BILLIARD TABLE AS A PLAYGROUND FOR A MATHEMATICIAN

A. B. KATOK

Lecture on March 10, 1999

The title of this lecture may be understood in two ways. Literally, in a somewhat light-hearted way: mathematicians play by launching billiard balls on tables of various forms and observe (and also try to predict) what happens. In a more serious sense, the expression "playground" should be understood as "proving ground": 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 a serious attention.

The literature concerning billiards is quite large, including research papers as well as monographs, textbooks, and popular literature. The booklets by G. A. Galperin and A. N. Zemlyakov [4] and by G. A. Galperin and N. I. Chernov [5] are written in an accessible manner, and touch upon a broad variety of questions. An introduction to problems related with billiards for a more advanced reader can be found in Chapter 6 of the book [9]. The next level is represented by a very well written book by S. Tabachnikov [14]. The book by the author and B. Hasselblatt [8] contains a fairly detailed modern exposition of the theory of convex billiards and twist maps. A serious but still accessible exposition of the theory of parabolic billiards in its modern state is contained in a survey paper by H. Masur and S. Tabachnikov which will appear in 2002 [11]. The volume [12] contains rich material on hyperbolic billiards and related questions. More specialized references will be given in the course of 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, a material point is moving freely, i.e., along straight lines with uniform speed, 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 in the long run.

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 can be dismissed as somewhat pathological; one can say that essentially there is a well-defined dynamical system.

The solution of the problem of motion of a ball depends on the domain. A nice feature of this problem and one of the reasons it attracted much interest 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 is related to the fact that if one attempts to somehow classify problems in the theory of dynamical systems, then they can be divided, in a somewhat tentative manner, into elliptic, parabolic, and hyperbolic ones (see Fig. 1). Thus, a billiard table is a proving ground

where one can test methods, conjectures, questions, arising in various fields of the theory of dynamical systems.

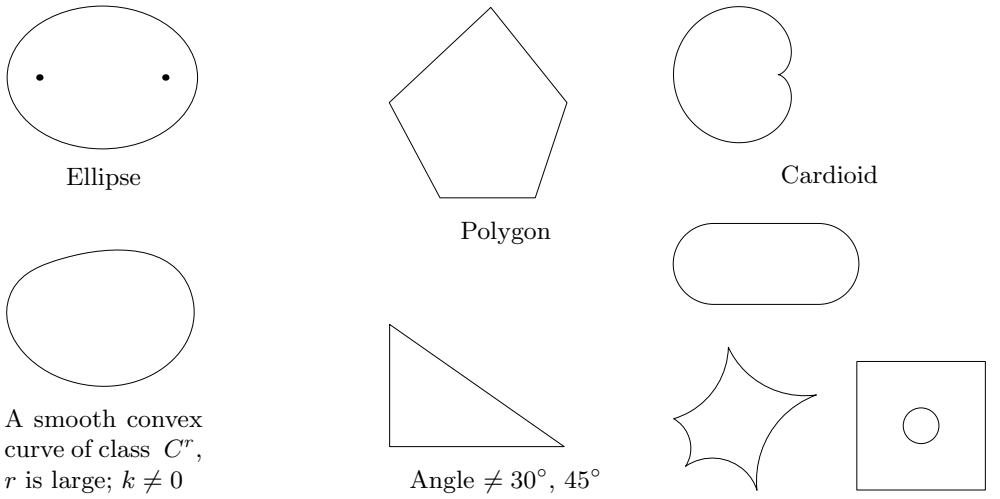

FIGURE 1. Elliptic, parabolic, and hyperbolic billiards

There is nothing new in using these words for expressing some sort of trichotomy. For example, the corresponding classification in the theory of partial differential equations is well known. But for dynamical systems, such a classification seemingly has not been been explored in a systematic fashion.

In the case of billiards, elliptic effects arise, for example, for an ellipse. While this semantic coincidence is not purely accidental, it does not extend to billiards inside a parabola or a hyperbola. In fact, a more general situation in which elliptic effects occur, is as follows: the curve is smooth (of a sufficiently high class of smoothness), convex, and its curvature vanishes nowhere. The study of the billiard problem inside such domains provides a good example for the demonstration of problems and results related to elliptic behavior of dynamical systems.

The parabolic situation arises for billiards inside polygonal domains. For simplicity one can even take a right triangle with angles other than 30° and 45°. A right 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 disk removed, the "stadium," and the cardioid.

The idea that there is at least a dichotomy in the behavior 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 XXth century is the book by Jürgen Moser "Stable and random motions 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 in dynamics, let me explain 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 appear:

(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 are recurrent (i.e., come back arbitrarily close to the initial position) and are stable, i.e., do not escape too far. 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, most orbits diverge from the origin and from each other with a polynomial speed. This is parabolic behavior.

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

Combinations of these three basic paradigms are also possible. For example, the situation when the spectrum contains a hyperbolic component and something else is called partially hyperbolic. In dynamics this situation plays a very important role.

It would be very naive to attempt to build a comprehensive description 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 such systems 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 close to their initial position (often called nontrivial recurrence). 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 nontrivial nature of returning.

This approach ignores a very essential class of problems in the theory of dynamical systems, for example, such things as the analysis of Morse–Smale systems, or effects related to bifurcations. These are situations when recurrence is simple, and the interesting phenomena are related, for example, to the way the phase space is divided into basins of attraction of several attracting periodic points or limit cycles. We are now concerned only with the part of dynamics which is related to recurrent behavior. We more or less ignore nonrecurrent behavior now.

To interpret phenomena of interest to us correctly, one must understand what a linearization of a dynamical system is. Let $f\colon M \to M$ be a 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 in itself only in the case of a fixed point. In the general case in dynamics one can consider the iterates Df_x^n. Introducing a Riemannian metric, one can speak about the asymptotic speed of growth of lengths of vectors. A Riemannian metric is naturally not unique, but on a compact manifold any two metrics differ by no more than by a bounded multiplicative factor, so the speed of growth of vectors is defined properly.

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 (sublinear growth).

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

The hyperbolic paradigm is the one which is best understood. It corresponds to the situation when the system splits and there is exponential growth in some directions, and exponential decay in others. When the time is reversed, these directions are interchanged.

Sometimes mixed situations occur. For example, one can take the direct product of two systems of different type. But as a meta-statement, 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 exercises dominating influence.

The following remark is also in order. When the dimension of the phase space is small, mixed behavior is impossible. For example, when the dimension of the phase space equals 2, partially hyperbolic behavior is impossible, because for hyperbolic behavior one needs at least one expanding and one contracting direction. For the same reason in small dimensions, elliptic or parabolic behavior occurs more often.

The purest example of elliptic behavior is the situation when one has a smooth isometry. In this case there is no orbit growth. The dynamics of smooth isometries on compact spaces can be understood rather easily. The phase space splits into invariant tori, and on each torus there is a translation (or a rotation, if one uses multiplicative notation). In particular, if the manifold itself is not a torus, then such a motion is not topologically transitive, i.e., it has no dense orbits.

Of course, this picture is a particular case of what is well known in Hamiltonian mechanics, namely, it corresponds to completely integrable Hamiltonian systems. Such systems provide a good example of the 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 usually parabolic, due to the twist in the direction transverse to the invariant tori. On the other hand, the space splits into invariant tori, and on each torus the analysis is carried out with elliptic methods.

This situation is typical. This is the reason why the elliptic paradigm is important. It is not common that the global behavior on the whole phase space is characterized by absence of growth. But frequently there are parts of 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 topological Markov chains and simple stochastic models. From the viewpoint of applications of dynamics, if a kind of hyperbolic behavior is established, then one can apply a 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 recurrence which exists more or less a priori. In linear systems hyperbolicity implies that orbits escape to infinity. But if there is no space to escape, if one necessarily has to return, then certain well understood types of complex behavior arise.

In contrast to elliptic and hyperbolic behavior, parabolic behavior is unstable, and is characterized by the absence of standard models. In the elliptic situation one has a universal model, namely, rotation on the torus (or some of its variations), and in the hyperbolic situation one has the Markov model which describes essential features of the asymptotic behavior. In the parabolic situation, it seems that there is no set of models to which everything is more or less reduced. Nevertheless, there are certain typical phenomena which occur in concrete classes of systems. Frequently the effects of moderate stretching can be replaced by that of cutting. For example, if one has a system which locally looks like an isometry but has discontinuities, then such a system is usually related with the parabolic paradigm.

A well known example is an interval exchange transformation. Its phase space is an interval. The transformation consists of cutting the interval into a fixed number of subintervals and permuting 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 easy to figure out that this system should be considered as a parabolic one: the number of segments grows under the iterations in a linear way. This linear growth results not from twisting but from cutting. But the resulting effect is approximately the same.

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

2. BILLIARDS IN SMOOTH CONVEX DOMAINS

George D. Birkhoff was the first to consider billiards systematically as models for problems of classical mechanics. Birkhoff considered billiards only in smooth convex domains; he did not think about billiards in polygons, or in nonconvex domains.

First of all, one can perform the reduction to the billiard map. 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 map. The phase space of the billiard map looks as follows. The outgoing vector v after reflection is characterized by the cyclic coordinate $\varphi \in S^1$ 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).

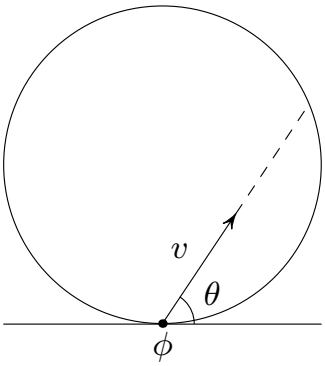

FIGURE 2. Vector coordinates after reflection

Thus the phase space of the billiard map is a cylinder. After the reflection we obtain a new point φ_1 and a new outgoing vector, which corresponds to the angle θ_1. The map $T(\varphi_0, \theta_0) = (\varphi_1, \theta_1)$ is what is called the billiard map. 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–line segments).

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

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

The billiard map possesses two important qualitative properties.

1) Preservation of area. The two-form

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

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

2) The twist property. Let us fix the coordinate φ_0 and change the coordinate θ. Then the coordinate φ of the image will change monotonically until it rounds the circle and comes back (Fig. 3). The image of a vertical line is twisted.

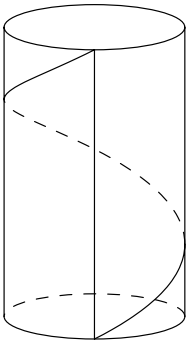

FIGURE 3. Twisting

These two properties allow us to find elements of elliptic behavior. The problems arising in connection with elliptic behavior are divided into two parts:

1) caustics,

2) Birkhoff orbits and Aubry–Mather sets.

Let us begin with the second topic. We want to find periodic orbits of the billiard system. There are different kinds of such orbits which differ not only by the period, but also by their combinatorics. For example, the two orbits of period 5 in Fig. 4 have different combinatorics. In the first case there is one rotation around the table, and in the second case there are two. These orbits are regular: the order of points on the orbit is preserved; it is the same as in a rotation. It is these (regular) orbits that are called Birkhoff orbits. This name is related to the fact that Birkhoff has proved a remarkable and relatively simple theorem about existence of regular orbits. This theorem was probably the starting point for applications of variational methods in dynamics.

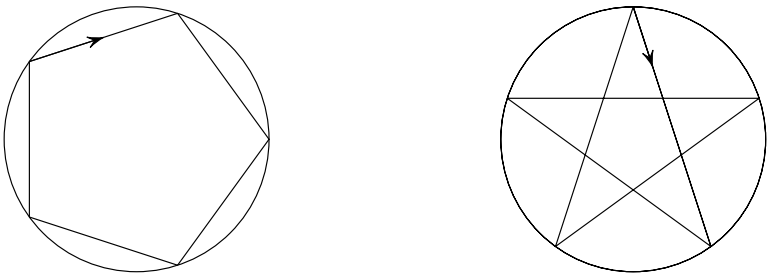

FIGURE 4. Orbits with equal periods and different combinatorics

Theorem (Birkhoff)**.** *For any two relatively prime 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 *states*, or more

precisely the states of type (p, q). All states form a finite-dimensional space. The length of a polygon is a naturally defined functional on the space of states. If we allow some vertices of a polygon to coincide, then the space of states becomes compact. Hence the length functional has a maximum.

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 not difficult to prove that the maximum value of the length cannot be achieved on the boundary, i.e., that the vertices of a polygon cannot coincide.

Thus, the longest polygon is a required periodic orbit. But this is still only the easy part of the theorem. One has to find another periodic orbit. This can be done in the following way. By cyclically renumbering of the vertices of the orbit we have found we obtain q different states for which the length functional achieves its maximum. 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 as little height as possible. In this case we need to pass through a saddle (Fig. 5), because if at the lowest point we were not in a saddle, then we could change the trajectory a little and decrease the loss of height. A saddle is also a critical point, i.e., the required second periodic orbit.

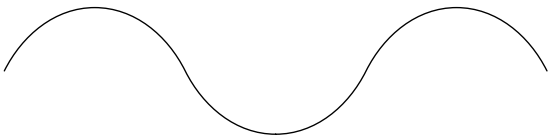

FIGURE 5. A saddle

If we do not lose height at all, then there is a whole family of periodic orbits. □

This proof shows concisely how one can replace one difficulty by another. The difficult (or at least subtle) point of this argument is in how to stay away from the boundary. This can be easily achieved if we just do not consider the boundary, and consider as states all inscribed q-gons going around the boundary p times. Evidently, the length function 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 reached on a properly ordered orbit. If, for instance, we consider globally maximal orbits which make two rotations during full round, then they do it in a correct order. Alternatively 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 what is called Poincaré's last geometric theorem: "If the bases of a cylinder rotate in different directions with area being preserved, 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 twist condition.

Birkhoff was extremely proud that he proved Poincaré's last geometric theorem. But he missed a very remarkable conclusion from his own elementary proof. Namely, let us look at what happens when one passes to the limit $p_n/q_n \to \alpha$, where α is an irrational number. Usually in dynamics such tricks do not work, because the asymptotic behavior is unstable with respect to initial conditions, and one cannot pass to the limit. But here, just because we are dealing with an 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 have to concentrate in certain places. It is not difficult to prove that these points always lie on a Lipschitz graph (i.e., on the graph of a function satisfying a 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. One can express this somewhat differently: 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 (for a subsequence), which is not surprising. But then any 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 its dynamics. It 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 by 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 the combinatorics will be the same, as needed, because on any finite segment the combinatorics stabilizes and is the same as under rotation by the angle α.

Thus, a closed invariant set on a circle arises (since topologically the limit graph of a Lipschitz function is a circle). This set has dynamics which preserves the order and exactly reproduces the order corresponding to the rotation by the angle α. From the time of Poincaré it has been known how this can happen: either the invariant set is the whole circle, the orbits are dense and the transformation 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 counterexamples (see, for example, Chapters 11 and 12 in [8]). A Denjoy counterexample looks as follows. Let us take a point on the circle and blow it up into an interval. Then its image and pre-image should be also blown up into intervals (Fig. 6), etc. To have convergence, one needs these intervals to become smaller and smaller. This is fairly easy to achieve topologically. As a result, one gets a transformation of the circle which contains an invariant Cantor set and which is semi-conjugate to a rotation (there exists a continuous mapping which makes it a rotation, but the intervals which were blown up shrink into points).

For transformations of a circle such a behavior is exotic, because by a theorem of Denjoy this is impossible in the class C^2 although it is possible in the class C^1. But for twist maps this is rather normal behavior (since, if it does not happen, we would get a whole circle). Thus, an interesting alternative arises. When one has accumulation of Birkhoff orbits on an invariant set (this is exactly the Aubry–Mather 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. For any rotation number one or the other of these two cases holds. 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 preclude the existence of other Cantor sets which have the same rotation number and on which the order

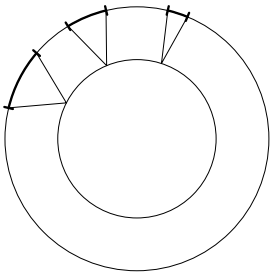

FIGURE 6. The Denjoy counterexample

is preserved. However, 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 said to be 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 an orbit segment which is tangent to an inner confocal ellipse. It turns out that this orbit remains tangent to this ellipse. The same is true for confocal hyperbolas. However, a hyperbola consists of two branches, but if an orbit segment is tangent to one of the branches of a hyperbola, then the orbit remains tangent to this hyperbola, and the segments tangent to the two branches will alternate.

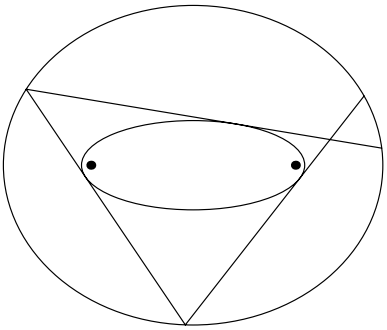

FIGURE 7. Confocal ellipses

This is a picture in the configuration space. And how will it look 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 unfolded). Namely, there are two orbits of period 2 corresponding to the large and small diameters of the ellipse. The " figure eight" corresponds to orbits passing through the foci of the ellipse (if an orbit segment passes through a focus, then the consecutive segments will alternately pass through both foci). The trajectories outside of the figure eight correspond to orbits tangent to ellipses. And the trajectories inside the figure eight correspond to the orbits tangent to hyperbolas.

Which orbits in this picture correspond to Birkhoff and Aubry–Mather orbits, and which do not? In other words, which of these orbits can be obtained by the Birkhoff's and Aubry–Mather constructions and what cannot be obtained? The orbits with rational rotation numbers tangent to ellipses are obtained by the Birkhoff construction, and the rest of the orbits tangent to ellipses

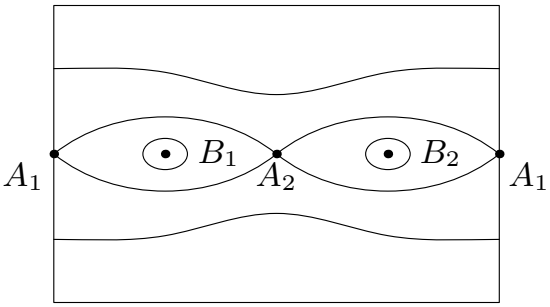

FIGURE 8. Trajectories in the phase space

are obtained by the Aubry–Mather construction. And hyperbolas cannot be obtained by such constructions. Indeed, for the rotation number $1/2$ one has only one minimal and one minimax orbit.

It is interesting to understand what happens in passing to the limit in the opposite fashion, 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 asymptotic curves. This is a rather general phenomenon, with the exception that one does not always obtain a full circle.

We have considered billiard tables of a very special shape. The following question is well known and is still open: "Describe all billiard tables for which at least a neighborhood of the upper and lower base of the cylinder is filled by invariant curves." In other words, when is the billiard system completely integrable? It is conjectured that this can happen only for ellipses.

Much more basic is the following question: when are at least some curves invariant? It is quite remarkable that necessary and sufficient conditions for the existence of at least one invariant curve are rather simple. Of course, we mean an invariant curve running 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, is it a curve or a Cantor set, then it is natural to ask when this limit is a curve. Let us assume that the curve which bounds the table is sufficiently smooth, for example, of class C^∞ (it suffices to require that the curve be of class C^6). In this case, a theorem proved by Vladimir Fedorovich Lazutkin (1941–2000) [10] states that an invariant curve exists when the curvature of the boundary does not vanish anywhere. Actually in this situation there are infinitely many invariant curves.

Lazutkin's proof is an adaptation for this case of the celebrated Kolmogorov theorem about perturbations of completely integrable Hamiltonian systems. Formally the Kolmogorov theorem does not cover this situation, because here we deal with behavior of a system with degeneracies. One must appropriately change the coordinates to apply the Kolmogorov method. 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 Mather. In fact, one can prove a stronger statement. Namely, if the boundary contains a flat point, then no Aubry–Mather set can pass through this point. (See, for example, §13.5 in [8].)

This argument is quite simple. A reflection with respect to a line changes the order of points (Fig. 9). If the point 1 is "ahead" of the point 2, then after the reflection the point 2 is ahead of

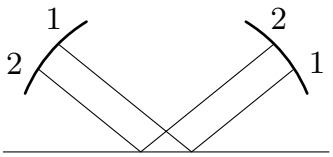

FIGURE 9. Reordering of points

the point 1. In Fig. 9 the lines are parallel, but the same effect takes place if they are not. 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 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 the ellipse. One can take an ellipse and throw around it a string whose length is greater than the length of the ellipse. Then one should stretch this string 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.

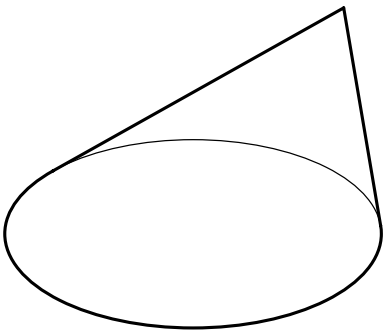

FIGURE 10. Construction of a confocal ellipse

The same construction works for an arbitrary curve. If one takes an arbitrary curve and a string longer than the curve, and then stretches the string 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 one chooses an astroid as the inner curve, then as a result one obtains a smooth table for which the astroid is a caustic (Fig. 11).

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

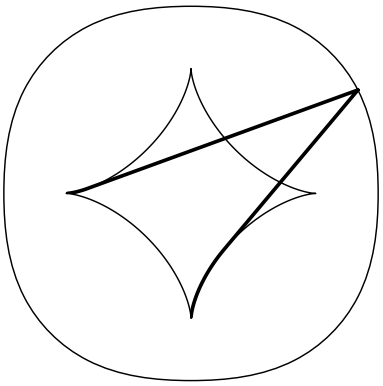

FIGURE 11. A billiard table for the astroid

For instance, let us consider the following question. Is it possible for a billiard to have an open set of periodic orbits? For a Hamiltonian twist 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

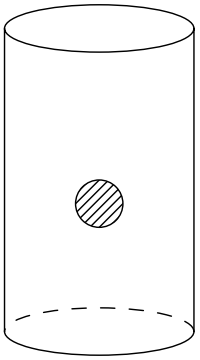

FIGURE 12. The cut out circle

to expect that for billiards nothing similar is possible. And this is not an idle question, because, for example, estimates of the 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 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 triangles with the angles 45° and 30°. The group generated by reflections with respect to the 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 one representative for each coset 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

(once identifications on the boundary are taken into account) can easily shown to be 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 then be identified by translations, and the billiard flow will thus be represented as the free motion of a particle on a (flat) torus: each tangent vector moves in its direction with velocity equal to one. This is a completely integrable system: the initial angle is a first integral, the phase space is foliated 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 genuinely parabolic billiard table is the right triangle with an angle $\frac{\pi}{8}$. 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 an octagon out of them (Fig. 13), then the motion turns into the parallel flow on this octagon, opposite sides being identified. The resulting 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.

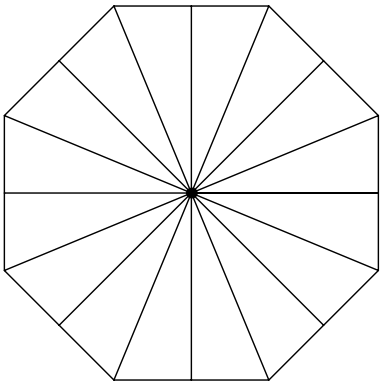

FIGURE 13. The simplest elliptic billiard

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

A similar construction works whenever the angles of the triangle are 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. The flow of parallel translations has a first integral, and this situation can be studied using powerful methods from Teichmüller 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 but countably many invariant manifolds (in this case, for a fixed value of the angle), the system is topologically transitive; and also on almost all invariant manifolds the system is uniquely ergodic, i.e., the invariant measure is unique. And in the

exceptional cases, when the invariant measure is not unique, 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 relatively prime 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 have the illusion that the structure of these flows is 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 a unique invariant measure (measures supported on fixed points are ignored). But, in contrast to the case of flows on a torus, the set of exceptional values of the parameter is uncountable. Recall that on a torus one has a simple dichotomy between the angles whose tangent is rational, when all the orbits are closed, and the angles whose tangent is irrational, when the invariant measure is unique and hence any orbit is uniformly distributed with respect to 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, these are several different homotopy 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. In addition to that, there exists a set of values of the parameter which has zero measure but positive Hausdorff dimension, for which the flow is quasi-minimal (i.e., any semi-trajectory which does not tend to a fixed point is dense), but there exists more than one nonatomic invariant measure.

A deeper inquiry reveals 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 an "anomalously good rational approximation" arises. In the case of linear flows on a torus, for Diophantine 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 angles, time averages can behave quite irregularly: at different times 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 quite low. Correspondingly, even smooth changes of time can essentially change large time dynamics: for example, eigenfunctions, even measurable, may 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 angles whose tangents are irrational Liouville numbers. Therefore it is not surprising that similar but more pronounced differences with the Diophantine situation appear: instead of slow convergence of averages to the integral with respect to 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 impressive 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 number of derivatives. The combination of unique ergodicity (uniqueness of an invariant measure) with the existence of an infinite set of independent invariant *distributions* is a hallmark of parabolic behavior in dynamics. The simplest example of this phenomenon, in which a 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 horocycle 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 decay of averages which is inversely 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. Consider, for example, 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 the 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 the radius T with the 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.,

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

It is also known that periodic orbits are dense in the phase space. The question of 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 are reduced to quadratic differentials on surfaces with a large number of symmetries (Veech surfaces), and on the other hand, for generic quadratic differentials. It is quite likely that there exist polygons with pathological behavior of the function $P(T)$. Note that our first nontrivial example of a billiard

in the right triangle with the angle $\frac{\pi}{8}$ and the hypotenuse 1 yields a Veech surface, and for it, one 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 a general kind. One has to admit that currently available methods of analysis are insufficient for a serious investigation of such systems. Indeed, a successful study of parabolic systems is related to two special situations:

(1) flows on surfaces discussed above, where the dimension of the phase space is very small (in addition to the dimension corresponding to orbits one has only one transverse direction), and
(2) flows on homogeneous spaces, where there is large local symmetry.

Two main open questions concerning arbitrary billiards are the description of the global complexity of behavior of trajectories and the asymptotic behavior of typical trajectories with respect to Lebesgue measure.

Let us begin with the second question. In some some sense a lot is known here, 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 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 is almost uniformly distributed in the phase space, one can show by rather standard category arguments that for a dense G_δ set in the space of parameters the billiard is ergodic in the whole phase space. However, this topologically ample set of billiards is very thin from the measure-theoretic point of view: not only its Lebesgue measure but also its Hausdorff dimension in the space of parameters equals zero. This set reminds one the set of numbers admitting a rational approximation with extremely high speed, like a triple exponential.

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 has been proved. The structure of singular invariant measures for irrational billiards is also not known.

Of course, a particular case of the last question is a 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 there exists at least one periodic billiard trajectory. As 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 the 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 homotopy type, and hence are very long. However, there are some special situations when periodic orbits with simple combinatorics arise which are preserved under small perturbations of the angles. A classical example is the orbit of period 3 formed by the bases of the altitudes in an arbitrary acute 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 maximal altitude traveled back and forth. 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 triangles.

The question of the density or at least the existence of an infinite number of parallel families of periodic orbits remains open.

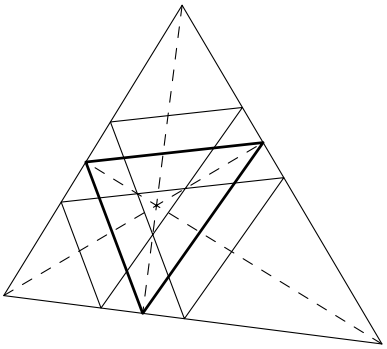

FIGURE 14. Orbits of periods 3 and 6 in an acute triangle

For an arbitrary right triangle the existence of periodic orbits was 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 then bounces 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 fairly easy consequence of preservation 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 the rational numbers. For an arbitrary obtuse 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 of the global complexity of the behavior of trajectories. The growth of the number of distinguishable trajectories with time can be estimated in different ways. The most natural way is related to coding. To each trajectory one assigns a sequence of symbols according to reflections from the sides of the polygon, so that each side is denoted by its own symbol. Of course, in this way one naturally encodes a billiard map, i.e., the return map of a billiard flow to the boundary. In order to obtain full information about the flow, one should also indicate the time between two consecutive reflections. The growth of complexity for the billiard map (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 as obvious that, conversely, each infinite periodic code corresponds to a family of parallel periodic orbits. These orbits close up either after one period or after two periods (as is the case with orbits of period 6 parallel to the Fagnano triangle of an acute triangle).

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

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

and

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

Note that in this case a positive proportion 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 a trajectory hits a vertex. It is also obvious that there is only a finite number of segments of trajectories of bounded length which hit vertices both in the positive and negative directions. By fairly obvious 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 $\le N$ edges (respectively, as the number of generalized diagonals of length $\le T$). As above, these quantities admit a quadratic estimate for rational polygons.

It is natural to believe 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 of much weaker 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 already mentioned, hyperbolic behavior is rather common and allows one to establish the basic elements of stochastic or "chaotic" behavior. The predominance of hyperbolic behavior is natural by analogy with linear systems. Indeed, a randomly chosen matrix most likely does not have any eigenvalues whose absolute value equals one. Even if one a priori restricts oneself to matrices with determinant 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 in billiards were found by Y. G. Sinai [13]. The simplest examples of a Sinai type billiard are a square with a circle cut out or a convex polygon whose sides are replaced by concave arcs (see Fig. 1). From the point of view of rigorous mathematical analysis, the second example turns out to be somewhat simpler than the first. Hyperbolic behavior in Sinai billiards is related to the phenomenon of scattering of light which is well known from geometric optics: a parallel or divergent beam of light becomes more divergent after reflection in a convex mirror. Not too complicated computations show that if the reflection is sufficiently regular, then the angle measure of a such a beam grows exponentially. This yields hyperbolicity of the linearized system.

Two technical difficulties arise in the analysis of scattering billiards.

Firstly, one must achieve sufficient regularity of reflections with respect to concave parts of the boundary. It is clear why in this respect the second example is better than the first: in the second example the time between two consecutive reflections is bounded. 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 small angles with them meet the obstacle only after a very long time. This phenomenon is called infinite horizon; conversely, boundedness of the time between two reflections corresponds to finite horizon. Infinite horizon implies nonuniformity 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 proving ground for various methods and tools of analysis of dynamics.

Indeed, nonuniform 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 nonuniform hyperbolicity is compatible with any topology. This fact, although predicted long ago, was established in full generality only recently by D. Dolgopyat and Y. Pesin [6].

The second difficulty in the analysis of dispersing billiards is related to the presence of singularities (discontinuities and unboundedness of derivatives) in the 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 points of tangency of trajectories with concave parts of the boundary. Naturally, 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 lead only to minor complications. Such singularities yield discontinuities of the first kind for functions representing the dynamics: a surface of discontinuity arises, and functions are smooth on both sides of the surface. Thus, the differential along a billiard trajectory that does not hit a discontinuity point behaves regularly. For trajectories tangent to the boundary from inside, the derivatives near these trajectories are unbounded, so 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 long-term behavior of trajectories is one of the central problems in mechanics. Here as well, billiards, and especially their multidimensional analogues, play the role of an important proving ground.

Scattering billiards are essential for the mathematical foundations of models of statistical physics. This is an important and interesting subject, which however we will not touch here. From the viewpoint of geometry, scattering billiards possess some defects, for example, unavoidable singularities on the boundary. However, if one considers billiards not on planar 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 may arise in ways other than scattering by concave parts of the boundary. The first answer to this question is given by the celebrated example of a "stadium," i.e., two semicircles connected by segments of their common tangents (see Fig. 15). This is an example of the so-called Bunimovich billiard [2], where hyperbolic behavior arises as a result of consecutive focusing of families 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 positions and velocities are taken into account, uniform exponential growth arises.

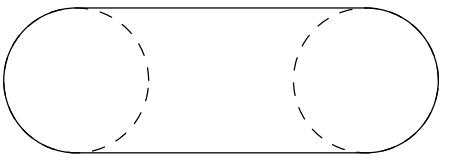

FIGURE 15

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 divergence and stochastic behavior of orbits. He discovered that if one adds little round "pockets" to a scattering billiard, then the billiard on the resulting table in which convex parts are followed by concave ones, exhibits exponential divergence of trajectories. Actually Bunimovich discovered a new important mechanism of hyperbolicity. However, he himself originally considered his work as just a little generalization of the results on scattering billiards. During Bunimovich's talk on a seminar at the Steklov Institute 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 was necessary. I pointed out to the speaker 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 the 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 could be weakened, so that, for example, in the case of stadium, the distance between the circles can be arbitrarily small.

Among Bunimovich billiards there are a lot of other interesting and rather simple shapes, but they all have the common property that the boundary contains only scattering parts, line segments and arcs of circles. The natural question of how essential this condition is puzzled the experts for about ten years. A technical difficulty is the following. Hyperbolicity is established with the help of a system of cones in tangent spaces to the points of the phase space, which are mapped into 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 sectors formed by a pair of lines intersecting at the origin. The invariant system of cones is the same in both scattering billiards and Bunimovich billiards. Geometrically, these cones are defined as the sets of infinitesimal dispersing beams 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 circular mirrors, the cone goes into itself, but one of its sides is left invariant. This is a typically parabolic effect, since this is exactly the 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 $x_1x_2 > 0$, i.e., it is the union of the first and the third quadrants on the plane. Its image is the cone $|x_1| > |x_2|$, $x_1x_2 > 0$ (see Fig. 16). After further iterations the image becomes thinner and thinner, 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* circular parts of the boundary (as in the case of the stadium). Since the trajectory which is reflected from the circular part under a very small angle, continues to do this many times, it looked like the explicit form of the iteration after reflection with respect to circular parts (integrability of the billiard in a circle) played an essential role. This is the way I explained to myself the rigidity of the Bunimovich conditions.

However, it turned out that one can overcome this difficulty. Billiards with convex parts of the boundary can be hyperbolic for a variety of reasons. As soon as the point of view based on the use of systems of invariant cones was fully understood, the problem of finding new classes of hyperbolic billiards became easy. Note that Bunimovich used another technique which is

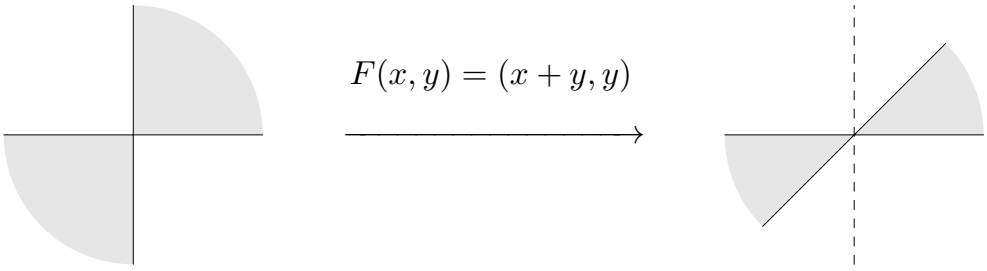

FIGURE 16. The action of the parabolic transformation on the cone

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 Jürgen Moser (1928–1999). An essential step was to introduce this method in nonuniformly hyperbolic situation. The author used this method for constructing examples of smooth systems with stochastic behavior on various manifolds. However, the crucial advance here is due to Maciej Wojtkowski. And again billiards turned out to be an ideal proving ground. Having understood the key role of systems of cones, Wojtkowski realized that the problem could be solved in reverse 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 design of billiards with nonvanishing Lyapunov exponents." While the square or torus with a removed circle is a quintessence of the phenomenon discovered by Sinai, and the stadium represents Bunimovich billiards, in the same way a typical example of Wojtkowski billiards is given by the cardioid (see Fig. 1). The importance of Wojtkowski's result for the theory of billiards is that he discovered classes of hyperbolic examples which are open in the C^2 topology, and thus this property does not depend on small variations of mirrors [11].

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 the 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 nonuniform hyperbolic behavior.

Important unsolved problems are related to the 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 not simply-connected boundaries are unknown.

What does hyperbolicity give? It allows one to show that in many cases a deterministic dynamical system behaves in many respects like a sequence of independent random variables. In a sense, this statement is true literally: under some (often easily verifiable) conditions, in addition to (even nonuniform) hyperbolicity, the phase space of a system preserving a finite volume can be divided into a finite number of sets $A_1, \dots, A_n$ of positive measure so that, first, almost every point of the phase space is uniquely determined by the sequence of visits 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 may be quite exotic, this property, which is naturally called the Bernoulli property, implies many important properties: convergence of time averages to the space average (ergodicity), decay 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*, Colloquium Publ. vol. 9, Amer. Math. Soc., Providence, 1927.

[2] L. A. Bunimovich. On the ergodic properties of nowhere dispersing billiards, *Comm. Mathemat. Phys.* 65 (1979), no. 3, 295–312.

[3] W. Wojtkowski. Invariant families of cones and Lyapunov exponents, *Erg. Theory and Dynam. Syst.* 5 (1985), 145–161. [See also Principles for the design of billiards with nonvanishing Lyapunov exponents. *Comm. Math. Phys.* 105 (1986), no. 3, 391–414.]

[4] G. A. Galperin, A. N. Zemlyakov. *Mathematical billiards*. Nauka, Moscow, 1990 ("Kvant" Library series, no. 77) (in Russian.)

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

[6] D. Dolgopyat, Y. Pesin. Every compact manifold carries a completely hyperbolic diffeomorphism. *Erg. Theory and Dynam. Syst.* 22 (2002), 409–435.

[7] V. J. Donnay. Using integrability to produce chaos: billiards with positive entropy, *Comm. Math. Phys.* 141 (1991), no. 2, 225–257.

[8] A. B. Katok, B. Hasselblatt. *Introduction to the modern theory of dynamical systems*, Cambridge University Press, 1995.

[9] I. P. Cornfeld, S. V. Fomin, Y. G. Sinai. *Ergodic theory*, Springer-Verlag, 1982.

[10] V. F. Lazutkin. Existence of caustics for the billiard problem in a convex domain, *Math. USSR-Izv.* 7 (1973), 185–214. 186–216.

[11] H. Masur, S. Tabachnikov. Rational billiards and flat structures. *Handbook in Dynamical Systems* Vol. 1A, 1015–1089, North-Holland, Amsterdam, 2002.

[12] D. Szász, ed., *Hard ball systems and the Lorentz gas,* Springer–Verlag, Berlin, 2000. (Encyclopedia of Mathematical Sciences, 101, Mathematical Physics, II.)

[13] Y. G. Sinai. Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards, *Uspekhi Mat. Nauk* (Russian Math. Surveys) 25 (1970), no. 2, 141–192.

[14] S. Tabachnikov. *Billiards*, Panoramas et Synthèses 1 (1995).