Physica 19D (1986) 311–333
North-Holland, Amsterdam

BILLIARDS IN POLYGONS

Eugene Gutkin*
*Department of Mathematics, University of Southern California**, Los Angeles, CA 90089-113, USA and Columbia University, Morningside Heights, New York, NY 10027, USA*

Received 20 November 1985

Review paper

We report on the state of the art for billiards in polygons.

1. Introduction

Let U be a closed plane domain with the boundary Γ. Assume that Γ is smooth everywhere except for a finite number of corners. Billiard in U is a natural dynamical system, its phase space $T(U)$ consists of unit vectors with foot points in U and looking inside U. Points of $T(U)$ obviously correspond to the dimensionless billiard balls moving inside U with the unit speed, hence the name. The billiard ball goes straight until it hits Γ where it bounces off according to the law that the angle of reflection is equal to the angle of incidence. The rule being inapplicable at the corners, we agree to stop the ball there.

Plane billiards have a number of interpretations and generalizations which we briefly mention here. First of all, billiards are Hamiltonian systems with two degrees of freedom, more precisely the billiard dynamics takes place on the energy level surface $H = \frac{1}{2}$ where $2H$ is the squared speed. One can obviously define n-dimensional billiards which take place in domains of $\mathbf{R}^n$, $n \geq 2$, and are examples of Hamiltonian systems with n degrees of freedom. Finally, billiards are particular examples of the geodesic flows on Riemannian manifolds with boundary (and, possibly, corners).

Behavior of a billiard naturally depends on the geometry of the billiard table. We distinguish the following 3 types of billiards: hyperbolic, elliptic and parabolic. The hyperbolic behavior is realized by the dispersing billiards of [24], e.g. polygons with smooth convex obstacles. It can also be realized by billiard tables with convex boundary, the so called Bunimovich billiards [4] and their significant generalization – Wojtkowski billiards [27]. Typical elliptic billiard tables are convex and have strictly positive curvature [19]. While the hyperbolic billiards have a strong mixing behavior, the properties of elliptic billiards are due to the fact that they are close to integrable Hamiltonian systems with two degrees of freedom (see, e.g. [5] and the bibliography there). Typical parabolic billiards are billiards in polygons. Referring the reader to the survey [14] for more information on hyperbolic and elliptic billiards, we now concentrate on billiards in polygons.

Let P be a bounded connected polygon (P is not necessarily convex, it can have obstacles inside and/or slits, see fig. 1) Billiards in some polygons

*Partially supported by NSF Grant DMS 84-03238.
**Current address.

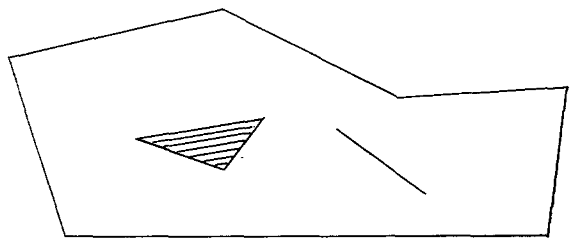


Fig. 1. A billiard table with an obstacle and a slit.

correspond to the "concrete" mechanical systems with two degrees of freedom. Consider, for instance, two point masses $m_1 \leq m_2$ confined to move on the unit interval. The masses interact by perfect elastic collisions with each other and with the ends of the interval. This mechanical system is realized by the billiard in the right triangle with the angle $\alpha = \arctan\sqrt{m_1/m_2}$ (cf. [18]). For every corner A of P consider the set of billiard trajectories that wind up at A some time in the future or in the past. This set is a countable union of one-parameter families of trajectories. Denote by B^t the billiard flow on the space $T(P)$ which is 3 dimensional. By the remark above, the set D of points in $T(P)$ for which the dynamics B^t is not defined for all times has dimension 2, thus D has measure 0 with respect to the invariant Lebesgue measure on $T(P)$ (it is a general fact that the Lebesgue measure is invariant under B^t, see, e.g. [18]). The billiard flow is well defined on $X = T(P) \setminus D$ and preserves the Lebesgue measure μ on X where $\mu(X) < \infty$. In this situation one is interested in the metric properties of B^t (cf. [18]), e.g. ergodicity, unique ergodicity, spectrum and mixing. The phase space $T(P)$ has a natural topology and although the flow B^t is not continuous due to the corners of P, it makes sense to study the topological dynamics of billiard flows, in particular the density of billiard trajectories (also called orbits) in $T(P)$. A trajectory in $T(P)$ determines a line in P, its foot line, which we also call a trajectory if there is no danger of confusion. Thus, we can talk about the density of billiard orbits in P which is called the spatial density. Finally, we will be interested in the periodic (also called closed) billiard orbits.

The purpose of this paper is to report on the state of the art for billiards in polygons. In the body of the paper we precisely formulate the questions and give some answers indicating the ideas of proofs and referring the reader to the literature for details. When the answers are not known (which is often the case) we formulate conjectures. The conjectures do not belong to the author, rather they reflect the opinion of workers in the area on what the answers are likely to be. The emphasis of this paper is not on the proofs but on the general picture illustrated by examples. Most of the results in the survey have been published earlier (we give references). The only exceptions are theorems 5 and 6 which belong to the author and appear in print for the first time here.

2. Elementary methods for billiards in polygons

2.1. *Unfoldings*

Let γ be a billiard trajectory in P. A part of γ between two consecutive reflections off the boundary ∂P is called the link of γ. Choose a link γ_0 of γ and reflect P about the side of ∂P where γ_0 ends. Denote the reflected P by P_1. The next link γ_1 of γ goes under the reflection into $\tilde{\gamma}_1$ which is the continuation of γ_0 in P_1 (see fig. 2). This is the method of reflecting the billiard table. Continuing the process for n steps we obtain the sequence $P_0 = P, P_1, \dots, P_n$ of polygons and the straight interval consisting of links γ_0 in P_0, $\tilde{\gamma}_1$ in $P_1, \dots, \tilde{\gamma}_n$ in P_n (see fig. 3). The polygons $P_0, P_1, \dots, P_n$ look like pieces of shishkebab on a skewer. By more common terminology, $P_0, P_1, \dots, P_n$ is the (finite) unfolding of the billiard along a finite piece of the trajectory γ. Continuing the unfolding to infinity we obtain the infinite sequence $P_0, P_1, \dots, P_n, \dots$ of polygons skewed on the ray $\gamma_0, \tilde{\gamma}_1, \dots, \tilde{\gamma}_n, \dots$. Unfolding along γ backwards we obtain the sequence

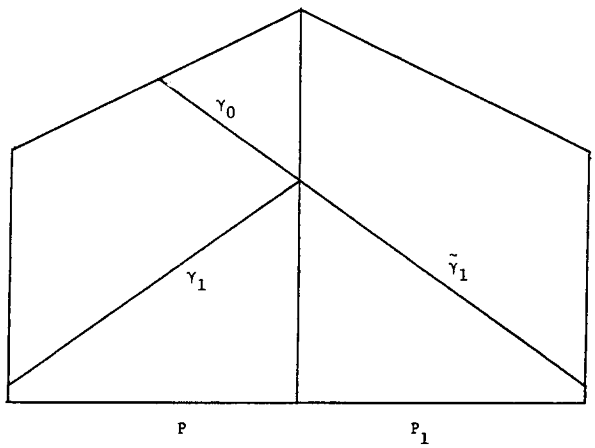


Fig. 2. First step in unfolding along a trajectory.

$P_0, P_{-1}, \ldots, P_{-n}, \ldots$ of polygons skewed on the ray $\gamma_0, \tilde{\gamma}_{-1}, \ldots, \tilde{\gamma}_{-n}, \ldots$. The two rays form the straight line l_γ and we have the infinite shishkebab $\{P_i, -\infty < i < \infty\}$ on the skewer l_γ.

If γ is a finite billiard trajectory, i.e. γ hits the corners of P both in the future and in the past, then $\tilde{\gamma}$ is a finite interval joining a vertex of P_{-m} with a vertex of P_n for some $m, n \geq 0$. Following [13] we call a finite billiard trajectory a generalized diagonal of P.

2.2. *Groups associated with polygonal billiards*

Denote by $O(\mathbf{R}^2)$ the group of motions of the Euclidean plane $\mathbf{R}^2$ and by $SO(\mathbf{R}^2)$ the subgroup of orientation preserving motions. Choose an origin 0 of $\mathbf{R}^2$, denote by U the unit circle around 0 and the group of rotations about 0. Then $SO(\mathbf{R}^2) = U \cdot \mathbf{R}^2$ which means that any orientation-preserving motion of the plane uniquely decomposes as the product of a parallel translation and a rotation about 0. The group $O(U)$ of motions of the circle imbeds into $O(\mathbf{R}^2)$ as the subgroup of 0 preserving motions. $O(U)$ consists of rotations and reflections and we denote by σ the reflection in the horizontal axis. We have decompositions $O(U) = \{1, \sigma\} \cdot U$ and $O(\mathbf{R}^2) = O(U) \cdot \mathbf{R}^2$.

If l is a line on the plane we denote by $s_l \in O(\mathbf{R}^2)$ the reflection about l (s_l sends any point of $\mathbf{R}^2$ into its mirror image with respect to l).

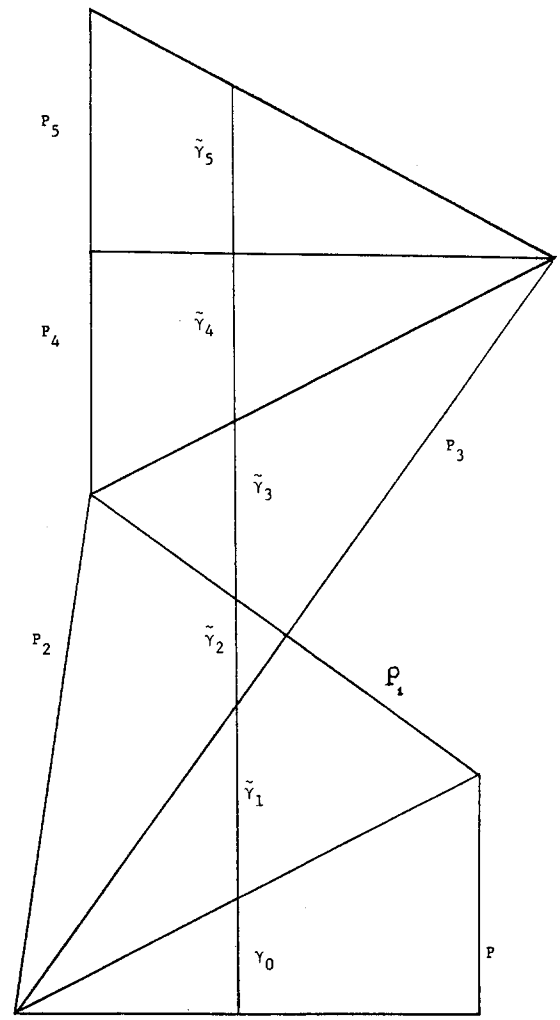


Fig. 3. Shishkebab or unfolding along trajectory.

Let $a_1, \ldots, a_n$ be the sides of P and denote by $s_1, \ldots, s_n$ the reflections about the lines $l_1, \ldots, l_n$ through the sides $a_1, \ldots, a_n$ respectively. Denote by G_P the subgroup of $O(\mathbf{R}^2)$ generated by $s_1, \ldots, s_n$. The decomposition $O(\mathbf{R}^2) = O(U) \cdot \mathbf{R}^2$ defines the homomorphism $O(\mathbf{R}^2) \to O(U)$. Denote by W_P the image of G_P under the homomorphism. To understand W_P we take the lines $\tilde{l}_1, \ldots, \tilde{l}_n$ through 0 parallel to $l_1, \ldots, l_n$ respectively. The reflections $\tilde{s}_1, \ldots, \tilde{s}_n$ about $\tilde{l}_1, \ldots, \tilde{l}_n$ leave the circle U invariant and induce reflections $\sigma_1, \ldots, \sigma_n$ of U. The group W_P is generated by

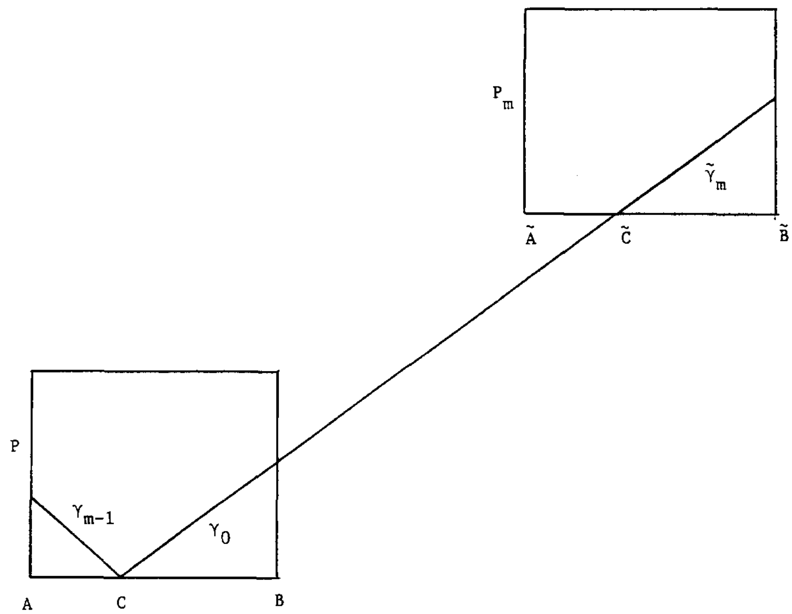


Fig. 4. Unfolding along periodic trajectory.

$\sigma_1, \ldots, \sigma_n$. Denote by T_P the kernel of the homomorphism $G_P \to W_P$, thus T_P consists of parallel translations. To illustrate the usefulness of these notions we prove the following:

Proposition 1 ([29]). The set of generalized diagonals of P is countable.

Proof. Let X be the set of polygons obtained by unfoldings along all billiard trajectories γ in P. Then X belongs to the set of polygons $\{gP: g \in G_P\}$ which is countable because the group G_P has a finite number of generators, thus countable. A generalized diagonal γ corresponds to an interval $\tilde{\gamma}$ joining the vertices of two polygons in X. Since X is countable, the set of these pairs of vertices is also countable, thus the set of such $\tilde{\gamma}$'s is countable which proves the proposition.

Let γ be a periodic trajectory in P with m links $\gamma_0, \ldots, \gamma_{m-1}$. Unfold P along γ starting with the link γ_0. After m steps we have polygons $P_0 = P$, $P_1, \ldots, P_{m-1}$, P_m and $m+1$ links $\gamma_0, \tilde{\gamma}_1, \ldots, \tilde{\gamma}_m$. Since γ is periodic, the end of γ_{m-1} coincides with the beginning of γ_0 and γ_0 is the reflection of γ_{m-1} (see fig. 4). Therefore, the element $g \in G_P$ that moves P into P_m moves γ_0 into $\tilde{\gamma}_m$ which belongs to the same line l_γ as γ_0. Thus, g preserves the line l_γ. If a motion g of $\mathbb{R}^2$ preserves a line l then g is either a parallel translation along l or a sliding reflection with the axis l, i.e., the composition of a parallel translation along l with the reflection s_l. In the former case g preserves the orientation, in the latter g reverses the orientation. The motion g that moves P into P_m is the product of m reflections, thus g preserves the orientation if and only if m is even. We say that a periodic trajectory is even (odd) if it has an even (odd) number of links. The discussion above proves the following.

Proposition 2. Let γ be a periodic trajectory of length L. Choose a link γ_0 of γ and let l be the line through γ_0. The unfolding along γ defines $g \in G_P$ which is the parallel translation along l by amount L if γ is even and g is the sliding (by the same amount) reflection about l if γ is odd.

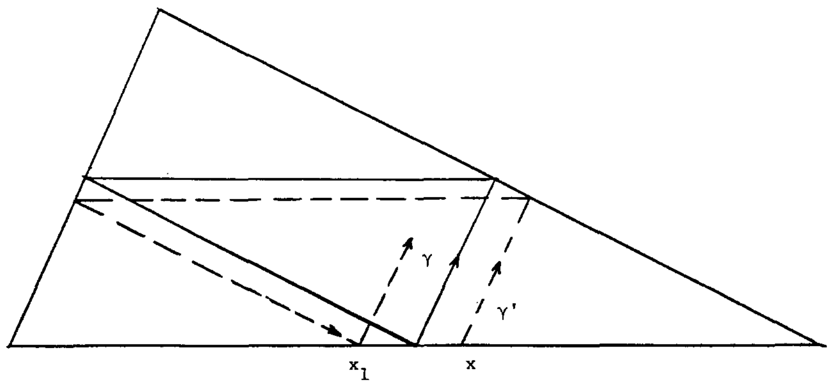


Fig. 5. Deformation of an odd periodic orbit.

Corollary 1. Let γ be a periodic trajectory with m links and of length L. If m is even, γ extends to a band of periodic trajectories of length L parallel to γ. Both boundaries of the band are unions of generalized diagonals. If γ is odd, every trajectory γ' starting close to γ and parallel to γ comes back after m reflections to the same edge, at the same distance from γ and in the same direction but on the opposite side of γ (see fig. 5).

Proof. It is convenient to denote γ by γ_0. Unfolding P along γ_0 we obtain the sequence $P = P_0, \ldots, P_{m-1}, P_m$, the line l_0 (see fig. 4.) and the motion g such that $P_m = gP$. Let the point x_0 on edge a of P be the starting point of γ_0 and let $y_0 = gx_0$ be the corresponding point on the edge $b = ga$ of P_m. By the discussion above, the periodicity of γ_0 implies that l_0 goes from x_0 to y_0. Let γ be the trajectory starting at $x \in a$ close to x_0 and parallel to γ_0. If γ is close enough to γ_0, unfolding along γ we obtain the line l through x parallel to l_0 and passing by the same sequence $P_0, \ldots, P_{m-1}$ of polygons. Let l intersect b at y. If γ_0 is even, by proposition 2, the quadrangle formed by a, b, l_0 and l is a parallelogram, thus $y = gx$, i.e., γ comes back to x in the same direction, whence γ is periodic. When γ_0 is odd, the quadrangle above is a trapezoid and the point $x_1 = g^{-1}y$ on the side a where γ returns after m reflections is symmetric to x with respect to x_0.

In the argument above we assumed that γ is so close to γ_0 that it hits the same edges as γ_0. Assume that γ_0 is even and start moving γ_0 to the right parallel to itself. Denote this deformation of γ_0 by γ_t, $t \geq 0$. For t small enough, γ_t hits the same edges as γ_0, thus, by the argument above, γ_t is periodic of the same length as γ_0. Increasing t we come to the moment $t = t_1$ such that $\gamma_1 = \gamma_{t_1}$ hits a corner A of P. Since γ_1 is the limit of periodic trajectories, starting off A it must come back to A, thus γ_1 is a generalized diagonal or a union of such in the unlikely event that γ_1 hits more than one corner. The same argument works for the deformation γ_t, $t \leq 0$ and we obtain γ_2 which is the other end of the band $\{\gamma_t\}$ of periodic trajectories. Corollary is proved.

By a family of periodic trajectories we will mean either a single odd trajectory γ or a band $\{\gamma_t\}$ of even periodic trajectories.

Corollary 2. The set of families of periodic trajectories is at most countable.

Proof. Proving proposition 2, we associated with any family of periodic trajectories an element $g \in G_p$. Different families define different elements g and the group G_p is countable.

2.3. *Birkhof–Poincaré mapping*

This mapping, also called the first return map, makes sense for billiards in general and has been very useful for their study (see, e.g. [5]). Assume

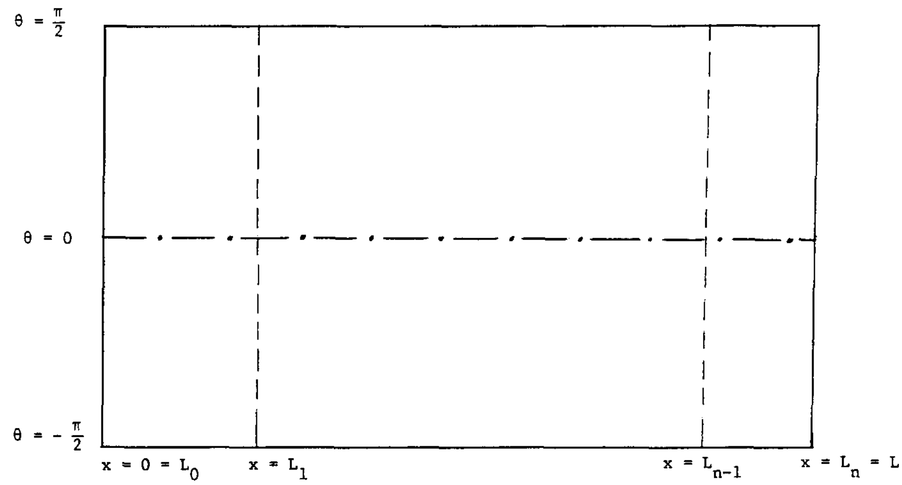


Fig. 6. Domain Ω of the Birkhof–Poincaré map F.

for simplicity that P is simply connected and enumerate the vertices of P counterclockwise $A_0, \ldots, A_{n-1}$, $A_n = A_0$. Denote by $a_i = [A_{i-1}, A_i]$, $i = 1, \ldots, n$ the edges of P and let $L = |a_1| + \cdots + |a_n|$ be the length of ∂P. The boundary ∂P with the length coordinate x is isomorphic to the circle of perimeter L. We parametrize the set $\Omega \subset T(P)$ of vectors with foot points in ∂P by coordinates $0 \le x \le L$ and $-\pi/2 \le \theta \le \pi/2$. The vector (x, θ) has the foot point $x \in a_i \subset \partial P$ and the angle θ between it and the inner normal to a_i. Set $L_k = |a_1| + \cdots + |a_k|$, $k = 0, \ldots, n$. The coordinate θ in (x, θ) is not well defined for $x = L_k$ because these x are the vertices of P and the angle θ can be measured with respect to any of the two normals. The set Ω with points (L_k, θ) excluded is isomorphic to the cylinder $[0, L) \times [-\pi/2, \pi/2]$ with deleted intervals $L_0 \times [-\pi/2, \pi/2], \ldots, L_{n-1} \times [-\pi/2, \pi/2]$ (see fig. 6).

The Birkhof–Poincaré map F: $\Omega \to \Omega$ is defined as follows. For $(x, \theta) \in \Omega$ we shoot the ball from x in direction θ. When the ball hits ∂P the first time and bounces off, it determines another point $(y, \eta) = F(x, \theta)$ to Ω. Besides the dotted lines on fig. 6, the mapping f is not well defined on the boundaries $\theta = \pm\pi/2$. Fix an edge $a_i = [A_{i-1}, A_i]$ and a vertex A_j, $j \neq i-1$, i. Points (x, θ), $x \in a_i$, such that the ball goes to the corner A_j (see fig. 7) form a curve in Ω on which F is not well defined. Each rectangle of Ω (see fig. 6) is divided by these curves into the domains of continuity of F. Thus, the set of discontinuities of F is the union of a finite number of curves in Ω, F is obviously invertible and F^{-1} is the Birkhof–Poincaré map for the billiard with the time reversed. The F-invariant Lebesgue measure is $\cos\theta \, d\theta \, dx$.

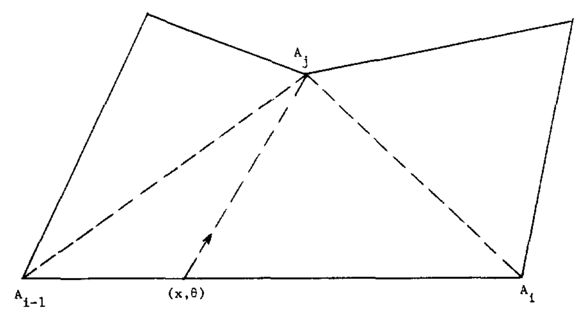


Fig. 7. Corners cause discontinuities of F.

Properties of the billiard flow B^t are easily translated into the properties of the mapping F. Another convenient set of coordinates (x, u), $u = \tan\theta$, on Ω make F piecewise projective and the curves of discontinuity become straight intervals [13].

3. Terminology and questions

The billiard flows are preserved by dilations, so we can assume that the perimeter of P is equal to one. The set of polygons with a fixed number of sides is a manifold with the natural topology and has Lebesgue measure. Thus, we can talk about typical polygons and any question about billiard flows can be stated for an individual and for a typical polygon. To state the first question we recall the basic notions of ergodic theory.

A flow B^t on the space X with a finite invariant measure μ is called ergodic if X has no nontrivial B^t-invariant measurable subset Y, i.e. for any such Y, either $\mu(Y)=0$ or $\mu(Y)=\mu(X)$.

Question 1. For which polygons P the billiard flow B^t is ergodic with respect to the Lebesgue measure on $T(P)$?

At the moment we have the class of rational polygons (π-rational vertex angles) for which the billiard flow B^t is not ergodic and the ergodic decomposition of B^t is known (see section 4). If P is not rational (i.e. irrational) almost nothing is known about the ergodicity of the billiard flow on P (see section 5, theorem 8). There are no examples of polygons of either kind and there are no conjectures on the sufficient conditions for ergodicity.

Question 1′. Is the billiard flow on P typically ergodic?

The conjecture is that the answer is yes. A result in this direction says that there is a "big" set of ergodic polygons (see section 5, theorem 8). With so little known about the ergodicity of billiard flows, other questions of ergodic theory have not been seriously considered yet, with few exceptions.

One exception is the entropy where it was proved in [2] by elementary methods that the entropy of billiards in polygons is zero (it also follows from the Pesin's theory of Lyapunov's exponents [28]). Ergodicity of the billiard flow of a polygon is equivalent to the ergodicity of its Birkhof–Poincaré map.

The topological analog of ergodicity (which is weaker than ergodicity) is the topological transitivity. A continuous flow (resp. a continuous selfmapping) on a topological space X is topologically transitive if it has a dense orbit. Under the assumptions above the flow (resp. the selfmapping) is called minimal if every nonperiodic orbit is dense in X. The minimality is obviously stronger than topological transitivity. The billiard flow B^t has finite orbits and semifinite ones, that is, orbits which are infinite only in the past or in the future. These orbits cause the discontinuities of the flow B^t. We modify the definition of minimality for the billiard flows and call a billiard flow B^t (resp. the Birkhof–Poincaré map F of a billiard) minimal (quasiminimal in the terminology of [29]) if every nonperiodic infinite or semiinfinite trajectory is dense in the phase space $T(P)$. The definition of transitivity remains valid for billiards. It is obvious that a billiard flow B^t is minimal (resp. transitive) if and only if the corresponding Birkhof–Poincaré map F is minimal (resp. transitive).

Question 2. When is the billiard flow of a given polygon topologically transitive?

Question 2′. Is the billiard flow of a typical polygon topologically transitive?

At a first glance the situations with questions 1 and 2 are similar, i.e. the answers are not known and there are no conjectures on which conditions on P should imply the transitivity of the billiard flow. However, much more is known about the density of trajectories than about the ergodicity for billiard flows. There are examples of irrational polygons with infinite nonperiodic trajectories which are not even spatially dense ([7], see section 5). On the other hand, in any polygon P almost all trajectories are spatially dense [2]. Of course, the flow of a rational polygon is not transitive and the transitive components are well known (see section 4). It was shown in [29] that there is a big set of polygons with the transitive billiard flow (see section 5). In view of these results it is reasonable to conjecture that the billiard flow of a polygon is typically transitive.

The most fascinating open problems for billiards in polygons have to do with the periodic trajectories. As we have seen in section 2, there is at most countable number of families of periodic trajectories.

Question 3. Does every polygonal billiard table have a periodic trajectory?

The answer is not known and conjectured to be yes. The only general result in this direction belongs to H. Masur [20] and says that every rational polygon has a periodic orbit. Besides, there are examples of periodic orbits in some polygons. The simplest example, [2], is the 3-link periodic orbit γ in an acute triangle P. The curve γ is the shortest billiard trajectory in P, it joins the bases of the perpendiculars drawn from the vertices of P to the opposite sides. The orbit γ disappears as soon as P becomes a right triangle. Attempts to show the existence of periodic trajectories by variational methods, following Birkhof, failed so far because of discontinuities of the billard flow. The techniques of [7] allow to construct "many" periodic orbits in some polygons (see section 5).

In the proof of proposition 2 we associated with any periodic orbit γ in P an element $g(\gamma) \in G_P$. If γ has m links then $g(\gamma) = s_1 \dots s_m$ where s_i are the reflections about the sides of P. It can be shown that for every polygon P there are constants c_1, $c_2 > 0$ such that the length $l(\gamma)$ of the orbit γ satisfies

$$c_1 m < l(\gamma) < c_2 m. \tag{1}$$

Since the group G_P has a finite number of words of a given length, it follows from (1) that the number of families of periodic orbits of length less than x is finite for any x. Denote this number by $p(x)$.

Problem 4. Find the asymptotics of $p(x)$ as x goes to infinity.

The only general (and very recent) result here belongs to A. Katok [13] and says that $p(x)$ grows subexponentially. Everybody seems to believe that $p(x)$ grows polynomially or at least has a polynomial upper bound. More precisely, the conjecture is that for any polygon P there is a constant c and a positive integer n such that

$$p(x) = cx^n + \mathcal{O}(x^{n-1}). \tag{2}$$

This is proved only for almost integrable polygons where $n = 2$ (see section 4, theorem 6) while for rational polygons there is an estimate

$$p(x) \le cx^n, \tag{3}$$

where $n = 2g(R)$, see theorem 5. Some workers in the area suggest that the exponent in (3) is at most 3, maybe even 2*.

4. Rational and almost integrable billiards

Definition 1. A polygon P is called rational if the angles between sides of P are rational multiples of π. If P is simply connected (i.e. has no obstacles) it suffices to require that the vertex angles of P are rational multiples of π. The billiard flow of a rational polygon is called a rational billiard.

As the reader no doubt noticed from section 3, rational billiards are the best understood. Besides being interested in their own right, they are useful for approximation of irrational billiards (see section 5). In this section we discuss rational billiards and answer the questions of section 3 in that context.

The group W_P of a rational polygon P is finite (this can be taken for the definition of a rational polygon). Any finite group of motions of the circle generated by reflections is the dihedral group D_N, $N \ge 1$. For $N > 1$ the group D_N is generated by

*H. Masur announced that he can prove $n = 2$ for rational polygons (private communication).

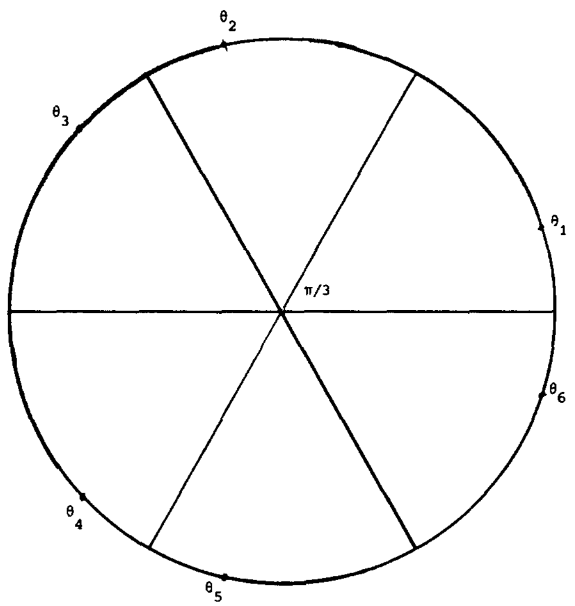


Fig. 8. Action of D_3.

the reflections σ_1, σ_2 in the lines l_1, l_2 meeting at the angle π/N. From the discussion of the group W_P in section 2 it is clear that N is the least common multiple of the denominators n_{ij} of the angles $\pi m_{ij}/n_{ij}$ between the sides a_i and a_j of P. If P is simply connected then N is the least common multiple of denominators n_i of the vertex angles $\pi m_i/n_i$ of P. The group D_N has $2N$ elements and the circle $0 \le \theta \le 2\pi$ is divided by the action of D_N into $2N$ intervals $\pi(i-1)/N \le \theta \le \pi i/N$, $i = 1, \dots, 2N$, (see fig. 8). Every θ, $0 \le \theta \le 2\pi$ is equivalent by the action of D_N to a unique θ_1, $0 \le \theta_1 \le \pi/N$, so the set of orbits of D_n is parametrized by $[0, \pi/N]$. The orbit of every θ, $0 < \theta < \pi/N$ has $2N$ elements, the orbits of 0 and π/N have N elements each.

When the "ball" (x, θ) flows under B^t, the direction θ does not change until the ball hits a side a of P, then θ instantaneously changes to $\sigma\theta$ where $\sigma \in W_P$ is the reflection about the direction of a. Thus, the set of directions of a billiard trajectory belongs to an orbit of D_N on the circle U of directions. And vice versa, for any θ, $0 \le \theta \le \pi/N$, the set $R_\theta \subset T(P)$ of elements (x, η) such that $\eta = w\theta$ for $w \in D_N$ is invariant under the flow B^t. Thus, the phase space $T(P)$ is decomposed into the one-parameter family of the flow-invariant surfaces R_θ, $0 \le \theta \le \pi/N$ (cf. [29]). In particular, rational billiards are not ergodic, not even topologically transitive.

Invariant surfaces R_θ are the level surfaces of the function ψ: $T(P) \to [0, \pi/N]$ which is defined by $\psi(x, \eta) = \theta$ where $0 \le \theta \le \pi/N$ and $\eta = w\theta$ for $w \in D_N$. Recall that the billiard flow comes from a Hamiltonian system with two degrees of freedom and let H be the Hamiltonian (one half of the squared speed). The function ψ is obviously independent of H and is a constant of the motion. If the billiard flow had no singularities we would have an integrable Hamiltonian system. Because of the singularities at the corners the Poisson bracket $\{H, \psi\}$ can be interpreted as the δ-function supported at the vertices of P. In view of the above we can call rational billiards the quasi-integrable Hamiltonian systems (cf. [22]) reserving the name almost integrable for an interesting subclass of rational polygons (see below).

We observe that the invariant surfaces R_θ for $0 < \theta < \pi/N$ are isomorphic to a surface R which can be geometrically constructed from the polygon P. To construct the surface R we take $2N$ copies of P and put them on the plane in a row without overlapping (see fig. 9). We label the polygons $P_1, \dots, P_{2N}$. Choose θ_1, $0 < \theta_1 < \pi/N$ and denote by $\theta_2, \dots, \theta_{2N}$ the elements of the D_N-orbit of θ_1 in the natural order (see fig. 8). Then we put inside P_i an arrow in direction θ_i for $i = 1, \dots, 2N$. Let a_j, $j = 1, \dots, n$ be the sides of P and denote by a_{ij} the j's side of P_i. Take the direction θ_i in P_i and reflect it in any side a_{ij} of P_i. We get another direction, say $\theta_{i'}$, of the set $\{\theta_1, \dots, \theta_{2N}\}$. Then glue the side a_{ij} of P_i with the side $a_{i'j}$ of $P_{i'}$ in an obvious way. For instance, for P on fig. 9, the hypotenuse of P_1 gets glued to the hypotenuse of P_2, the horizontal side of P_1 gets glued to the horizontal side of P_{12} and the vertical side of P_1 is glued to the vertical side of P_6.

It is clear that the surface R obtained from $P_1, \dots, P_{2N}$ after the gluings does not depend on

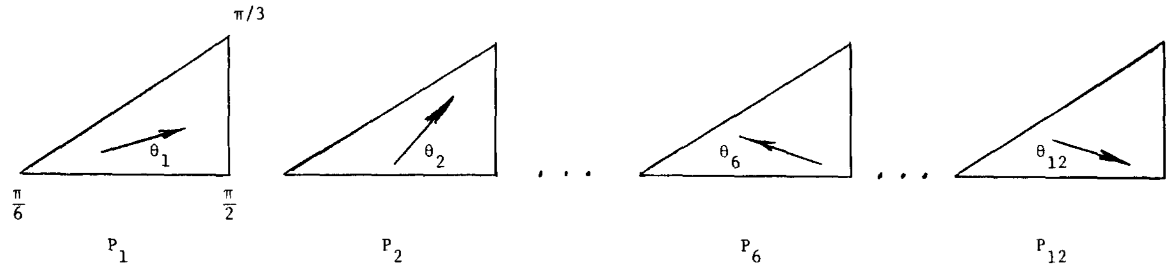


Fig. 9. Construction of surface R.

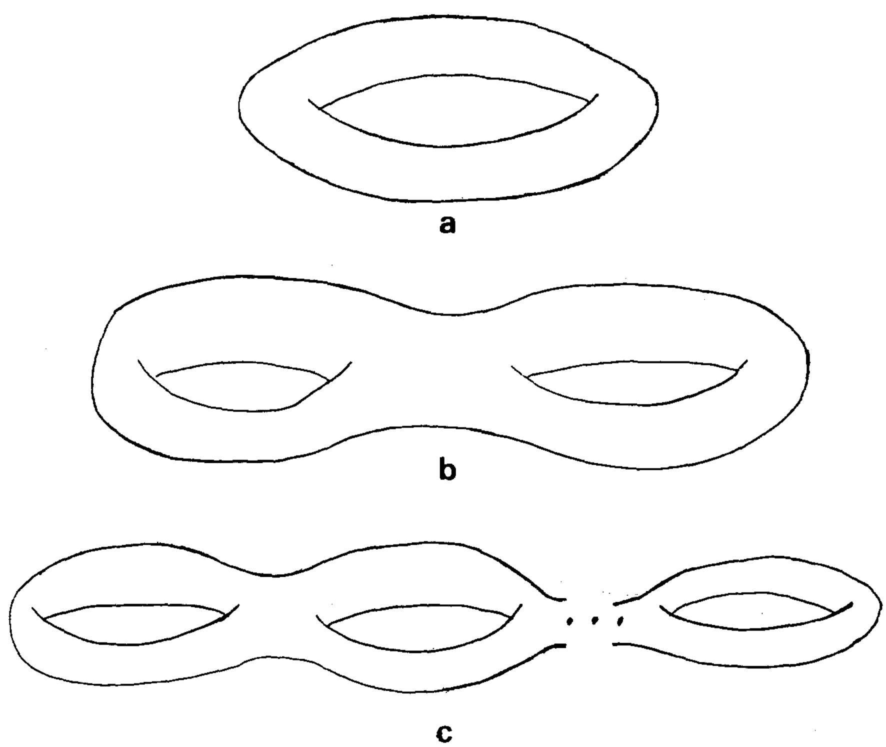


Fig. 10. Closed orientable surfaces. a) R = torus (doughnut), $g(R) = 1$; b) R = pretzel, $g(R) = 2$; c) R = pretzel with g holes, $g > 2$

the choice of θ_1, which can be eliminated from the recipe altogether. It is also clear that R is closed (has no boundary) because every side of every polygon is glued to some other side. It is a little bit harder to show that R is orientable. First, we observe that, by our recipe, even numbered polygons P_i get glued to the odd numbered ones. Orient all $2N$ polygons P_i counterclockwise. The gluings preserve the orientation of the sides. Change the orientation of all even-numbered polygons to the opposite. Now every time we glue two sides together they have opposite orientations which gives us an orientation of the surface R.

The topological type of a closed orientable surface R is determined by its genus $g(R)$. The surface of genus g looks like a pretzel with g holes (fig. 10).

The Euler number χ of the surface of genus g is given by $\chi = 2 - 2g$. Our construction of R is convenient for the calculation of its Euler number.

Let for simplicity P be a simply connected polygon with vertex angles $\pi m_i/n_i$, $i=1,\dots,p$, and let N be the least common multiple of n_i. The usual counting of faces, edges and vertices gives (cf. [10], proposition 5)

$$g(R) = 1 + \frac{N}{2}\sum_{i=1}^{p}\frac{m_i-1}{n_i}. \tag{4}$$

For instance, (4) shows that the surface R corresponding to the triangle on fig. 9 has genus 1, i.e. it is a torus (doughnut).

To realize the billiard flow in direction θ on R we use our construction of R with $\theta_1=\theta$ and instead of the single arrow in direction θ_i we put in P_i the vector field of such arrows. Each arrow of the field corresponds to the billiard ball. Let the ball follow the arrow and come to the edge a_{ij} of P_i which is glued to the edge $a_{i'j}$ of $P_{i'}$. The ball crosses the edge a_{ij} and finds itself in $P_{i'}$ where it now goes in the direction $\theta_{i'}$ and so on. Another way to describe the flow B_θ^t is to say that each P_i comes with the vector field X_{θ_i} (unit vectors in direction θ_i). Gluing the polygons P_i together into the surface R we glue the vector fields X_{θ_i} together into the vector field X_θ on R which is the generating vector field of the flow B_θ^t.

The vector field X_θ has isolated singularities at the vertices of R. Computing the indices of the singularities and using the well known formula that the Euler number is the sum of these indices one again obtains formula (4) (this approach was used in [22]).

Using the idea of reflecting the billiard table instead of reflecting the ball we can construct concrete realizations of the surface R. To do it, we return to fig. 9, take the polygon P_1 erase it from the list on fig. 9 and shoot the ball in direction θ_1

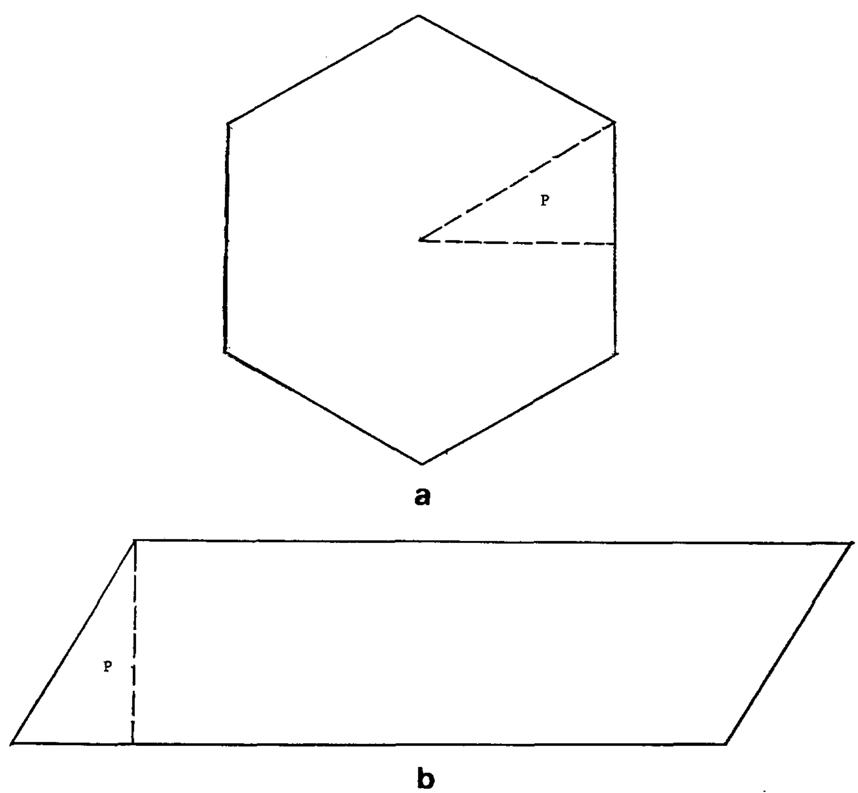


Fig. 11. Two possible polygons $\tilde{R}$ for the $\pi/6$, $\pi/3$, $\pi/2$ triangle; R = torus. a) $\tilde{R}$ = regular hexagon. b) $\tilde{R}$ = parallelogram.

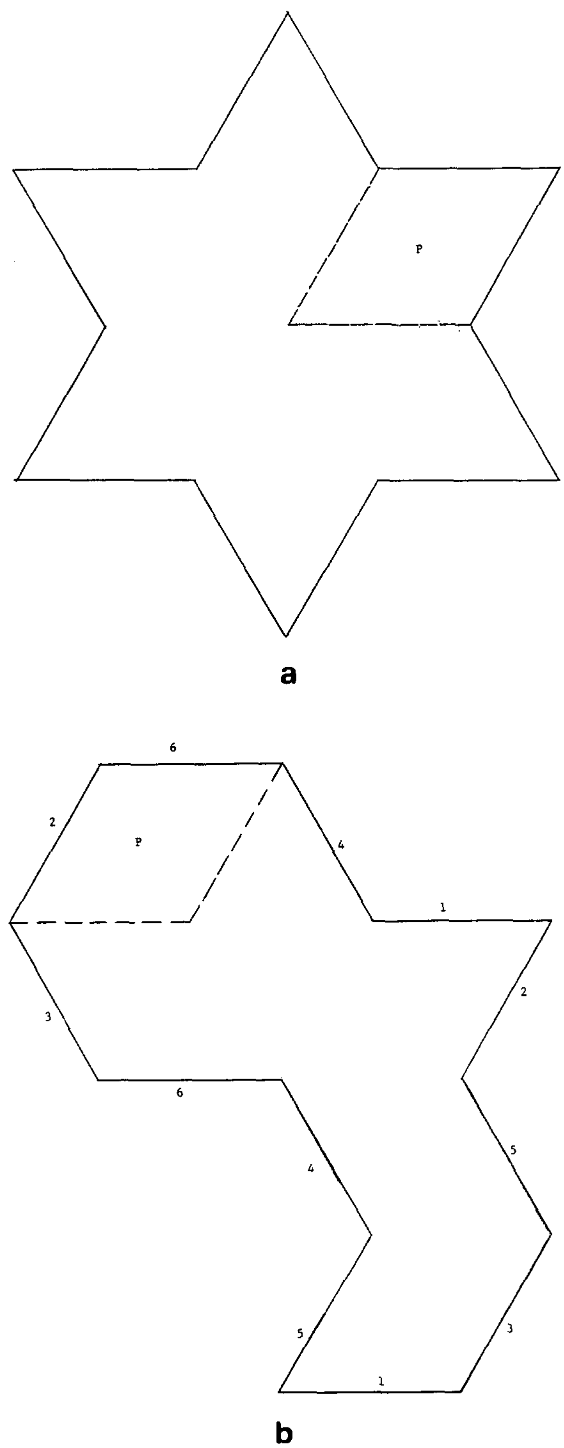


Fig. 12. Two $\tilde{R}$ for the diamond; R = pretzel. a) $\tilde{R}$ = six vertex star. b) another $\tilde{R}$. Identified sides are labelled by the same numbers.

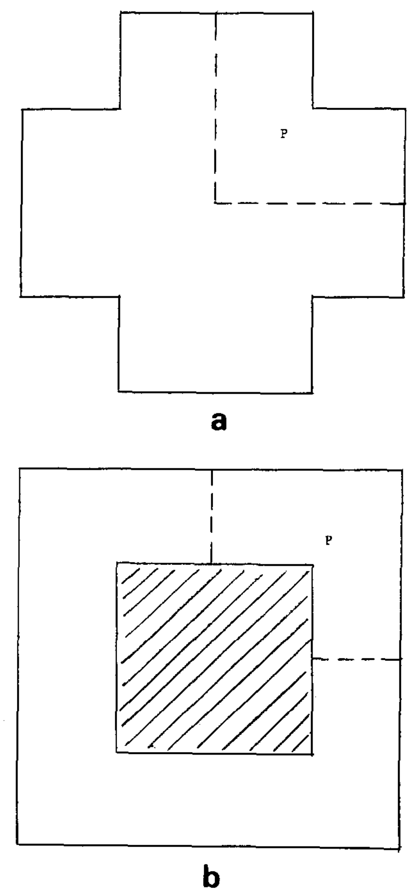


Fig. 13. Polygons $\tilde{R}$ for the gnomon; R = pretzel. a) $\tilde{R}$ = Maltese cross; b) $\tilde{R}$ = window.

until it hits a side, say a_{ij}. Then we reflect P_1 on the plane about a_{ij}, join the reflected P_1 (identified as $\tilde{P}_{1'}$ to the figure under construction and erase $P_{1'}$ from the list. Now repeat the whole process with $\tilde{P}_{1'}$ shooting in the same direction θ but not necessarily along the same line. Each time we come to a side a_{ij} we check if the polygon $P_{i'}$ is still on the list. If it is, we do the reflection, join the new polygon $\tilde{P}_{i'}$ to the figure and strike $P_{i'}$ out of the list. If it is not, we do not reflect P_i but instead find $\tilde{P}_{i'}$ in the figure, identify the sides a_{ij} and $a_{i'j}$ which are already drawn and transfer the ball to $\tilde{P}_{i'}$. We keep repeating the procedure until all the polygons P_i are struck out of the list. At this point, we have drawn a polygon $\tilde{R}$ on the plane which is the union of $\tilde{P}_i$, $i = 1, \ldots, 2N$ (if there are overlappings between $\tilde{P}_i$ we consider them as belonging to different copies of the plane). The polygon $\tilde{R}$ has an even number of sides and they are divided into pairs, each side of any pair differs from the other one by a parallel translation. The surface R is obtained by identifying the sides of $\tilde{R}$ in each pair. The polygon $\tilde{R}$ is by no means unique because of the choices we had to make drawing it. On fig. 11 we have drawn two possible $\tilde{R}$ corresponding to the triangle P from fig. 9. On fig. 12 we have two $\tilde{R}$ for the diamond and on fig. 13 two $\tilde{R}$ for the gnomon (these do not exhaust all the possibilities).

The advantage of the polygon $\tilde{R}$ is that the flows B_θ^t have a simple realization on it. The ball is traveling straight in direction θ until it reaches a side a of $\tilde{R}$. The sides of $\tilde{R}$ come in pairs, let b be the pair of a and let g be the parallel translation that identifies a with b. Once the ball hits a, it gets instantaneously transferred to b by g and it starts anew in the same direction from there. This is the flow B_θ^t. Notice that we have a one-parameter family of flows and that all B_θ^t are obtained from any one of them, say from B_0^t by rotating the field X_0 of directions by the angle θ. This operation is called the rotation of flows and is well defined for flows on closed orientable surfaces. Another useful observation is that the flows B_θ^t can be defined for any polygon Q if its sides are divided into pairs $a_1, b_1, \ldots, a_m, b_m$ such that b_i differs from a_i by a parallel translation g_i, $i = 1, \ldots, m$. Notice that although any rational polygon P defines several polygons $\tilde{R}$ with this property, not every such polygon Q can be obtained this way. It is clear that polygons $\tilde{R}$ have π-rational angles, so, for instance, parallelograms with π-irrational angles are not obtained this way.

It is useful to know the structure of the flow lines of flows B_θ^t. Again, let $\pi m_i/n_i$, $i = 1, \ldots, p$, be the vertex angles between the sides of P. Using the construction of R illustrated by fig. 9 we see that each vertex A_i of P with the angle $\pi m_i/n_i$ gives rise to N/n_i singular points A_{ij} of B_θ^t on R. Each A_{ij} has $2m_i$ equally spaced throngs and the flow lines in a small neighborhood of such singular point are shown in fig. 14. Singular points of this type are called multisaddles, the flow lines coming into (resp. going from) the singular point are called the incoming and the outgoing separatrices. Thus, a vertex A_i of P with the angle $\pi m_i/n_i$ gives rise to N/n_i multisaddles on R (for any B_θ^t) with m_i incoming and m_i outgoing separatrices. Varying θ we do not change the position of multisaddles A_{ij} but uniformly rotate the separatrices around A_{ij}. A simple calculation shows that the index of a multisaddle A with $2m$ separatrices (see fig. 14) is equal to $1 - m$. The index formula for the Euler number gives

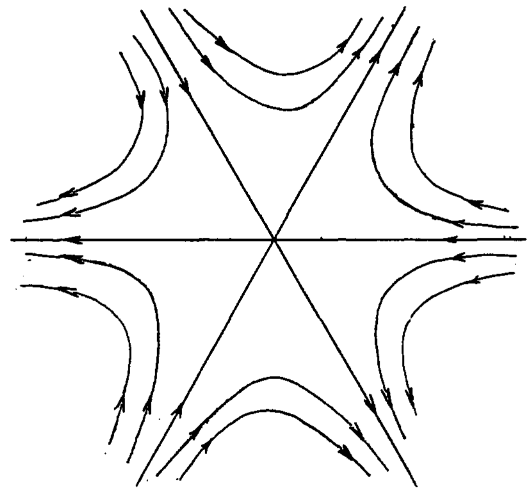


Fig. 14. Singular point with $m = 3$.

$$\chi(R) = 2 - 2g(R) = \sum_{i=1}^{p} \frac{N}{n_i}(1 - m_i), \tag{5}$$

which implies (4). Since $R_\theta \approx R$ for $0 < \theta < \pi/N$, by (4), the topology of R_θ for $\theta \neq 0, \pi/N$ is determined by the angles of P. The surfaces R_0, $R_{\pi/N}$ are called exceptional invariant surfaces. Their topology, which is not determined by the angles of P, is discussed in [9].

Now we will consider the Birkhof–Poincaré map F for rational billiards. Since the billiard flow B^t decomposes into the family B_θ^t of flows, $0 \leq \theta \leq \pi/N$, the mapping F must decompose into the

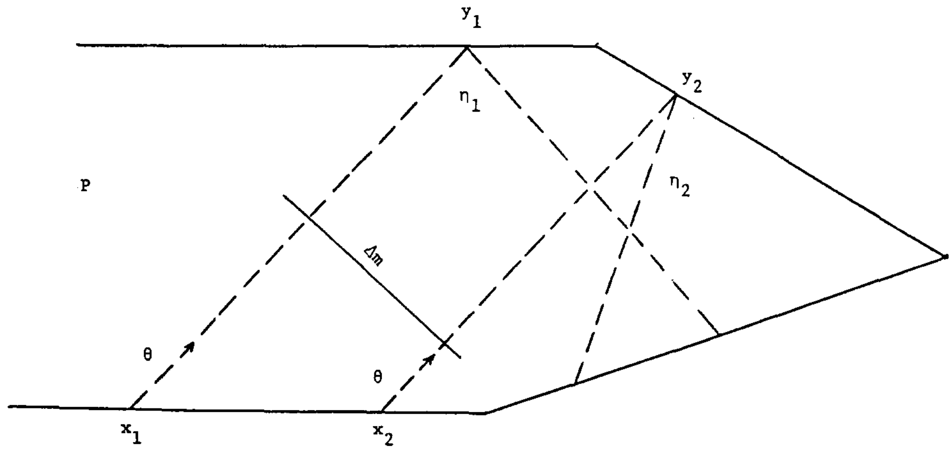


Fig. 15. Birkhof–Poincaré map; $\Delta m = (x_2 - x_1)\sin\theta$.

one-parameter family $F_\theta : \Omega_\theta \to \Omega_\theta$ mappings where $\Omega_\theta = \Omega \cap R_\theta$ is the set of vectors with footpoints on ∂P with directions equivalent to θ and $F_\theta = F|\Omega_\theta$. Using the description of F introduced in section 2 we will show that $\{F_\theta\}$ is a family of interval exchanges. We define the interval exchanges first (see, e.g. [18]). Given $0 < d_1 < \cdots < d_{n-1} < 1$ and a permutation w of $\{1,\ldots,n\}$, we define a transformation T of $I = [0,1)$. Consider I as the union of intervals $I_1 = [0, d_1), \ldots, I_n = [d_{n-1}, 1)$. Let $i_1 = w^{-1}(1), \ldots, i_n = w^{-1}(n)$. We take the interval I_{i_1} and physically pull it to the beginning of $[0,1)$. Then we pull I_{i_2} next to it and so on. If I is a piece of wire then the transformation T consists in the cutting of I into pieces $I_1, \ldots, I_n$, rearranging them according to the permutation w and welding them together again. This transformation is an interval exchange. Identifying I with the circle we can think of T as an interval exchange on the circle. An exchange of two intervals is given by one number $0 < \alpha < 1$ and it is simply the rotation by angle α.

Recall that in coordinates (x, θ) the set Ω is $[0, L) \times (-\pi/2, \pi/2)$ and $F(x, \theta) = (y, \eta)$ where η locally depends only on θ (see fig. 15). It is easy to check that F preserves $\mathrm{d}m = \sin\theta\, \mathrm{d}x$ which is interpreted as the mass element carried by the flow in direction θ. Fix a direction θ and for each side a_i of P let $-\pi/2 < \theta_{i1} < \cdots < \theta_{iN} < \pi/2$ be the directions D_N-equivalent to θ (they depend on i because a_i determines the angle of reference). The set Ω_θ is the union

$$a_1 \times \{\theta_{11}, \ldots, \theta_{1N}\} \cup \cdots \cup a_p \times \{\theta_{pl}, \ldots, \theta_{pN}\}$$

of horizontal intervals (see fig. 16). The mapping F preserves the set Ω_θ which is a union of pN horizontal intervals (they can be glued into one interval) and the length element $\mathrm{d}m$ on Ω_θ. The positive orientation of P induces an orientation of Ω_θ (each interval is oriented from left to the right) and, as we see from fig. 15, F reverses the orientation. Thus, the restriction F_θ of F to Ω_θ is an interval exchange with the flipping of intervals. Multiplying F_θ by the trivial orientation reversing map $J : (x, \theta) \to (L - x, \theta)$ we obtain an honest interval exchange. Dividing by the total length m_θ of Ω_θ which is

$$m_\theta = |a_1|(\sin\theta_{11} + \cdots + \sin\theta_{1N}) + \cdots + |a_p|(\sin\theta_{pl} + \cdots + \sin\theta_{pN}),$$

we normalize F_θ to an interval exchange on $[0,1)$ (with flipping). For a fixed polygon P the parameters of F_θ, i.e. the number of exchanged intervals, the permutation w and the lengths of intervals depend only on θ. The obvious upper bound on the number of exchanged intervals is p^2N. Now we state the first theorem about rational billiards.

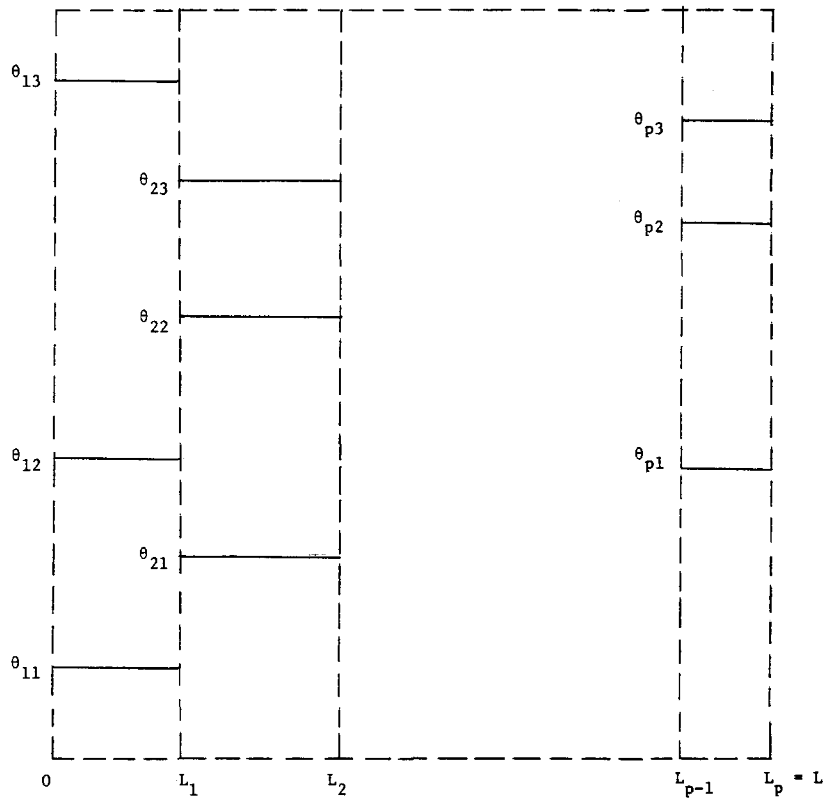


Fig. 16. Poincaré map for rational billards.

Theorem 1. Let P be rational polygon with p sides and let D_N be the corresponding dihedral group. The billiard flow of B^t decomposes into the one-parameter family B_θ^t, $0<\theta<\pi/N$, of flows on the closed orientable surface R of genus g given by (4). The flows B_θ^t are called the billiard flows in direction θ, they are obtained from any one of them by rotations. For all but a countable number of values of θ the billiard flow in the direction θ is minimal. The Birkhof–Poincaré map F of the flow B^t decomposes into the one-parameter family of interval exchanges (with flipping) F_θ, $0<\theta<\pi/N$.

Proof. The only assertion that was not proved in the previous discussion is the minimality of B_θ^t for all but a countable number of θ's. Since the flow B_θ^t and the interval exchange F_θ are minimal or not simultaneous, this can be proved using a minimality criterion for interval exchanges [2]. We will prove it here using the previous discussion.

A saddle connection for a flow F^t on a surface S is a separatrix which is outgoing for one multisaddle and incoming for another. Following [29] we will use the fact that a flow B_θ^t on R is not minimal if and only if it has either a periodic orbit or a saddle connection (and necessarily the latter if $g(R)>1$). This can be proved for the billiard flows quite elementarily [6], but it also holds in general for flows on closed orientable surfaces [21]. Saddle connections for B_θ^t correspond to the generalized diagonals of P and, by proposition 1, their number is countable. Thus, the set of their directions is countable.

Remark 1. Invariant surfaces of rational billiards were introduced apparently independently in [6], [29], [16] and [22].

Remark 2. The billiard flows B_0^t and $B_{\pi/N}^t$ respectively are discussed in [9], their directions are parallel to at least one side of P. Although interesting geometrically, they are irrelevant for the ergodic theory of rational billiards.

The surface R constructed from a rational polygon P has a canonical complex structure [10] or, as is customary to say, a conformal structure. This conformal structure is flat, i.e. $R = \mathbb{C}/\Gamma$ where Γ is a discrete group of translations, if and only if $g(R) = 1$, otherwise the conformal structure on R is hyperbolic (cf. [10]), i.e. $R = D/\Gamma$ where D is the unit disc and Γ is a discrete group of isometries of D in the Poincaré metric. The elements of Γ preserve the orientation and have no fixed points on D. This conformal structure on R, irrelevant so far, becomes important in the proof of the following.

Theorem 2 ([17], see also [31]). For Lebesgue-almost all-directions θ the billiard flow B_θ^t is ergodic.

The actual theorem of [17] says that for almost all θ the flow B_θ^t is uniquely ergodic which is stronger than Lebesgue-ergodic. We only indicate some ideas of the proof that makes use of the Teichmuller theory. The conformal structure makes R an element of the Teichmuller space T_g of all such structures on the surface of genus g. The authors of [17] relate the unique ergodicity of billiard flows B_θ^t on R with the behavior of a certain flow on T_g and use the available results of the Teichmuller theory.

Denote by $E \subset [0, \pi/N]$ the set (of full measure) of directions θ such that the flow B_θ^t is (uniquely) ergodic. Theorem 2 can be reformulated as follows.

Theorem 2′. The decomposition of a rational billiard B^t into ergodic components is $B^t = \int_E B_\theta^t \, \mathrm{d}\theta$.

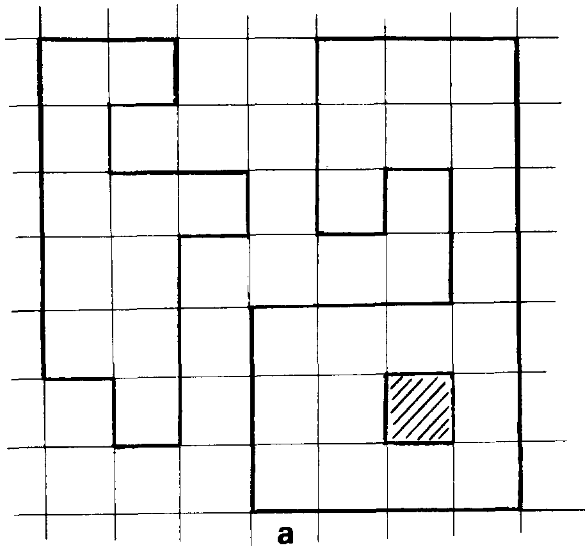
What about other ergodic properties of the flows B_θ^t? A. Katok has shown [30] that B_θ^t are not mixing. If the polygon P is almost integrable (see definition 2) then, as follows from the results of [10], the flows B_θ^t are not even weakly mixing.

It was mentioned earlier that rational billiards are only quasi-integrable because of the singularities which are due to their corners. It is clear from the previous discussion that a vertex A of P with the angle $\pi m/n$ causes singularities of B_θ^t if and only if $m > 1$. If the numerators of all vertex angles of P are equal to 1 then the billiard flow has no singularities, i.e. it is integrable and, by formula (4), its invariant surfaces R_θ are tori which agrees with the Arnold–Liouville's theorem (see, e.g. [1]). We call the polygons P satisfying $m_i = 1$ condition integrable. It is well known that the only integrable polygons are: i) rectangles, ii) equilateral triangle, iii) $\pi/2$, $\pi/4$, $\pi/4$ triangle; iv) $\pi/2$, $\pi/3$, $\pi/6$ triangle. There is a larger class of billiards which are very close to integrable by their properties.

Definition 2 [10]. A polygon P is called almost integrable if the group G_p generated by the reflections in the sides of P (see section 2) is a discrete subgroup of $\mathrm{O}(\mathbb{R}^2)$. The corresponding billiards B^t are called almost integrable.

The only infinite discrete groups of motions of the plane generated by reflections are the groups $G_1, \ldots, G_4$ corresponding to the four integrable billiards $\Delta_1, \ldots, \Delta_4$ discussed above. Denote by $L_1, \ldots, L_4$ the lattice, i.e. the pattern of lines, obtained by tiling the plane with the copies of the polygon $\Delta_1, \ldots, \Delta_4$ respectively. For instance, L_1 is the rectangular lattice and L_2 is the lattice of equilateral triangles (see fig. 17). Since the group G_P of an almost integrable polygon P is isomorphic to one of $G_1, \ldots, G_4$, the polygon P must be drawn on the corresponding lattice L_1, L_2, L_3, or L_4. We say that a polygon P is drawn on the lattice L if the vertices of P belong to the set of lattice vertices and the sides of P belong to the lines of the lattice (see fig. 17).

Thus, for any lattice L_1, L_2, L_3 or L_4 there is a countable number of almost integrable polygons



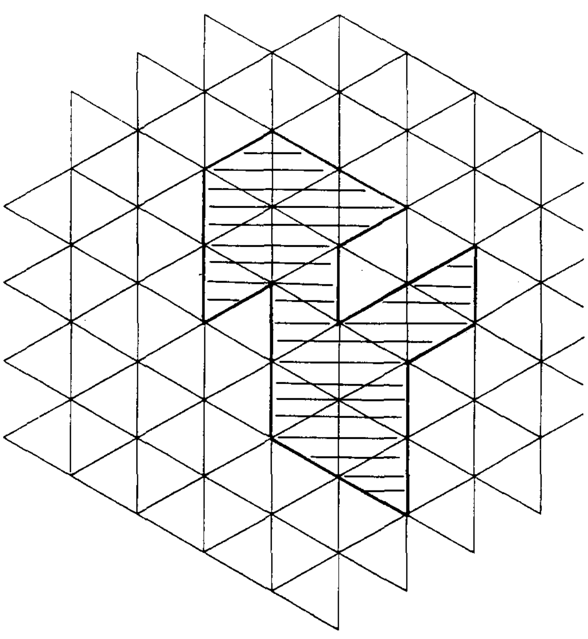
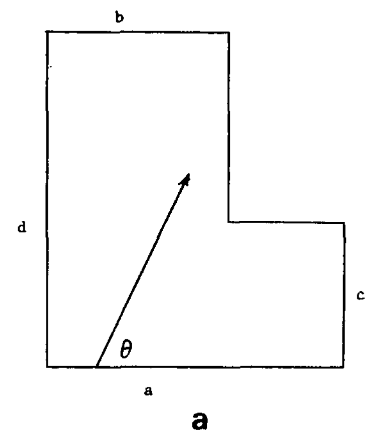


Fig. 17a) Polygons on square lattice. 17b) Polygon on the equilateral triangle lattice (shaded).



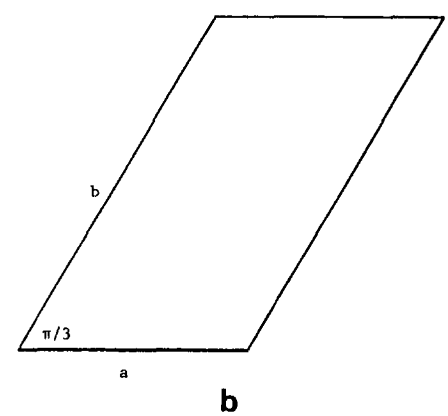


Fig. 18. a) Gnomon is almost integrable if and only if b/a and d/c are rational; b) parallelogram is almost integrable if and only if b/a is rational.

drawn on it. It is not hard to show that a polygon P is almost integrable if and only if its angles are multiples of $\pi/6$ or $\pi/4$ and its side lengths satisfy certain rationality conditions. Fig. 18 illustrates this.

Given an almost integrable polygon P, let Δ be the corresponding integrable polygon with its lattice L and let e_1, e_2 be a basis of L. A billiard direction θ in P is called rational if for a vector $v = a_1e_1 + a_2e_2$ in direction θ the ratio a_1/a_2 is rational. The definition does not depend on the choice of e_1, e_2 and v. For instance, for an almost integrable gnomon the condition is that $\tan\theta$ is rational (see fig. 18). The set of rational directions is countable.

Theorem 3 [10]. Let P be an almost integrable polygon and let B_θ^t be the billiard flow in direction

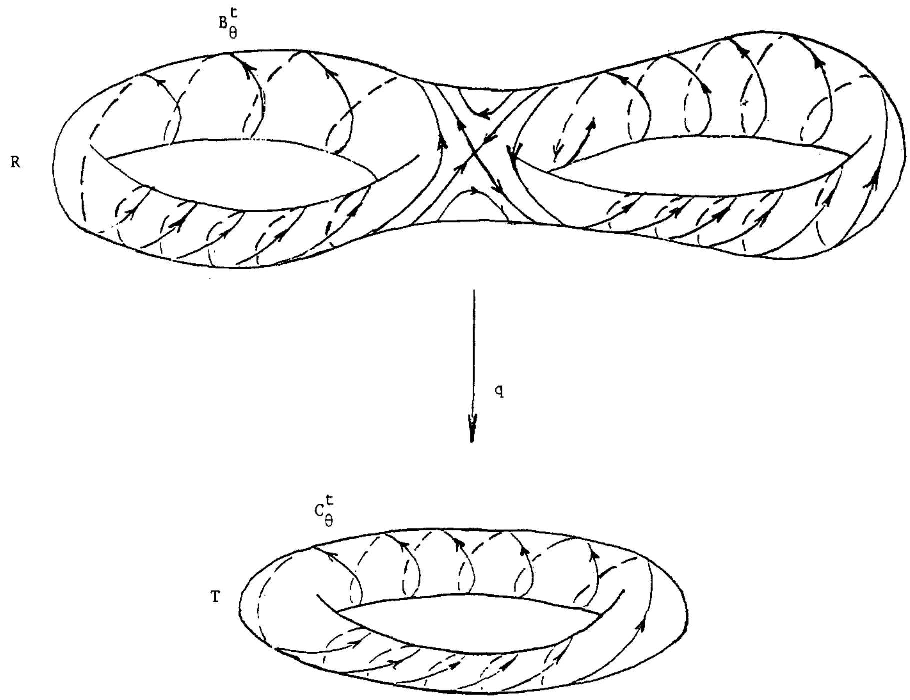


Fig. 19. Projection of flows.

θ. The following conditions are equivalent:

i) θ is irrational;
ii) B_θ^t is minimal;
iii) B_θ^t is ergodic;
iv) B_θ^t is aperiodic, i.e. $B_\theta^t \neq 1$ for any t.

We will only sketch the proof here. Denote by T the torus defined by the integrable polygon Δ corresponding to P and by C_θ^t the flow in direction θ on T (these flows are called linear or constant). There is a (branched) covering q: $R \to T$ such that the billiard flow B_θ^t projects onto the flow C_θ^t on T (see fig. 19). It is a classical theorem that the linear flow C_θ^t is ergodic if θ is irrational and periodic if θ is rational. Although the coverings of linear flows can, in general, be very complicated [23], these particular coverings are special, and using a theorem of Veech pertaining to the situation (Veech, private communication, the proof is based on an idea from [26]) we prove that B_θ^t is ergodic for irrational θ. For rational θ the flow C_θ^t is periodic, therefore, its covering B_θ^t is also periodic.

Equivalence of assertions ii) and iii) of the theorem is also contained in the results of [32].

Now we switch over to the questions about periodic billiard trajectories.

Theorem 4 [20]. Every rational billiard has at least one periodic orbit.

The proof uses the conformal structure on R and the Teichmuller theory.

Theorem 5. Let P be a rational polygon and let h be the genus of the corresponding surface R. The counting function $p(x)$ of the families of periodic orbits satisfies

$$p(x) \leq cx^{2h} \tag{6}$$

for some constant c.

Proof. It suffices to estimate the number of bands of periodic orbits of length less than or equal to x with an even number k of links. By earlier discussion, to any such orbit corresponds a parallel translation $g \in G = G_P$, we denote by G_0 the subgroup of parallel translations of G and by $G_0(k)$ the set of elements of G_0 corresponding to periodic orbits with k links. We have seen in the earlier discussion that the estimate (6) is equivalent to

$$|G_0(k)| \leq c_2 k^{2h} \tag{7}$$

for some constant c_2. Elements of $G_0(k)$ are words of length k in the canonical generators $s_1, \ldots, s_p$ of G.

As we mentioned earlier, $R = \mathrm{D}/\Gamma$ where Γ is a discrete group of automorphisms of D and there is a developing map ψ: $D \to \mathbf{C}$ which is equivariant and sends Γ onto G_0 (cf. [10]). The group Γ is the fundamental group of R, thus it has $2h$ generators $\gamma_1, \ldots, \gamma_{2h}$. The parallel translations $a_1 = \psi(\gamma_1), \ldots, a_{2h} = \psi(\gamma_{2h})$ generate G_0 and, since they commute, the number $W(k)$ of words in $a_1, \ldots, a_{2h}$ of length less than or equal to k is estimated by $|W(k)| \leq c_1 k^{2h}$. There is a constant d such that any element $g \in G_0$ of length k in $s_1, \ldots, s_p$ is a word of length no more than dk in $a_1, \ldots, a_{2h}$, that is $G_0(k) \subset W(dk)$. Thus

$$|G_0(k)| \leq |W(dk)| \leq c_1(dk)^{2h} = \left(c_1 d^{2h}\right) k^{2h}$$

which proves (7) and, therefore, (6).

Theorem 5 has an explanation in terms of the polygons $\tilde{R}$ introduced earlier. The sides of $\tilde{R}$ are divided into pairs $a_1, b_1, \ldots, a_{\tilde{n}}, b_{\tilde{n}}$ where a_i and b_i differ by a parallel translation $g_i \in G_0$, $i = 1, \ldots, \tilde{n}$. The elements g_i generate G_0. By the argument above, $p(x) \leq cx^{\tilde{n}}$. The difference between this estimate and (6) is that $\tilde{n}$ has no intrinsic meaning since it depends on the choice of $\tilde{R}$.

For almost integrable billiards the estimate (6) can be improved.

Theorem 6. Let P be an almost integrable polygon and let Δ be the corresponding integrable one. Let h be the genus of the surface R corresponding to P. Denote by $|P|$, $|\Delta|$ the areas of P and Δ respectively. Then there is a constant c depending on P, $1 \leq c \leq |P|/|\Delta|$ such that

$$p(x) = c\frac{\pi h}{|P|}x^2 + \mathcal{O}(x). \tag{8}$$

Proof. The group G_P which in the proof we denote by G is generated by reflections in the sides of Δ, the translation subgroup G_0 is discrete and uniform on the plane, i.e. G_0 has 2 generators. By the argument of theorem 5, this gives the estimate $p(x) \leq cx^2$. Let us first find the asymptotics $p_\Delta^{\mathrm{ev}}(x)$ of the number of even periodic orbits for the integrable polygon Δ. The surface T corresponding to Δ is a torus and we can take $\tilde{T}$ to be a parallelogram glued from $2N$ copies of Δ. Identify G_0 with the lattice generated by this parallelogram. A family $[\gamma]$ of periodic orbits on Δ of length x gives a vector of G_0 of the same length, any vector of G_0 corresponds to a family of periodic orbits and two vectors a_1, a_2 correspond to the same family if and only if $a_2 = wa_1$ and $w \in D_N$ (see fig. 20). Thus

$$p_\Delta^{\mathrm{ev}}(x) = \#\{a \in G_0,\ \|a\| \leq x\}/2N.$$

It is well known that the number of points of a lattice G_0 inside the disc of radius x is equal up to $\mathcal{O}(x)$ to πx^2 divided by the area of a fundamental parallelogram of G_0. Since $|\tilde{T}| = 2N|\Delta|$, we have

$$p_\Delta^{\mathrm{ev}}(x) = \frac{\pi}{|\Delta|}x^2 + \mathcal{O}(x) \tag{9}$$

which is a special case of (8).

Let γ be a periodic orbit in Δ with an odd number of links and let g be the corresponding element of G. Then g is a sliding reflection with an axis l and the direction of l is a reflecting direction of D_N. Obviously, the axis l of g passes

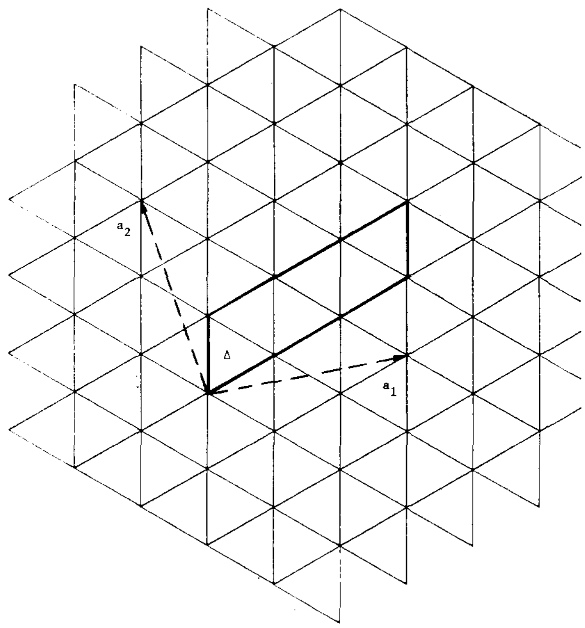


Fig. 20. Counting of periodic trajectories in equilateral triangle.

through Δ. By discreteness of G, there is only a finite number of such lines. For a given axis l the set of sliding reflections in G about l is obtained from one of them, say g_0, by translations along l. Let g_0 be the shortest sliding reflection about l and let t_0 be the shortest translation along l (in G). Then the set in question is $\{g_0 \cdot nt_0 : n \in \mathrm{Z}\}$. Thus, the set of elements $g \in G$ corresponding to odd periodic orbits is "one dimensional", therefore $p_\Delta^{\mathrm{odd}}(x) = \mathcal{O}(x)$ and we have

$$p_\Delta(x) = p_\Delta^{\mathrm{ev}}(x) + p_\Delta^{\mathrm{odd}}(x) = \frac{\pi}{|\Delta|}x^2 + \mathcal{O}(x). \quad (10)$$

Now we discuss the growth function $p(x)$ for the polygon P. The same argument as above shows that $p^{\mathrm{odd}}(x) = \mathcal{O}(x)$, so it suffices to consider the even periodic orbits. Realizing B^t on R and recalling the projection $q: R \to T$ we see that a periodic orbit γ on R projects onto a periodic orbit $\tilde{\gamma} = q(\gamma)$ on T. Orbits γ and $\tilde{\gamma}$ define the same parallel translation $g \in G_0$. Denote by $G_0(P)$ the subset of G_0 obtained this way. If $g \in G_0$ is arbitrary and if $\tilde{\gamma}$ is the corresponding periodic orbit on T, its lifting γ on R is not necessarily periodic, but γ iterated d times where d is the order of the covering $q: R \to T$ is periodic. Thus, $G_0(P)$ contains the subgroup dG_0 of G_0. If γ is a periodic orbit on R, its projection $q(\gamma)$ on T generates one band of periodic orbits. On the other hand, translating γ parallel to itself we obtain several bands of periodic orbits, passing from one band to another each time we hit a corner. By results of [21], the number of bands obtained this way is equal to h, the genus of R. Thus, we have at most h different bands of periodic orbits on R that project into the same orbit on T. The degree d of the covering $q: R \to T$ is equal to $|P|/|\Delta|$. We conclude from this discussion that

$$hd^{-1}p_\Delta^{\mathrm{ev}}(x) \le p^{\mathrm{ev}}(x) \le hp_\Delta^{\mathrm{ev}}(x) \quad (11)$$

which implies, by (10), and since $p^{\mathrm{odd}}(x) = \mathcal{O}(x)$,

$$h\frac{\pi}{|P|}x^2 + \mathcal{O}(x) \le p(x) \le h\frac{\pi}{|\Delta|}x^2 + \mathcal{O}(x). \quad (12)$$

Formula (8) immediately follows.

Theorems 1–6 answer the questions of section 3 in the context of rational and almost integrable billiards. In the next section we summarize what is known for irrational and "typical" billiards.

5. Irrational billiards

By irrational billiard we mean that at least one angle of the polygon P is not a rational multiple of π. Recall that there is a natural topology on the set of polygons with a fixed number of sides. A subset X of a topological space is called G_δ if X is an intersection of a countable number of open sets.

Theorem 7 [29]. There is a G_δ set X dense in the space of all polygons such that the billiard in any $P \in X$ is transitive.

Proof. Consider, for simplicity of exposition, the space $\mathcal{P}$ of simply connected polygons with p

sides. For any polygon P from $\mathscr{P}$ the interior $T(P)_{\text{int}}$ of the phase space $T(P)$ is homeomorphic to the product $D \times S^1$ of the open unit disc and the circle and we can choose the homeomorphisms h_P: $D \times S^1 \to T(P)_{\text{int}}$ so that they depend continuously on P. Choose a countable basis B_k, $k = 1, 2, \dots$ of open sets in $D \times S^1$. For instance, we can choose B_k to be balls in the natural metric on $D \times S^1$. The set $\{B_k\}$ is a basis means that any ball in $D \times S^1$ contains B_k for some k. For any n denote by $X_n \in \mathscr{P}$ the set of polygons P that contain a billiard trajectory passing through all $h_P(B_k)$ for $k = 1, \dots, n$. The set X_n is open by construction, so $X = \cap X_n$ is a G_δ set. If P belongs to X then there is a billiard trajectory γ in P passing through $h_P(B_k)$ for all k, i.e. γ is dense in $T(P)$. Thus, the billiard in P is transitive. It remains to show that X is dense in $\mathscr{P}$. Denote by Y_m the set of rational polygons P such that the least common denominator N of the angles of P satisfies $N \geq m$. It is easy to see that any invariant surface R_θ of a polygon P in Y_m is $1/m$-dense in the phase space $T(P)$. Thus, for any n there exists m such that the invariant surface R_θ for any $P \in Y_m$ intersects all $h_P(B_k)$ for $k = 1, \dots, n$. If R_θ has a dense trajectory γ (which is true for almost all θ) then γ intersects all $h_P(B_k)$ for $k = 1, \dots, n$, thus, $Y_m \subset X_n$. Since Y_m is everywhere dense in $\mathscr{P}$ for any m, we see that X_n is dense in $\mathscr{P}$ for any n, hence X is a dense subset of $\mathscr{P}$.

We say, for brevity, that a polygon P is ergodic if the billiard in P is ergodic.

Theorem 8 [15], [17]. In the space of polygons with a given number of sides there is a dense G_δ set of ergodic polygons.

The proof relies on theorem 2 and goes along the lines of the proof of theorem 7.

Now we come to the estimation of the growth rate of the counting function $p(x)$ for periodic orbits on P. Let $G = G_P$ be the group associated with P and let G_0 be the subgroup of translations in G. The approach of theorem 5 is based on the fact that for rational polygons P the group G has a polynomial growth, which implies that G_0 has a finite number of generators. This approach completely fails for irrational polygons since for them G has an exponential growth and G_0 has an infinite number of generators [11]. The following theorem shows that $p(x)$ nevertheless grows subexponentially, i.e. $\lim_{x \to \infty} \log p(x)/x = 0$.

Theorem 9 [13]. For any polygon P the number of families of periodic orbits of length less than x grows subexponentially.

The proof is based on the study of the Birkhof–Poincaré map for the billiard on P.

On the other hand, the existence of periodic orbits is not proved for irrational billiards. We conclude the paper by describing some results from [7] on the periodic and other billiard orbits. The author of [7] constructs billiard orbits which are neither periodic nor spatially dense. He shows that for any $n \geq 3$ there exist n-gons with such orbits.

Here is the main idea of the proof. Let P be a polygon and let $\Omega = \partial P \times S^1$ be the set on which the Birkhof–Poincaré map F acts. We want to have a subset Y of Ω such that any $y \in Y$ returns to Y after some number of iterations of F. Denote by $T: Y \to Y$ the first return map. Assume, furthermore, that T is an exchange of two intervals. Besides, we want Y to have the form $Y = X \times \{\theta\}$ where X is an interval on ∂P. Then F restricted to Y corresponds to the emitting from X in direction θ the beam of trajectories which after some reflections come back to X and induce an exchange of two intervals. Galperin claims without proof [7] that any interval exchange can be realized in this way by a convex polygonal billiard.

Consider the centrally symmetric hexagon shown on fig. 21. The hexagon has the property that the perpendicular to AC′ (resp. BC) from A (resp. B) bysects the angle BAC (resp. ABC′). Denote by α (resp. β) one half of the angle BAC (resp. ABC′). It is elementary to show that the beam X of trajectories parallel to AB is invariant and that F

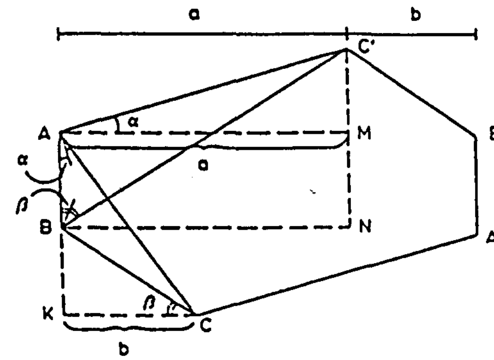


Fig. 21. Billiard in hexagon induces an exchange of two intervals.

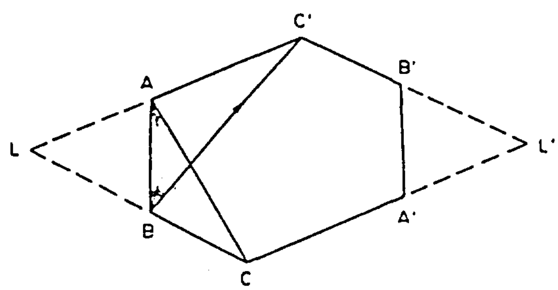


Fig. 22. Parallelogram L C L′ C′ has a trajectory filling the hexagon A B C A′ B′ C′.

induces on X the exchange of two intervals a and b (see fig. 21) with $a/b = (\cot 2\alpha - \tan\beta)/(\cot 2\beta - \tan\alpha)$. Obviously a/b is irrational for almost all hexagons H of this type. If a/b is irrational for H then any infinite trajectory γ from the family X is spatially dense in H. Let P be the parallelogram L C L′ C′ obtained from H as shown on fig. 22. Then γ is a billiard trajectory in P whose closure is H. Cutting P outside of H we obtain polygons P' with any number of sides ≥ 4 and with this property.

The same idea works for the right triangles with small acute angles but the construction is much more complicated (cf. [7], fig. 8).

Let now P be a polygon with a family X of trajectories realizing an exchange T of two intervals a and b and let a/b be irrational. Denote by θ_0 the direction of X and by H the closure in P of any infinite trajectory γ from X. Using elementary facts about continued fractions Galperin shows in [7] that γ can be approximated as close as we like by periodic orbits. More precisely, for almost all $x \in H$ and for any $\varepsilon > 0$ there exists θ, $|\theta - \theta_0| < \varepsilon$, such that the trajectory through (x, θ) is periodic. This theorem provides us with examples of polygons with lots of periodic orbits. Such are, for instance, almost all hexagons from fig. 21.

References

[1] V.I. Arnold, Mathematical methods of classical mechanics (Springer, Berlin, 1978).

[2] C. Boldrighini, M. Keane and F. Marchetti, Billiards in polygons, Ann. Prob. 6 (1978) 532–540.

[3] M. Boshernitzan, A condition for minimal interval exchange maps to be uniquely ergodic, to appear in Duke Math. J., 1985.

[4] L.A. Bunimovich, On the ergodic properties of nowhere dispersing billiards, Comm. Math. Phys. 65 (1976) 295–312.

[5] Y. Colin De Verdiere, Sur les longueurs des trajectories periodiques d'un billiard, 122–139, Seminaire sud-rhodanien de geometrie. III, Travaux en Cours (Hermann, Paris, 1984).

[6] R.H. Fox and R.B. Kershner, Concerning the transitive properties of geodesics on a rational polyhedron, Duke Math. J. 2 (1936) 147–150.

[7] G.A. Galperin, Nonperiodic and not everywhere dense billiard trajectories in convex polygons and polyhedrons, Comm. Math. Phys. 91 (1983) 187–211.

[8] G.A. Galperin, Elastic collisions of particles on a line, Russ. Math. Surveys 33 (1978) 211–212.

[9] E. Gutkin, Geodesic flows on rational polyhedra, preprint.

[10] E. Gutkin, Billiards on almost integrable polyhedral surfaces, Ergod. Th. and Dynam. Sys. 4 (1984) 569–584.

[11] E. Gutkin, private communication.

[12] F.S. Henyey and N. Pomphrey, The autocorrelation function of a pseudointegrable system, Physica 6D (1982) 78–94.

[13] A. Katok, The growth rate for the number of saddle connections for a polygonal billiard, preprint.

[14] A. Katok and M. Wojtkowski, Billiards, in preparation.

[15] A. Katok, Ergodicity of generic irrational billiards, Proc. MSRI conf. 2-manifolds, Oct. 1984, p. 13.

[16] M. Keane, Coding problems in ergodic theory, Proc. Int. School of Math. Phys., Univ. Camerino, 1974.

[17] S. Kerckhoff, H. Masur and J. Smillie, Ergodicity of billiard flows and quadratic differentials, preprint.

[18] I.P. Kornfeld, S.V. Fomin and Y.G. Sinai, Ergodic Theory (Springer, Berlin, 1982).

[19] V.F. Lasutkin, The existence of caustics for a billiard problem in a convex domain, Math. of USSR-Izv. 7 (1973) 185–214.

[20] H. Masur, Rational billiards have periodic orbits, preprint.

[21] A. Mayer, Trajectories on the orientable surfaces, Mat. Sbornik 12 (1943) 71–84 (in Russian).

[22] P.T. Richens and M.V. Berry, Pseudointegrable systems in classical and quantum mechanics, Physica 2D (1981) 495–512.

[23] E.A. Sataev, On the number of invariant measures for flows on orientable surfaces, Math. of USSR-Izv. 9 (1976) 813–830.

[24] Y.G. Sinai, Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards, Russ. Math. Surveys 25 (1970) 137–189.

[25] Y.G. Sinai, Billiard trajectories in a polyhedral angle, Russ. Math. Surveys 33 (1978) 229–230.

[26] W. Veech, Finite group extensions of irrational rotations, Isr. J. Math. 21 (1975) 240–259.

[27] M. Wojtkowski, Principles for the design of billiards with nonvanishing Lyapunov exponents, preprint.

[28] A. Katok and J.M. Strelcyn, Smooth maps with singularities: invariant manifolds, entropy and billiards, Université Paris-Nord, Prepublications math., n 56.

[29] A.N. Zemlyakov and A.B. Katok, Topological transitivity of billiards in polygons, Math. Notes 18 (1975) 760–764.

[30] A. Katok, Interval exchange transformations and some special flows are not mixing, Isr. J. Math. 35 (1980) 301–310.

[31] S. Kerkhoff, H. Masur and J. Smillie, A rational billiard flow is uniquely ergodic in almost every direction, Bull. AMS 13 (1985) 141–142.

[32] M. Boshernitzan, Rank two interval exchange maps, in preparation.