiard transformation is discontinuous in this case.

For the dual billiards, like for the inner ones, the shape of the table imposes restrictions on the position of invariant curves of the corresponding dual billiard map.

Theorem 3. *If a dual billiard curve γ has a point of infinite curvature (say, a corner) then there are no invariant curves sufficiently close to γ.*

This dual billiard analog of Mather's Theorem 2.9.4 was stated in Ph. Boyland's MSRI talk in 1989; see [Gu-K 1] for a proof. The next figure illustrates the idea of Boyland's proof.

fig. 88

Let x be sufficiently close to a point of infinite curvature. Note that T is a reflection in a point in the linear approximation. Let u be a tangent vector at x in the direction of the tangent line to the dual billiard curve, and v be its image under the derivative of the map T^2. In the (α, r) coordinates, u is vertical and v is deviated from the vertical in the negative sense.

Suppose there is an invariant curve C through x; according to Birkhoff's theorem (Section 2.9) it is a graph: $C = \{(\alpha, r(\alpha))\}$. Since T deviates to the right, consecutive images of of the vertical vector u under the derivative of T are contained in the cones, spanned by an upward vertical vector and a tangent vector to C with a positive first component. In particular, v cannot have a negative first component.

Alternatively the result can be deduced from the "dual billiard mirror equation" of Lemma 4.1.6 (see [Gu-K 1]). The work of Gutkin and Katok contains explicit estimates for the size of the domain free from invariant curves in the spirit of Theorem 2.9.6.

Ph. Boyland also constructed a differentiable dual billiard curve γ with a crash orbit of the dual billiard map, i.e. an orbit that converges to a point on γ. However if γ is C^1 and the derivative of its curvature is bounded such an example is impossible (compare Halpern's construction discussed at the end of Section 1.7).

We now turn to the question of integrability. Clearly the dual billiard transformation with respect to an ellipse is integrable: its invariant curves are concentric homothetic ellipses. The dual billiard counterpart to Birkhoff's conjecture (Section 2.4) says that this is the only integrable case. This conjecture is not proved; a partial result in this direction is the following theorem from [Ta 3].

Theorem 4. *Let two smooth strictly convex curves be given such that the corresponding dual billiard transformations commute in their common domain (that is sufficiently far away from both curves). Then the curves are concentric homothetic ellipses.*

In conclusion of this section we consider the areas of circumscribed polygons corresponding to periodic orbits of the dual billiard map. Call the set of these areas the area spectrum of a dual billiard curve. An interesting problem is to what extent the area spectrum determines the curve, in particular, is it true that it uniquely determines the curve up to an area-preserving affine transformation. Another question to ask is whether the area spectrum is related to the spectrum of some differential operator in a way similar to the relation between the length spectrum and the spectrum of the Laplace operator (Section 2.7). Notice that such an operator, if it exists, must be invariant under area-preserving affine transformations.

fig. 89

To formulate the next result, which is the dual billiard analog of Theorem 2.7.1 and which is proved in a similar way, we make a short digression to discuss affine length (see, e.g. [Sp]). Let γ be a smooth strictly convex plane curve. Fix a point $x \in \gamma$ and let u be a tangent vector to γ at x. Consider a tangent vector field to γ that takes the value u at x, and let x_ϵ be the point of γ to which the flow of the field for time ϵ takes x. The area $A(u, \epsilon)$ of the segment, bounded by γ and the chord xx_ϵ, is of the third order in ϵ. Hence

$$B_x(u) = \lim_{\epsilon \to 0} \frac{A(u, \epsilon)}{\epsilon^3}$$

is a cubic form on the tangent bundle to γ, and $B^{1/3}$ is a 1-form. The integral of this 1-form over the curve is called its affine length. By the very definition the affine length is invariant under area-preserving affine transformations of the plane. Also, by definition, the affine length has the degree of homogenuity 2/3, that is it is multiplied by $k^{2/3}$ under a dilation with the coefficient k (in down-to-earth terms, one measures the affine length in $meters^{2/3}$). An equivalent definition, more convenient for computations: parametrize γ by a parameter t, $a \leq t \leq b$, so that $[\gamma', \gamma''] = 1$ ($[\ ,\ ]$ being the cross-product). Then the affine length of γ is $b - a$, and the function $[\gamma'', \gamma''']$ is called the affine curvature.

It is interesting that the isoperimetric inequality for convex curves in affine geometry goes in the "wrong direction":

$$(Affine\ Length)^3 \leq 8\pi^2 Area,$$

with the equality for ellipses only (see [B-Z]).

We now formulate a result from [Ta 1]. Let A_n denote the area of a simple polygon, circumscribed about a smooth strictly convex dual billiard curve γ, corresponding to an n-periodic orbit of the dual billiard map.

Theorem 5. *The asymptotic expansion holds:*

$$A_n \sim a_0 + \frac{a_1}{n^2} + \frac{a_2}{n^4} + ... + \frac{a_i}{n^{2i}} + ...,$$

where a_0 is the area bounded by γ, and

$$a_1 = \frac{1}{24}(Affine\ Length\ of \gamma)^3.$$

In view of the affine isoperimetric inequality, $3a_1 \leq \pi^2 a_0$, one can thus recognize an ellipse by its area spectrum.

4.3 Poncelet's Theorem

Poncelet's theorem is one of the most beautiful results of classical projective geometry. Since it is close to both billiards and dual billiards we choose to discuss it here. Given two nested ellipses $\gamma_0,\ \gamma_1$ in the plane one plays the game illustrated in the figure: choose a point $x \in \gamma_1$, draw a tangent line to γ_0 through it, find the intersection y with γ_1, and iterate, taking y as a new starting point. The statement is that if x returns back after a number of iterations, then every point of γ_1 will return back after the same number of iterations.

fig. 90

We will give a few proofs of this theorem; the reader, interested in its history and the classical proofs, is referred to [B-K-O-R].

Let us start with the dual billiard approach ([Ta 4]). Consider an ellipse in the plane, and take its interior as the Klein-Beltrami model of the hyperbolic plane. The distance between points is given by $dist\ (x, y) = |\log[x, y, b, a]|$, where [ ] denotes the cross-ratio, and straight lines are represented by chords of the ellipse. The hyperbolic metric determines an area form with infinite total area.

fig. 91

Let γ be a strictly convex smooth closed curve in the hyperbolic plane. One defines the dual billiard transformation T of its exterior in the same way as in Section 4.1 (using hyperbolic distances, of course). The arguments of Section 4.1 still apply, so T is area preserving.

Given two nested ellipses γ_0 and γ_1 consider the one-parameter family of conics γ_t, called a pencil, which pass through the four intersection points $\gamma_0 \cap \gamma_1$ (imaginary in our case). The curves γ_t foliate the annulus between γ_0 and γ_1. Choose an ellipse from the pencil, which contains γ_1, and call it γ_∞. Let T be the dual billiard map with respect to γ_0 in the hyperbolic plane, modelled in the interior of γ_∞.

Lemma 1. *T is integrable: its invariant curves are ellipses from the pencil.*

Proof. Let l be a line in the plane; its intersections with conics from a pencil γ_t define an involution on l. We claim that this involution is a projective transformation of the line (Desargues' theorem – see [Be 1]).

fig. 92

Indeed, applying a projective transformation of the plane, we make the conics γ_t concentric. Take the center as the origin. Then the pencil consists of conics

$$\gamma_t = \{x \mid < A_t x,\ x >= 1\},$$

with $A_t = A + tE$ where A and E are selfadjoint operators. Let l be tangent to γ_0 at point x and u be a tangent vector to γ_0 at x. Parametrize l by a parameter s so that points of l are $x + su$. The intersection $l \cap \gamma_t$ is given by

$$< (A + tE)\ (x + su),\ x + su >= 1.$$

Since $< Ax,\ x >= 1$ and $< Ax,\ u >= 0$ the previous equation is rewritten as

$$s^2 < (A + tE)u,\ u > +2st < Ex,\ u > +t < Ex,\ x >= 0.$$

It follows that

$$\frac{1}{s_1} + \frac{1}{s_2} = -2\frac{< Ex,\ u >}{< Ex,\ x >}$$

independently of t; here s_1 and s_2 are the roots. Thus the correspondence $s_1 \leftrightarrow s_2$ is fractional-linear, that is projective.

To finish the proof consider the next figure. The above involution sends x to y and a to b, preserving c. Since it preserves the cross-ratio, $dist\ (x, c) = dist\ (c, y)$. *Q.E.D.*

fig. 93

Back to Poncelet's theorem. Since T is integrable, its restriction to each invariant curve, γ_1 in particular, is a translation in an appropriate affine coordinate therein (Section 1.8). Thus if some orbit is n-periodic, so is any other orbit.

Two other results readily follow. First, consider a number of ellipses from one pencil: $\gamma, \gamma', \gamma''$, etc, and let γ contain all the others. Pick a point $x \in \gamma$, draw a tangent line to γ', find its intersection with γ, draw a tangent line to γ'', etc. Then if x returns back after a number of iterations, any initial point of γ does (this generalization was known to Poncelet). This is again a consequence of the fact that T, T', T', etc, are translations in the affine parameter on γ.

Secondly, consider three nested ellipses $\gamma, \gamma', \gamma''$ from a pencil, and identify the interior of the outer one γ with the hyperbolic plane. Then the dual billiard transformations T' and T'' commute. This follows from Corollary 1.8.3.

fig. 94

Another elegant proof of Poncelet's theorem is contained in [Ki]. One constructs a smooth invariant measure on the outer ellipse Γ, invariant under Poncelet's map $x \to y$ (see figure 95).

Applying an affine transformation we may assume that Γ is a circle. Let the desired measure be $f(x)\,dx$, where dx is the Lebesgue measure on Γ.

fig. 95

Denote by $R_\gamma(x)$ and $L_\gamma(x)$ the length of the right and left tangent segments from x to γ. Consider a point x_1, ϵ-close to x. Since the segment xy makes equal angles with Γ, the proportion holds:

$$\lim_{\epsilon\to 0}\frac{Arc\ x_1x}{Arc\ y_1y}=\frac{R_\gamma(x)}{L_\gamma(y)}.$$

The measure of the x-arc is its length times $f(x)$, and likewise for y. So the measure is invariant if and only if

$$\frac{f(y)}{f(x)}=\frac{R_\gamma(x)}{L_\gamma(y)}.$$

If γ happens to be a circle the right and left tangent segments are equal: $R_\gamma(x)=L_\gamma(x)$. Denote this common value by $D_\gamma(x)$. If γ is not a circle, let A be an affine transformation that takes γ to one. We have

$$\frac{R_\gamma(x)}{L_\gamma(y)}=\frac{R_{A\gamma}(Ax)}{L_{A\gamma}(Ay)}=\frac{D_{A\gamma}(Ax)}{D_{A\gamma}(Ay)}.$$

Thus $f(x)=1/D_{A\gamma}(Ax)$ is the desired function.

Finally, choose a coordinate t in which $f(x)\,dx=dt$. Then Poncelet's map is a translation $t\to t+c$, and Poncelet's theorem follows.

A variation of this argument is found in [Ko 1]. If both ellipses are circles, one proceeds as above. We claim that there exists a projective transformation of the plane that takes the given ellipses γ and Γ to two circles. First, apply a projective transformation to make the ellipses concentric. Then apply an affine transformation to make Γ into a circle. Let the plane be the horizontal plane in 3-space, and consider the sphere whose equator is Γ. Let δ be its meridian whose plane contains the major axis of γ. For a point $x\in\delta$ consider the tangent plane to the sphere at the point, antipodal to x. Project γ and Γ onto this plane from x.

fig. 96

This projection carries Γ to a circle (any circle on a sphere goes to a circle under a stereographic projection), and γ – to an ellipse γ', one of whose axes lies in the plane of δ and another is perpendicular to it. If x is the south pole, the major axis of γ' lies in the plane of δ, and if x is close to the equator this major axis is perpendicular to this plane. By continuity there exists a position of x such that γ' is a circle.

Finally we outline the proof by Ph. Griffiths and J. Harris via algebraic geometry ([G-H 1,2]). First, complexify the situation: both ellipses γ and Γ are considered as conics in the complex plane $\mathbf{C}^2$. Consider the incidence set E that consists of pairs: (a tangent line to γ, a point of Γ on this

line). E is an algebraic curve in $\gamma^* \times \Gamma = (\mathbf{CP}^1)^* \times \mathbf{CP}^1$ (the asterisk denotes the dual space), nonsingular because the intersection $\gamma \cap \Gamma$ is transversal. The projection of E to Γ is two-fold except for the four points of intersection $\gamma \cap \Gamma$. This makes it possible to compute the Euler characteristic of E, which proves to be equal to 0. Hence E is a torus.

fig. 97

Poncelet's correspondence is the product of two involutions on E: the one, which interchanges the two points of intersection of a tangent line with Γ, and the one, which interchanges the two tangent lines to γ from a point. These involutions are induced by automorhisms of the universal covering $\tilde{E} = \mathbf{C}$. Each one is of the form $z \to -z+c$, and their composition is a parallel translation. Thus Poncelet's correspondence is a translation of the torus E, and Poncelet's theorem follows.

4.4 Polygonal Dual Billiards

A comparison of Chapters 2 and 3 clearly shows that billiards in polygons differ drastically from the ones in smooth strictly convex domains; likewise polygonal dual billiards are worlds apart from smooth stricly convex ones. To start with, if a dual billiard curve γ is a convex polygon, the dual billiard transformation T and its inverse are defined off the set U which consists of the union of straight lines containing the extensions of the sides of γ, as well as its forward and backward orbits under T (this difficulty is analogous to the one encountered in the case of billiards: one has to ignore the trajectories that hit a corner of a billiard table). Let V be the exterior of γ with the set $V_0 = \cup_{n=-\infty}^{\infty} T^n(U)$ deleted. Since V_0 is a countable union of lines the set V has full measure.

The set V consists of two pieces: V_f – the set of periodic points, and V_∞ – the set of points with infinite orbits. Let $x \in V_f$, that is $T^n(x) = x$ for some n. The map T^n is locally a parallel translation (if n is even) or a reflection in a point (if n is odd). It follows that a point, sufficiently close to x, is either n-periodic (for an even n) or $2n$-periodic (for an odd n). The set V_f is open and its connected components are open polygons bounded by segments from V_0. The set V_∞ may be empty: for example, it happens when the vertices of γ belong to a lattice (see below). It is still unknown whether V_f may be empty for any polygon.

In analogy with billiards in polygons (Section 3.3) call a periodic orbit of the dual billiard transformation stable if an arbitrary small perturbation of the dual billiard polygon leads to a deformation of this orbit but not to its destruction. Enumerate the vertices of a polygon counterclockwise. Let x be a $2n$-periodic point, and let $S(x) = (i_1, ..., i_{2n})$ be the sequence of vertices in which x makes successive reflections. The next statement is parallel to Lemma 3.3.1.

Lemma 1. *The orbit of x is stable if and only if the symbols in the list $(i_1, ..., i_{2n})$ can be partitioned in pairs of equal symbols, so that each symbol from every pair once appears at an even position and once at an odd one.*

Proof. Choose an origin and let V_i be the position vector of the i-th vertex. Then $T^{2n}x$ is

its parallel translation through the vector

$$2(-V_{i_1} + V_{i_2} - V_{i_3} + ... - V_{i_{2n-1}} + V_{i_{2n}}) = 0.$$

This relation persists under an arbitrary deformation of the vertices if and only if the terms cancel pairwise. *Q.E.D.*

As in Section 3.3 it follows that for a generic dual billiard polygon all periodic trajectories are stable.

The procedure of unfolding a billiard trajectory has a dual billiard analog too: choose a point x outside of γ as a point of reference, draw a supporting line to γ and reflect γ at the vertex that lies on this line (we assume that the line does not contain a side). Iterating this reflection one obtains a "necklace" of polygons, congruent to γ, around x. This necklace is periodic if and only if the T-orbit of x is periodic, and it is bounded if and only if so is the T-orbit of x. The shape of the necklace is, roughly, that of the polygon γ^* from Section 4.2.

fig. 98

J. Moser ([Mo 6]) asked for which polygons all orbits of the dual billiard transformation are bounded. A sufficient condition was found by A. Shaidenko and F. Vivaldi ([S-V]) and by R. Kolodziej ([Ko 2]); see also [Gu-S]. This condition is formulated in terms of the motion whose trajectories are the curves γ^*, defined in Section 4.2. If γ is a polygon the vector function $v(\alpha)$ is piecewise constant, each vector $v(\alpha)$ is a diagonal of γ, and γ^* is also a polygon.

fig. 99

Let γ^* have N sides (actually N is even because γ^* is centrally symmetric), and let $t_1, t_2, ..., t_N$ be the values of time it takes a point, moving along the vector field V from Section 4.2, to cross the sides of γ^* (said otherwise, t_i is the ratio of the length of the i-th side of γ^* and the length of the corresponding vector v). Since γ^* is defined up to a dilation, these numbers are defined up to a common factor. Call a polygon *quasirational* if $(t_1 : t_2 : ... : t_N) \in \mathbf{QP}^{N-1}$. The property of a polygon to be quasirational is affine invariant.

Lemma 2. *Any lattice polygon (i.e. a polygon whose vertices belong to a lattice in the plane) is quasirational.*

Proof. Without loss of generality assume that the vertices of γ are rational points. Choose an origin O in the plane. The polygon γ^* can be constructed as follows. Draw the lines through O parallel to the sides of γ. These lines have rational slopes, and they partition the plane into cones; the value of the vector function $v(\alpha)$ is constant in each cone. Choose a rational point A_1 on one of the lines as the first vertex of γ^*. To construct the next vertex A_2 draw the line through A_1 in the direction of the corresponding vector v until its intersection with the next line through O. Since v

is a rational vector, A_2 is a rational point too. Therefore the vector A_1A_2 is a rational multiple of v, which means that $t_1 \in \mathbf{Q}$. Repeat the argument to conclude that $t_2 \in \mathbf{Q}$, etc. *Q.E.D.*

fig. 100

Notice that a regular polygon is quasirational as well: by symmetry $t_1 = t_2 = ... = t_N$. A regular n-gon is a lattice polygon only when $n = 3, 4, 6$. Thus the class of quasiregular polygons is bigger than that of lattice polygons.

Theorem 3. *All orbits of the dual billiard map, corresponding to a quasirational polygon, are bounded.*

There are two approaches to the proof of this theorem. Shaidenko and Vivaldi, as well as Kolodziej, show that outside every circle containing γ there exists an invariant set of the dual billiard transformation that separates γ from infinity. This set is a union of polygons, and it plays the role similar to that of invariant curves in the smooth case (see figure 101). Gutkin and Simanyi gave another proof based on the unfolding procedure. We outline their argument.

fig. 101

Proof. Start as in the proof of the previous lemma: choose an origin O and partition the plane into the cones $C_1, ..., C_{2N}$ by lines parallel to the sides of γ. Consider the set P of polygons, not containing O, obtained from γ either by a parallel translation or by a reflection in a point. Let $p \in P$ be a polygon such that O does not belong to the extension of any of its sides. Define the necklace transformation W as the reflection of p in its vertex that lies on the left supporting line to p from O. Call the vertex of reflection the head of p and denote it by $h(p)$.

fig. 102

One wants to show that if γ is quasirational then all orbits of W are bounded. The polygon $p \in P$ is uniquely characterized by the following data: the position of its head and the value of the function $\epsilon(p) = \pm 1$, according to whether p is a parallel translation of γ or its central symmetric image. To each cone C_i a vector v_i corresponds, the value of the locally constant function $v(\alpha)$ in this cone. Thus if p and $W(p)$ belong to C_i then the vector from $h(p)$ to $h(W(p))$ equals v_i. The sides of the polygon γ^* are parallel to the vectors v_i.

Let $R_1, ..., R_{2N}$ be the boundary rays of the cones. Consider a ray R_i, and let $S_i^{\pm}$ be the union of the heads of polygons $p \in P$ which intersect R_i and for which $\epsilon(p) = \pm 1$, correspondingly. Then S_i^+ and S_i^- are semi-infinite strips, bounded by the rays R_i and $R_i + v_i$, and truncated by certain polygonal lines. Let S_i be the disjoint union of S_i^+ and S_i^-, and let $F_i : S_i \to S_{i+1}$ be the mapping induced by W.

fig. 103

Consider the collection of parallel rays $R_{i+1}, R_{i+1} - v_i, R_{i+1} - 2v_i, R_{i+1} - 3v_i, ...$; their intersection with $S_i^{\pm}$ partition the strips into parallelograms. Let a_i be the vector of the side of such parallelogram along the ray R_i, and let $b_i = a_i + v_i$ be the vector of the other side of a parallelogram. It follows that the map F has the periodicity property: $F_i(x + 2a_i) = F_i(x) + 2b_i$.

fig. 104

Suppose that γ is quasirational. Without loss of generality assume that all numbers t_i are integers. Then the triangle whose sides are a_i and b_i is similar to the triangle OA_iA_{i+1}, where A_i and A_{i+1} are the vertices of γ^*. Thus $OA_{i+1} = t_ib_i = t_{i+1}a_{i+1}$, and

$$F_i(x + 2t_ia_i) = F_i(x) + 2t_ib_i = F_i(x) + 2t_{i+1}a_{i+1}.$$

fig. 105

It follows that the first return map $\Phi : S_1 \to S_1$, which is a composition of the maps F_i, is also periodic:

$$\Phi(x + 2t_1a_1) = \Phi(x) + 2t_1a_1.$$

Let $\Pi^{\pm}$ be the union of the first $2t_1$ parallelograms, counting from the origin, entirely contained in $S_1^{\pm}$. Then $S_1^{\pm}$ is an infinite union of parallelograms, congruent to $\Pi^{\pm}$, and an additional polygon, neighbouring O. Let Π be the disjoint union of Π^+ and Π^-. The map Φ induces a transformation T of Π such that for $x \in \Pi$

$$\Phi(x) = T(x) + i(x)2t_1a_1,$$

with an integer $i(x) \geq -1$. Replace Φ by Φ^{-1} to conclude that $i(x) \leq 1$. The same argument applies to any iteration of Φ:

$$\Phi^n(x) = T^n(x) + i_n(x)2t_1a_1; \quad i_n(x) \in \{-1, 0, 1\}.$$

Thus the Φ-orbit of a point from Π stays a bounded distance from Π; by periodicity all orbits of Φ are bounded. *Q.E.D.*

Corollary 4. *If the vertices of a dual billiard polygon γ belong to a lattice then each orbit of the dual billiard map is periodic.*

Proof. We know from Lemma 2 that γ is quasirational. By Theorem 3 all orbits are bounded. The group of motions of the plane generated by the reflections in vertices of γ is discrete. Each orbit of the dual billiard map is contained in an orbit of this group. Being discrete and bounded it is finite. *Q.E.D.*

Consider a polygon such that all orbits of the dual billiard map are bounded. Let $x \in V_\infty$ be a point with an infinite orbit, and let $S(x)$ be its encoding sequence of vertices in which it undergoes consecutive reflections.

Lemma 5. *$S(x)$ is an aperiodic sequence. The set of points y with $S(y) = S(x)$ is nowhere dense. The point x belongs to the closure of the set V_0.*

Proof. Let $S(x)$ be periodic with the period $2n$. Then the vector from x to $T^{2n}x$ is a nonzero vector, otherwise x is a periodic point; call this vector v. This vector is determined by the sequence of vertices in the segment of $S(x)$ of length $2n$, so $T^{2nk}x = x + kv$. This means that the orbit of x is unbounded. The first statement follows.

To prove the second one, let U be the largest connected open set consisting of points y with $S(y) = S(x)$. Since T is area preserving and its orbits are bounded, there exists an n such that $T^{2n}U \cap U$ is not empty. One can connect any two points of the set $T^{2n}U \cup U$ by a path that does not intersect V_0, hence they all are encoded by the same sequence. This contradicts the maximality of U.

The third statement readily follows. Any neighbourhood W of x contains a point y with $S(y) \neq S(x)$. Let n be the first time y and x reflect in distinct vertices of the dual billiard polygon. Then $T^{n-1}x$ and $T^{n-1}y$ are separated by a segment from V_0. Therefore V_0 intersects W. *Q.E.D.*

We now turn to a more detailed study of the simplest quasirational polygon that fails to be a lattice one, namely, an (affine) regular pentagon. We follow [Ta 1,2].

Consider the following computer picture. The white regions, which are actually regular pentagons or decagons, consist of periodic orbits. The black "web" is the set of points with infinite orbits or, equivalently, the closure of the set V_0. Distinct infinite orbits constitute "web necklaces" around the pentagon. Notice self-similarity of the web clearly seen on a blow-up.

fig. 106

We shall study the dual billiard transformation inside its smallest invariant region shown in the next figure. The situation outside of this region is similar, due to the periodicity involved in the proof of Theorem 3.

fig. 107

Theorem 6. *The Hausdorff dimension of the set V_∞ of points with infinite orbits equals*

$$\frac{\log 6}{\log(\sqrt{5}+2)}.$$

Each orbit is dense in V_∞ (that is, the system (V_∞, T) is minimal). A connected component of the set V_f of periodic points is either a regular decagon or a regular pentagon. The periods of

even-periodic (that is, generic) points equal

$$\frac{10}{7}(8 \cdot 6^{n-1} + (-1)^n); \quad n = 1, 2, ...$$

in the former case, and

$$\frac{10}{7}(6^n + (-1)^{n-1}) : \quad n = 1, 2, ...$$

in the latter one.

Proof. The exterior of the pentagon is the union of its five identical exterior angles; identify them by the rotation about the center of the pentagon through $2\pi/5$. Let V be the part of one of these five exterior angles inside the invariant set under consideration, and let T be the transformation of V induced by the dual billiard map.

fig. 108

Denote by A and B the points on the bisector ON such that the clockwise rotation about A through $3\pi/5$ sends the point K to M, and the clockwise rotation about B through $\pi/5$ sends M to K. Denote these rotations by a and b, respectively. The first observation is made by an inspection of the figure 108: T is a piecewise isometry whose restriction to the triangle OKL is a, and whose restriction to the triangle LMN is b.

Assign to a point its itinerary which is a word in the characters a and b that encodes which of the two rotations are applied to the point under consecutive iterations of T. Two points with the same itineraries belong to the same connected component of the set $V - V_0$ (see the proof of Lemma 5).

Denote by D the composition of the dilation, centered at O, that takes B to A, with the reflection in the line AB. Then D decreases distances by the factor of $\lambda = \sqrt{5} - 2$, and it sends N to N_1, M to K_1, K to M_1, etc. Inspecting figure 108 one makes a crucial observation: if $x \in \triangle OKL$ then $DT(x) = T^7 D(x)$; and if $y \in \triangle LMN$ then $DT(y) = T^3 D(y)$. More specifically,

$$Da(x) = aababaaD(x); \quad Db(y) = aaaD(y).$$

Consider figure 108 again. One sees two periodic domains in it: a big T-invariant regular decagon and two big regular pentagons interchanged by T. Apply D and T in all orders to these polygons to obtain new regular decagons and pentagons, which are also periodic due to the above made observation. Given a periodic orbit σ define its rank to be the maximal n for which there exists $x \in \sigma$ such that $D^{-n}(x)$ is still inside the quadrilateral $OKNM$. Periodic orbits of rank zero constitute the above mentioned big regular decagon and the two big regular pentagons.

We claim that all periodic orbits are obtained from those of rank zero by applying compositions of D and T. Indeed, if σ is periodic of rank n let x be its point inside $OK_1N_1M_1$ such that $D^{-n}(x) \in OKNM$. Then, again by the above observation, $D^{-1}(x)$ is periodic, and the rank of its

orbit is $n-1$. Then one proceeds inductively. One also concludes that T acts transitively on the set of periodic decagons of a fixed rank, as well as on that of periodic pentagons of a fixed rank.

It follows that the sets V_f and V_∞ are self-similar. Namely, let W_1 and W_2 be the parts of V_∞ in $\triangle OKL$ and $\triangle LMN$, respectively. W_2 consists of two parts λ-homothetic to W_1; and W_1 consists of five parts λ-homothetic to W_1 and three parts λ-homothetic to W_2 (see figure 108 again). We apply the general techniques from [Fa] to find the Hausdorff dimension. Let d be the equal Hausdorff dimensions of W_1 and W_2, and w_1 and w_2 be their d-volumes. Then

$$w_1 = 5\lambda^d w_1 + 3\lambda^d w_2, \quad w_2 = 2\lambda^d w_1.$$

Hence $\lambda^d = 1/6$ and

$$d = \frac{\log 6}{\log(\sqrt{5}+2)}.$$

To show that each T-orbit is dense in V_∞ let $x, y \in V_\infty$ and $\epsilon > 0$. We want to show that $T^n(x)$ is ϵ-close to y for some n. Choose N be big enough so that there exist periodic polygons X and Y of the same rank and with the same number of sides (i.e. two decagons or two pentagons), such that

$$diamX < \epsilon/2,\ diamY < \epsilon/2,\ dist(x,X) < \epsilon/2,\ dist(y,Y) < \epsilon/2.$$

Since T acts transitively on periodic polygons of the same rank and kind there exists an n such that $T^n X = Y$, and the itineraries of X and x coincide up to the n-th place. Then $T^n(x)$ is ϵ-close to y.

Finally we compute the periods. Let α_n and β_n be the number of a's and b's in the itinerary of a periodic polygon of rank n (we consider this itinerary as a finite sequence). The above discussed self-similarity implies:

$$\alpha_{n+1} = 5\alpha_n + 3\beta_n, \quad \beta_{n+1} = 2\alpha_n.$$

For the decagons $\alpha_1 = 1$ and $\beta_1 = 0$, and for the pentagons $\alpha_1 = \beta_1 = 1$. Solving the linear recurrences one finds:

$$\alpha_n = \frac{1}{7}(6^n + (-1)^{n-1}), \quad \beta_n = \frac{2}{7}(6^{n-1} + (-1)^n),$$

for the decagons, and

$$\alpha_n = \frac{1}{7}(9\cdot 6^{n-1} + 2\cdot(-1)^n), \quad \beta_n = \frac{1}{7}(3\cdot 6^{n-1} + 4\cdot(-1)^{n-1})$$

for the pentagons. Thus in the former case

$$p_n = \alpha_n + \beta_n = \frac{1}{7}(8\cdot 6^{n-1} + (-1)^n),$$

and in the latter one

$$q_n = \alpha_n + \beta_n = \frac{2}{7}(6^n + (-1)^{n-1}).$$

The first numbers are always odd, so the return map for a periodic decagon is the reflection in its center; thus the period of a generic point equals $2p_n$. The return map for a periodic pentagon is the identity, so the period of every point is q_n. These numbers are the periods for the map induced by the dual billiard transformation in one of the exterior angles of the pentagon. Thus the periods for the dual billiard map are five times as big. *Q.E.D.*

Itineraries of points provide a symbolic description of the dynamics on the set V_∞. Namely one considers the closure of the set of sequences obtained from

$$...aababaa\ aababaa\ aaa\ aababaa\ aaa\ aababaa\ aababaa...$$

in the space of all sequences in the symbols a and b. This sequence is invariant under the substitution

$$D(a) = aababaa, \quad D(b) = aaa,$$

and T acts as a shift: $T(c_n) = (c'_n)$ with $c'_n = c_{n+1}$. The infinite sequence is aperiodic although the ratio of the symbols a's to the symbols b's in it equals 3.

Regular n-gons with $n \geq 7$ have not been studied yet, but the situation appear to be qualitatively the same. The next figure shows infinite orbits of the dual billiard map for some regular n-gons ($n = 8, 9, 10, 11, 12, 17, 20, 25$).

fig. 109

4.5 Higher-Dimensional Dual Billiards

This section is concerned with higher-dimensional dual billiards. We follow [Ta 1,2]. The definition of the billiard transformation in higher-dimensional setting does not cause any trouble – it goes exactly along the lines of the two-dimensional case. Consider a smooth closed hypersurface M in a linear space. How does one define the dual billiard transformation associated with M? The obvious difficulty is that, given a point in space, there are too many tangent lines through it to M.

As a motivation consider again higher-dimensional billiards (compare to Section 1.5). Let $\Gamma^n \subset \mathbf{R}^{n+1}$ be a hypersurface. Reflection in Γ is the billiard transformation of the set of rays N^{2n} in $\mathbf{R}^{n+1}$. The set of rays is a symplectic manifold and the billiard transformation preserves the symplectic structure. To Γ a hypersurface $\Sigma \subset N$ corresponds that consists of rays tangent to Γ. A characteristic curve on Σ consists of rays tangent to Γ along a geodesic line on it. Let a ray r hit Γ at x, and let r_1 be the reflected ray. The rays r and r_1 generate a 2-plane, and the rays, that lie in this plane and pass through x, form a line in N. This line is tangent to Σ and has the characteristic direction at the point of tangency. Also we have the condition: the angle of incidence equals the angle of reflection.

A loose translation to the language of dual billiards yields the following definition.

Definition. Let $M^{2n-1} \subset \mathbf{R}^{2n}$ be a smooth closed hypersurface in linear symplectic space. Two points x and y are in the dual billiard relation T with respect to M if the line xy is tangent

to M, has the characteristic direction at the point of tangency, and the segment xy is bisected by the tangency point.

Let us remark that this definition still makes sense when M is a front, that is a singular hypersurface with a well-defined tangent hyperplane at each point. Identify $\mathbf{R}^{2n}$ with $\mathbf{C}^n$. Then the characteristic direction at a point of M is obtained from the normal one by multiplying by $\sqrt{-1}$. Thus x and y are in the dual billiard relation if there exists a point $z \in M$ such that $zx = \sqrt{-1}\, N_z,\ zy = -\sqrt{-1}\, N_z$ for some normal vector N_z to M at z. One can slightly generalize. Let α be an angle. Define the relation T_α: x and y are in T_α if there exists $z \in M$ such that

$$zx = \exp(\sqrt{-1}\alpha)\ N_z, \quad zy = \exp(-\sqrt{-1}\alpha)\ N_z$$

for some normal vector N_z. The dual billiard relation T is a particular case of T_α for $\alpha = \pi/2$.

Theorem 1. *T and T_α are symplectic relations.*

Proof. Let ω be the linear symplectic structure in $\mathbf{R}^{2n}$, and let $G \subset \mathbf{R}^{2n} \times \mathbf{R}^{2n}$ be the graph of T. We want to show that G is a Lagrangian submanifold, the symplectic structure in the product space being $\omega_1 \ominus \omega_2$, where ω_i are the linear symplectic structures in the factors. Let x, y be Darboux coordinates in the first copy of $\mathbf{R}^{2n}$, so that $\omega_1 = dx \wedge dy$, and $\overline{x}, \overline{y}$ be Darboux coordinates in the second one. Consider the cotangent bundle $T^*\mathbf{R}^{2n}$ with its canonical symplectic structure; let q_1, q_2 be space and p_1, p_2 be momentum coordinates. Then the linear map

$$q_1 = \frac{x + \overline{x}}{2},\ q_2 = \frac{y + \overline{y}}{2},\ p_1 = \overline{y} - y,\ p_2 = x - \overline{x}$$

is a symplectomorphism of $T^*\mathbf{R}^{2n}$ to $(\mathbf{R}^{2n} \times \mathbf{R}^{2n},\ \omega_1 \ominus \omega_2)$.

The graph $G \subset T^*\mathbf{R}^{2n}$ is contained in the set

$$\{(q, p) \in T^*\mathbf{R}^{2n} \mid q \in M,\ p \in Ann\ \ T_qM\}.$$

This set is the conormal bundle of $M \subset \mathbf{R}^{2n}$, and therefore a Lagrangian submanifold.

Similarly, the graph G_α of T_α in $T^*\mathbf{R}^{2n}$ is contained in the set of $(q, p) \in T^*\mathbf{R}^{2n}$ such that

$$q \text{ is the end-point of a normal vector } N \text{ to } M,\ p \text{ is the covector } \cot\alpha < N, \cdot > \text{ at } q.$$

This set is the graph of the differential of the following function in $\mathbf{R}^{2n}$:

$$f(q) = \cot\alpha \times \text{distance}^2(q, M),$$

and therefore is a Lagrangian submanifold. *Q.E.D.*

Notice that the above considered sets are bigger than the graphs of T (or T_α), namely they are the unions of the graphs of T and of T_{-1} (or of T_α and of T_α^{-1}).

So far we have defined the dual billiard relation T and included it into a family of relations T_α. To make it into a map we need M to be strictly convex (just as in the case of the plane). Notice that an orientation of M induces an orientation of its characteristic lines, which will be refered to as characteristic rays.

Lemma 2. *Let M be closed and strictly convex. Then for any point x outside of M there exists a unique point $y \in M$ such that yx is the characteristic ray to M at y. Likewise, given an angle $0 \leq \alpha < \pi/2$, for any point x outside of M there exists a unique point $y \in M$ such that $yx = \exp(\sqrt{-1}\ \alpha)\ N_y$ for some outward normal vector N_y to M at y.*

Proof. Let N_y be the unit outward normal vector to M at y. Consider the map F_α from $M \times [0, \infty)$ to the exterior of M:

$$F_\alpha(y,t) = y + t\ \exp(\sqrt{-1}\ \alpha)\ N_y.$$

This map has degree one so it is onto. Assume $F_\alpha(y_1, t_1) = F_\alpha(y_2, t_2)$. Let $N_i = t_i\ N_{y_i}$ and $T_i = \sqrt{-1}\ N_i,\ i = 1, 2$. Then $y_1 + \cos\alpha\ N_1 + \sin\alpha\ T_1 = y_2 + \cos\alpha\ N_2 + \sin\alpha\ T_2$, and thus $y_2 - y_1 = \cos\alpha\ (N_1 - N_2) + \sin\alpha\ (T_1 - T_2)$.

fig. 110

Notice that $\omega(N_i, T_i) > 0$. Let v be an outward vector at y_i with normal component N. Then $\omega(v, T_i) = \omega(N, T_i)$ since T_i belongs to the kernel of ω, restricted to the tangent hyperplane at y_i. Thus $\omega(v, T_i) > 0$. Consider the vector $v = y_2 - y_1$. At the point y_2 it has the outward direction, so $\omega(v, T_2) > 0$. Likewise $\omega(v, T_1) < 0$. Subtract the latter inequality from the former to get: $\omega(v, T_2 - T_1) > 0$. Substitute the above obtained expression for v and use the fact that ω is skew-symmetric: $\cos\alpha \times \omega(N_2 - N_1, T_2 - T_1) < 0$. If $\alpha = \pi/2$, which is the case with the dual billiard relation, this is a contradiction. Otherwise $\omega(N_2 - N_1, T_2 - T_1) < 0$, which is a contradiction again since $\omega(u, \sqrt{-1}\ u) \geq 0$ for any vector u. *Q.E.D.*

Thus in the presence of symplectic structure one defines "the tangent line" to a hypersurface at a point as the characteristic line, and the characteristic half-lines to a strictly convex hypersurface foliate its exterior, just as in the plane. One defines the dual billiard transformation of the exterior of M: given a point x outside of M, find the unique point $y \in M$ such that yx is the characteristic ray to M at y, and define $T(x)$ to be the reflection of x in y. Likewise one defines the transformation T_α. Notice that T commutes with linear symplectic transformations, while T_α is invariant under a smaller group of motions.

The next problem to address concerns periodic points of the dual billiard transformation T.

Theorem 3. *Let M be a smooth closed strictly convex hypersurface in linear symplectic space, and T the corresponding dual billiard transformation. For any odd prime k there exist k-periodic orbits of T.*

Proof. We use the method developed in [Gi]. Consider the product $(\mathbf{R}^{2n})^k \times (\mathbf{R}^{2n})^k$ with the symplectic structure $\omega_1^k \ominus \omega_2^k$. Let $z_1, ..., z_k$ be coordidinates in the first, and $\overline{z_1}, ..., \overline{z_k}$ in the second factor (so that each z and $\overline{z}$ is a vector in $\mathbf{R}^{2n}$). Consider two submanifolds:

$$G^k = \{\overline{z_i} = T^{\pm 1} z_i\} \quad \text{and} \quad C = \{\overline{z_i} = z_{i+1}\}$$

where $i = 1, ..., k$ and $k+1$ is understood to be equal to 1. The intersection $G^k \cap C$ consists of k-periodic orbits of the relation T, that is of chains $z_1, ..., z_k$ such that $z_{i+1} = T^{\pm 1} z_i$.

As in Theorem 1, $(\mathbf{R}^{2n})^k \times (\mathbf{R}^{2n})^k$ is symplectomorphic to $(T^*\mathbf{R}^{2n})^k$. Let $q_1, ..., q_k$ be space and $p_1, ..., p_k$ be momenta coordinates. Then G^k is the conormal bundle to $M^k = M \times ... \times M$ (k times) $\subset (\mathbf{R}^{2n})^k$. C is a linear subspace in $(T^*\mathbf{R}^{2n})^k$, and a direct computation shows that for odd k it is the graph of the differential of the following quadratic function:

$$\Phi(q_1, ..., q_k) = \sum_{i<j} (-1)^{i+j}\ \omega(q_i, q_j),$$

where the points q_i are treated as vectors in linear space.

Thus the points of $G^k \cap C$ are critical points of the restriction of Φ on M^k. Since M^k is compact this intersection is not void. By Morse theory there are at least $k+1$ points in this intersection, and at least 2^k for M in general position. However some of the critical points correspond to "fake' orbits with $z_i = T z_{i_1},\ z_{i+1} = T^{-1} z_i$ for some i. If $\overline{q} = (q_1, ..., q_k)$ is the corresponding critical point of Φ then $q_{i-1} = q_i$. Consider a tangent vector ξ to M^k at $\overline{q}$ all of whose components vanish, except for $(i-1)$-th and i-th, and these two are equal to a non-zero tangent vector to M at q_i. Skew-symmetry of ω implies that $d^2\Phi(\xi) = 0$, so the second differential of Φ is not positive at "fake' periodic orbits. It follows that maxima (or minima) of Φ correspond to k-periodic orbits of T , maybe, multiple ones. Since k is prime this multiplicity equals one. *Q.E.D.*

We would like to make several comments on this theorem. First, it seems plausible that the dual billiard transformation has k-periodic orbits for every $k \geq 3$. There are two difficulties in proving it. On the one hand, if k is even, the above generating finction for the linear subspace C does not work because the projection of C on the diagonal has a kernel. This difficulty can be overcome by applying the symplectic reduction or by a small perturbation of the projection, as in [Gi]. On the other hand, if k is not prime one should be able to distinguish between "faithful" orbits and multiple orbits of smaller periods. This is a common problem in Morse theory. Another question of interest is whether one can improve the result even for prime k: are there more than just one periodic orbit?

Secondly, the functional Φ in the plane case is different from the area functional on circumscribed polygons. For example, 5-periodic points correspond to extrema of Φ shown in the next figure. Similarly in the billiard problem there exists a functional other than perimeter, responsible for periodic trajectories, which might prove useful.

fig. 111

Another remark concerns the relation between periodic orbits of the dual billiard transformation and closed characteristic lines on M. In a sense, the dual billiard map is a discretization of the characteristic flow on M. It is known that M always possesses a closed characteristic (see, e.g. [Sik]), and it is conjectured that there are at least n of them ($2n$ being the dimension of

the ambient space). The question is whether one can use periodic trajectories of the dual billiard map to approximate closed characteristics. More specifically, does there exist a sequence of such periodic trajectories with the periods going to infinity that accumulates to any closed characteristic? Closed characteristics are extrema of the symplectic area functional on closed curves in M (that is, the integral of the Liouville 1-form over a curve or the integral of the symplectic form over a film, bounded by the curve). Let $\gamma : [0,1] \to \mathbf{R}^{2n}$ be a smooth curve; partition it by the points $q_i = \gamma(i/(2N+1),\ i = 0, ..., 2N$. One expects the function $\Phi(q_0, ..., q_{2N})$ to approximate the symplectic area of a film, bounded by γ.

Lemma 4. $\Phi(q_0, ..., q_{2N}) \to \int_0^1\ \omega(\gamma(t), \gamma'(t))\ dt\ \ as\ \ N \to \infty$.

The proof is a straightforward estimation, and we omit it.

As in the plane case, the dual billiard transformation associated to an ellipsoid is integrable. Up to a linear symplectic map, any ellipsoid is given by the equation $\sum a_i\ (x_i^2 + y_i^2) = 1$ where (x_i, y_i) are Darboux coordinates. Let $x_i = r_i \cos\alpha_i,\ y_i = r_i \sin\alpha_i$. In these polar coordinates the dual billiard map acts as follows: $T(r_i, \alpha_i) = (r_i, \beta_i)$ where $\beta_i = \alpha_i + \tan^{-1}(a_i/t)$ and t is the root of the equation

$$\sum_i \frac{a_i r_i^2}{1 + r_i^2/t^2} = 1.$$

Thus the radii r_i are integrals. Are there other integrable dual billiards in higher dimensions?

An interesting question is whether orbits of the dual billiard transformation may escape to infinity or "fall" on the hypersurface M. In the plane, KAM theory provided invariant curves which prohibit both events. In higher-dimensional situation, even if KAM theory is applicable, it would provide invariant tori which no more separate the space, and therefore do not ensure stability. Moreover, for a generic hypersurface M the dual billiard map is not close to an integrable one either near M or at infinity. At a vicinity of M its trajectories are approximated by the leaves of the characteristic foliation on M, and at infinity - by those of the characteristic foliation of another hypersurface, constructed from M by symmetrization and polar dualization as in Section 4.2 (we will not work it out in detail here).

In conclusion we mention a higher-dimensional analog of polygonal dual billiards, namely, dual billiards associated to convex polyhedra. Identify $\mathbf{R^{2n}}$ with $\mathbf{C^n}$ and let J be the multiplication by $\sqrt{-1}$. Let P be the boundary of a convex polyhedron. Given a point $x \in P$ let C_x be its dual cone, that is the convex hull of the outward normal vectors to $(2n-1)$-dimensional faces of P that contain x. If x belongs to the interior of a k-dimensional face then C_x is $(2n-k)$-dimensional. The exterior of P is the union of these dual cones. Rotate each cone C_x by J about x; an argument similar to that of Lemma 2, shows that the union of $JC_x,\ x \in P$, covers the exterior of P.

Consider a face F such that the restriction of ω to it is degenerate. Then the intersection of the rotated space JF (about one of its points) with the normal complement of F (at this point) is not trivial. Hence $dim\ (\cup_{x\in F}\ JC_x) < 2n$, and this union is contained in the closures of other cones. We ignore such faces, in particular, odd-dimensional ones. Consider the prisms $\cup_{x\in F}\ JC_x$

for the faces F such that the restriction of ω to F is nondegenerate. Their interiors are disjoint and their union V over all such faces covers the exterior of P except for a codimension one set. We define the dual billiard transformation of V: given a point $x \in V$, find the unique point $y \in P$ such that $x \in JC_y$, and reflect x in y. Therefore the dual billiard map can be thought of as reflecting in even-dimensional faces of a convex polyhedron in linear symplectic space. The study of this map is an interesting problem. In particular, for which polyhedra are all the orbits bounded or finite?

5. Hyperbolic Billiards

The last chapter concerns chaotic billiards. We start with a brief introduction to hyperbolicity; we consider two examples: a torus automorpism and the geodesic flow on a surface of constant negative curvature. Section 2 is devoted to dispersing billiards; the main mechanism of ergodicity is the divergence of nearby trajectories. Section 3 deals with chaotic billiards with convex arcs. We start with the famous "stadium" and proceed to several sufficient geometrical conditions for hyperbolicity. The last section contains a discussion on Bolzmann's Hypothesis, an estimate of the growth rate of periodic trajectories in dispersing billiards and an application of Poincaré's recurrence theorem to the billiard under a curve lower-asymptotic to the horizontal axis.

5.1 Introducing Hyperbolicity

This section is a very sketchy introduction to hyperbolic dynamics, whose particular case – hyperbolic billiards – is the topic of this chapter. Instead of formulating results in their generality we look at two examples (see [Si 1, Ni, Man]) , in which the situation appears particularly clear, and which serve as models for numerous more involved results concerning hyperbolic billiards.

Let A be an automorphism of a two-dimensional torus $\mathbf{T}^2 = \mathbf{R}^2/\mathbf{Z}^2$. Lifted to $\mathbf{R}^2$, the linear transformation A is given by a 2×2 matrix. To act on the torus it must preserve the integer lattice, so all the entries are integers; and being a diffeomorphism, the determinant equals ± 1. Assume that A has real eigenvalues λ_1 and λ_2 with $|\lambda_1| > 1,\ |\lambda_2| < 1$. Denote the corresponding eigenspaces by E^u and E^s ("u" for unstable and "s" for stable).

Observe that the slopes of both spaces are irrational. Indeed, otherwise this space will project to the torus as a closed circle. The diffeomorphism A, restricted to this circle, would be either an expansion by $|\lambda_1|$ or a contraction by $|\lambda_2|$, which is impossible since a circle is compact. Thus the projections of E^u and E^s to the torus are dense immersed lines. By a parallel translation of the eigenspaces in $\mathbf{R}^2$ and then projecting to the torus one obtains two transversal foliations W^u and W^s therein, invariant under A.

fig. 112

Equivalently, the leaf of W^s through a point x consists of points y, asymptotic to x;

$$W^s = \{y \mid dist\ (A^n(x), A^n(y)) \to 0 \ \text{ as } \ n \to \infty\}.$$

Likewise,

$$W^u = \{y \mid dist\ (A^n(x), A^n(y)) \to 0 \ \text{ as } \ n \to -\infty\}.$$

We claim that periodic points of A are dense in the torus. Indeed, any point with rational coordinates is periodic. Let d be the least common denominator of the two coordinates of a rational point x. Consider the entries of the matrix A modulo d. Since $\mathbf{GL}(2, \mathbf{Z}_d)$ is a finite group $A^n = Id\ mod\ d$ for some n. Hence $A^n(x) - x \in \mathbf{Z}^2$, which means that x is a periodic point.

Associate a partition of the torus into two parallelograms with A. To fix ideas, let

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix},$$

so that $\lambda_1 = \frac{1+\sqrt{5}}{2},\ \lambda_2 = \frac{1-\sqrt{5}}{2}$.

fig. 113

The edges of the two rectangles lie on the expanding and contracting eigenlines of A. The image of the first rectangle is contained in the second one, while the image of the second one

intersects both. One can encode the orbit of a point x by a sequence of symbols 1 and 2, according as $A^n(x)$ belongs to the first or to the second rectangle. In this encoding no two consecutive 1's appear; otherwise any combination of the symbols is possible (we ignore the ambiguity over which symbol to attach to a point on a side of a rectangle). In particular, one easily constructs a sequence that contains all blocks of 1's and 2's with no two consecutive 1's; to this sequence a dense orbit of A corresponds. This symbolic description also makes it possible to compute the number of periodic orbits of period n, which amounts to counting the number of n-length segments of 1's and 2's with no two consecutive 1's. This number grows exponentially in n.

This partition of the torus and the corresponding symbolism is a particular case of a general technique, introduced by Adler and Weiss and by Sinai in the 60-s, called the method of Markov partitions (see [Bow, Si ?]).

Our second example is the geodesic flow on a closed surface of constant negative curvature. The universal cover of such a surface is the hyperbolic plane, which we consider in Poincaré's upper half-plane model. The metric is given by $(dx^2 + dy^2)/y^2$, and straight lines are represented by circles or straight lines perpendicular to the x-axis, which is the line at infinity. Consider a one-parameter family of asymptotic lines passing through one point at infinity. Applying a hyperbolic motion we may assume that it consists of parallel vertical lines.

fig. 114

Consider two points that move along two asymptotic geodesics with unit velocities, starting at height 1. The distance between the points along horizontal line at time t equals

$$\int_{x_1}^{x_2} \frac{dx}{y} = \frac{x_2 - x_1}{y}, \quad \text{while} \quad t = \int_1^y \frac{du}{u} = \log y.$$

Hence this distance decreases as e^{-t}.

Horizontal lines are perpendicular to the family of vertical asymptotic geodesics. In general, curves perpendicular to asymptotic geodesics are circles, that are tangent to the x-axis at a fixed point. They are called horocycles. Given a unit tangent vector v in the hyperbolic plane there are two horocycles, perpendicular to it, which pass through its foot point. Extend v to a unit normal vector fields along these horocycles; these framed horocycles are two curves in the unit tangent bundle of the hyperbolic plane, which pass through v and which are transversal to the geodesic flow. Thus the unit tangent bundle carries two one-dimensional horocyclic foliations. It follows from the above considerations that the geodesic flow exponentially contracts the leaves of one foliation and exponentially expands the leaves of the other (this can be also deduced from the formulas found in Section 3.5).

fig. 115

Projecting to the unit tangent bundle of the compact surface one obtains the stable and

unstable one-dimensional foliations W^s and W^u therein. These foliations are analogous to the ones from the previous example. The existence of the two foliations makes it possible to prove the following theorem due to Hopf and Hedlund.

Theorem 1. *The geodesic flow on a compact surface of constant negative curvature is ergodic.*

Proof. Let g_t be the geodesic flow. Given a continuous function $f(x)$, by Birkhoff's ergodic theorem

$$\lim_{T\to+\infty}\frac{1}{T}\int_0^T f(g_tx) = \overline{f_+}(x), \quad \lim_{T\to-\infty}\frac{1}{T}\int_0^T f(g_tx) = \overline{f_-}(x)$$

almost everywhere, and $\overline{f_+}(x) = \overline{f_-}(x)$ on a set of full measure. One wants to prove that $\overline{f_\pm}(x)$ are almost everywhere constant. It is enough to prove this locally, that is in a neighbourhood of a point.

Since f is uniformly continuous and since points of a leaf of the stable foliation are asymptotic in positive time with respect to the flow, $\overline{f_+}(x)$ exists and has a constant value on such a leaf, provided it exists at one of its points. Likewise, $\overline{f_-}(x)$ is constant on a leaf of the unstable foliation. Since $\overline{f_+}(x)$ is g_t-invariant, it is constant on a two-dimensional surface which is the union of the images of a stable curve under the flow. One can choose a stable curve so that $\overline{f_+}(x) = \overline{f_-}(x)$ almost everywhere on it. Then $\overline{f_+}(x) = \overline{f_-}(x) = const$ on a subset U of full measure of the above constructed surface. Finally, the union of unstable curves through the points of U has full measure and $\overline{f_+}(x) = \overline{f_-}(x) = const$ thein.

Said otherwise, two points belong to the same ergodic component if they can be connected by a broken line that consists of stable curves, unstable curves and trajectories of the geodesic flow (such broken lines are called Hopf chains). Any two points of a neighbourhood in the unit tangent bundle can be connected by a Hopf chain, so the geodesic flow is ergodic. *Q.E.D.*

The study of geodesic flows on Riemannian manifolds of negative curvature goes back to Hadamard, Morse, Hopf and Hedlund. The situation with a compact manifold of negative curvature is similar to the one with a surface of constant negative curvature: almost all trajectories of the geodesic flow are dense; the flow is ergodic and mixing; periodic trajectories are dense, and the number of closed geodesics of length not greater than l is finite and grows exponentially with l (see [An]).

Thus the dynamics in the two examples enjoys similar stochastic properties. The basic reason is the existence of a decomposition of the tangent space at each point into contracting and expanding subspaces, invariant under the dynamics. Let F be a measure-preserving diffeomorphism of a compact Riemannian manifold M (the measure being equivalent to the Riemannian volume). By the Oseledets multiplicative ergodic theorem ([Os]) at almost every point $x \in M$ there exists a DF-invariant measurable decomposition $T_xM = \oplus H_i(x)$ such that

$$\lim_{n\to\pm\infty}\frac{1}{n}\log\,||DF_x^n(v)|| = \chi_i(x)$$

uniformly over unit vectors $v \in H_i(x)$. The numbers $\chi_i(x)$ are called Lyapunov exponents. The reader recognizes the spaces $H_i(x)$ as the stable and unstable eigenspaces of a torus automorphism,

and the Lyapunov exponents as the corresponding eigenvalues. The diffeomorphism F is called hyperbolic if none of the Lyapunov exponents vanishes (similar definitions are given for vector fields, in which case the direction of the field necessarily contributes a zero exponent).

A general theory of hyperbolic systems was constructed by Ya. Pesin (see [Pes 1, 2; Pol]). A particular result of this theory is the following Pesin's formula that expresses metric entropy in terms of the Lyapunov exponents.

Theorem 2.

$$h(F) = \int_M (\sum \chi_i(x)\ dim H_i\ (x))\ d\mu,$$

where the summation is taken over positive Lyapunov exponents.

Another expression for the entropy is

$$h(F) = \int_M \log\ det\ (DF^u_x)\ d\ \mu,$$

where DF^u is the restriction of the differential of F on the unstable subspace $E^u = \oplus H_i(x)$ over $\chi_i(x) > 0$.

Since the billiard transformation has singularities Pesin's theory is not directly applicable to it. A relevant generalization of this theory to hyperbolic systems with singularities is contained in the book by A. Katok and J.-M. Strelcyn [K-S].

5.2 Dispersing and Semi-Dispersing Billiards

Before we start our disscusion of hyperbolic billiards let us refer to three surveys of this topic: [Bu 1, Si 3, G-C].

The first class of stochastic billiards had been discovered by Ya. Sinai (see [Si 4]). These billiards are bounded by piecewise smooth curves whose smooth components are strictly convex inwards, and which intersect transversally. The following figure shows three examples: a simply connected plane billiard, a plane billiard with a hole and a billiard on a torus with a hole.

fig. 116

A parallel beam of rays, after a reflection in a side, becomes dispersing. For this reason these billiards are called dispersing (they are also known as Sinai's billiards). Each consecutive reflection forces the beam to further diverge. Reversing the time direction one obtains a beam that exponentially converges. This situation is typical for hyperbolic dynamics. Dispersing billiards also exist in higher-dimensional spaces; they are bounded by hypersurfaces strictly convex inwards.

fig. 117

An intuitive explanation why one may expect chaotic dynamics in dispersing billiards was proposed by V. Arnold (see [Ar 2]). Consider the billiard on a torus with a hole. Its double

(see Section 1.1) is a sphere with two handles with a Riemannian metric of negative curvature (supported in a neighbourhood of the hole). The billiard flow is the limit of the geodesic flow on this surface, which enjoys stochastic properties; thus one expects the billiard flow to be stochastic as well.

Recently I. Babenko ([Bab 2]) analyzed this construction rigorously. At a neighbourhood of the hole, the metric of the surface is $(1 + |u|\ k(s))^2\ ds^2 + du^2$, where s is the length parameter on the boundary curve, $k(s)$ is its curvature, and $|u|$ is the distance to the curve. The difficulty arises from the fact that this metric is not smooth along the curve. One lifts the billiard flow to the universal cover which is the hyperbolic plane, and studies "reachable sets" on the circle at infinity, that is sets which can be reached from a given set by moving along billiard trajectories. These sets are Cantor-like, and the topological entropy of the original billiard is estimated in terms of their Hausdorff dimensions.

Back to dispersing billiards. Sinai constructed local stable and unstable curves through almost all points of the phase space, i.e., curves whose points are exponentially asymptotic for positive or negative time. We remark that these curves cannot be too long, due to singularities. Namely a stable or unstable curve breaks when the trajectory of one of its points hits a corner or becomes tangent to the boundary.

fig. 118

Consider a one-parameter family of rays that undergo reflections at the same boundary components. Let γ be a curve perpendicular to the rays (a wave front). Being framed by the unit vectors of the rays it determines a curve $\overline{\gamma}$ in the unit tangent bundle of the plane. We will investigate how the curvature of γ changes under the billiard flow g_t. For $x \in \gamma$ denote by $\chi(x)$ the curvature of γ at x. Consider the evolution of the front until the first reflection in the boundary. Let t_1 be the distance to the boundary from x along the ray through it, and let χ_1^- be the curvature of $g_{t_1}(\gamma)$ at the point $g_{t_1}(x)$ immediately before reflection. Approximating γ by its osculating circle at x one finds:

$$\frac{1}{\chi(x)} - \frac{1}{\chi_1^-} = t_1.$$

fig. 119

Consider now how the curvature of a front changes under reflection. Let k_1 be the (positive) curvature of the boundary at the point where the ray through x hits it, ϕ_1 be the angle between this ray and the boundary, and χ_1^+ be the curvature of the front immediately after reflection. In this setting the mirror equation from Lemma 2.4.2 reads as follows:

$$\chi_1^- - \chi_1^+ = \frac{2k_1}{\cos\phi_1}.$$

Combining the two equations we find:

$$\chi(x) = \cfrac{1}{t_1 + \cfrac{1}{\cfrac{2k_1}{\cos\phi_1} + \chi_1^+}}.$$

Iterating reflections we get

$$\chi(x) = \cfrac{1}{t_1 + \cfrac{1}{\cfrac{2k_1}{\cos\phi_1} + \cfrac{1}{t_2 + \cfrac{1}{\cfrac{2k_2}{\cos\phi_2} + \chi_2^+}}}},$$

etc.

Suppose that the rays constitute a local stable curve, so that the rays become asymptotically parallel. Then $\chi_i^+ \to 0$ and we arrive at a continued fraction expression for the curvature of its front:

$$\chi(x) = \cfrac{1}{t_1 + \cfrac{1}{\cfrac{2k_1}{\cos\phi_1} + \cfrac{1}{t_2 + \cfrac{1}{\cfrac{2k_2}{\cos\phi_2} + \cfrac{1}{t_3 + ...}}}}}$$

All terms here are positive, and the criterion for convergence of such a continued fraction is that

$$\sum (t_i + \frac{2k_i}{\cos\phi_i}) = \infty$$

(see [Kh]). This condition holds because the trajectory of x makes a finite number of reflections at any finite time interval. Notice that χ is a function on the unit tangent bundle. We remark that the continued fraction essentially gives a solution of the Jacobi equation for a geodesic flow.

Thus a local stable curve $\overline{\gamma}^s$ through a point of the phase space is uniquely determined by the curvature of γ, given by the above continued fraction. Reversing the framing of γ one constructs an unstable curve $\overline{\gamma}^u$. A similar argument applies to the higher-dimensional case in which an operator-valued continued fraction gives the second quadratic form $B(x)$ of a local stable manifold.

A detailed analysis shows that local stable and unstable curves pass through almost every point of the phase space. Thus dispersing billiards are hyperbolic dynamical systems.

To evaluate the metric entropy of a dispersing billiard one uses the second formula for entropy from the previous section. Let V be the phase space consisting of unit inward vectors with the footpoints on the boundary, with the coordinates (α, t) introduced in Section 1.2; $T : V \to V$ be the billiard map, and $\nu = \sin\alpha \; d\alpha \; dt$ – the invariant measure. Given a unit tangent vector $x \in V$ denote by $t(x)$ the length of its free path until it hits the boundary.

Theorem 1.

$$h(T) = \int_V \log\ |1 + t(x)\ \chi(x)|\ d\nu(x).$$

This formula, discovered by Sinai ([Si 4]), holds not only for dispersing, but also for other hyperbolic billiards which we discuss in the next section (see [Ch 1, Ch-M]). The entropy of the billiard flow in the unit tangent bundle M is

$$h(g_t) = \int_M \chi(x)\ d\mu(x),$$

where μ is the Louiville measure. These formulas also have higher-dimensional analogs:

$$h(T) = \int_V \log\ det\ (I + t(x)\ B(x))\ d\nu(x),$$

$$h(g_t) = \int_M tr\ B(x)\ d\mu(x).$$

Proof. We need to evaluate the rate of expansion of the unstable direction under the billiard transformation in the Euclidean metric $dl^2 = d\alpha^2 + dt^2$. What we know is how the curvature of the front corresponding to an unstable curve changes during a free path:

$$\frac{1}{\chi} - \frac{1}{\chi_1} = t \quad \text{or} \quad \frac{\chi_1}{\chi} = 1 + t\chi.$$

fig. 120

Let $d\rho$ be the length form of the front; since curvature is inverse proportional to length

$$\frac{d\rho(Tx)}{d\rho(x)} = 1 + t(x)\ \chi(x).$$

In terms of the (α, t) coordinates $d\rho = \sin\alpha\ dt$ (see the figure). Let

$$J(x) = \frac{dl(x)}{d\rho(x)} = \frac{1}{\sin\alpha}\sqrt{1 + \frac{d\alpha^2}{dt^2}}$$

on an unstable curve. Then

$$\frac{dl(Tx)}{dl(x)} = \frac{d\rho(Tx)}{d\rho(x)}\ \frac{J(Tx)}{J(x)} = (1 + t(x)\ \chi(x))\frac{J(Tx)}{J(x)}.$$

Since the measure ν is T-invariant

$$\int_V \log\ J(tx)\ d\nu(x) = \int_V \log\ J(x)\ d\nu(x),$$

and therefore

$$\int_V \log\ \left|\frac{dl(Tx)}{dl(x)}\right|\ d\nu(x) = \int_V \log\ |1 + t(x)\ \chi(x)|\ d\nu(x).$$

Here one makes use of integrability of the function

$$\log\ (1 + \frac{d\alpha^2}{dt^2}),$$

which is proved in [Ch 1]. *Q.E.D.*

To apply the method of Hopf chains to dispersing billiards it is necessary to control the size of local stable and unstable manifolds. This was achieved by Sinai ([Si 4]), and the result is known as the Fundamental Theorem of the theory of hyperbolic billiards. We formulate it in a slightly weakened form.

Theorem 2. *If the trajectory of a phase point of a dispersing billiard is never tangent to the boundary nor hits a corner then, given a local unstable curve of length l in a neighbourhood of this point, the probability to find a local stable curve of length Cl, where C is an arbitrary constant, tends to 1 as the radius of the neighbourhood goes to zero. The statement holds true if one interchanges the words "stable" and "unstable".*

A consequence of the Fundamental Theorem is the following result.

Theorem 3. *A dispersing billiard is ergodic.*

Other proofs and modifications of these theorems are found in [Bu-Si 1, Si 5, Ch-Si]. It follows from Theorem 3 and hyperbolicity that the billiard transformation is mixing and even has K-property. A two-dimensional dispersing billiard is also isomorphic to a Bernoulli shift ([G-O]).

Some of these results hold for semi-dispersing billiards, which are billiards whose boundary components have non-negative inward curvature.

fig. 121

Straight components of the boundary are called neutral; unlike dispersing ones they do not change converging properties of incoming trajectories: after a reflection a parallel beam remains parallel. An example of a semi-dispersing billiard is a three-dimensional torus with two deleted non-parallel cylinders, studied in [KSS 1]; we will discuss its physical interpretation in Section 4.

Stable and unstable manifolds for semi-dispersing billiards are constructed in [Ch 2]; unlike the case of dispersing billiards their dimensions may be less than one half of the dimension of the phase space. A version of the Fundamental Theorem for semi-dispersing billiards is given in [KSS 2], see also [Ch 3].

As the example of a torus automorphism suggests, Markov partitions are very useful in the study of hyperbolic dynamics. Markov partitions for dispersing billiards were first constructed in [Bu-Si 2], and further improved in [B-C-S 1], see also [Kr-Tr] and [Ch 4]. The construction is much more delicate than that for torus automorphisms. An element of a Markov partition is a "rectangle" consisting of intersections of local stable curves through a subset of a given unstable curve and local unstable curves through a subset of a given stable curve. These sets are typically

Cantor-like, and the number of elements in a partition is necessarily countable, due to singularities of the billiard transformation. Using Markov partitions Bunimovich, Chernov and Sinai proved the following theorem.

Theorem 4. *Periodic orbits of a two-dimensional hyperbolic billiard are dense in the phase space. The number of periodic orbits of period not greater than N is bounded bellow by e^{CN} for some constant C and sufficiently great N.*

The same authors also used Markov partitions to study strong statistical properties of two-dimensional hyperbolic billiards, such as the central limit theorem and time decay of correlation functions – see [Bu-Si 3, B-C-S 2].

5.3 Hyperbolic Billiards with Focusing Arcs

Chaotic behavior in billiards discussed in the previous section was due to scattering of trajectories upon reflection in the boundary. Quite a different chaos producing mechanism was discovered by L. Bunimovich in 1974 ([Bu 2]), who demonstrated that there exist convex ergodic billiards. His billiards are bounded by arcs of circles (focusing components of the boundary) and straight segments (neutral components), with the condition that the circle, containing a focusing component, lies strictly inside the billiard table. Some examples are shown in the figure.

fig. 122

The first billiard is the famous "stadium" bounded by two half-circles and two segments; it is a C^1-smooth curve. The last billiard does not belong to the described class, but the dynamics in it reduces to that in its double – the union of the domain with the one symmetric to it with respect to the segment; this double already satisfies the above condition.

Since the pioneering work by Bunimovich the conditions on focusing components, that ensure hyperbolic dynamics, had been relaxed in the works of many authors, and up to now several constructions of such billiards are known. In all these examples the billiard curves are at most C^1-smooth. A convex ergodic billiard cannot be too smooth in view of Lazutkin's theorem (Section 2.8); the least smoothness prohibiting ergodicity is not known. A numerical study of a transformation of the billiard inside a circle to the one inside a stadium is found in [B-S].

A heuristic explanation of how focusing arcs may contribute to hyperbolicity goes as follows. Consider an incoming infinitesimal beam of parallel rays. After a reflection it focuses, and then, before the next reflection, it may defocus. If the time between focusing and the next reflection exceeds that between the previous reflection and focusing, that is if the beam spends more time diverging than converging, then close trajectories will diverge in the phase space, similarly to dispersing billiards.

fig. 123

Bunimovich applied the technique involving continued fractions to billiards with focusing arcs. An obvious difficulty in proving that such continued fractions converge is that not all their terms are positive any more. Bunimovich proved the following result concerning the class of billiards described at the beginning of the section (as well as some other billiards with focusing arcs – see [Bu 2, 3, 4]).

Theorem 1. *The billiard transformation is ergodic and isomorphic to a Bernoulli shift.*

Here we describe an approach due to M. Wojtkowski ([Wo 2]). It was further developed by R. Markarian ([Mar 1, 2]), V. Donnay ([Don 2]); see also [Ch-M].

Let T be the billiard transformation of the phase space V consisting of unit inward tangent vectors with the footpoints on the boundary. Being a two-dimensional area preserving transformation T has two Lyapunov exponents $\lambda_+ \geq 0$ and $\lambda_- \leq 0$ with $\lambda_-(x) = -\lambda_+(x)$. Let $C(x),\ x \in V$ be a measurable field of closed sectors in the tangent bundle of V, i.e. $C(x)$ is a closed sector in T_xV defined for almost all x and depending on x measurably. We say that T preserves $C(x)$ if $DT(C(x)) \subset C(Tx)$ almost everywhere; T strictly preserves $C(x)$ if the inclusion is strict almost everywhere; and T eventually strictly preserves $C(x)$ if T preserves $C(x)$ and for almost every x there exists a positive integer $n(x)$ such that $DT^{n(x)}(C(x))$ is strictly inside $C(T^{n(x)}x)$.

The next statement is Wojtkowski's projective criterion for hyperbolicity that holds for piecewise differentiable invertible area preserving mappings ([Wo 2, 3]).

Theorem 2. *If there exists a measurable field of sectors eventually strictly preserved by T then the Lyapunov exponent λ_+ is almost everywhere positive.*

To be able to apply this criterion to billiards we need to describe the action of the differential of the billiard transformation on the projectivized tangent bundle $\mathbf{P}TV$. Let $x = (t, v) \in V,\ t$ is a point on a billiard curve $\gamma,\ v$ – a unit vector, and let r be the corresponding ray. A tangent vector $\nu \in T_xV$ determines an infinitesimal family of rays that includes r. Let f_0 be the signed distance from t to the point on r in which this family focuses in the linear approximation; the sign is chosen according to the direction of r. Since f_0 depends on the direction of ν only, one may take it as a projective coordinate in $\mathbf{P}T_xV$.

At each nonflat point t of γ consider the disc $D(t)$ which is obtained from the osculating disc of γ at t by the dilation, centered at t, with the coefficient $1/2$. Denote by d the length of the segment of an incoming ray r inside $D(t)$. Let k be the curvature of γ at t, and θ – the angle between γ and r. Then

$$\frac{k}{\sin\theta} = (sgn\ k)\ \frac{1}{d}.$$

fig. 124

Let $Tx = y,\ x = (t, v),\ y = (t_1, v_1)$, and denote by L the distance between the points t and t_1. Given a tangent vector $\nu \in T_xV$ set $\nu_1 = DT(\nu)$, and let f_1 be the corresponding projective

coordinate in $\mathbf{P}T_yV$, i.e. the signed distance from t_1 to the point on r_1 in which the outcoming infinitesimal family of rays focuses. Notice that $f_0 - L$ is the signed distance from t_1 to the focusing point on r. The mirror equation from Section 2.4 reads:

$$\frac{1}{f_1} + \frac{1}{L - f_0} = \frac{2k_1}{\sin\theta_1} \quad \text{or} \quad \frac{1}{f_1} + \frac{1}{L - f_0} = (sgn\ k_1)\ \frac{2}{d_1}.$$

Thus

$$f_1 = \frac{(f_0 - L)\ d_1}{2(f_0 - L)\ (sgn\ k_1)\ + d_1}$$

is the projective action of the differential of the billiard map.

Now define a field of sectors $C(x)$ in TV. For a focusing (convex outward) boundary component $C(x) \subset T_xV$, $x = (t, v)$ is defined by the condition that the focusing point of the respective infinitesimal family of rays lies in $D(t)$. Equivalently, C is defined by the inequality $0 \leq f \leq d$. For a dispersing component C is defined by the condition that the focus of the infinitesimal family of rays lies on the exterior side of the boundary: $-\infty \leq f \leq 0$. Neutral components are irrelevant: one can unfold a billiard trajectory that hits such a component and proceed to another copy of the billiard table (the method familiar from Chapter 3). In other terms, one deals with a smaller phase space of unit vectors over the nonflat boundary components, the transformation being the first return map of T.

fig. 125

What are the conditions on the billiard curve ensuring the preservation of the field of sectors $C(x)$? Consider the segment of a trajectory between two consecutive reflections in nonflat boundary components γ and γ_1 in points t and t_1; let L be its length, d and d_1 the lengths of its pieces inside the discs $D(t)$ and $D(t_1)$. If γ and γ_1 are both dispersing then the field C is strictly preserved along this trajectory segment: diverging rays remain diverging. It follows that dispersing billiards are hyperbolic. If γ is focusing and γ_1 is dispersing then the condition $d < L$ implies the strict preservation of the field C. Indeed, if $d < L$ then $f_0 - L < 0$ for all $f_0 \in [0, d]$. Then the formula for the projective action implies that $f_1 < 0$. The same condition works when γ is dispersing and γ_1 is focusing. Finally, if both curves γ and γ_1 are focusing then the condition $d + d_1 < L$ implies the strict preservation of C. Indeed, in this case

$$\frac{1}{f_1} = \frac{2}{d_1} - \frac{1}{L - f_0}.$$

If $d + d_1 < L$ then $L - f_0 \geq L - d > d_1$, hence $1/f_1 > 1/d_1$, or $f_1 < d_1$. Also $2(L - f_0) - d_1 \geq 2L - 2d - d_1 > d_1 > 0$, hence $f_1 > 0$. If $d + d_1 \leq L$ then the field of cones is preserved but not necessarily strictly preserved.

fig. 126

Thus if the above conditions holds the corresponding billiard is hyperbolic. These conditions can be enforced in all cases but one by moving nonflat pieces of the boundary sufficiently far apart to make L big enough; the only remaining case is that of $\gamma = \gamma_1$ being a focusing curve. Then the condition $d + d_1 < L$, which should hold for each chord of the curve, imposes geometrical restrictions on it. We will describe such curves a little later; and now, as a matter of example, consider the billiard in a stadium.

fig. 127

Given a chord of a circle the relation $d + d_1 = L$ holds – see the figure. Therefore as long as a billiard trajectory reflects in one of the two stadium's semicircles the field of sectors is preserved but not strictly preserved. When a trajectory goes from one semicircle to another (possibly with intermediate reflections in the flat pieces) we have the inequality $d + d_1 < L$, L being the length of the trajectory segment. In such a case the field of sectors is strictly preserved. Since almost every trajectory visits both semicircles, the field of sectors is eventually strictly preserved, and the billiard is hyperbolic. Notice that ergodicity does not follow from this argument; to establish it a more involved analysis is needed.

Now we describe convex arcs satisfying Wojtkowski's condition.

Lemma 3. *The inequality $d + d_1 < L$ hold for every chord of a smooth convex arc γ if and only if its radius of curvature $r(t)$ is a strictly concave function of the length parameter t:*

$$\frac{d^2 r}{dt^2} \leq 0.$$

Proof. Choose a coordinate system as shown. Let $\phi(t)$ be the angle between the curve and the x-axis. Then

$$\frac{dx}{dt} = \cos\ \phi(t), \quad \frac{dy}{dt} = \sin\ \phi(t) \quad \text{and} \quad \frac{1}{r(t)} = \frac{d\phi}{dt}.$$

Also $d = -r(t_0)\ \sin\ \phi(t_0),\ d_1 = r(t_1)\ \sin\ \phi(t_1)$.

fig. 128

We have

$$L = \int_{t_0}^{t_1} \frac{dx(t)}{dt}\, dt = \int_{t_0}^{t_1} \cos\ \phi(t)\, dt = \int_{t_0}^{t_1} \frac{d \sin\ \phi(t)}{dt}\, r(t)\, dt$$

$$= r(t_1)\ \sin\ \phi(t_1) - r(t_0)\ \sin\ \phi(t_0) - \int_{t_0}^{t_1} \sin\ \phi(t)\, \frac{dr}{dt}\, dt.$$

Hence

$$L - d - d_1 = -\int_{t_0}^{t_1} \sin\ \phi(t)\, \frac{dr}{dt}\, dt = -\int_{t_0}^{t_1} \frac{dy(t)}{dt}\, \frac{dr}{dt}\, dt$$

$$= -y(t_1)\ \frac{dr}{dt}(t_1) + y(t_0)\ \frac{dr}{dt}(t_0) + \int_{t_0}^{t_1}\ y(t)\ \frac{d^2r}{dt^2}\ dt = \int_{t_0}^{t_1}\ y(t)\ \frac{d^2r}{dt^2}\ dt,$$

because $y(t_1) = y(t_0) = 0$. Since $y(t) < 0$ for $t \in [t_0, t_1]$ the necessity follows. If

$$\frac{d^2r}{dt^2} > 0$$

at some point t then, choosing t_0 and t_1 sufficiently close to t, one gets $L - d - d_1 < 0$. *Q.E.D.*

Notice that the condition

$$\frac{d^2r}{dt^2} < 0$$

is open in C^4 topology.

Here are some examples of the curves satisfying the condition

$$\frac{d^2r}{dt^2} \leq 0 :$$

an arc of a circle; an arc of a logarithmic spiral $r(t) = at + b$; an arc of a cycloid $r^2(t) = -a^2t^2 + b^2$; an arc of an ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \quad a < b$$

on which $|x| \leq a/\sqrt{2}$.

Wojtkowski formulated the following principles for design of hyperbolic billiards:

-any focusing component of the boundary should satisfy the inequality

$$\frac{d^2r}{dt^2} < 0;$$

-any focusing component should be sufficiently far away from any other component;

-if two components meet at a vertex then the internal angle between them should be greater than π if both components are focusing, not less than π if one is focusing and another dispersing, and greater than $\pi/2$ if one is focusing and another neutral.

(The last condition deals with reflections near a vertex.)

Bunimovich billiards satisfy these conditions. Here are some other examples.

fig. 129

The first curve is the cardioid; the second is an elliptic stadium whose curves are great arcs of an ellipse. The third is a unit square with a hole in the shape of an astroid $|x|^{2/3} + |y|^{2/3} = a^{2/3}$. If $a \leq \sqrt{2}/4$ this billiard is hyperbolic.

A more general class of convex curves that may be used to build hyperbolic billiards are absolutely focusing arcs. These are C^4-smooth convex curves whose total rotation (i.e. the integral of curvature) does not exceed π, and such that any incoming infinitesimal parallel beam focuses between each two consecutive reflections and focuses again after the last reflection (see [Bu 5, Don

2, Ch-M]). Bunimovich proved that a C^4-small perturbation of an arc preserves its property to be absolutely focusing. Donnay showed that for any smooth convex arc there exists $\alpha > 0$ such that the arc is absolutely focusing for rays that make the angle with the arc less than α. This implies that any sufficiently short smooth strictly convex arc is absolutely focusing.

fig. 130

Arcs satisfying Wojtkowski's condition $d + d_1 < L$ are absolutely focusing. So are arcs for which Markarian's condition holds: $d_1(L + L_1) < L_1 L$ (see [Mar 1]). A half-ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \quad x \geq 0$$

is absolutely focusing if and only if $a/b < \sqrt{2}$ ([Don 2, Bu 5]).

One may use absolutely focusing arcs instead of the ones satisfying Wojtkowski's condition, and construct hyperbolic billiards according to the above principles. An example is an elliptic stadium whose curves are small arcs of an ellipse; unlike the previous elliptic stadium here

$$\frac{d^2 r}{dt^2} > 0.$$

fig. 131

In conclusion we mention a possibility to design higher-dimensional hyperbolic billiards with focusing boundary components as discussed in [Bu 6]. However a serious difficulty arises in this case: linearly stable billiard trajectories cannot be destroyed by merely moving focusing pieces of the boundary far apart (see [Wo 4]).

5.4 Miscellanea

We saw in Section 3.6 that the dynamics of point-masses in the line reduces to a billiard in a polyhedral angle or a polyhedron. A more realistic physical model for gas deals with elastic balls, say n identical unit balls in 3-space. The configuration space S of this system is the subset of $\mathbf{R}^{3n}$, corresponding to the positions of the ball's centers, in which the inequlities hold:

$$(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 \geq 4; \quad i \neq j, \quad i, j = 1, ..., n.$$

These inequalities say that the balls do not penetrate each other. Thus the space S is the complement of a union of cylinders. The dynamics of the system of balls is that of the billiard in S. Since a cylinder is a convex, but not strictly convex body, the billiard is semi-dispersing.

The space S is kind of a higher-dimensional angle with faces convex inwards. Similarly to the case of a polyhedral angle a billiard trajectory eventually escapes from S (due to G. Galperin and

L. Vaserstein - see [G-C]). Hence the number of collisions of finitely many elastic balls in space is finite. Unlike the case of point-masses in the line this number is not proved to be uniformly bounded.

A model of gas in a closed vessel is a collection of elastic balls inside a compact domain, say, a rectangular box. The famous Bolzmann's Hypothesis of statistical physics states that this system is ergodic on a constant energy level in the phase space. A slight variation of this model is a system of elastic balls in a torus. In a torus the total momentum and the center of mass are invariant; Bolzmann's Hypothesis concerns a level surface of these invariants in this case. Bolzmann's Hypothesis is not proved; the best known results are due to A. Kramli, N. Simanyi and D. Szasz ([K-S-S 3, 4]) who proved it for 3 and 4 balls in a torus of any dimension.

Another popular model of statistical physics is the Lorentz gas which describes the motion of electrons in metals. This is the billiard in a domain in $\mathbf{R}^n$ with a number of disjoint convex bodies (scatterers) removed. In particular, when the domain is the plane and the scatterers are periodically positioned identical discs one obtains the billiard in a torus with a circular hole.

L. Bunimovich, C. Liverani, A. Pellegrinotti and Yu. Sukhov constructed a class of systems with an infinite number of elastic balls in which K-property holds on each component of a constant energy manifold ([B-L-P-S]). The following figure shows one such system. The "walls" are convex inward, and there are small scatterers inside the domain. The balls are too big to sqeeze through the narrow passages between the walls, but they can collide with each other and the scatterers. After a collision between two balls each ball first hits a wall or a scatterer, and only then may collide with another ball again.

fig. 132

Another result we want to mention here is an estimate from above for the number P_n of n-periodic trajectories in a semi-dispersing billiard in any dimension, due to L. Stojanov ([Sto 2]). Let s be the number of smooth boundary components.

Theorem 1. *In a dispersing billiard*

$$P_n \leq s(s-1)^n(s-2) \quad \textit{for} \quad n \geq 3.$$

In particular,

$$\lim_{n\to\infty} \sup \frac{\log P_n}{n} \leq s-1.$$

This result complements the exponential estimate from below mentioned in Section 5.1. The proof is based on the observation that the length functional, whose critical points are periodic trajectories, is convex (in the plane it follows from a comparison of a boundary component with an appropriate ellipse – see the figure). It follows that in a dispersing billiard there cannot exist two distinct trajectories which consecutively reflect in the same boundary components, from which the theorem easily follows. In semi-despersing billiards trajectories that reflect in the same pieces of

the boundary appear in strips; each trajectory in a strip has the same length (compare to the case of polygons in Section 3.3).

fig. 133

We conclude this section with the following problem, discussed in [Ki]. A billiard table is bounded by the positive x-axis and a smooth curve $y = f(x) > 0$ which is lower-asymptotic to the x-axis, that is $\liminf_{x\to\infty} f(x) = 0$. A billiard ball is shot from a point on the y-axis in this billiard; is it possible that the ball will forever stay in the cusp of the table and never escape to the left half-plane?

fig. 134

Let V_x be the set of unit vectors with the footpoints on the vertical segment through $(x, 0)$, equiped with its usual measure $\mu = \sin\alpha \; d\alpha \; dt$ from Section 1.2. We introduce the escape-set $E \subset V_0$ which consists of phase points that never return to V_0. The first observation one makes concerns the case when the area below the curve is finite. Since the billiard flow is measure preserving it follows from Poincaré's recurrence theorem that E is a nullset.

If, in addition, the curve $y = f(x)$ is convex then the escape-set is actually empty. Acute angles made by a billiard trajectory with the x-axis increase with each bounce in this case ($\beta > \alpha$ in figure 135). Hence the trajectory of a point in E must monotonically go to the right. Given $v \in E$ consider another vector w which makes a smaller angle with the horizontal direction, and whose footpoint is to the right of the footpoint of v. After a reflection in the upper and then in the lower boundaries the ball w is still to the right of v and still makes a smaller angle with the x-axis. Iterating we see that w escapes as well. Hence E cannot be a nullset unless it is void.

fig. 135

Finally let the table have infinite area. We claim that E is still a nullset. Assume that it has a positive measure. By dropping to a positive measure subset we may assume that for all $v \in E$

$$\limsup_{t\to\infty} \; x\text{-coordinate of } \gamma_t(v) = +\infty,$$

where γ_t is the billiard flow. Indeed, otherwise there will be a subset U of positive measure in E such that the trajectories of its points stay in a bounded part of the billiard table to the left of some vertical line; as before, one may apply Poincaré's recurrence theorem to U. Thus for any $x > 0$ each point $v \in E$ will eventually reach V_x, i.e. $\gamma_t(v) \in V_x$ for some $t > 0$. Let $T : E \to V_x$ be the corresponding billiard transformation which sends a vector to the first point of V_x on its forward trajectory. Since T is measure preserving $\mu\,(V_x) \geq \mu\,(E)$. However the curve $y = f(x)$ is lower-asymptotic, hence $\mu\,(V_x)$ can be made arbitrarily small by choosing x big enough. This is a contradiction. As J. King puts it, one cannot squeeze a gallon into a bottle with a pint-sized neck.

fig. 136

It follows that in a billiard, bounded by a curve lower-asymptotic in both directions, almost every trajectory which starts at the y-axis will visit it again infinitely many times. Such billiards, which are models for the motion of a charge in an electro-magnetic field of a special configuration, were studied by A. Leontovich in early 60-s.

References

[Am] E. Amiran. Caustics and Evolutes for Convex Planar Domains. J. Diff. Geom., 28, (1988), 345-357.

[A-M] K. Anderson, R. Melrose. The Propagation of Singularities along Gliding Rays. Invent. Math., 41, (1977), 23-95.

[An] D. Anosov. Geodesic Flows on Closed Riemannian Manifolds with Negative Curvature. Proc. Steklov Inst. Math., 90, (1967).

[Ar 1] V. Arnold. Symplectic Geometry and Topology. Ceremade Preprint No 9410, 1994.

[Ar 2] V. Arnold. Mathematical Methods of Classical Mechanics. Springer-Verlag, 1978.

[Ar 3] V. Arnold. Contact Geometry and Wave Propagation. Monogr. 34 de l'Enseign. Math.

[Ar 4] V. Arnold. First Steps of Symplectic Topology. Russ. Math. Surv., 41, No 6, (1986), 1-21.

[Ar 5] V. Arnold. Ordinary Differential Equations. Springer-Verlag, 1992.

[A-G] V. Arnold, A. Givental. Symplectic Geometry, 1-136. Encycl. of Math. Sci., Dynamical Systems, 4, Springer-Verlag, 1990.

[A-K-N] V. Arnold, V. Kozlov, A. Neishtadt. Mathematical Aspects of Classical and Celestial Mechanics. Encycl. of Math. Sci., Dynamical Systems, 3, Springer-Verlag, 1988.

[Arn] P. Arnoux. Ergodicité Générique des Billards Polygonaux. Sém. Bourbaki, No 696, (1987-88).

[A-M-S-T] P. Arnoux, C. Mauduit, I. Shiokawa, J.-I. Tamura. Complexity of Sequences Defined by Billiards in the Cube. Bull. SMF, 122, F.1, (1994), 1-12.

[Bab 1] I. Babenko. Periodic Trajectories in Three-Dimensional Birkhoff Billiards. Math. of USSR, Sbornik, 71, (1992), 1-13.

[Bab 2] I. Babenko. On the Behavior of Trajectories of Scattering Billiards on the Flat Torus. Math. of USSR, Sbornik, 72, (1992), 207-220.

[Ban] V. Bangert. Mather Sets for Twist Maps and Geodesics on Tori. Dynamics Reported, 1, (1988), 1-56.

[Bar] Yu. Baryshnikov. Indices for Extremal Embeddings of 1-Complexes. Adv. in Soviet Math., 1, 137-144, Amer. Math. Soc., 1990.

[Ben] D. Bennequin. Topologie Symplectique, Convexité Holomorphe et Structure de Contact. Sém. Bourbaki, No 725, 1990.

[B-S] G. Benettin, J.-M. Strelcyn. Numerical Experiments on a Billiard. Stochastic Transition and Entropy. Phys. Rev., A 17, (1978), 773-786.

[Be 1] M. Berger. Geometry. Springer-Verlag, 1987.

[Be 2] M. Berger. Sur les Caustiques de Surfaces en Dimension 3. C. R. Acad. Sci., 311, (1990), 333-336.

[Be 3] M. Berger. Unpublished Manuscript.

[Be 4] M. Berger. La Mathématique du Billard. Pour la Science, 163, Mai 1991, 76-85.

[Bia] M. Bialy. Convex Billiards and a Theorem by E. Hopf. Math. Zeitschrift, 214, (1993), 147-154.

[B-P] M. Bialy, L. Polterovich. Hamiltonian Systems, Lagrangian Tori and Birkhoff's Theorem. Math. Ann., 292, (1992), 619-627.

[Bi 1] G. Birkhoff. Dynamical Systems. Amer. Math. Soc. Colloquium Publ., 9, 1927.

[Bi 2] G. Birkhoff. Surface Transformations and Their Dynamical Applications. Acta Math., 43, (1922), 1-119.

[Bi 3] G. Birkhoff. Sur Quelques Courbes Fermeés Remarquables. Bull. Soc. Math. de France, 60, (1932), 1-26.

[B-K-M] C. Boldrighini, M. Keane, F. Marchetti. Billiards in Polygons. Ann. of Prob., 6, (1978), 532-540.

[Bol] S. Bolotin. Integrable Birkhoff Billiards. Moscow Univ., Vestnik, 2, (1990), 33-36.

[B-K-O-R] H. Bos, C. Kers, F. Oort, D. Raven. Poncelet's Closure Theorem. Expos. Math., 5, (1987), 289-364.

[Bos] M. Boshernitzan. Billiards and Rational Periodic Directions in Polygons. Amer. Math. Monthly, 99, (1992), 522-529.

[B-G-K-T] M. Boshernitzan, G. Galperin, T. Kruger, S. Troubetzkoy. Some Remarks on Periodic Billiard Orbits in Rational Polygons. Preprint.

[Bo] J.-B. Bost. Tores Invariants des Systèmes Dynamique Hamiltoniens. Sém. Bourbaki, No 639, 1985.

[B-G] J. Bruce, P. Giblin. Curves and Singularities. Cambridge Univ. Press, 1984.

[Br] N. de Bruijn. Sequences of Zeroes and Ones, Generated by Special Production Rules. Indag. Math., 43, (1981), 27-37.

[Bu 1] L. Bunimovich. Systems of Hyperbolic Type with Singularities. 173-203. Encycl. of Math. Sci., Dynamical Systems, 2, Springer-Verlag, 1989.

[Bu 2] L. Bunimovich. On the Ergodic Properties of Certain Billiards. Fuct. Anal. Appl., 8, (1974), 254-255.

[Bu 3] L. Bunimovich. On the Ergodic Properties of Nowhere Dispersing Billiards. Comm. Math. Phys., 65, (1979), 295-312.

[Bu 4] L. Bunimovich. A Theorem on Ergodicity of Two-Dimensional Hyperbolic Billiards. Comm. Math. Phys., 130, (1990), 599-621.

[Bu 5] L. Bunimovich. On Absolutely Focusing Mirrors. Springer Lect. Notes Math., 1514, 62-82, Springer-Verlag, 1992.

[Bu 6] L. Bunimovich. Many-Dimensional Nowhere Dispersing Billiards with Chaotic Behavior. Physica D, 33, (1988), 58-64.

[B-L-P-S] L. Bunimovich, C. Liverani, A. Pellegrinotti, Yu. Sukhov. Ergodic Systems of n Balls in a Billiard Table. Comm. Math. Phys., 146, (1992), 357-396.

[Bu-Si 1] L. Bunimovich, Ya. Sinai. The Fundamental Theorem of the Theory of Scattering Billiards. Math. USSR, Sbornik, 19, (1973), 407-423.

[Bu-Si 2] L. Bunimovich, Ya. Sinai. Markov Partitions for Dispersing Billiards. Comm. Math. Phys., 73, (1980), 247-280.

[Bu-Si 3] L. Bunimovich, Ya. Sinai. Statistical Properties of Lorentz Gas with Periodic Configuration of Scatterers. Comm. Math. Phys., 78, (1981), 479-497.

[B-C-S 1] L. Bunimovich, N. Chernov, Ya. Sinai. Markov Partitions for Two-Dimensional Hyperbolic Billiards. Russ. Math. Surv., 45, No 3, (1990), 105-152.

[B-C-S 2] L. Bunimovich, N. Chernov, Ya. Sinai. Statistical Properties of Two-Dimensional Hyperbolic Billiards. Russ. Math. Surv., 46, No 4, (1991), 47-106.

[B-Z] Yu. Burago, V. Zalgaller. Geometric Inequalities. Springer-Verlag, 1988.

[C-S] S.-J. Chang, K.-J. Shi. Billiard Systems on Quadric Surfaces and the Poncelet Theorem. J. Math. Phys., 30, (1989), 798-804.

[Che] A. Chenciner. Le Dynamique au Voisinage d'un Point Fixe Elliptique Conservatif; de Poincaré et Birkhoff à Aubry et Mather. Sém. Bourbaki, No 622, (1983-84).

[Ch 1] N. Chernov. A New Proof of Sinai's Formula for Entropy of Hyperbolic Billiards. Applications to Lorentz Gas and Bunimovich Stadium. Funct. Anal. Appl., 25, No 3, 91991), 50-69.

[Ch 2] N. Chernov. Construction of Transversal Fibers for Multidimensional Semi-Dispersing Billiards. Fuct. Anal. Appl., 16, (1982), 35-46.

[Ch 3] N. Chernov. On Local Ergodicity in Hyperbolic Systems with Singularities. Fuct. Anal. Appl., 27, No 1, 60-64.

[Ch 4] N. Chernov. Statistical Properties of the Periodic Lorentz Gas. Multidimensional Case. J. Stat. Phys., 74, No 1/2, (1994), 11-53.

[Ch-M] N. Chernov, R. Markarian. Entropy of Non-Uniformly Hyperbolic Plane Billiards. Bol. Soc. Bras. Mat., 23, No 1-2, (1992), 121-135.

[Ch-Si] N. Chernov, Ya. Sinai. Ergodic Properties of Certain Systems of Two-Dimensional Discs and Three-Dimensional Balls. Russ. Math. Surv., 42, No 3, (1987), 181-207.

[C-H-K] B. Cipra, R. Hanson, A. Kolan. Periodic Trajectories in Right Triangle Billiards. Preprint, 1994.

[Col] Y. Collin de Verdière. Sur les Longueurs des Trajectoires Périodiques d'un Billard. Géometrie Symplectique et de Contact: Autour du Théoreme de Poncaré-Birkhoff, 122-139, Travaux en Cours, Hermann, 1984.

[C-Z] C. Conley, E. Zehnder. The Birkhoff-Lewis Fixed Point Theorem and a Conjecture of V. I. Arnold. Invent. Math., 73, (1983), 33-49.

[Con] J. Connett. Trapped Reflections? Amer. Math. Monthly, 99, (1992), 178-179.

[C-F-S] I. Cornfeld, S. Fomin, Ya. Sinai. Ergodic Theory. Springer-Verlag, 1982.

[C-F-G] H. Croft, K. Falconer, R. Guy. Unsolved Problems in Geometry. Springer-Verlag, 1991.

[C-S] H. Croft, H. Swinnerton-Dyer. On the Steihaus Billiard Table Problem. Proc. Camb. Phil. Soc., 59, (1963), 37-41.

[Da] M. Day. Polygons Circumscribed about Closed Convex Curves. Trans. Amer. Math. Soc., 62, (1947), 315-319.

[D-G-S] M. Denker, Ch. Grillenberger, K. Sigmund. Ergodic Theory on Compact Spaces. Lect. Notes in Math., 527, Springer-Verlag.

[Don 1] V. Donnay. Perturbations of Elliptic Billiards. Preprint.

[Don 2] V. Donnay. Using Integrability to Produce Chaos: Billiards with Positive Entropy. Comm. Math. Phys., 141, (1991), 225-257.

[Do 1] R. Douady. Une Démonstration Directe de l'Equivalence des Théorémes de Tores Invariants pour Difféomorphismes et Champs de Vecteures. C. R. Acad. Sci., 295, (1982), 201-204.

[Do 2] R. Douady. Thése de Troisieme Cycle. Univ. Paris 7, 1982.

[D-G-K-R] B. Drizzi, B. Grammaticos, A. Kalliterakis, A. Ramani. Integrable Curvilinear Billiards. Phys. Lett. A., 115, (1986), 25-28.

[El] D. Elliott. M. L. Urquhart. J. Aust. Math. Soc., 8, (1968), 129-133.

[Fa] K. Falconer. The Geometry of Fractal Sets. Camb. Univ. Press, 1985.

[F-K] R. Fox, R. Kershner. Geodesics on a Rational Polyhedron. Duke Math. J., 2, (1936), 147-150.

[Fr] M. Frantz. A Focusing Property of the Ellipse. Amer. Math. Monthly, 101, (1994), 250-258.

[F-T] D. Fuchs, S. Tabachnikov. Segments of Equal Areas. Quantum, 2, (1992).

[Fu] H. Furstenberg. Recurrence in Ergodic Theory and Combinatorial Number Theory. Princeton Univ. Press, 1981.

[G-O] G. Gallavotti, D. Ornstein. Billiards and Bernoulli Schemes. Comm. Math. Phys., 38, (1974), 83-101.

[Ga] G. Galperin. Nonperiodic and not Everywhere Dense Billiard Trajectories in Convex Polygons and Polyhedrons. Comm. Math. Phys., 91, (1983), 187-211.

[G-C] G. Galperin, N. Chernov. Billiards and Chaos. Math. and Cybernetics, 5, (1991) (in Russian; see also a forthcoming book G. Galperin, N. Chernov, A. Zemlyakov. The Mathematics of Billiards, Cambridge Univ. Press, which combines [G-C] and [G-Z]).

[G-K-T] G. Galperin, T. Kruger, S. Troubetzkoy. Local Instability of Orbits in Polygonal and Polyhedral Billiards. Comm. Math. Phys., to appear.

[G-S-V] G. Galperin, A. Stepin, Ya. Vorobets. Periodic Billiard Trajectories in Polygons: Generating Mechanisms. Russ. Math. Surv., 47, No 3, (1992), 5-80.

[G-Z] G. Galperin, A. Zemlyakov. Mathematical Billiards. Nauka, Moscow, 1990 (in Russian).

[Ga-Zv] G. Galperin, A. Zvonkin. In preparation.

[Gi] A. Givental. Periodic Maps in Symplectic Topology. Funct. Anal. Appl., 23, No 4, (1989), 37-52.

[G-H 1] Ph. Griffiths, J. Harris. A Poncelet Theorem in Space. Comm. Math. Helv., 52, (1977), 145-160.

[G-H 2] Ph. Griffiths, J. Harris. On Cayley's Explicit Solution to Poncelet's Porism. L'Enseign. Math., 24, (1978), 31-40.

[Gro] M. Gromov. Lectures on Symplectic Topology. Univ. of Maryland, 1992.

[Gru] P. Gruber. Convex Billiards. Geom. Dedicata, 33, (1990), 205-226.

[G-M] V. Guillemin, R. Melrose. A Cohomological Invariant of Discrete Dynamical Systems. Christoffel Cent. Vol., 672-679, Birkhauser, 1981.

[G-S 1] V. Guillemin, S. Sternberg. Symplectic Technics in Physics. Cambridge Univ. Press, 1984.

[G-S 2] V. Guillemin, S. Sternberg. Geometric Asymptotics. Amer. Math. Soc. Math. Surv., 14, 1977.

[Gu 1] E. Gutkin. Billiard Tables of Constant Width and Dynamical Characterization of the Circle. Penn. State Workshop Proc., Oct. 1993.

[Gu 2] E. Gutkin. Billiards in Polygons. Physica D, 19, (1986), 311-333.

[Gu 3] E. Gutkin. Billiards on Almost Integrable Polyhedral Surfaces. Ergod. Th. and Dyn. Syst., 4, (1984), 569-584.

[Gu-H] E. Gutkin, N. Haydn. Topological Entropy of Generalized Polygon Exchanges. Preprint, 1993.

[Gu-K 1] E. Gutkin, A. Katok. Caustics for Inner and Outer Billiards. Preprint, 1993.

[Gu-K 2] E. Gutkin, A. Katok. Weakly Mixing Billiards. Springer Lect. Notes in Math., 1345, (1989), 163-176.

[Gu-Kn] E. Gutkin, O. Knill. Billiards that Share a Common Caustic. Preprint, 1994.

[Gu-S] E. Gutkin, N. Simanyi. Dual Polygonal Billiards and Necklace Dynamics. Comm. Math. Phys., 143, (1991), 431-450.

[Guy] R. Guy. A Quater Century of Monthly Unsolved Problems, 1969-1993. Amer. Math. Monthly, 100, (1993), 945-949.

[Ha] B. Halpern. Strange Billiard Table. Trans. Amer. Math. Soc., (1977), 297-305.

[He 1] M. Herman. Sur les Curbes Invariantes par les Difféomorphismes de l'Anneau. Astérisque, 103-104, 1983.

[He 2] M. Herman. Sur la Conjugaison Différentiable des Difféomorphismes du Cercle à une Rotation. Publ. Math. IHES, 49, (1979), 5-234.

[H-CV] D. Hilbert, S. Cohn-Vossen. Geometry and Imagination. Chelsea Publ., NY, 1952.

[Hu] A. Hubacher. Instability of the Boundary in the Billiard Ball Problem. Comm. Math. Phys., 108, (1987), 483-488.

[Hub] P. Hubert. Complexite de Suites Definies par des Trajectoires de Billard. To appear in Bull. SMF.

[Ig] P. Iglesias. Sur les Géodésiques qui Coupent un Convex en Courbure Ngative ou Nulle. Ann. Fac. Sci. Toulouse, ser. 6, 1, No 1, (1992), 39-42.

[Ka 1] A. Katok. Periodic and Quasi-Periodic Orbits for Twist Maps. Lect. Notes in Phys., 179, (1983), 47-65.

[Ka 2] A. Katok. The Growth Rate for the Number of Singular and Periodic Orbits for a Polygonal Billiard. Comm. Math. Phys., 111, (1987), 151-160.

[Ka 3] A. Katok. Interval Exchange Transformations and Some Special Flows are not Mixing. Isr. J. of Math., 35, (1980), 301-310.

[K-S] A. Katok, J.-M. Strelcyn. Invariant Manifolds, Entropy and Billiards; Smooth Maps with Singularities. Springer Lect. Notes in Math., 1222, Springer-Verlag, 1986.

[K-Z] A. Katok, A. Zemlyakov. Topological Transitivity of Billiards in Polygons. Math. Notes, 18, (1975), 760-764.

[Ke] M. Kean. Coding Problems in Ergodic Theory. Proc. Int. School of Math. Phys., Univ. Camerino, 1974.

[Ki] J. King. Four Problems in Search of a Measure. Preprint; to appear in Amer. Math. Monthly.

[K-M-S] S. Kerckhoff, H. Masur, J. Smillie. Ergodicity of Billiard Flows and Quadratic Differentials. Ann. of Math., 124, (1986), 293-311.

[Kh] A. Khovanski. Applications of Continued Fractions and their Generalizations to Problems of the Theory of Approximations. Groningen, Noordhoff, 1963.

[Kl] W. Klingenberg. Lectures on Closed Geodesics. Springer-Verlag, 1978.

[Ko 1] R. Kolodziej. The Rotation Number of Some Transformation Related to Billards in an Ellipse. Stud. Math., 81, (1985), 293-302.

[Ko 2] R. Kolodziej. The Antibilliard Outside a Polygon. Bull. Pol. Acad. Sci., 37, (1989), 163-168.

[K-T] V. Kozlov, D. Treshchev. Billiards. A Genetic Introduction to the Dynamics of Systems with Impacts. Translations of Math. Monographs, 98, Amer. Math. Soc., 1991.

[K-K-S 1] A. Kramli, N. Simanyi, D. Szasz. Ergodic Properties of Semi-Dispersing Billiards. Two Cylindrical Scatterers in the 3-D Torus. Nonlinearity, (1989), 311-326.

[K-K-S 2] A. Kramli, N. Simanyi, D. Szasz. A "Transversal" Fundamental Theorem for Semi-Dispersing Billiards. Comm. Math. Phys., 129, (1990), 535-560.

[K-K-S 3] A. Kramli, N. Simanyi, D. Szasz. Three Billiard Balls on the n-Dimensional Torus is a K-Flow. Ann. Math., 133, (1991), 37-72.

[K-K-S 4] A. Kramli, N. Simanyi, D. Szasz. The K-Property of Four Billiard Balls. Comm. Math. Phys., 144, (1992), 107-148.

[Kr-Tr] T. Kruger, S. Troubetzkoy. Markov Partitions and Shadowing for Non-Uniform Hyperbolic Systems with Singularities. Ergod. Th. and Dyn. Syst., 12, (1992), 487-508.

[La 1] V. Lazutkin. The Existence of Caustics for a Billiard Problem in a Convex Domain. Math. USSR, Izvestija, 7, (1973), 185-214.

[La 2] V. Lazutkin. Asymptotics of the Eigenvalues of the Laplacian and Quasimodes. A Series of Quasimodes Corresponding to a System of Caustics Close to the Boundary of the Domain. Math. USSR, Izvestija, 7, (1973), 439-466.

[La 3] V. Lazutkin. Convex Billiard and Eigenfunctions of the Laplace Operator. Leningrad Univ. Press, 1981 (in Russian).

[LeC] P. Le Calvez. Propriétés Dynamiques des Difféomorphismes de l'Anneau et du Tore. Astérisque, 204, 1991.

[L-T] Ph. Levallois, M. Tabanov. Séparation des Séparatrices du Billiard Elliptique pour une Perturbation Algébrique et Symétrique de l'Ellipse. C.R. Acad. Sci., 316, (1993), 589-592.

[Lev] Ph. Levallois. Non-Integrabilite des Billards Definis par Certaines Perturbations Algébriques d'une Ellipse et du Flot Geodesique de Certaines Perturbations Algébriques d'un Ellipsoide. Thése de Doctorat, Univ. Paris 7, 1993.

[Ma-M] R. Mackay, J. Meiss. Linear Stability of Periodic Orbits in Lagrangian Systems. Phys. Lett., 98 A, (1983), 92-94.

[Man] A. Manning. Dynamics of Geodesic and Horocycle Flows on Surfaces of Constant Negative Curvature. Ergodic Theory, Symbolic Dynamics and Hyperbolic Spaces, T. Bedford, M. Keane, C. Series (ed.), Oxford Univ. Press, 1991.

[Mar 1] R. Markarian. Billiards with Pesin Region of Measure One. Comm. Math. Phys., 118, (1988), 87-97.

[Mar 2] R. Markarian. Nonuniform Hyperbolicity, Quadratic Forms and Billiards. Preprint.

[M-Me] S. Marvizi, R. Melrose. Spectral Invariants of Convex Planar Regions. J. Diff. Geom., 17, (1982), 475-502.

[M 1] H. Masur. Closed Trajectories for Quadratic Differentials with an Application to Billiards. Duke Math. J., 53, (1986), 307-314.

[M 2] H. Masur. Hausdorff Dimension of the Set of Nonergodic Foliations of a Quadratic Differential. Duke Math. J., 66, (1992), 387-442.

[M 3] H. Masur. The Growth Rate of Trajectories of a Quadratic Differential. Ergod. Th. and Dyn. Syst., 10, (1990), 151-176.

[Ma 1] J. Mather. Glancing Billiards. Ergod. Th. and Dyn. Syst., 2, (1982), 397-403.

[Ma 2] J. Mather. Non-Existence of Invariant Circles. Ergod. Th. and Dyn. Syst., 4, (1984), 301-309.

[Ma 3] J. Mather. Variational Construction of Orbits of Twist Diffeomorphisms. Journ. A. M. S., 4, (1991), 207-263.

[Me 1] R. Melrose. Equivalence of Glancing Hypersurfaces. Invent. Math., 37, (1976), 165-192.

[Me 2] R. Melrose. Equivalence of Glancing Hypersurfaces 2. Math. Ann., 255, (1981), 159-198.

[M-T] I. Monroe, S. Tabachnikov. Asymptotic Dynamics of the Dual Billiard Map. An Example of a Semicircle. UARK Preprint, 1992.

[M-H] M. Morse, G. Hedlund. Symbolic Dynamics 2. Sturmian Trajectories. Amer. J. of Math., 62, (1940), 1-42.

[Mo 1] J. Moser. Various Aspects of Integrable Hamiltonian Systems. Progr. in Math., 8, Birkhauser, 1980.

[Mo 2] J. Moser. Geometry of Quadric and Spectral Theory. Chern Symp., 147-188, Springer-Verlag, 1980.

[Mo 3] J. Moser. Lectures on Hamiltonian Systems. Courant Inst. Math. Sci., 1968.

[Mo 4] J. Moser. Stable and Random Motions in Dynamical Systems. Ann. of Math. Stud., 77, 1973.

[Mo 5] J. Moser. Recent Developments in the Theory of Hamiltonian Systems. SIAM Rev., 28, No 4, (1986), 459-485.

[Mo 6] J. Moser. Is the Solar System Stable? Math. Intell., 1, (1978), 65-71.

[Mo-S] J. Moser, C. Siegel. Lectures on Celestial Mechanics. Springer-Verlag, 1971.

[Mo-V] J. Moser, A. Veselov. Discrete Versions of Some Classical Integrable Systems and Factorization of Matrix Polynomials. Comm. Math. Phys., 139, (1991), 217-243.

[Ni] Z. Nitecki. Differentiable Dynamics. MIT Press, 1971.

[Os] V. Oseledets. The Multiplicative Ergodic Theorem. the Lyapunov Characteristic Numbers of Dynamical Systems. Trans. Mosc. Math. Soc., 19, 91968), 197-231.

[Pei] R. Peirone. Reflections Can Be Trapped. Amer. Math. Monthly, 101, (1994), 259-260.

[Pes] Ya. Pesin. Characteristic Lyapunov Exponents and Smooth Ergodic Theory. Russ. Math. Surv., 32, No 4, (1977), 55-114.

[P-P] Ya. Pesin, B. Pitskel. Topological Pressure and the Variational Principle for Noncompact Sets. Fuct. Anal. Appl., 18, (1984), 307-318.

[Pet] K. Petersen. Ergodic Theory. Cambridge Univ. Press, 1983.

[Por] H. Poritsky. The Billiard Ball Problem on a Table with Convex Boundary – an Illustrative Dynamical Problem. Ann. of Math., 51, (1950), 446-470.

[Pol] M. Pollicott. Lectures on Ergodic Theory and Pesin Theory on Compact Manifolds. Cambridge Univ. Press, 1993.

[Pos] J. Poschel. Integrability of Hamiltonian Systems on Cantor Sets. Comm. Pure Appl. Math., 35, (1982), 653-695.

[Ra] J. Rauch. Illuminations of Bounded Domains. Amer. Math. Monthly, 85, (1978), 359-361.

[R-B] R. Richens, M. Berry. Pseudointegrable Systems in Classical and Quantum Mechanics. Physica D, 2, (1981), 495-512.

[Ru] T. Ruijgrok. Periodic Orbits in Triangular Billiards. Acta Phys. Polon., 22, (1991), 955-981.

[Ry] M. Rychlik. Periodic Points of the Billiard Ball Map in a Convex Domain. J. Diff. Geom., 30, (1989), 191-205.

[Sa] L. Santalo. Integral Geometry and Geometric Probability. Addison-Wesley, 1976.

[S-T] M. Senechal, J. Taylor. Quasicrystals: The View from les Houches. Math. Intell., 12, No 2, (1990), 54-64.

[Se] C. Series. The Geometry of Markoff Numbers. Math. Intell., 7, No 3, (1985), 20-29.

[Sev] M. Sevryuk. Estimate of the Number of Collisions of N Ellastic Particles on a Line. Theor. Math. Phys., 96:1, (1993), 818-826.

[S-V] A. Shaidenko, F. Vivaldi. Global Stability of a Class of Discontinuous Dual Billiards. Comm. Math. Phys., 110, (1987), 625-640.

[Sik] J.-C. Sikorav. Systemes Hamiltoniens et Topologie Symplectique. Univ. de Pisa, 1990.

[Si 1] Ya. Sinai. Introduction to Ergodic Theory. Princeton Univ. Press, 1976.

[Si 2] Ya. Sinai. Billiard Trajectories in a Polyhedral Angle. Russ. Math. Surv., 33, No 1, (1978), 229-230.

[Si 3] Ya. Sinai. Hyperbolic Billiards. Proc. ICM, Kyoto 1990, 249-260.

[Si 4] Ya. Sinai. Dynamical Systems with Elastic Reflections. Ergodic Properties of Dispersing Billiards. Russ. Math. Surv., 25, No 2, (1970), 137-189.

[Si 5] Ya. Sinai. Ergodic Properties of a Lorentz Gas. Funct. Anal. Appl., 13, (1979), 192-202.

[Sin] R. Sine. A Characterization of the Ball in $\mathbf{R}^3$. Amer. Math. Monthly, 83, (1976), 260-261.

[S-K] R. Sine, V. Kreinovic. Remarks on Billiards. Amer. Math. Monthly, 86, (1979), 204-206.

[Sp] M. Spivak. Differential Geometry. Publish or Perish, 1979.

[Sto 1] L. Stojanov. Note on the Periodic Points of the Billiards. J. Diff. Geom., 34, (1991), 835-837.

[Sto 2] L. Stojanov. An Estimate from Above of the Number of Periodic Orbits for Semi-Dispersing Billiards. Comm. Math. Phys., 124, (1989), 217-227.

[Str] K. Strebel. Quadratic Differentials. Springer-Verlag, 1984.

[Ta 1] S. Tabachnikov. On the Dual Billiard Problem. Adv. in Math., to appear.

[Ta 2] S. Tabachnikov. Dual Billiards. Russ. Math. Surv., 48, No 6, (1993), 75-102.

[Ta 3] S. Tabachnikov. Commuting Dual Billiards. Geom. Dedicata, to appear.

[Ta 4] S. Tabachnikov. Poncelet's Theorem and Dual Billiards. L'Enseign. Math., 39, (1993), 189-194.

[Tu] Ph. Turner. Convex Caustics for Billiards in $\mathbf{R}^2$ and $\mathbf{R}^3$. Convexity and Related Combinatorial Geometry, 85-105, Dekker, 1982.

[V 1] W. Veech. Teichmuller Curves in Moduli Space, Eisenstein Series and an Application to Triangular Billiards. Invent. Math., 97, (1989), 553-583.

[V 2] W. Veech. The Billiard in a Regular Polygon. Geom. and Funct. Anal., 2, (1992), 341-379.

[Ve 1] A. Veselov. Integrable Maps. Russ. Math. Surv., 46, No 5, (1991), 3-45.

[Ve 2] A. Veselov. Confocal Surfaces and Integrable Billiards on the Sphere and in the Lobachevsky Space. J. Geom. Phys., 7, (1990), 81-107.

[Wa] P. Walters. An Introduction to Ergodic Theory. Springer-Verlag, 1982.

[Wo 1] M. Wojtkowski. Two Applications of Jacobi Fields to the Billiard Ball Problem. J. Diff. Geom., to appear.

[Wo 2] M. Wojtkowski. Principles for the Design of Billiards with Nonvanishing Lyapunov Exponents. Comm. Math. Phys., 105, (1986), 391-414.

[Wo 3] M. Wojtkowski. Invariant Families of Cones and Lyapunov Exponents. Ergod. Th. and Dyn. Syst., 5, (1985), 145-161.

[Wo 4] M. Wojtkowski. Linearly Stable Orbits in 3 Dimensional Billiards. Comm. Math. Phys., 129, (1990), 319-328.