A. B. Katok

Billiard table as a playground for a mathematician

Lecture on March 10, 1999

The title of this lecture can be understood in two ways. Literally, in a somewhat facetious sense: mathematicians are playing by launching billiard balls on tables of various forms and observing (and also trying to predict) what happens. In a more serious sense, the expression "playground" should be understood as "testing area": various questions, conjectures, methods of solution, etc. in the theory of dynamical systems are "tested" on various types of billiard problems. I hope to demonstrate convincingly that at least the second interpretation deserves serious attention.

The literature concerning billiards is rather large, including scientific papers as well as monographs, textbooks, and popular literature. Short brochures by G. A. Galperin and A. N. Zemlyakov [4] and by G. A. Galperin and N. I. Chernov [5] are written in a rather accessible manner, and touch a broad circle of questions. An introduction to problems related with billiards for a more advanced reader is contained in Chapter 6 of the book [9]. The next level is represented by a very well written book of S. Tabachnikov [14], whose publication in Russian is unfortunately delayed. The book by the author and B. Hasselblatt [8] contains a rather detailed modern exposition of the theory of convex billiards and twisting maps. A serious but rather accessible exposition of modern state of the theory of parabolic billiards is contained in a survey paper by H. Masur and S. Tabachnikov which will be published (in English) in spring 2002 [11]. The collection of papers [12] contains rich material on hyperbolic billiards and related questions. More special references will be given below during the exposition.

1 Elliptic, parabolic, and hyperbolic phenomena in dynamics

The problem of motion of a billiard ball is stated in a very simple way. One has a closed curve $\Gamma \subset \mathbb{R}^2$. Inside the domain $\mathcal{B}$ bounded by the curve, one has a uniformly moving point which covers segments of straight lines, and when the point meets the curve, it is reflected according to the rule "the angle of incidence equals the angle of reflection." The problem is to understand the nature of this motion at a large time.

Here we have a dynamical system which is in general not everywhere defined. For example, if in a domain with a piecewise smooth boundary the point hits an angle, then it is unclear how to continue the trajectory. There are also more delicate effects: for some initial conditions it is possible that during a finite time an infinite number of hits of the boundary occurs and the motion cannot be continued. But these effects are pathological; one can say that we have a dynamical system.

The solution of the problem of motion of a ball depends on the domain. One of the reasons of interest to this problem is that the formal description of the motion is very simple and only the essential part is to be investigated. The second, more serious reason has been already mentioned. It is related to the fact that if one attempts to somehow classify problems of theory of dynamical systems, then, in a somewhat rough manner, they can be divided into elliptic, parabolic, and hyperbolic ones (see Fig. 1). Thus, a billiard table is a testing area on which one can test methods, conjectures, questions, arising in various fields of theory of dynamical systems.

There is nothing new in using these words for expressing some trichotomy. The corresponding classification in theory of partial differential equations is well known. But for dynamical systems, such a classification seemingly has not been carried out systematically.

In the case of billiards, elliptic effects arise, for example, for an ellipse. This coincidence is not completely accidental, but it cannot be extended to billiards inside a parabola or a hyperbola. A more general situation in which elliptic effects occur, is as follows: the curve is smooth (of a sufficiently large class of smoothness), convex, and its curvature nowhere vanishes. The study of the billiard problem inside such domains gives a good example for demonstration of problems and results related to elliptic behavior of dynamical systems.

In a parabolic situation, the domain is a usual polygon. For simplicity one can even take a right-angled triangle whose angles differ from 30° and 45°. A right-angled triangle with the angle $\pi/8$ already gives an example of a dynamical system with parabolic behavior.

The hyperbolic situation is well represented by three examples (see Fig. 1): a square with a small disk removed, a "stadium," and a cardioida.

The idea about at least a dichotomy which exists in the theory of dynamical systems has been widely accepted in the last years. One of the most remarkable books on the theory of dynamical systems written in the second half of the twentieth century is the book by Yorgen Moser "Stable and random motion in dynamical systems." "Stable" means elliptic effects, "random" means hyperbolic effects. Parabolic effects are not discussed in Moser's book.

To give some idea on the nature of the trichotomy arising here, let me explain

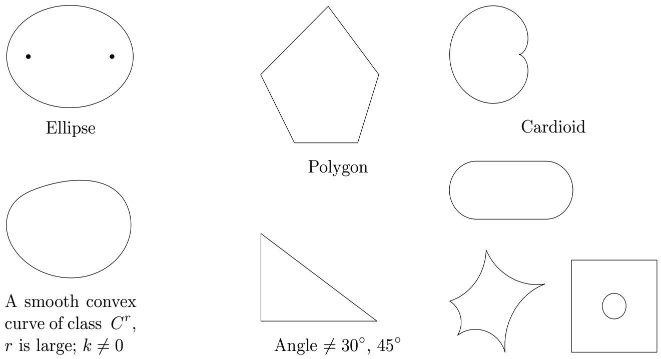

Ellipse
Polygon
Cardioid
A smooth convex curve of class C^r, r is large; $k \neq 0$
Angle $\neq 30^\circ$, 45°

Figure 1. Elliptic, parabolic, and hyperbolic billiards

the origin of these terms. For linear mappings the corresponding trichotomy is well known. For a linear mapping $L\colon \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ three main kinds of behavior are possible:

(1) Stable behavior. It arises when all the eigenvalues λ_i have absolute value 1 and there are no nontrivial Jordan blocks: $\mathrm{Sp} L \subset S^1$. In this situation all the orbits come back and are stable. This is elliptic behavior.

(2) Again $|\lambda_i| = 1$, but there are nontrivial Jordan blocks. A Jordan block has an eigenvector, therefore there are stable orbits. But in this situation, typical is the polynomial growth of distance between orbits. This is parabolic behavior.

(3) Hyperbolic behavior: $\mathrm{Sp} L \cap S^1 = \varnothing$. In this situation the distance between any two orbits exponentially grows either in the positive or in the negative direction.

Combinations of these three paradigms are also possible. For example, a rather important situation is what is called partially hyperbolic behavior, when the spectrum contains a hyperbolic component and something else. This is a very important paradigm in dynamics.

It would be very naive to attempt to construct a concept of nonlinear differential dynamics based only on these three models. What is the subject of nonlinear differential dynamics? It is the analysis of asymptotic behavior of smooth systems for which one has the notion of local infinitesimal behavior of the system and, on the other hand, due to compactness of the phase space, there is the phenomenon of returning of orbits arbitrarily closely to their initial position. Roughly speaking, dividing nonlinear dynamical systems into elliptic, parabolic, and hyperbolic ones corresponds to the situations when the linear behavior, which is more or less approximated by these three types, is combined with the nontrivial nature of returning.

This approach ignores a very essential part of problems of theory of dynamical systems, for example, such things as analysis of Morse–Smale systems or effects related with bifurcations. These are situations when returning is simple, and interesting phenomena are related, for example, to how the phase space is divided into basins of attraction to several attracting points or limit cycles. All of this is ignored. We are now speaking only on the part of dynamics which is related to recurrent behavior. Nonrecurrent behavior is more or less ignored by us now.

To interpret phenomena interesting to us correctly, one must understand what is a linearization of a dynamical system. Let $f\colon M \to M$ be a given map acting on the phase space. We assume that the phase space is a smooth object, hence one can speak about the action on tangent vectors. For any point $x \in M$ one has the linear map $Df_x\colon T_xM \to T_{f(x)}M$. Such a map is interesting by itself only in the case of a fixed point. But in the general case in dynamics one can consider the iterations Df_x^n. Introducing a Riemannian metric, one can speak about the asymptotic speed of growth of length of vectors. A Riemannian metric can be introduced not uniquely, but on a compact manifold any two metrics differ no more than by a multiplicative constant, therefore the speed of growth of vectors is defined correctly.

The elliptic behavior arises when in the linearized system either there is no growth of length of vectors at all, or it is slower than a linear one (a sublinear growth).

A Jordan block of minimal size 2 already corresponds to a linear growth. The parabolic behavior is a subexponential growth (usually a polynomial one).

The hyperbolic paradigm is the most well understood one. It corresponds to the situation when the system splits and in some directions one has exponential growth, and in the other directions one has exponential decay. When the time is reversed, these directions are exchanged with one another.

Sometimes mixed situations occur. For example, one can take the direct product of two systems of different type. But as a metastatement, one can say that the hyperbolic paradigm dominates: if there is a nontrivial hyperbolic effect and something else, then usually the behavior of the system can be understood based on its hyperbolic part. This is not true literally. For example, this is not true for the direct product of dynamical systems. But for a typical dynamical system the hyperbolic behavior dominates everything else.

One can also make the following interesting remark. When the dimension of the phase space is small, mixed behavior is impossible. For example, when the dimension of the phase space equals 2, the partially hyperbolic behavior is impossible, because for the hyperbolic behavior one needs at least one extending and one compressing direction. By the same reason in small dimensions the elliptic or parabolic behavior occurs more often.

The most pure example of elliptic behavior is the situation when one has a smooth isometry. In this case there is no growth. For smooth isometries dynamics can be understood rather easily. If on a compact phase space there is a smooth isometry, then the phase space splits into invariant tori, and on each torus a parallel translation arises (or a rotation, if one uses multiplicative notation). In particular, if the manifold itself is not a torus, then such motion is not transitive.

Of course, it is a particular case of what is well known in Hamiltonian mechanics, namely, it corresponds to completely integrable Hamiltonian systems. This is a good example of interaction of paradigms, because, if one looks at a completely integrable system naively, then it should be attributed to the parabolic paradigm. Indeed, the linear part of a completely integrable system is parabolic, because in the direction transversal to the invariant tori one has a twisting. On the other hand, the space splits into invariant tori, and on each torus analysis is carried out with elliptic methods.

This situation is typical. This is the reason why the elliptic paradigm is important. It is a rather rare case that the global behavior on the whole phase space is characterized by absence of growth. But rather frequently one has some elements inside the phase space, where the behavior can be described by means of the elliptic paradigm.

The hyperbolic situation is the most well studied. In a sense, it is the only universal paradigm of complex behavior in dynamics. It can be well understood with the help of Markov chains and simple stochastic models. From the viewpoint of applications of dynamics, if the hyperbolic behavior is established, then one can apply a rather powerful machinery which makes it possible to study the behavior of nonlinear systems. All this arises due to the interaction of a certain behavior of the linearized system with more or less *a priori* existing returning. In linear systems hyperbolicity is followed just by running away of the system to infinity. But if there is no space to run away, if one necessarily has to return, then the above mentioned and well understood types of complex behavior arise.

In contrast to elliptic and hyperbolic behavior, parabolic behavior is, firstly, unstable, and, secondly, it is characterized by absence of standard models. In the elliptic situation one has a universal model, namely, rotation on the torus (or some of its avatars), and in the hyperbolic situation one has the Markov model which describes everything. In the parabolic situation, seemingly, one even cannot say that there is a set of models to which everything is more or less reduced. Nevertheless, there are rather typical phenomena which occur in concrete classes of systems. One of these phenomena consists in that frequently the effect of moderate stretch can be replaced by the effect of cutting. For example, if one has a system which locally looks like an isometry but has discontinuities, then such a system is concerned with the parabolic paradigm.

A well known example is exchange of segments. We cut a segment into pairs and exchange them according to a permutation given *a priori*. Locally this system looks like an elliptic one, but there is the effect of cutting. It is rather easy to realize that this system should be considered as a parabolic one: during the iterations the number of segments grows in a linear way. This linearity is not the result of twisting, but it is the result of cutting. But the effect is approximately the same.

Thus, parabolic behavior is frequently related to the presence of moderate singularities in systems. So it is not occasional that a polygon was drawn on the picture illustrating parabolic behavior.

2 Billiards in smooth convex domains

George K. Birkhoff was the first to consider billiards systematically as models for problems of classical mechanics. Birkhoff considered billiards only in smooth convex domains. Of course, he did not think about billiards in polygons, and all the more in nonconvex domains.

First of all, one can perform the reduction to the billiard mapping. The initial dynamical system for a billiard is a system with continuous time. But the trajectory inside the billiard table can be easily reconstructed if one knows what happens at the moments of reflection. Therefore it suffices to consider the so-called billiard mapping. The phase space of the billiard mapping looks as follows. The vector v outcoming after reflection is characterized by the cyclic coordinate $\phi \in S^1$ which fixes the position of the point on the curve Γ, and the angle $\theta \in [0, 2\pi)$ between the tangent vector and the vector v (Fig. 2).

The phase space of the billiard mapping is a cylinder. After the reflection we obtain a new point ϕ_1 and a new outcoming vector, which corresponds to the angle θ_1. The map $T(\phi_0, \theta_0) = (\phi_1, \theta_1)$ is what is called the billiard mapping. It maps the open cylinder into itself; by continuity it can be extended to the closed cylinder. The points with $\theta = 0$ are fixed (we assume that the curve Γ does not contain straight segments).

Exercise 1. Show that the presence of straight segments in the boundary implies discontinuity of the billiard mapping.

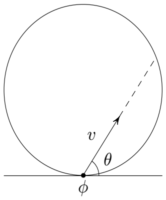


Figure 2. Vector coordinates after reflection

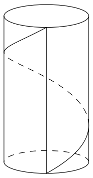


Figure 3. Twisting

Exercise 2. Find conditions under which the billiard mapping is differentiable (one or infinitely many times) on the boundary of the cylinder.

The billiard mapping possesses two important qualitative properties.

1) Conservation of area. The element of area

$$dA = \sin\theta d\theta d\phi = d\alpha d\phi, \qquad \text{where} \quad \alpha = \cos\theta,$$

is conserved. (To introduce coordinates in which area is conserved, one should take $\cos\theta = \alpha$ instead of θ.)

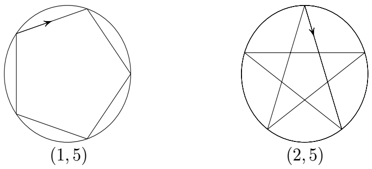
2) Twisting. Let us fix the coordinate ϕ_0 and change the coordinate θ. Then the coordinate ϕ of the image will change monotonously until it passes all the circle and comes back (Fig. 3). The image of a vertical line is twisted.

These two properties allow one to realize that one has elliptic behavior. The problems arising in connection with elliptic behavior are divided into two parts:

1) caustics,

2) Birkhoff orbits and Aubry–Maser sets.

Let me begin with the second part. We want to find periodic orbits of the billiard system. Periodic orbits can be various. They differ not only by the period, but also by some combinatorics. For example, two orbits of period 5 in Fig. 4 have different combinatorics. In the first case there is one rotation, and in the second case there are two. These orbits are regular: the order of



(1, 5) (2, 5)

Figure 4. Orbits with equal periods and different combinatorics

points on the orbit is conserved; it is the same as in the rotation. It is these (regular) orbits that are called Birkhoff orbits. This name is related to the fact that he has proved a remarkable and relatively simple theorem on existence of regular orbits. Seemingly, this theorem was the starting point of application of variational methods in dynamics.

Theorem 1 (Birkhoff). *For any two coprime numbers p and q there exist at least two periodic orbits of the type (p, q).*

Sketch of the proof. The proof uses only convexity and smoothness. Consider various inscribed polygons with the required combinatorial properties. Let us call such polygons by states. The states form a finite-dimensional space. On the space of states there is the functional of length. If we allow the vertices of polygons to coincide, then we obtain a compact space. Hence the length functional has maxima.

Any extremal point of the length functional is a billiard orbit (if this point is not on the boundary). This is a local statement. It is easy to check that the derivative of the length vanishes if and only if the angles are equal. The linear part of variation of the functional depends just on the difference of the angles.

It is easy to prove that the maximum cannot be achieved on the boundary, i.e., the vertices of a polygon cannot coincide.

Thus, the longest polygon is a required periodic orbit. But this is still only the easiest part of the theorem. One has to find another periodic orbit. This can be done in the following way. Cyclic renumeration of the vertices of the orbit we have found gives q maxima. Let us deform one of these maxima into another one. If we go from one maximum to another one, then we have to go down. Let us try to lose a smallest possible height. In this case we need to pass a saddle (Fig. 5), because if at the lowest point we were not on a saddle, then we could change the trajectory a little and decrease the lack of height. A saddle is also a critical point, i.e., the required periodic orbit.

If we do not lose height at all, then in this case one has a whole family of periodic orbits. □

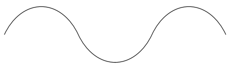


Figure 5. A saddle

This proof shows concisely how one can change a difficulty by another one. The difficult point of this argument is in how to keep aside from the boundary. This can be easily achieved if we just do not consider the boundary, and consider all states. Evidently, the function in question is bounded: any edge is no longer than the diameter of the curve. One can omit the condition of ordering of points and then prove that the global maximum is necessarily approached on a correctly ordered orbit. If, for instance, we consider globally maximal orbits which make two rotations during a full round, then they do it in a correct order. And instead we can prove that one can avoid the boundary inside an ordered family.

The importance of the Birkhoff theorem is in that we immediately find infinitely many periodic orbits.

Now an interesting story begins on how Birkhoff missed an important discovery.

Birkhoff presents his variational argument, and then he says that in exactly the same manner one can purely topologically prove the so-called last geometric theorem of Poincaré: "If the bases of a cylinder rotate in different directions with area being conserved, then such a diffeomorphism has at least two fixed points." Moreover, if the angles of rotation on the upper and lower bases are different, then for any rational angle of rotation one can find a corresponding periodic orbit, even without the condition of twisting.

Birkhoff was extremely proud to prove the last geometric theorem of Poincaré. But he missed a very remarkable conclusion of his own elementary proof. This conclusion is the following. Let us look what happens in passing to the limit $p_n/q_n \to \alpha$, where α is an irrational number. Usually in dynamics such tricks would not work, because the asymptotic behavior is unstable with respect to initial data, and one cannot pass to the limit. But here, just because we are dealing with the elliptic situation, a simple but surprising phenomenon arises. If we consider a Birkhoff orbit on the cylinder, then it consists of a finite number of points. If the number q_n is large, then the number of points will be also large and they will be strongly condensed. It is rather easy to prove that these points always lie on a Lipschitz graph (i.e., on the graph of a function satisfying the Lipschitz condition). The Lipschitz constant here is fixed, it does not depend on the length of an orbit. The set of Lipschitz functions with a given Lipschitz constant is compact, hence one can pass to a limit. In a

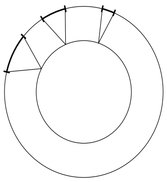


Figure 6. The Denjoy counterexample

somewhat different way, one can say the following: let us take finite orbits and consider their limit in the Hausdorff topology. In the Hausdorff topology closed subsets of a compact set form a compact set, hence the limit exists, which is not surprising. But the limit is an invariant set which is a subset of a Lipschitz graph, because in the Hausdorff topology subsets of Lipschitz graphs form a closed subset.

We still don't know what is the geometry of the obtained graph, but we know what is its dynamics. Its dynamics is the same as the one of the rotation on the angle α, nothing else can occur. Indeed, the order in which the points are transposed under rotation on the angle α, is uniquely determined by the orders in which the points are transposed under rotations on the angles p_n/q_n approximating the angle α. Hence on any finite segment in the limit combinatorics will be the same as needed, because on any finite segment combinatorics is stabilized and is the same as under rotation on the angle α.

Thus, a closed invariant set on a circle arises (since topologically the limit Lipschitz graph is a circle). This set has a dynamics which preserves the order and exactly reconstructs rotation on the angle α. From the times of Poincaré it is known when this is possible: either the invariant set is the whole circle, the orbits are dense and the transform is conjugate to a rotation of the circle (this is, of course, elliptic behavior, at least in the topological sense), or the circle contains an invariant Cantor set which arises in the so-called Denjoy counterexample (see, for example, Chapters 11 and 12 in [8]). The Denjoy counterexample looks as follows. Let us take a point on the circle and blow it up into an interval. Then its image and preimage should be also blown up into intervals (Fig. 6), etc. To have convergence, one needs these intervals to be smaller and smaller. This is rather easy to make topologically. As a result, one gets a transform of the circle which contains an invariant Cantor set and which is half-conjugated to a rotation (there exists a continuous mapping which makes it a rotation, but these intervals shrink into points).

For transforms of a circle such a behavior is exotic, because by Denjoy's theorem this is impossible in the class C^2, it is possible only in the class C^1.

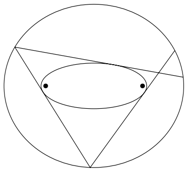


Figure 7. Confocal ellipses

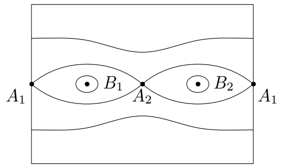


Figure 8. Trajectories in the phase space

But for twisting maps this is rather normal behavior (of course, if it did not happen that we get a full circle). Thus, an interesting alternative arises. When one has accumulation of Birkhoff orbits on an invariant set (this is exactly the Aubry–Maser set), this invariant set is either a Cantor set (possibly with some additions) or the whole circle. The case of the whole circle is called a caustic. One of these two cases holds always and for any rotation number. The corresponding Cantor set is unique, i.e., if one removes from the invariant set the wandering part corresponding to separate wandering points, then the remaining Cantor set is unique. But this does not obstruct the existence of other Cantor sets which have the same rotation number and on which the order is preserved, but this order, although compatible with the cyclic order, would not be compatible with the order on this set. The Cantor set constructed as the limit of Birkhoff orbits of maximal length is special; it is called minimal. It is the set of minimal energy.

The next question is the following: does it happen that one obtains the whole circle? The answer to this question is illustrated by Fig. 7. Let us take a big ellipse as a billiard table and consider the orbit which is tangent to an inner confocal ellipse. It turns out that this orbit will be tangent to this ellipse further. The same is true for confocal hyperbolas. However, a hyperbola consists of two branches, but if an orbit is tangent to one of the branches of a hyperbola, then it will be further tangent to this hyperbola, and the branches tangent to the orbit will alternate, one after another.

This is a picture in the configuration space. And what will be in the phase space? The picture in the phase space is also well known, it looks very much like the picture for a pendulum (Fig. 8; in this figure the cylinder is unwound). Namely, there are two orbits of period 2 corresponding to the large and small diameters of the ellipse. An "eight" trajectory corresponds to orbits passing through the foci of the ellipse (if an orbit passes through a focus, then it will further pass through the foci, one after another). The trajectories situated outside this eight correspond to orbits tangent to ellipses. And the trajectories inside the eight correspond to the orbits tangent to hyperbolas.

Which of these orbits correspond to Birkhoff and Aubry–Maser orbits, and which do not correspond to them? In other words, which of these orbits can be obtained by Birkhoff's and Aubry–Maser's constructions and which cannot be obtained? The orbits with rational rotation numbers tangent to ellipses are obtained by Birkhoff's construction, and the rest of the orbits tangent to ellipses are obtained by Aubry–Maser's construction. And hyperbolas cannot be obtained by such constructions. Indeed, for the rotation number $1/2$ one has one minimal orbit and one minimax orbit.

It is interesting to understand what happens in passing to the inverse limit, i.e., when we pass from irrational to rational numbers. In the situation under consideration the answer is rather simple. We obtain an invariant circle, but it is not fully covered by Birkhoff orbits. It consists of Birkhoff orbits and asymptotical curves. This is a rather general phenomenon, with the exception that not always one obtains a full circle.

We have considered billiard tables of a rather special form. The following rather famous question has not yet got a definite answer: "What can be other billiard tables for which at least a neighborhood of the upper and lower base of the cylinder is fibered into invariant curves?" In other words, when the system is completely integrable? It is assumed that this can happen only for an ellipse.

Much more fundamental is the following question: when at least some curves are conserved? It is rather remarkable that necessary and sufficient conditions of the existence of at least one invariant curve are rather simple. Of course, we mean an invariant curve passing around the cylinder. Only such a curve can arise as the limit of Birkhoff orbits. It is not difficult to prove that if one has an invariant curve on which the order is preserved and the rotation number equals α, then such a curve is unique if α is irrational, and it is the limit of Birkhoff orbits.

If we are interested in the question: what arises as the limit of Birkhoff orbits, whether it is a curve or a Cantor set, then it is natural to ask when it is a curve. Let us assume that the curve which bounds the table is sufficiently smooth, for example, of the class C^∞ (it suffices to require that the curve be of the class C^6). In this case, a theorem proved by Vladimir Fedorovich

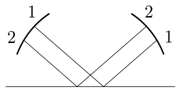


Figure 9. Reordering of points

Lazutkin (1941–2000) [10] states that an invariant circumference exists when the curvature of the boundary vanishes nowhere. Actually in this situation there are infinitely many invariant circumferences.

Lazutkin's proof is an adaptation for this case of the celebrated Kolmogorov theorem on perturbations of Hamiltonian systems. Formally the Kolmogorov theorem does not cover this situation, because here we deal with behavior of a degenerating system. One must appropriately change the coordinates to apply anything. The nonzero curvature is needed just to make this change of coordinates possible.

If the curvature vanishes, then there are no invariant curves. This much more simple fact has been proved by John Maser. In fact, one can prove a stronger statement. Namely, if the boundary contains a flat point, then no Aubry–Maser set can pass through this point. And on an invariant circle one must have not only points corresponding to Birkhoff orbits but also points corresponding to Aubry–Maser orbits. (See, for example, Section 13.5 in [8].)

This argument is rather simple. A reflection with respect to a line changes the order of points (Fig. 9). If first is the point 1 and second is the point 2, then after a reflection first is the point 2 and second is the point 1. In Fig. 9 the lines are parallel, but the same effect holds if the lines are different. Thus, after a reflection with respect to a line the order of points must change. Infinitesimally the same happens during reflection with respect to a curve at a point with the zero curvature.

Here some interesting geometric effects arise. Consider the inverse problem: how to construct a billiard table for which caustics exist? To this end, one can use a construction which is well known for the case of an ellipse. One can take an ellipse and throw on it a lace whose length is greater than the length of the ellipse. Then one should stretch this lace and draw a curve (Fig. 10). As a result one obtains a confocal ellipse. For the larger ellipse the smaller one will be a caustic.

The same construction works for an arbitrary curve. If one takes an arbitrary curve and a lace longer than the curve, and then stretches the lace and draws a new curve, then for the new curve the initial curve will be a caustic.

Sometimes a nonsmooth inner curve yields a smooth billiard table. For example, if for the inner curve one chooses an astroida, then as a result one

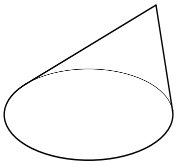


Figure 10. Construction of a confocal ellipse

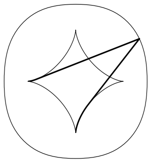


Figure 11. A billiard table for the astroid

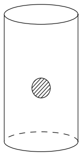


Figure 12. The cut out circle

obtains a smooth table for which the astroida is a caustic (Fig. 11).

It is rather clear why elliptic billiards can be considered as a testing area. Firstly, they give examples of twisting maps. Billiards give some geometric intuition which can be developed and then used for arbitrary twisting maps, and twisting maps cover many interesting cases besides billiards. And, secondly, billiards give an example of a standard difficulty in dynamics. How can one take into account the Lagrangian structure? The picture on a cylinder, where one has the coordinate and the impulse, is a Hamiltonian picture, it is a picture in the phase space. And a billiard is a Lagrangian structure. The Lagrangian structure is not invariant, it is related with division into coordinates and impulses.

For instance, let us consider the following question. Can a billiard have an open set of periodic orbits? For a Hamiltonian twisting map an example is constructed very easily. One should cut out a small circle from the cylinder, make a rational rotation, and then glue the circle back (Fig. 12). There are no Hamiltonian obstructions. However, there are some reasons to expect that for billiards nothing similar is possible. And this is not an idle question, because, for example, estimates of remainder terms in the Weyl asymptotics for eigenfunctions of the Laplace operator depend on the assumption that in a billiard the set of periodic orbits has the zero measure. This is proved only for orbits of period 3.

We will finish the discussion of elliptic effects by describing a natural "bridge" to the parabolic case.

Consider a convex polygon P possessing the property that the group generated by reflections with respect to its sides generates a "covering" of the plane. In other words, the images of P under the action of elements of this group cover the plane and if two such images intersect each other, then they coincide. There are just a few such polygons: rectangles, right triangles, right-angled triangles with the angles 45° and 30°. The group generated by reflections with respect to sides of such a polygon contains a normal subgroup of finite index consisting of parallel translations. In the four cases the indices equal respectively $4, 6, 8$, and 12. Taking representatives of congruence classes of the subgroup of translations and acting by them on the initial polygon, we obtain a fundamental domain for the subgroup of translations, and this domain is a torus. Let us make a partial unfolding of the billiard flow by means of the chosen fundamental domain, i.e., instead of reflecting the trajectory let us reflect the polygon. Some pairs of parallel sides will be then identified by translations, and the billiard flow will be thus represented as the free motion of a particle on a (flat) torus: each tangent vector moves in its direction with the velocity equal to one. This is a completely integrable system: the initial angle is a first integral, the phase space is fibered into invariant tori, and on each torus a flow of isometries acts. Each such flow is a standard elliptic system.

3 Parabolic behavior: billiards in polygons

A simplest parabolic billiard table is a right-angled triangle with the angle $\pi/8$. When a trajectory meets the boundary, let us reflect the triangle instead of reflecting the trajectory. In this concrete case everything stops rather soon. If one takes 16 copies of the triangle and makes from them an octagon (Fig. 13), then the motion turns into the parallel flow on this octagon, the opposite sides being identified. The obtained object is a Riemann surface (in this case of genus 2) with a quadratic differential. When the vertices of the octagon are glued together, one obtains the angle 6π. To resolve this singularity one should take the cubic root. Then one can obtain a Riemann surface with a field of directions. The field of directions has one singular point which is a saddle with 6 separatrices. This field of directions can be realized by means of a quadratic differential.

This flow has a first integral, it is the angle (in the octagon the direction of movement is preserved). This first integral has singularities.

Exercise 3. Analyze, in a similar manner, billiards in a right hexagon and a "gnomon."

Such a construction holds in all cases when the angles of the triangle are

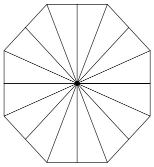


Figure 13. The simplest elliptic billiard

commensurable with π. In this case one can construct a Riemann surface with a quadratic differential from a finite number of copies of the billiard table. This flow has a first integral, and what is obtained can be studied using rather powerful methods from Teichmuller theory. As a result one achieves a rather good understanding of what is going on. Here one meets typical parabolic effects. For example, on all invariant manifolds (in this case, for the fixed value of the angle), the system is topologically transitive; and also on almost all invariant manifolds the system is strictly ergodic, i.e., the invariant measure is unique. And in the exceptional cases, when the invariant measure is nonunique, the number of invariant measures does not exceed the genus of the surface. These are typical parabolic effects; the invariant measure is not always unique, but usually the number of nontrivial invariant measures is finite.

Thus, a billiard system in a polygon with the angles commensurable with π, namely $\pi p_i/q_i$, where p_i and q_i are coprime integers, generates a one-parameter family of flows on some surface whose genus is determined by the geometry of the polygon and arithmetic properties of the numbers p_i/q_i. One should not fall into the illusion that the structure of these flows is rather simple. For example, the genus of the surface (and hence, in typical cases, the number of fixed points of the flow) is proportional to the least common multiple of the denominators q_i.

Nevertheless, these one-parameter families possess more complicated versions of some properties of the family of linear flows on a torus (which, as it was explained above, correspond to billiards in rectangles and some simple triangles). As I have already mentioned, for almost all values of the first integral the flow has the unique invariant measure (point supported measures corresponding to equilibrium states are not taken into account). But, in contrast to the case of flows on a torus, the set of exceptional values of the parameter is noncountable. Recall that on a torus one has a simple dichotomy between the slope angles whose tangents are rational, when all the orbits are closed, and the angles whose tangents are irrational, when the invariant measure is unique and hence any orbit is uniformly distributed with respect to the Lebesgue measure.

In the case of families of flows generated by quadratic differentials on surfaces of genus greater than one (in particular, for the families of flows arising from billiards in rational polygons), the situation is more complex. Still there is a countable number of “rational” values of the parameter for which all the trajectories are closed. Note that, in contrast to the case of the torus, there are several different homotopic types of closed orbits. The number of such types can be estimated using the simple argument that the orbits from different families do not intersect each other and hence their number does not exceed the genus of the surface. Besides that, there exists a set of values of the parameter which has zero measure but positive Hausdorff dimension, for which the flow is quasiminimal (i.e., any semitrajectory which does not tend to a fixed point is dense), but there exist more than one non-atomic invariant measure.

A more deep consideration shows that this difference is a reflection of the dichotomy between *Diophantine* irrational numbers or vectors, for which the speed of rational approximation is not very high, and *Liouville* numbers or vectors, for which “anomalous good approximation” arises. In the case of linear flows on a torus, for Diophantine slope angles, time averages for sufficiently smooth functions converge very rapidly. Moreover, Diophantine flows are rather stable: time changes and even small nonlinear perturbations preserving the rotation number of such flows can be “straightened.” For Liouville slope angles, time averages can behave rather irregularly: from time to time they can be very close to the integral or rather far from it, so that the speed of convergence over some sequences of moments of time is very high, and over other sequences it is rather low. Respectively, even smooth changes of time can essentially change large time dynamics: for example, eigenfunctions, even measurable, can disappear, and the flow becomes weakly mixing.

For flows arising from quadratic differentials and for billiards in rational polygons, the values of parameters for which there is more than one invariant measure, correspond to the slope angles with irrational Liouville tangents. Therefore it is not surprising that similar but more bright phenomena arise: instead of slow convergence of averages to the integral by the Lebesgue measure, there is no convergence at all. On the other hand, for a set of values of the parameter of full measure corresponding to slope angles with Diophantine tangents, one has similar though much more complex stability phenomena. They were discovered and studied during the last five years by the young mathematician Giovanni Forni; his papers constitute one of the most bright modern achievements in the theory of dynamical systems. The central observation due to Forni is that although the invariant measure is unique, there are also invariant distributions (generalized functions), i.e., invariant continuous linear functionals defined on smaller spaces of functions than all continuous functions. For functions of a given class of smoothness, the space of invariant distributions is finite-dimensional, but the dimension tends to infinity with the growth of smoothness class. Combination of strict ergodicity (uniqueness of an invariant measure) with the existence of an infinite set of independent invariant *distributions* is rather typical for dynamical systems with parabolic behavior. The simplest example, in which full investigation can be carried out with the help of elementary Fourier analysis, is the affine mapping of the two-dimensional torus

$$(x, y) \mapsto (x + \alpha, x + y) \pmod 1,$$

where α is an irrational number. A more interesting example which is studied by means of the theory of infinite-dimensional unitary representations of the group $\mathrm{SL}(2, \mathbb{R})$, is the oricyclic flow on a surface of constant negative curvature.

Returning to flows on surfaces, note that according to Forni's results, invariant distributions determine the speed of convergence of time averages. Roughly speaking, one has some typical power speed; if the first group of invariant distributions vanishes, then this speed increases, and this happens several times, until one obtains the maximal possible speed of decreasing of averages which is inverse proportional to time. Vanishing of a sufficient number of invariant distributions also guarantees that the flow obtained by a change of time can be straightened.

Even in the case of polygons with rational angles the description of the billiard is not completely reduced to considering separately the flows on invariant manifolds. For example, let us consider the question on the growth of the number of periodic trajectories of length no greater than T as a function of T. Of course, periodic orbits arise in families which consist of "parallel" orbits of equal length. Hence one should count the number $P(T)$ of such families. In the case of a billiard in a rectangle (which, as we mentioned several times, is reduced to the geodesic flow, i.e., the free motion of a particle on a flat torus), this problem amounts, after a suitable renormalization, to counting the number of points with integer coordinates in the circle of radius T with center at the origin. Therefore,

$$\lim_{T\to\infty} \frac{P(T)}{\pi T^2} = 1.$$

For general rational billiards, the growth of $P(T)$ is also quadratic, i.e.,

$$0 < \liminf_{T\to\infty} \frac{P(T)}{T^2} \leqslant \limsup_{T\to\infty} \frac{P(T)}{T^2} < \infty.$$

Besides that, it is known that periodic orbits are dense in the phase space. The question on existence of the limit $\frac{P(T)}{T^2}$ as $T \to \infty$ for an arbitrary rational rectangle still remains open. The positive answer is obtained, on the one hand, for some special polygons which amount to quadratic differentials on surfaces with a large number of symmetries (Veech surfaces), and on the other hand, for generic quadratic differentials. It is rather likely that there exist polygons with pathological behavior of the function $P(T)$. Note that our first nontrivial example of a billiard in a right-angled triangle with the angle $\pi/8$ and the hypotenuse 1 yields a Veech surface, and for it we can find $\lim_{T\to\infty}\frac{P(T)}{T^2}$.

For billiards in polygons in which not all angles are commensurable with π, surprisingly little is known. Such billiards are good examples of parabolic systems of sufficiently general kind. One has to say that the methods of analysis available now are insufficient for serious investigation of such systems. Indeed, successful study of parabolic systems is related with two special situations:

(1) flows on surfaces discussed above, where the dimension of the phase space is very small (in complement to the dimension corresponding to orbits one has only one transversal direction), and

(2) flows on homogeneous spaces, where one has large local symmetry.

Two main open questions concerning arbitrary billiards, are the description of global complexity of behavior of trajectories and the asymptotic behavior of typical trajectories with respect to the Lebesgue measure. Let us begin with the second question. Here a lot is known, and at the same time very little. If one fixes the type of the billiard table (for instance, convex polygons with a given number of edges), then the angles are the natural parameters in the space of such billiards. Billiards with angles commensurable with π, for which, as it was explained above, a lot is known, form a dense set in this space. Starting with ergodicity of rational billiards on most invariant submanifolds and taking into account the fact that for large denominators each such manifold almost uniformly covers the phase space, one can show by rather standard categorical arguments that for a dense G_δ in the space of parameters the billiard is ergodic in the whole phase space. However, this topologically ample set of billiards is rather thin from the metric point of view: not only its Lebesgue measure but also its Hausdorff dimension in the space of parameters equals zero. This set reminds one of the set of numbers admitting a rational approximation with extremely high speed, like a triple exponent. It is assumed that for typical Diophantine values of the vector of angles the billiard is ergodic. Up to now no serious approaches to this problem are known. Also more subtle statistical properties, such as mixing, are not known for any irrational billiards including the Liouville situation described above for which ergodicity is proved. The structure of singular invariant measures for irrational billiards is also not known.

Of course, a particular case of this last question is description of periodic trajectories, since each such trajectory generates a singular ergodic invariant measure. On the one hand, it is unknown whether for an arbitrary polygon

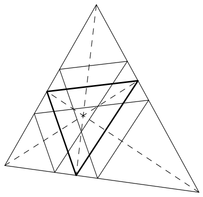


Figure 14. Orbits of periods 3 and 6 in an acute triangle

there exists at least one periodic billiard trajectory. As it was mentioned above, for rational polygons there are infinitely many such trajectories and they are dense in the phase space. However, one has not succeeded in passing to a limit for irrational polygons, even of a special kind. The problem here is that as the denominator increases, invariant manifolds become surfaces of a very high genus and periodic orbits have very complicated homotopical type, and hence are very long. However, there are some special situations when periodic orbits with simple combinatorics arise which are conserved during small perturbations of angles. A classical example is the orbit of period 3 formed by the bases of altitudes in an arbitrary acute-angled triangle. Of course, this orbit admits a variational description. But, in contrast to Birkhoff orbits in convex billiards, the triangle formed by the bases of altitudes has the minimal perimeter among all inscribed triangles. And the maximal and minimax triangles degenerate into the double maximal altitude. The orbit of period 3 thus constructed is surrounded by a family of parallel orbits of period 6 (see Fig. 14). Note that these are the only periodic orbits whose existence is known for all acute-angled triangles. The question on the density or at least the existence of an infinite number of parallel families of periodic orbits remains open.

For an arbitrary right-angled triangle the existence of periodic orbits has been proved just a few years ago. Unfortunately, these orbits are somewhat disappointing. These are trajectories which are reflected orthogonally from one of the sides and after a finite number of reflections return to the same side also in the orthogonal direction. Evidently, such an orbit is reflected then back and repeats its way in the backward direction. This is an example of orbits with stable combinatorics. It turns out that for almost any initial position the orbit orthogonal to a side returns back to this side in the orthogonal direction and therefore is periodic. This is a rather easy consequence of conservation of measure and of the fact that possible directions of an orbit form the unique trajectory of the infinite dihedral group generated by reflections with respect to the two nonperpendicular sides of the triangle. This argument can be generalized to some polygons "close" to rational ones, i.e., those for which the values of angles modulo π lie in a one-dimensional space over rational numbers. For an arbitrary obtuse-angled triangle this argument cannot be applied, and the existence of even one periodic orbit is unknown.

The existence of periodic orbits is closely related to the question on the global complexity of the behavior of trajectories. The growth of the number of distinguishable trajectories with the time can be estimated in different ways. The most natural way is related to coding. To each trajectory one assigns a sequence of symbols in correspondence with the reflections with respect to sides of the polygon, so that each side is denoted by its own symbol. Of course, in this way one naturally encodes billiard maps, i.e., returning maps of the billiard flows to the boundary. In order to obtain full information on the flow, one should also indicate the time between two consecutive reflections. The growth of complexity for the billiard mapping (respectively, flow) is given by the function $S(N)$ (respectively, $\mathcal{S}(T)$) equal to the number of different codes of length n (respectively, to the number of different codes arising for segments of trajectories of length T). Obviously, each family of parallel periodic orbits generates an infinite periodic code, and it is almost also obvious that, vice versa, each infinite periodic code corresponds to a family of parallel periodic orbits. These orbits can be closed either after one period or after two periods (as orbits of period 6 parallel to the Faniano triangle in an acute-angled triangle).

In the case of polygons with rational relative to π angles, both functions admit a quadratic estimate:

$$0 < \liminf_{N\to\infty} \frac{S(N)}{N^2} \leqslant \limsup_{N\to\infty} \frac{S(N)}{N^2} < \infty$$

and

$$0 < \liminf_{T\to\infty} \frac{\mathcal{S}(T)}{T^2} \leqslant \limsup_{T\to\infty} \frac{\mathcal{S}(T)}{T^2} < \infty.$$

Note that in this case a positive part of all admissible codes is realized by periodic trajectories.

An alternative way of describing complexity is to compute the number of ways in which codes can change. Obviously, the code changes when the trajectory hits an angle. It is also obvious that there is only a finite number of segments of trajectories of bounded length which hit angles both in the positive and negative directions. By rather evident reasons, such singular trajectories are called *generalized diagonals* of a polygon. Define $D(N)$ (respectively, $\mathcal{D}(T)$) as the number of generalized diagonals with $\leqslant N$ edges (respectively, as the number of generalized diagonals of length $\leqslant T$). As above, these quantities admit a quadratic estimate for rational polygons.

It is natural to assume that for arbitrary polygons the growth of trajectories should not be much more rapid than for rational ones, since the local geometric structure of the billiard flow is the same in both cases. However, the only known fact in this direction consists in much more weak subexponential estimates:

$$\lim_{N\to\infty}\frac{\log S(N)}{N}=\lim_{N\to\infty}\frac{\log D(N)}{N}=\lim_{T\to\infty}\frac{\log \mathcal{S}(T)}{T}=\lim_{T\to\infty}\frac{\log \mathcal{D}(T)}{T}=0.$$

4 Hyperbolic behavior: billiards of Sinai, Bunimovich, Wojtkowski and other authors

As we have already mentioned, hyperbolic behavior is rather common and allows one to establish the basic elements of stochastic or “chaotic” behavior. The domination of hyperbolic behavior is natural by analogy with linear systems. Indeed, a randomly chosen matrix most likely has no eigenvalues whose absolute value equals one. Even if one *a priori* restricts oneself to matrices with the determinant equal to one, this is still true for matrices of size 3×3 or more. Although this analogy cannot be literally transferred to nonlinear systems, it at least shows the importance of the hyperbolic paradigm.

Historically the first examples of hyperbolic behavior of billiards were found by Ya. G. Sinai [13]. The simplest examples of a Sinai type billiard are, firstly, a square with a circle cut out, and, secondly, a convex polygon whose sides are replaced by arcs convex inwards (see Fig. 1). From the point of view of rigorous mathematical analysis, the second example turns out to be somewhat more easy than the first one. Hyperbolic behavior in Sinai billiards is related to the phenomenon of scattering of light well known from geometric optics: a parallel or divergent flow of light becomes more divergent after reflection with respect to a convex mirror. Not too complicated computations show that if the reflection is sufficiently regular, then the angle measure of a sheaf grows exponentially. This yields hyperbolicity of the linearized system.

In the analysis of scattering billiards two technical difficulties arise.

Firstly, one must achieve sufficient regularity of reflections with respect to convex inwards parts of the boundary. It is clear why to this extent the second example is better than the first one: in the second example the time between two consecutive reflections is bounded. And in the first example there are periodic trajectories parallel to the sides of the square which do not meet the obstacle at all. Of course, such trajectories form a set of zero measure, but trajectories which form very little angles with them meet the obstacle only in a very large time. This phenomenon is called infinite horizon; respectively, boundedness of the time between two reflections corresponds to finite horizon.

Infinite horizon implies non-uniformity of hyperbolic estimates over the phase space. Although this yields essential technical complications in proofs of ergodicity, mixing and other stochastic properties, this also confirms the role of billiards as an important testing area for various methods and tools of analysis of dynamics. Indeed, non-uniform hyperbolicity is much more common than uniform one. For example, global uniform hyperbolic behavior for classical conservative systems imposes restrictions on the topology of the phase space. But non-uniform hyperbolicity is compatible with any topology. This fact, although predicted rather long ago, has been established in full generality only recently by D. Dolgopyat and Ya. Pesin [6].

The second difficulty in the analysis of dispersing billiards is the presence of singularities (discontinuities and unboundedness of derivatives) in a system. Here is the difference between these systems and billiards in smooth convex domains, considered above, where the billiard mapping is smooth. Singularities arise at tangent points of trajectories with convex inwards parts of the boundary. Of course, they also arise when a trajectory hits an angle. Singularities of the second type arise also in parabolic billiards, and in the case of scattering billiards they yield not significant complications. Such singularities yield discontinuities of the first kind for functions representing dynamics: a surface of discontinuity arises, and functions are smooth on both sides of the surface. Thus, the differential along a billiard trajectory which does not hit a discontinuity point behaves rather regularly. For trajectories tangent to the boundary from inside, the derivatives near these trajectories are unbounded, so that the discontinuities are more serious. Note that elastic collisions and more complex effects of this kind naturally arise in many important problems of classical mechanics, for example, in the problem of n bodies. The influence of such phenomena on the large time behavior of trajectories is one of the central problems in mechanics. Here also billiards, and especially their multidimensional analogs, play the role of an important testing area.

Scattering billiards are rather essential for the mathematical background of models of statistical physics. This is an important and interesting subject, which however we will not touch here. From the view point of geometry, scattering billiards possess some defects, for example, inavoidable singularities on the boundary. However, if one considers billiards not on plain domains but on domains on a flat torus, this defect can be avoided. For example, the billiard on a torus with a circle removed is a classical example of a Sinai billiard. Nevertheless, it is interesting to know how hyperbolic behavior can arise in other ways than scattering by convex inwards parts of the boundary. The first answer to this question is given by a rather celebrated example of a "stadium," i.e., two semicircumferences connected by segments of common tangent lines (see

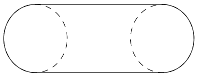


Figure 15

Fig. 15). This is an example of the so-called Bunimovich billiards [2], where hyperbolic behavior arises as a result of consecutive focusing of sheaves of orbits. From the point of view of configuration space this picture dramatically differs from the case of scattering billiards; however, in the phase space, where both coordinates and velocities are taken into account, uniform exponential growth arises.

Bunimovich billiards were discovered in an interesting way. In the beginning of the 70's L. A. Bunimovich, who was then a graduate student of Sinai, was working on extending the class of billiards with exponential running away and stochastic behavior of orbits. He discovered that if one adds little "pockets" to a scattering billiard, then the billiard on the thus obtained table in which convex parts are followed by concave ones, possesses exponential running away of trajectories. Actually Bunimovich discovered a new important mechanism of hyperbolicity. However, he himself firstly considered his work as just a little generalization of the results on scattering billiards. During Bunimovich's talk on a seminar at MIAS[1] directed by D. V. Anosov and the author, the natural question on the mechanism of hyperbolicity arose, and in particular on whether the presence of any scattering components is necessary. I draw the speaker's attention to the fact that his arguments seemingly did not imply this necessity, and proposed a stadium as a model for verification of this conjecture. The rest of Bunimovich's geometric conditions were satisfied, at least if full circles did not intersect each other (see Fig. 15). After thinking a little Bunimovich said that his arguments should hold in this case, and in the next version of his paper he stated conditions which did not require the presence of scattering components. Moreover, it turned out that the initial geometric conditions can be weakened, so that, for example, in the case of a stadium, the distance between the circles can be arbitrarily small.

Among Bunimovich billiards there are a lot of other interesting and rather simple forms, but they all have the common property that the boundary can contain, except scattering parts, only segments of straight lines and arcs of circumferences. The natural question, how essential is this condition, has been studied by specialists for about ten years. A technical difficulty is the following. Hyperbolicity is established with the help of a system of cones in tangent spaces

[1] Mathematical Institute of the Academy of Sciences (Moscow).

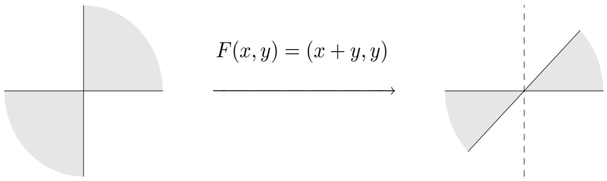


Figure 16. The action of the parabolic transformation on the cone

to points of the phase space, which are transformed to themselves under the action of dynamics. For simplicity and geometric visualization, it is better to think about the billiard mapping rather than the flow. In this case the phase space is two-dimensional, and the cones in question are the interior parts of two opposite angles formed by a pair of lines intersecting at the origin. The system of cones which is invariant, both in scattering billiards and Bunimovich billiards is the same. Geometrically, these cones are defined as the sets of infinitesimal dispersing pieces of trajectories. For hyperbolicity it is necessary that the cone together with its boundary be mapped strictly inside the corresponding cone in the image. Of course, this holds in the case of scattering billiards already after one reflection. And in the case of flat and circumference mirrors, the cone goes into itself, but one of its sides is left invariant. This is a typically parabolic effect, since in this way unipotent matrices act. Let us take, for instance, the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. The cone in question is defined by the condition $x_1 x_2 > 0$, i.e., it is the union of the first and the third quadrant on the plane. Its image is the cone $|x_1| > |x_2|$, $x_1 x_2 > 0$ (see Fig. 16). After further iterations the image becomes more and more thin, but it is still “glued” to the horizontal axis. In order to get hyperbolicity, Bunimovich uses a geometric condition which yields strict invariance of cones after the reflection with respect to *different* circumference parts of the boundary (as in the case of a stadium). Since the trajectory which is reflected from the circumference part under a very small angle, continues to do this many times, it seemed that the explicit form of the iteration after reflection with respect to circumference parts (integrability of the billiard in a circle) played an essential role. This way I explained for myself rigid Bunimovich’s conditions.

However, it turned out that one can overcome this difficulty. Billiards with convex parts of the boundary can be hyperbolic by many reasons. As soon as the method based on the use of systems of invariant cones was invented, the problem of finding new classes of hyperbolic billiards became easy. Note that Bunimovich used another technique which is formally equivalent to the system of invariant cones, but is much less visual. The pioneers of the use of the method of invariant cones in dynamics were V. M. Alekseyev (1932–1980) and Yorgen Moser (1928–1999). An essential step was introducing this method in non-uniformly hyperbolic situation. The author used this method for constructing examples of smooth systems with stochastic behavior on various manifolds. However, the most essential progress here is due to Maczei Wojtkowski. And again billiards turned out to be an ideal testing area. Having understood the key role of systems of cones, Wojtkowski realized that the problem could be solved in the inverse order, namely, to find classes of billiard tables corresponding to a given system of cones. The preprint of his key paper on this subject [3] was called "Principles for the design of billiards with nonvanishing Lyapunov exponents." As a square or torus with a removed circle is a quintessence of the phenomenon discovered by Sinai, and the stadium symbolizes Bunimovich billiards, in the same way a typical example of Wojtkowski billiards is given by cardioida (see Fig. 1). The importance of Wojtkowski's result in the theory of billiards is in that he discovered classes of hyperbolic examples which are open in C^2 topology, and thus [11] this property does not depend on small variations of mirrors.

As I have already mentioned, constructing new classes of hyperbolic billiards became possible with the use of the method of invariant cones. As an example of flexibility of this method, let us mention the following result due to Victor Donnay [7]: any sufficiently small piece of a convex curve is a part of the boundary of a piecewise smooth convex hyperbolic billiard. Note also that the use of the method of invariant cones allowed one to obtain many remarkable examples of classical dynamical systems with non-uniform hyperbolic behavior.

Important unsolved problems are related with existence of hyperbolic billiards with smooth (at least twice differentiable) boundary. Note that the boundary of the stadium is differentiable, but the curvature (and hence the second derivative) is discontinuous. Even twice differentiable examples with nonconvex or non-simply-connected boundary are unknown.

What does hyperbolicity give? It allows one to show that in many cases a deterministic dynamical system behaves in many respects as a sequence of independent random quantities. In a sense, this statement is true literally: under some (often rather easily checked) conditions, in complement to (even non-uniform) hyperbolicity, the phase space of a system conserving finite volume can be divided into a finite number of sets $A_1, \dots, A_n$ of positive measure so that, firstly, each point of the phase space is encoded by the sequence of hits to these sets at positive and negative moments of time, and, secondly, these sets are completely independent with respect to the dynamics F, i.e.,

$$\operatorname{vol}\left(\bigcap_{k=0}^{n} F^k(A_{i_k})\right) = \prod_{i=0}^{n} \operatorname{vol} A_{i_k}.$$

Although these sets are of exotic nature, this property, which is naturally called the Bernoulli property, implies many important properties: convergence of time averages to the space average (ergodicity), decrease of correlation (mixing), asymptotic independence of the future from the past (the K-property, or the Kolmogorov property).

References

[1] G. D. Birkhoff. *Dynamical Systems* (New York: Amer. Math. Soc., 1927).

[2] L. A. Bunimovich. On the ergodic properties of nowhere dispersing billiards. *Comm. Math. Phys.* **65** (3) (1979), 295–312.

[3] W. Wojtkowski. Invariant families of cones and Lyapunov exponents. *Ergodic Theory Dynam. Systems*, **5** (1985), 145–161.

[4] G. A. Galperin and A. N. Zemlyakov. *Mathematical Billiards* (Moscow: Nauka, 1990) [in Russian].

[5] G. A. Galperin and N. I. Chernov. *Billiards and Chaos* (Moscow: Znanie, 1991) [in Russian].

[6] D. Dolgopyat and Ya. Pesin. Every compact manifold carries a completely hyperbolic diffeomorphism. To appear in *Ergodic Theory Dynam. Systems.*

[7] V. J. Donnay. Using integrability to produce chaos: Billiards with positive entropy. *Comm. Math. Phys.* **141** (2) (1991), 225–257.

[8] A. B. Katok and B. Hasselblatt. *Introduction to Modern Theory of Dynamical Systems* (Cambridge: Cambridge Univ. Press, 1995).

[9] I. P. Kornfeld, Ya. G. Sinai, and S. V. Fomin. *Ergodic Theory* (Moscow: Nauka, 1980) [in Russian].

[10] V. F. Lazutkin. Existing of caustics for the billiard problem in a convex domain. *Izv. Akad. Nauk SSSR, Ser. Mat.*, **37** (1) (1973), 186–216.

[11] H. Masur and S. Tabachnikov. Rational billiards and flat structures. To appear in *Handbook in Dynamical Systems 1A* (Amsterdam: Elsevier).

[12] *Hard Ball Systems and the Lorentz Gas*, ed. D. Szász (Berlin: Springer-Verlag, 2000).

[13] Ya. G. Sinai. Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards. *Usp. Mat. Nauk*, **25** (2) (1970), 141–192.

[14] S. Tabachnikov. Billiards. *Panor. Synth.*, **1** (1995).