Markov partitions for two-dimensional hyperbolic billiards

1990 Russ. Math. Surv. 45 105

(http://iopscience.iop.org/0036-0279/45/3/R02)

Uspekhi Mat. Nauk 45:3 (1990), 97–134 Russian Math. Surveys 45:3 (1990), 105–152

Markov partitions for two-dimensional hyperbolic billiards

L.A. Bunimovich, Ya.G. Sinai, and N.I. Chernov

Contents

§1. Fundamental definitions and statements of results

1.1. Let Q be a bounded connected closed domain in the plane $\mathbb{R}^2$ or on the two-dimensional torus Tor^2 with the Euclidean metric. We assume that the boundary ∂Q consists of finitely many smooth (class C^3) non-intersecting curves Γ_i, $1 \leqslant i \leqslant d$, which are either closed or have end points in common.

A *billiard dynamical system* (or simply, a *billiard*) in Q is generated by the inertial motion of a material point inside Q. The point moves rectilinearly inside Q with unit velocity. When hitting the boundary ∂Q, the point is reflected by the law "the angle of incidence equals the angle of reflection".

Billiards serve as model systems in many problems of physics (see the survey [6]). In particular, such a popular model in statistical mechanics as the Lorentz gas belongs to the class of billiard systems under consideration [6], [12], [19].

The ergodic properties of billiard systems are determined by the structure of the boundary ∂Q, more precisely, by its geometric curvature at the regular points. We equip each component Γ_i with the unit normal vectors $n(q)$, $q \in \Gamma_i$, directed to the inside of Q. Then at each point $q \in \Gamma_i$ the curvature $\varkappa(q)$ of Γ_i at q has a well-defined sign. We assume it to be of constant sign on each curve Γ_i, including the end points. Correspondingly, we call a component Γ_i *scattering* if $\varkappa(q) > 0$ for all $q \in \Gamma_i$ (in this case Γ_i is convex, seen from the interior of Q), *focussing* if $\varkappa(q) < 0$ for all $q \in \Gamma_i$ (that is, Γ_i is convex as viewed from the outside), and *neutral* if $\varkappa(q) \equiv 0$ on Γ_i (in this case Γ_i is a segment of a straight line).

Definition. A billiard in Q is called *scattering* if all components of ∂Q are scattering (Fig. 1).

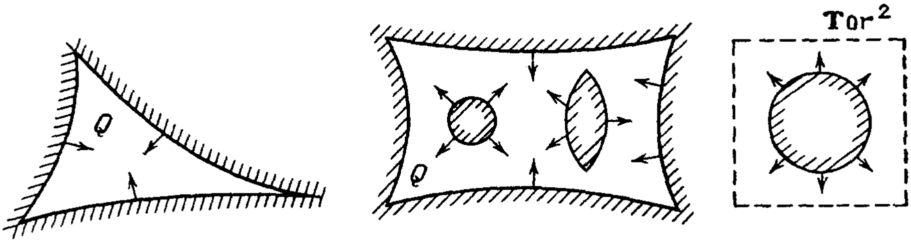


Fig. 1

The stochastic properties of two-dimensional scattering billiards have been studied rather exhaustively. They are ergodic, are K-systems (see [10]), and in a number of cases the B-property has been proved for them [21].

Everywhere in this article (except in §6) we consider scattering billiards only. Such billiards have a very strongly expressed hyperbolic behaviour (for more detail see §2). Certain classes of billiards with focussing and neutral boundary components are also hyperbolic. One such class is considered in §6.

The phase space of a billiard system in Q is the unit tangent bundle M over Q, that is, $M = \{x = (q, \upsilon) : q \in Q, \|\upsilon\| = 1\}$, or $M = Q \times S^1$. We denote by π the natural projection of M on Q: $\pi x = q$. It can easily be seen that M is a three-dimensional manifold with boundary $\partial M = \bigcup\limits_i \pi^{-1}\Gamma_i = \bigcup\limits_i M^{(i)}$. We put

$$S_0 = \{x = (q, v) \colon q \in \partial Q, \ (v, n(q)) = 0\},$$
$$V_0 = \bigcup_{i \neq j} \pi^{-1}(\Gamma_i \cap \Gamma_j), \quad R_0 = S_0 \cup V_0.$$

The one-parameter group of shifts along the trajectories of a billiard system is denoted by $\{S^t\}$, $-\infty < t < \infty$. The systems considered are discontinuous. In particular, the flow $\{S^t\}$ is defined only on the subset $M' = \{x : S^t x \notin V_0$ for all $t\}$. We will assume that the trajectories of points $x \in M \backslash M'$ are defined only up to the moment of hitting the set V_0.

The Liouville measure on M has the form $d\mu = \text{const}\ dq d\upsilon$, where dq and $d\upsilon$ are the Lebesgue measures on Q and S^1, respectively, and const is a normalizing factor. It is well known that $\{S^t\}$ preserves the measure μ and that $\mu(M') = 1$ (see, for example, [7]).

The ergodic properties of flows are often studied using special representations of the flows, in other words, using the successor maps of a certain cross-section [14]. In the case of billiards a cross-section can easily be constructed, using the natural boundary ∂M of the phase space. We put

$M_1 = \{x = (q, v) : q \in Q, (v, n(q)) \geqslant 0\}$. Clearly, M_1 is a two-dimensional manifold with boundary $\partial M_1 = R_0$. Let us put $M_1^{(i)} = M_1 \cap M^{(i)}$, $1 \leqslant i \leqslant d$.

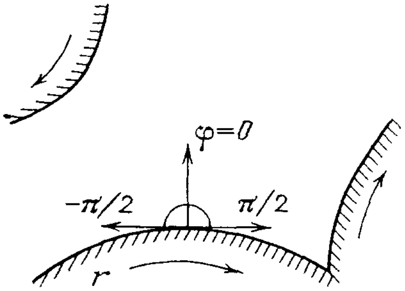


Fig. 2

On each $M_1^{(i)}$ we introduce natural coordinates (r, φ), where r is the length parameter on the curve Γ_i and φ is the angle between the vectors v and $n(q)$, $-\pi/2 \leqslant \varphi \leqslant \pi/2$. The coordinate r is measured from some point $q_0 \in \Gamma_i$ and increases when moving along the boundary ∂Q in such a way that the domain Q remains on the left (Fig. 2). With these coordinates $M_1^{(i)}$ is either a rectangle (when Γ_i is not closed) or a cylinder (when Γ_i is closed).

For a phase point $x \in M$ we denote by $\tau^+(x)$ and $\tau^-(x)$ the first positive and negative moments of hitting the boundary ∂M: $\tau^+(x) = \min\{\tau > 0 : S^\tau x \in \partial M\}$, $\tau^-(x) = \max\{\tau < 0 : S^\tau x \in \partial M\}$. It can be shown [10] that $|\tau^\pm(x)| < \infty$ almost everywhere on M and everywhere on M_1. The successor map $T : M_1 \to M_1$ is defined by $Tx = S^{\tau^+(x)+0}x$, $x \in M_1$. It is known [7] that T preserves the measure $d\nu = \text{const} \cos \varphi \, d\tau \, d\varphi$ (as before, const is a normalizing factor).

We put $M_1' = M_1 \cap M'$. The map T is one-to-one on M_1'. The images $T^n x$ of the remaining points $x \in M_1 \backslash M_1'$ are not defined for all $n \in \mathbb{Z}$, but only up to the moment of hitting the set V_0. In the language of ergodic theory $\{S^t\}$ is called a *special flow*, constructed from the automorphism T, base space M_1, and function $\tau^+(x)$ (see [7]).

The maps T and T^{-1} are piecewise smooth on M_1. The set of singular points for T (respectively, T^{-1}) coincides with $R_0 \cup T^{-1}R_0$ (respectively, $R_0 \cup TR_0$). We put $S_k = T^k S_0$, $V_k = T^k V_0$, $R_k = T^k R_0$ for $-\infty < k < \infty$, and also $R_{k,m} = \bigcup_k^m R_i$ for $-\infty \leqslant k < m \leqslant \infty$ (here and in the sequel, $TV_0 = \{Tx : x = (q, v), q \in \Gamma_i \cap \Gamma_j$ for $i \neq j$ and v pointing inside $Q\}$; the set $T^{-1}V_0$ is similarly defined). Then the set of singular points of $T^{\pm n}$, $n \geqslant 1$, coincides with $R_{-n,0}$ (respectively, $R_{0,n}$). It is not difficult to see that the set $R_{-\infty,\infty}$ consists of countably many smooth (C^1) curves, which will simply be called *discontinuity curves* in the sequel. Finally, the *multiplicity* of a point $x \in M_1$ is the (possibly infinite) number of discontinuity curves passing through it or ending at it.

A basic instrument in the study of ergodic properties of hyperbolic systems consists of *locally stable* and *unstable manifolds* (LSM and LUM, for short). In our case a LSM is defined as a smooth (C^1) curve $\gamma^s \subset M_1$ (regarded without end points) such that T^n is continuous on γ^s for all $n \geqslant 1$, while the length of the image $T^n\gamma^s$ tends to 0 as $n \to \infty$. The LUM γ^u is similarly defined, with T^n replaced by T^{-n}.

For two-dimensional scattering billiards it has been proved [10] that ν-almost every point $x \in M_1$ has a LSM and a LUM passing through x. We denote by $\gamma^u(x)$ (respectively, $\gamma^s(x)$) the maximal smooth segment of the LUM (respectively, LSM) passing through x. The sets $\Gamma^s(x) = \bigcup_{n \geqslant 0} T^{-n}\gamma^s(T^n x)$ and $\Gamma^u(x) = \bigcup_{n \geqslant 0} T^n\gamma^u(T^{-n}x)$ will be called the *globally stable and unstable manifolds* (abbreviated to GSM and GUM), respectively (see [14]). For $x, y \in M_1$ we put $[x, y] = \gamma^u(x) \subset \gamma^s(y)$. In §2 we will prove that the set $[x, y]$ consists of at most one point. If $A \subset \gamma^s(x)$, $B \subset \gamma^u(y)$ and the point $[x', y']$ is defined for all $x' \in A$, $y' \in B$, then we put $[A, B] = \{[x', y'] : x' \in A, y' \in B\}$.

1.2. Markov partitions.

In M_1 we introduce special subsets, called parallelograms.

Definition. A subset $U \subset M_1$ is called a *parallelogram* if for any pair of points $x, y \in U$ the point $[x, y]$ is defined and belongs to U.

In the definition of a parallelogram it is usual to require $\nu(U) > 0$. We discard this requirement, and call a parallelogram of positive measure *non-degenerate*, and one of zero measure *degenerate*.

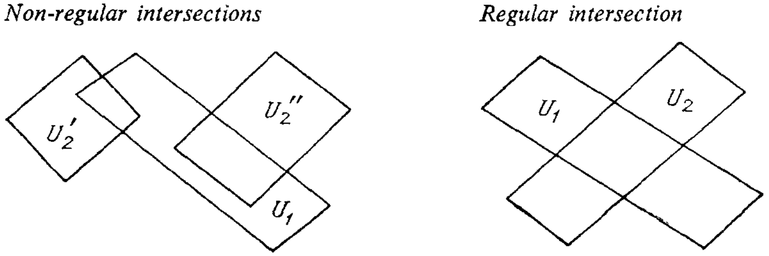
The term "parallelogram" for elements of a Markov partition was introduced in [16] (in which a Markov partition was constructed for an automorphism of the two-dimensional torus with elements of the form of the "present" parallelograms). In [3] the term "rectangle" was used for the same purpose (see also [8]).

For a subset $A \subset M_1$ we put $\gamma^u_A(x) = \gamma^u(x) \cap A$ and $\gamma^s_A(x) = \gamma^s(x) \cap A$. Subsets $A_1 \subset \gamma^u(x_1)$ and $A_2 \subset \gamma^u(x_2)$ are called *canonically isomorphic* [18] if for any point $y \in A_1$ the LSM $\gamma^s(y)$ intersects A_2, and conversely. A similar definition is given for subsets of LSM. In a parallelogram U, for all $x \in U$ the sets $\gamma^u_U(x)$ are canonically isomorphic. This is true also for the sets $\gamma^s_U(x)$, $x \in U$. Therefore any parallelogram U can be represented as $U = [\gamma^s_U(x), \gamma^u_U(y)]$, where x, y are arbitrary points of U. In other words, the parallelogram U has the structure of a direct product.

In the definition of a parallelogram it is not assumed that $\gamma^u_U(x)$, $\gamma^s_U(x)$ are connected, open, or closed. In certain smooth systems they can be chosen to be connected [14], and then U is a curvilinear quadrangle. However, in discontinuous systems (including billiards) the LUM and LSM can be arbitrarily short in a neighbourhood of any point of the phase space.

Therefore the parallelograms U and sets $\gamma^u_U(x)$, $\gamma^s_U(x)$ for $x \in U$ have a rather complicated structure. In our case they are totally disconnected sets of Cantor type (see §5).

The intersection of an ordered pair of parallelograms U_1 and U_2 is called *regular* if it is non-empty and if $U_1 \cap U_2 = [\gamma^s_{U_2}(x), \gamma^u_{U_1}(x)]$ for any point $x \in U_1 \cap U_2$. The corresponding intersections are schematically drawn in Fig. 3.


Non-regular intersections

Regular intersection

Fig. 3

Let η be a finite or countable covering (mod 0) of the space M_1 of closed parallelograms $\{U_i\}$ such that $\nu(U_i \cap U_j) = 0$ for $i \neq j$ and $\nu(M_1 \setminus \bigcup_i U_i) = 0$ (that is, ν-almost every point $x \in M_1$ is covered by exactly one parallelogram). For those $x \in M_1$ that are covered by exactly one element $U_i \in \eta$ we will denote this element by $U(x)$.

Definition. A *Markov partition* is a covering η such that: a) every parallelogram $U \in \eta$ lies in a connected domain $V(U) \subset M_1$ on which the maps T and T^{-1} are continuous; b) for ν-almost every point $x \in M_1$ the parallelograms $U(x)$ and $TU(T^{-1}x)$ intersect regularly.

The first Markov partitions for Anosov Y-systems were (in a somewhat different form) introduced in [8]. Later they were constructed for the more general A-systems of Smale [2], [3]. An exhaustive exposition of the corresponding theory is given in the books [3], [14]. In all these cases the partition η is finite, and its elements have a comparatively simple structure.

The importance of Markov partitions lies in the fact that they allow one to construct a convenient symbolic representation of the automorphism T.

We recall how one constructs in the general case a symbolic representation of an automorphism T of a measure space (M, ν) using an arbitrary finite or countable measurable partition $\eta = \{U_i\}$, $1 \leqslant i \leqslant N$ $(N \leqslant \infty)$. For any point $x \in M$ we define a two-sided sequence of indices $\sigma = \sigma(x) = \{\dots, \sigma_{-1}, \sigma_0, \sigma_1, \dots, \sigma_n, \dots\}$, where $T^n x \in U_{\sigma_n}$, $-\infty < n < \infty$. The secquence $\sigma(x)$ is called a *coding* of the point x. We denote by Σ the space of two-sided sequences $\sigma = \{\sigma_n\}$, $1 \leqslant \sigma_n \leqslant N$, and let θ be the left shift on Σ, that is, $\theta\sigma = \sigma'$, where $\sigma'_n = \sigma_{n+1}$. We consider the map $\Phi : M \to \Sigma$ mapping x to $\sigma(x)$, and let $\Sigma_\Phi = \Phi(M) \subset \Sigma$. It can easily be verified that $\Phi \circ T = \theta \circ \Phi$, that is, Φ is the union of the automorphism T and the shift θ.

The invariant measure ν in M induces a measure ν_Σ in Σ by the formula $\nu_\Sigma(A) = \nu(\Phi^{-1}A)$, $A \subset \Sigma$. In this manner the measure space (Σ, ν_Σ) becomes a realization space of a stationary stochastic process with a finite or countable number of states. Therefore problems on the study of stochastic properties of dynamical systems become problems in probability theory or in the one-dimensional statistical mechanics of lattice systems.

Among the simple partitions, Markov partitions are distinguished by the fact that the symbolic dynamics (Σ_Φ, θ) obtained via them allows a relatively simple description: it is (mod 0) a topological Markov chain [2]. Namely, we introduce the intersection matrix $\Pi = ||\pi_{ij}||$ by

$$\pi_{ij} = \begin{cases} 1 & \text{if } U_j \cap TU_i \text{ is a regular intersection,} \\ 0 & \text{otherwise.} \end{cases}$$

The space Σ_Π of sequences $\{\sigma_n\}$ satisfying the condtion $\pi_{\sigma_n \sigma_{n+1}} \equiv 1$ for all $n \in \mathbb{Z}$, on which the left shift θ is defined, is a *topological Markov chain* (TMC). It can easily be shown [3] that for any sequence $\sigma \in \Sigma_\Pi$ the intersection $\bigcap_{n=-\infty}^{\infty} T^{-n} U_{\sigma_n}$ consists of precisely one point.

Theorem [3], [8]. *If η is a finite Markov partition of a space M with elements of sufficiently small diameter, then*

1. $\Sigma_\Phi \supseteq \Sigma_\Pi$;
2. $\nu_\Sigma (\Sigma_\Phi \setminus \Sigma_\Pi) = 0$;
3. *the map Φ^{-1} is well defined on Σ_Π and is continuous on this set.*

This theorem can be transferred without difficulty to countable Markov partitions, as was noted in [18].

Thus, any sequence σ satisfying the "scattered transitions" conditions $\pi_{\sigma_n \sigma_{n+1}} \equiv 1$ is the coding of some trajectory of the automorphism T.

1.3. The fundamental results of this paper.

***Theorem* 1.1.** *Let the domain Q generating the two-dimensional scattering billiard satisfy the following conditions:*

A. *All interior angles formed by the intersection of the smooth components of the boundary ∂Q are strictly positive.*

B. *The multiplicity of all points of the space M_1 is uniformly bounded above by a constant $K_0 = K_0(Q) < \infty$ (this condition holds for a domain Q in general position).*

Then for any $\varepsilon > 0$ there is a countable Markov partition of M_1 with elements of diameter less than ε.

The first Markov partitions for two-dimensional scattering billiards were constructed in [18] (see the correction in [20]). In the present paper the construction in [18] is simplified at several points. Moreover, we give it in a more general setting.

Unlike smooth systems, for which there is a finite Markov partition, in the case of billiards one cannot expect this. The reason is that the LSM and

LUM can be arbitrarily short, and therefore there must exist elements of the Markov partition that have arbitrarily small dimensions.

Moreover, the presence of a countable Markov partition does not allow one to immediately obtain the consequences holding for a smooth system with finite Markov partition. (Such are: an estimate for the rate of decrease of correlations [19], the central limit theorem, and asymptotics of the number of periodic trajectories [20].) In our case the symbolic dynamics (Σ_Π, θ) obtained does have certain additional properties (these are partly described in [18]). The derivation of these properties requires an additional study of the Markov partition constructed (a number of statements in §5 are concerned with this). A more precise treatment will be given in a forthcoming publication of the authors.

Remark 1.2 [3], [18]. It suffices to construct a Markov partiton η for $T_1 = T^m$ for some $m \geqslant 1$, since then

$$\eta \vee T\eta \vee \ldots \vee T^{m-1}\eta$$

is a Markov partition for T.

The proof of Theorem 1.1 is presented in §§3–5. In §§3 and 4 we construct a so-called pre-Markov partition (an intermediate stage in the construction of a Markov partition). In §3 we impose one additional restriction: $|\tau^{\pm}(x)| \leqslant \mathrm{const}(Q) < \infty$ for all $x \in M_1$ (such systems are called *billiards with finite horizon*). In §4 we study billiards with infinite horizon and generalize the results of §3 to this case. In §5 we construct a Markov partition starting from a pre-Markov partition. In §6 the results of Theorem 1.1 are transferred to a class of non-scattering billiards with hyperbolic behaviour. Finally, in §7 we derive, as a corollary of Theorem 1.1, an exponential lower bound for the number of periodic trajectories of the automorphism T. We hope that the asymptotics of the number of periodic points can be studied more completely by using the Markov partition constructed.

The authors express their gratitude to A. Krámli and D. Szász, who indicated an inaccuracy in the handwritten text.

§2. General properties of two-dimensional scattering billiards

The main content of §2 lies in the description of the geometric structure of the discontinuity curves, the LSM, and the LUM in the space M_1. A number of statements are published for the first time.

2.1. Increasing and decreasing curves.

A smooth (C^1) curve γ in M_1 is called *increasing* (*decreasing*) if it is given by an equation $\varphi = \varphi(r)$ and $d\varphi/dr \geqslant 0$ ($d\varphi/dr \leqslant 0$). Such curves will be called *monotone*. They have the important property of semi-invariance.

Lemma [10]. *If γ is an increasing (decreasing) curve and T (T^{-1}) is continuous on γ, then $T\gamma$ ($T^{-1}\gamma$) is also an increasing (decreasing) curve.*

This property is, in essence, equivalent to the condition of semi-invariance of a system of stable and unstable cones in the tangent space, introduced in [31].

A curve γ is called *m-increasing* (*m-decreasing*) for $m \geqslant 1$ if T^{-m} (T^m) is continuous on γ and $T^{-m}\gamma$ ($T^m\gamma$) is increasing (decreasing). Clearly, the m-increasing (m-decreasing) curves do not intersect $R_{0,m}$ ($R_{-m,0}$). A curve γ is called *neutral* if γ is a segment in R_0.

Let an increasing or decreasing curve γ be given by an equation $\varphi = \varphi(r)$, $r_1 \leqslant r \leqslant r_2$. We denote by $l(\gamma)$ its length in the metric $ds^2 = dr^2 + d\varphi^2$. Also, we define the *p-length* of γ by the formula [21]

$$p(\gamma) = \int_{r_1}^{r_2} \cos\varphi \, d\varphi. \tag{2.1}$$

At each point $x = (r, \varphi)$ of the curve γ we define the quantities

$$\chi_\gamma^\pm(x) = \frac{1}{\cos\varphi}\left(\frac{d\varphi}{dr} \pm \varkappa(r)\right), \tag{2.2}$$

where $\varkappa(r)$ denotes the curvature of ∂Q (see §1).

Remark 2.1. Since the components of ∂Q are smooth, we have $0 < \varkappa_{\min} \leqslant \varkappa(r) \leqslant \varkappa_{\max} < \infty$ for all points $r \in \partial Q$.

We disclose the clear geometrical meaning of the quantities introduced in (2.1)–(2.2). The points of γ generate an "outgoing" pencil of trajectories $\{S^t y\}$, $y \in \gamma$, $t > 0$. We consider an orthogonal cross-section $\sigma(x)$ of this pencil, passing through an arbitrary point $x = (r, \varphi) \in \gamma$ (Fig. 4), and its equipment by normal vectors, directed along the motion of the pencil. Then the p-length element of γ at x equals the length element of $\sigma(x)$ at x, while $\chi_\gamma^+(x)$ equals the curvature of $\sigma(x)$ at x. Similarly, $\chi_\gamma^-(x)$ equals the curvature of the orthogonal cross-section of the "incoming" pencil of trajectories $\{S^t y\}$, $y \in \gamma$, for $t < 0$.

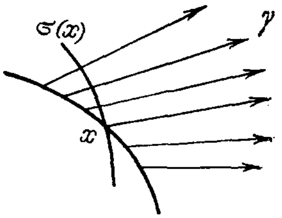


Fig. 4

2.2. The hyperbolic structure of the automorphism T.

First we study the properties of expansion and contraction.

***Lemma* 2.2** [21]. *If T (T^{-1}) is continuous on the increasing (decreasing) curve γ, then* $p(T^{\pm 1}\gamma) = \int_\gamma (1 + \chi_\gamma^\pm(x)\tau^\pm(x))\cos\varphi\, d\varphi$ *("$+$" corresponds to an increasing curve, "$-$" to a decreasing curve).*

Note that for decreasing curves $\chi_\gamma^-(x) < 0$ and $\tau^-(x) < 0$. Therefore increasing (decreasing) curves expand under the action of T (T^{-1}). It is important to estimate the coefficient of expansion. From (2.2) and Remark 2.1 it follows that for increasing (decreasing) curves γ the estimate $\chi_\gamma^+(x) > \varkappa_{\min} > 0$ (respectively, $\chi_\gamma^-(x) < -\varkappa_{\min} < 0$) holds. Condition A of Theorem 1.1 readily implies that for certain constants $m_0 = m_0(Q)$ and $\tau_0 = \tau_0(Q) > 0$ one has the following result.

***Assertion* 2.3** [11], [14]. *For any point $x \in M_1$ there are among the m_0 first reflections of its trajectory on the boundary ∂Q two neighbouring reflections between which the segment of the trajectory has length at least τ_0.*

In other words, the trajectories of a billiard cannot undergo arbitrarily small reflections while remaining in a small neighbourhood of a break point of ∂Q.

Statements 2.2 and 2.3 imply the following result.

***Lemma* 2.4.** *If T^{m_0} (T^{-m_0}) is continuous on the increasing (decreasing) curve γ, then* $\frac{p(T^{m_0}\gamma)}{p(\gamma)} \geqslant \Lambda_0$, $\frac{p(T^{-m_0}\gamma)}{p(\gamma)} \geqslant \Lambda_0$, *where* $\Lambda_0 = 1 + \varkappa_{\min}\tau_0 > 1$ *is a constant for Q.*

Thus, the expansion and contraction in the tangent space to M_1 under the action of T^{m_0} is of a uniform nature.

For $m \geqslant 1$ we denote by Λ_m the minimal coefficient of expansion of increasing (decreasing) curves under the action of T^m (T^{-m}).

***Corollary* 2.5.** $\Lambda_m \geqslant \Lambda_0^{[m/m_0]}$.

Finally we study the angles between stable and unstable directions in the tangent space to M_1.

Let γ be an increasing curve and let T^m, $m \geqslant 1$, be continuous on γ. Then for each point $x \in \gamma$ we have [10], [21]

$$(2.3)\quad \chi^+_{T^m\gamma}(x_m) = \frac{2\varkappa_m}{\cos\varphi_m} + \cfrac{1}{\tau_m + \cfrac{1}{\cfrac{2\varkappa_{m-1}}{\cos\varphi_{m-1}} + \cfrac{1}{\tau_{m-1} + \ddots + \cfrac{1}{\tau_1 + (\chi_\gamma^+(x))^{-1}}}}},$$

using the notations $x_i = (r_i, \varphi_i) = T^i x$, $\varkappa_i = \varkappa(r_i)$, $\tau_i = \tau^+(x_{i-1})$. A similar formula holds for decreasing curves. A description of the hyperbolic properties of billiard systems using continued fractions appeared in [10]. Formula (2.3) allows one to prove the following estimate for the derivative $d\varphi/dr$ for monotone curves:

Lemma **2.6.** *For any l-increasing (l-decreasing) curve* $\varphi = \varphi(r)$ *we have*

$$| d\varphi/dr | \geqslant \varkappa (r) \geqslant \varkappa_{\min}.$$

The same estimate holds for the smooth components of the sets R_1 *and* R_{-1}.

Lemma **2.7.** *For any* m_0*-increasing* (m_0*-decreasing*) *curve* $\varphi = \varphi(r)$ *lying in* $M^{(i)}$ *we have*

$$| d\varphi/dr | \leqslant \text{const} (Q) (d (r))^{-1/2},$$

where $d(r)$ *is the distance from the point* $r \in \Gamma_i$ *to the nearest end point of* Γ_i *(for closed curves* Γ_i *we may put* $d(r) \equiv 1$).

Thus, in a neighbourhood of V_0 the angles between stable and unstable directions are not separated from zero. Therefore, if $V_0 \neq \varnothing$, then the billiard system is only non-uniformly hyperbolic (see also [22]).

Lemmas 2.6 and 2.7 readily imply that for m_0-increasing and m_0-decreasing curves γ the following relation holds:

$$(2.4) \qquad \sqrt{1 + \varkappa_{\min}^2}\, p (\gamma) \leqslant l (\gamma) \leqslant \text{const} (Q) \sqrt{p (\gamma)}.$$

For a monotone curve γ and points $a, b \in \gamma$ we will denote by $\gamma(a, b)$ the segment of γ from a to b.

2.3. Discontinuity curves.

Lemma [18]. *For* $k \geqslant 1$ *the smooth components of the set* $R_{1,k}$ ($R_{-k,-1}$) *are increasing (decreasing) curves.*

Lemma **2.8.** *For* $k \geqslant 1$ *any smooth component of the set* $R_{1,k}$ ($R_{-k,-1}$) *lies on some continuous (not necessarily smooth) monotone curve in* $R_{0,k}$ ($R_{-k,0}$) *whose end points belong to* R_0. *For any integer* $l < k$ *the set* $R_{l,k}$ *partitions* M_1 *into curvilinear polygons, whose interior angles do not exceed* 180°.

The proof is by induction over l, k using simple geometric analysis.

In Fig. 5 the typical structure of discontinuity curves is depicted.

We denote by $\mathcal{O}_{\varepsilon,m}^+$ ($\mathcal{O}_{\varepsilon,m}^-$) the union of all l-increasing and l-decreasing curves of p-length not exceeding ε and ending on $R_{0,m}$ ($R_{-m,0}$).

Remark 2.9 [18]. In billiards with finite horizon the set $R_{l,k}$ consists, for all integral $l < k$, of finitely many smooth components.

Remark 2.10. In billiards with finite horizon there is, for any $1 \leqslant m < \infty$, a $\delta_0 = \delta_0(m) > 0$ such that any increasing (decreasing) curve of p-length not exceeding δ_0 intersects $R_{-m,0}$ $(R_{0,m})$ at K_0 points at most.

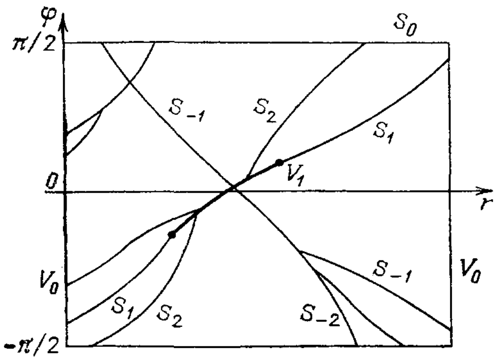


Fig. 5

2.4. For two-dimensional scattering billiards the LUM and LSM have been constructed and well described in [10], [22]. They are solutions of ordinary differential equations $d\varphi/dr = B^u(r, \varphi)\cos\varphi + \varkappa(r)$ (for the LUM) and $d\varphi/dr = -B^s(r, \varphi)\cos\varphi - \varkappa(r)$ (for the LSM). Here

$$B^s(r, \varphi) = \cfrac{1}{\tau_1 + \cfrac{1}{\cfrac{2\varkappa_1}{\cos\varphi_1} + \cfrac{1}{\tau_2 + \cfrac{1}{\cfrac{2\varkappa_2}{\cos\varphi_2} + \cfrac{1}{\cdots}}}}} \tag{2.5}$$

is a continued fraction in which $(r_n, \varphi_n) = T^n(r, \varphi)$, $\varkappa_n = \varkappa(r_n)$, and $\tau_n = \tau^+(T^{n-1}(r, \varphi))$. The quantity $B^u(r, \varphi)$ is similarly defined, using the semitrajectories $T^n(r, \varphi)$ for $n \leqslant 0$. It is easily seen that all elements of the continued fraction (2.5) are positive and $\sum_n \tau_n = \infty$. This implies [15] that $B^s(r, \varphi)$ is defined for all points $x \in M_1 \setminus R_{-\infty,0}$. Similarly, $B^u(r, \varphi)$ is defined for all $x \in M_1 \setminus R_{0,\infty}$. Moreover, $B^u(r, \varphi)$ and $B^s(r, \varphi)$ are continuous as functions of $x = (r, \varphi)$ in their domains of definition [10].

These properties of the LSM and LUM allow one to prove the following.

***Lemma* 2.11.** *If $\{\gamma_n\}$ is a sequence of k_n-increasing (k_n-decreasing) curves converging in the metric of C^0 to a continuous curve γ, and if $k_n \to \infty$ as $n \to \infty$, then γ is a LUM (LSM).*

Thus, for large k the k-increasing (k-decreasing) curves approximate the LUM (LSM). In particular, LUM (LSM) are increasing (decreasing) curves, so that for arbitrary $x, y \in M_1$ the set $[x, y]$ consists of at most one point.

It is well known that for $x \in M_1$ the length of the LUM $\gamma^u(x)$ (the length of the LSM $\gamma^s(x)$) depends on how close the trajectory $T^n x$ for $n < 0$ ($n > 0$) can approach the boundary ∂M_1. To give a strict statement, we define $d^+(x)$ ($d^-(x)$) for $x \in M_1$ as the minimal p-length of l-increasing (l-decreasing) curves joining x with the set $R_{0,1}$ ($R_{-1,0}$). We denote by M_2^+ (M_2^-) the set of points $x \in M_1$ with the following property: for any $\lambda < 1$ there is a $c^+(x, \lambda) > 0$ ($c^-(x, \lambda) > 0$) such that $d^+(T^n x) \geqslant c^+(x, \lambda)\lambda^{-n}$ ($d^-(T^n x) \geqslant c^-(x, \lambda)\lambda^n$) for all $n \leqslant 0$ ($n \geqslant 0$). Roughly speaking, the set M_2^+ (M_2^-) consists of the points whose trajectories in the past (future) approach the boundary sufficiently slowly (slower than an arbitrary exponential).

***Lemma* 2.12.** *Let $x \in M_2^+$ ($x \in M_2^-$). Then*

a) *the p-length of the segments of the LUM $\gamma^u(x)$ (of the LSM $\gamma^s(x)$) from x to its end point is at least $A_0 c^+(x, \lambda_0)$ ($A_0 c^-(x, \lambda_0)$);*

b) $\gamma^u(x) \subset M_2^+$ ($\gamma^s(x) \subset M_2^-$).

Here $\lambda_0 = \Lambda^{-1/m_0}$ and $A_0 = \Lambda_0^{-2}$.

***Lemma* 2.13.** *If $x \in M_2^+$ ($x \in M_2^-$), then the ends of the LUM $\gamma^u(x)$ (of the LSM $\gamma^s(x)$) belong to the set $R_{0,\infty}$ ($R_{-\infty,0}$).*

The proof of the lemmas follows immediately from the construction of the LUM and LSM [22], and we omit it.

We put $M_2 = M_2^+ \cap M_2^-$. Clearly, $T^{-1}M_2^+ \subseteq M_2^+$, $TM_2^- \subseteq M_2^-$, and $TM_2 = T^{-1}M_2 = M_2$. Lemma 2.12b implies that if $x, y \in M_2$ and the point $[x, y]$ is well defined, then $[x, y] \in M_2$ (in this sense M_2 has the structure of a direct product).

It is well known that $\nu(M_2) = 1$ (see, for example, [21]). This follows directly from the Borel–Cantelli lemma and the estimate

$$\nu\,\{x\colon\, d^{\pm}(x) \leqslant \varepsilon\} \leqslant \text{const}\,(Q)\,\varepsilon \tag{2.6}$$

(for billiards with finite horizon this estimate follows from Remark 2.9; in the case of infinite horizon the proof can easily be obtained from the estimate given in §4).

Finally, we note that two LUM (LSM) cannot intersect, but may have end points in common, lying on discontinuity curves in $R_{0,\infty}$ ($R_{-\infty,0}$).

2.5. Regular partitions.

In the sequel we denote by int A and clos A the interior and the closure of a set $A \subset M_1$, respectively, and also put $\mathscr{F}(A) = \text{clos (int } A)$.

Definition. A finite or countable covering ξ (mod 0) of the space M_1 by sets $\Delta_1, \Delta_2, \ldots$ such that:

a) $\nu(M_1 \setminus \bigcup\limits_i \Delta_i) = 0$ and $\nu(\Delta_i \cap \Delta_j) = 0$ for $i \neq j$;

b) each Δ_i is the unon of finitely or countably many closed domains in M_1 and $\Delta_i = \mathscr{F}(\Delta_i)$;

c) each connected component $A \subset \Delta_i$ has piecewise smooth boundary, consisting of finitely many monotone or neutral curves,
is called a *regular partition*.

For each $\Delta_i \in \xi$ we denote by $\partial^u\Delta_i$ ($\partial^s\Delta_i$) the union of all increasing (decreasing) components of $\partial\Delta_i$, and we put $\partial^u\xi = \bigcup_i \partial^u\Delta_i$, $\partial^s\xi = \bigcup_i \partial^s\Delta_i$, $\partial\xi = \partial^u\xi \bigcup \partial^s\xi \bigcup R_0$.

If ξ and ζ are regular partitions, then $\xi \bigvee \zeta$ denotes the regular partition consisting of the sets $\Delta = \mathscr{F}(\Delta' \bigcap \Delta'')$, $\Delta' \in \xi$, $\Delta'' \in \zeta$.

One of the methods for constructing a regular partition with connected elements is the specification of its boundary $\partial\xi$. Namely, a fnite or countable system Γ of monotone curves is called *consistent* if the end points of each curve lie either on two other curves from Γ or on R_0. Then together with R_0 the curves $\gamma \in \Gamma$ partition M_1 into connected components, whose closures form a regular partition.

§3. Construction of a pre-Markov partition

We recall that in §3 we consider billiards with finite horizon. By Remark 1.2, to prove Theorem 1.1 it suffices to construct a Markov partition for $T_1 = T^m$ for some $m \geqslant 1$. In the sequel we put $m_1 = m+m_0$. The construction of the Markov partition will depend on a small parameter $0 < \varepsilon < \varepsilon_0(m)$. The quantities m and $\varepsilon_0(m)$ are chosen during the construction process.

3.1. The initial partition ξ_0.

The first step is the construction of a regular partition of M_1; this can be done rather arbitrarily. We only have to take care that the dimensions of its elements and their positions satisfy certain very weak restrictions. The partition ξ_0 looked for is given by a finite system of consistent curves, forming $\partial\xi_0$ as described in 2.5.

***Proposition* 3.1.** *In M_1 we can choose a finite system of m_1-increasing curves $\Gamma_0^+ = \{\gamma_i^+\}$, $1 \leqslant i \leqslant I_0^+$, and a finite system of m_1-decreasing curves $\Gamma_0^- = \{\gamma_i^-\}$, $1 \leqslant i \leqslant I_0^-$, such that*

a) *the p-lengths of the curves in $\Gamma_0^+ \bigcup \Gamma_0^-$ lie between the bounds $\lambda_1\varepsilon$ and $\lambda_1^{-1}\varepsilon$;*

b) *the curves in Γ_0^+ (Γ_0^-) lie outside $\mathscr{O}^+_{\lambda_1\varepsilon, m_1}$ ($\mathscr{O}^-_{\lambda_1\varepsilon, m_1}$);*

c) (*consistency*) *the end points of each curve in Γ_0^+ (Γ_0^-) lie on two curves in Γ_0^- (Γ_0^+);*

d) *any m_0-increasing (m_0-decreasing) curve of p-length $\lambda_1^{-1}\varepsilon$ intersects at least one curve $\gamma \in \Gamma_0^-$ ($\gamma \in \Gamma_0^+$) in such a way that the point of intersection divides γ into segments of p-lengths at least $\lambda_1\varepsilon$.*

Here $\lambda_1 = \lambda_1(Q) \in (0, 1)$ is a constant not depending on m.

Note that requirement d) implies a definite density of the filling of M_1 by each system of curves Γ_0^+, Γ_0^-.

Proof. We choose a finite $c_1\varepsilon$-net $\{x_i^+\}$, $1 \leqslant i \leqslant \hat{I}_0^+$ ($\{x_i^-\}$, $1 \leqslant i \leqslant \hat{I}_0^-$) in the set $M_1 \setminus \mathcal{O}_{2\varepsilon, m_1}^+$ ($M_1 \setminus \mathcal{O}_{2\varepsilon, m_1}^-$). Here $c_1 = \min\{1, \varkappa_{\min}\}/2$, while the $c_1\varepsilon$-net is chosen in the sense of the metric $ds^2 = dr^2 + d\varphi^2$. Through each point x_i^+ (x_i^-) we draw an arbitrary m_1-increasing (m_1-decreasing) curve $\hat{\gamma}_i^+$ ($\hat{\gamma}_i^-$) which is divided by the point x_i^+ (x_i^-) into two segments of p-length ε. We denote by $\hat{a}_i^\pm$ and $\hat{b}_i^\pm$ the end points of the curves $\hat{\gamma}_i^\pm$ (Fig. 6). We mark off $\hat{\gamma}_i^\pm$ points $\hat{a}_{i,1}^\pm$ and $\hat{b}_{i,1}^\pm$ such that $p(\hat{\gamma}_i^\pm(\hat{a}_i^\pm, \hat{a}_{i,1}^\pm)) = p(\hat{\gamma}_i^\pm(\hat{b}_i^\pm, \hat{b}_{i,1}^\pm)) = \varepsilon/10$. By Remark 2.10, for sufficiently small $\varepsilon_0(m)$ each segment $\hat{\gamma}_i^\pm(\hat{a}_i^\pm, \hat{a}_{i,1}^\pm), \hat{\gamma}_i^\pm(\hat{b}_i^\pm, \hat{b}_{i,1}^\pm)$ intersects at most K_0 discontinuity curves in $R_{-m,0}$ ($R_{0,m}$). Therefore there is a point $a_i^\pm$ on the segment $\hat{\gamma}_i^\pm(\hat{a}_i^\pm, \hat{a}_{i,1}^\pm)$ and a point $b_i^\pm$ on the segment $\hat{\gamma}_i^\pm(\hat{b}_i^\pm, \hat{b}_{i,1}^\pm)$ not belonging to the set $\mathcal{O}_{5c_2\varepsilon, m_1}^+ \cup \mathcal{O}_{5c_2\varepsilon, m_1}^-$, where $c_1 = (100K_0)^{-1}$. We make an additional construction: through a_i^+ on $\hat{\gamma}_i^+$ we draw an m_1-decreasing curve of p-length $c_2\varepsilon$ with end points at a_i^+ and at a certain point $a_{i,1}^+$. Further, through $a_{i,1}^+$ we draw an m_1-increasing curve of p-length $c_2\varepsilon$ with end points at $a_{i,1}^+$ and at a certain point $a_{i,2}^+$. Finally, through $a_{i,2}^+$ we draw an m_1-decreasing curve up to the intersection with $\hat{\gamma}_i^+$ at a certain point $a_{i,3}^+$ lying on the segment $\hat{\gamma}_i^+(a_i^+, x_i^+)$ (Fig. 7). For sufficiently small $\varepsilon_0(m)$ this construction can be done in such a way that the p-lengths of the curve constructed and of the curve $\hat{\gamma}_i^+(a_i^+, a_{i,3}^+)$ lie between the bounds $c_2\varepsilon/2$ and $2c_2\varepsilon$. A similar "loop" of three additional curves is constructed around the point b_i^+ (Fig. 7). These "loops" are also constructed around the points a_i^-, b_i^- on the curves $\hat{\gamma}_i^-$, $1 \leqslant i \leqslant \hat{I}_0^-$ (in this case the monotonicity of each curve is replaced by the opposite). We put $\gamma_i^\pm = \hat{\gamma}_i^\pm(a_i^\pm, b_i^\pm)$, $1 \leqslant i \leqslant \hat{I}_0^\pm$.

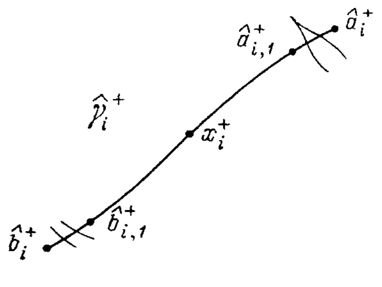


Fig. 6

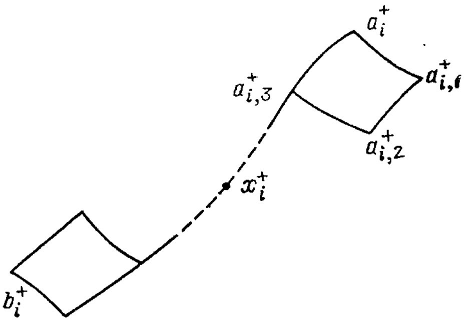


Fig. 7

The system $\Gamma_0^\pm$ looked for consists of the curves $\gamma_i^\pm$, $1 \leqslant i \leqslant \hat{I}_0^\pm$, and of all increasing (decreasing) curves occurring in the above constructon of the loops in the neighbourhoods of the end points of the curves $\gamma_i^\pm$. It is easy to compute that $I_0^\pm = 7\hat{I}_0^\pm$.

Assertions a) — c) in Proposition 3.1 can be verified immediately. We prove assertion d). We assume that $\lambda_1^{-1} \geqslant 10K_0$. By Remark 2.10, for sufficiently small $\varepsilon_0(m)$ any m_0-increasing curve γ of p-length $\lambda_1^{-1}\varepsilon$ intersects at most K_0 discontinuity curves in $R_{-m_1,0}$. Therefore there is a segment $\tilde{\gamma}$ of p-length 5ε on this curve, not intersecting $R_{-m_1,0}$. We choose a point $x_0 \in \tilde{\gamma}$ not in $\mathcal{O}^+_{2\varepsilon,m_1}$. Then there is a point x_i^-, $1 \leqslant i \leqslant \hat{I}_0^-$, with distance to x_0 at most $c_1\varepsilon$. It is not difficult to verify that the curves γ_i^- and $\tilde{\gamma}$ intersect, and that the point of intersection divides the first curve into two segments of p-lengths not less than $\lambda_1\varepsilon$. The case of an m_0-decreasing curve γ can be treated similarly. The propostion has been proved.

Note that we can put $\lambda_1 = (200K_0)^{-1}$.

3.2. The pre-Markov partition ξ.

The next step consists in replacing the smooth curves in $\Gamma_0^\pm$ by segments of the LUM and LSM near to them. The system of LUM and LSM obtained must be consistent with the dynamics T^m; namely, the image of each LUM (LSM) under the action of T^{-m} (T^m) must fall inside some other LUM (LSM) in this system.

We give precise definitions. Let ξ be a regular partition all boundary components of which are discontinuity curves or segments of LUM and LSM.

Definition. The partion ξ is called *pre-Markov* for T^m if

$$(3.1) \qquad T^m(\partial^s\xi) \subseteq \partial\xi \quad \text{and} \quad T^{-m}(\partial^u\xi) \subseteq \partial\xi.$$

If this definition is formally generalized to smooth hyperbolic systems, then (since there are no discontinuity curves) relation (3.1) gives a Markov partition ξ. Hence a pre-Markov partition ξ for a billiard may be constructed similarly to the construction of a Markov partition for two-dimensional smooth systems [2], [9].

First of all it is necessary to "provide" a sufficiently large coefficient of expansion (contraction) for the automorphism $T_1 = T^m$ (for this reason we make the transition to T_1). We choose m so large that

$$(3.2) \qquad \Lambda_m \geqslant (\hat{c}\lambda_1^2)^{-1} > 1$$

for some $\hat{c} < 1/2$ whose value will be given below.

We now turn to the construction. We consider an arbitrary curve $\gamma_0 \in \Gamma_0^-$. Its end points a_1, a_2 lie on certain curves $\gamma_1, \gamma_2 \in \Gamma_0^+$. By Proposition 3.1b) the curves γ_r ($r = 1, 2$) can be extended on both sides from a_r over p-distance $\lambda_1\varepsilon$ while preserving the m_1-increase property. We mark off two points a_1', a_2' on γ_1, γ_2, respectively, lying on one side of the curve γ_0 and such that $p(\gamma_r(a_r, a_r')) = \hat{c}\lambda_1\varepsilon$, $r = 1, 2$ (Fig. 8a). It is easily seen that we can either draw from a_1' an m_1-decreasing curve intersecting $\gamma_2(a_2, a_2')$, or, conversely, we can draw from a_2' an m_1-decreasing curve up to the intersection with $\gamma_1(a_1, a_1')$. Without loss of generality we assume the first can be done. Then we can also draw from any point of $\gamma_1(a_1, a_1')$ an m_1-decreasing curve up to the intersection with $\gamma_2(a_2, a_2')$; moreover, the p-length of this curve will not exceed $2\lambda_1\varepsilon$.

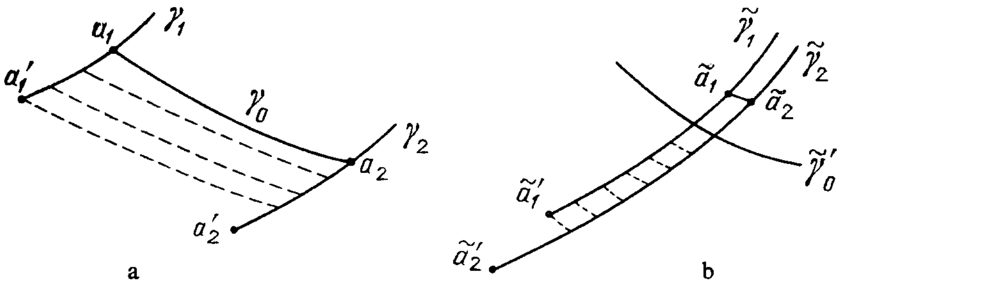


Fig. 8

We consider the curves $\widetilde{\gamma}_r = T_1\gamma_r$, $r = 0, 1, 2$, and the points $\tilde{a}_r = T_1 a_r$, $\tilde{a}_r' = T_1 a_r'$, for $r = 1, 2$ (Fig. 8b). By (3.2), $p\,(\tilde{\gamma}_0\,(\tilde{a}_1, \tilde{a}_2)) \leqslant \hat{c}\lambda_1^3\varepsilon$ and $p\,(\tilde{\gamma}_r\,(\tilde{a}_r, \tilde{\boldsymbol{a}}_r')) \geqslant \lambda_1^{-1}\varepsilon$. Moreover, through each point of $\tilde{\gamma}_1\,(\tilde{a}_1, \tilde{a}_1')$ we can draw an m_0-decreasing curve of p-length not exceeding $2\hat{c}\lambda_1^3\varepsilon$ up to the intersection with $\tilde{\gamma}_2\,(\tilde{a}_2, \tilde{a}_2')$ (this is guaranteed by the nearness of $\tilde{\gamma}_1\,(\tilde{a}_1, \tilde{a}_1')$ and $\tilde{\gamma}_2\,(\tilde{a}_2, \tilde{a}_2')$). By Proposition 3.1d) there is a curve $\tilde{\gamma}_0' \in \Gamma_0^-$ intersecting both segments $\tilde{\gamma}_r\,(\tilde{a}_r, \tilde{a}_r')$, $r = 1, 2$. We also assume that among all these curves $\tilde{\gamma}_0'$ is nearest to the points $\tilde{a}_1, \tilde{a}_2$. Let $\tilde{\gamma}_0''$ denote the segment on $\tilde{\gamma}_0'$ contained between the curves $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$. Then $\hat{\gamma}_0 = T_1^{-1}\tilde{\gamma}_0''$ is a $2m$-decreasing curve whose end points lie on $\gamma_1\,(a_1, a_1')$ and $\gamma_2\,(a_2, a_2')$.

Such a curve $\hat{\gamma}_i$ can be constructed for each curve $\gamma_i \in \Gamma_0^-$. The set of curves $\{\hat{\gamma}_i\}$ will be denoted by Γ_1^-. Next, each curve $\gamma_i^+ \in \Gamma_0^+$ is elongated or shortened from each of its end points so that it ends on the curves $\hat{\gamma}_{i_1}, \hat{\gamma}_{i_2} \in \Gamma_1^-$ if earlier it ended on $\gamma_{i_1}, \gamma_{i_2} \in \Gamma_0^-$ (when elongated, the m_1-increase property is preserved; this is guaranteed by conditon b) in Proposition 3.1).

A similar procedure (with T_1 replaced by T_1^{-1}) is performed for each curve $\gamma_i \in \Gamma_0^+$, giving a new system of curves Γ_1^+ consisting of $2m$-increasing curves. The systems $\Gamma_1^\pm$ satisfy the relation

$$\forall\tilde{\gamma} \in \Gamma_1^\pm \;\; \exists\gamma \in \Gamma_0^\pm\colon \; T_1^{\mp 1}\tilde{\gamma} \subset \gamma$$

(here and everywhere in §3 the inclusion "$\subset$" means that one curve lies strictly in the interior of the other). Between the curves $\tilde{\gamma} \in \Gamma_1^\pm$ and $\gamma \in \Gamma_0^\pm$ there is a natural relation, under which corresponding curves remain at a distance at most $\hat{c}\lambda_1\varepsilon$ from each other.

To the systems of curves $\Gamma_1^\pm$ we again apply the above described procedure, as a result of which each curve $\gamma \in \Gamma_1^\pm$ is replaced by a curve $\hat{\gamma}$ near to it. The system of curves $\{\hat{\gamma}\}$ obtained by this step is denoted by $\Gamma_2^\pm$. Recurrently repeating this procedure gives a sequence of systems of curves $\Gamma_n^\pm$, $n = 1, 2, \ldots$.

***Lemma* 3.2.** *The curves $\gamma \in \Gamma_n^+$ ($\gamma \in \Gamma_n^-$) are $(n+1)m$-increasing ($(n+1)m$-decreasing). They are subject to the condition*

$$\forall\hat{\gamma} \in \Gamma_n^{\pm}\ \exists\gamma \in \Gamma_{n-1}^{\pm}:\ T_1^{\mp1}\hat{\gamma} \subset \gamma.$$

Between the curves $\hat{\gamma} \in \Gamma_n^{\pm}$ and $\gamma \in \Gamma_{n-1}^{\pm}$ there is a natural relation, under which corresponding curves remain at a distance at most $(\bar{c}\lambda_1)^n\varepsilon$ from each other.

The lemma can be proved by induction.

***Lemma* 3.3.** *The systems of curves $\Gamma_n^{\pm}$ have limits as $n \to \infty$ (in the metric of C^0). We denote these by $\Gamma_\infty^{\pm}$. The limit system Γ_∞^+ (Γ_∞^-) consists of the segments of LUM (LSM) satisfying the conditions*

a) $\forall\gamma \in \Gamma_\infty^{\pm}\ \exists\gamma' \in \Gamma_\infty^{\pm}:\ T_1^{\mp1}\,\gamma \subset \gamma'$;

b) *between the curves $\hat{\gamma} \in \Gamma_\infty^{\pm}$ and $\gamma \in \Gamma_0^{\pm}$ there is a natural correspondence, under which corresponding curves remain at a distance of at most $2\bar{c}\lambda_1\varepsilon$ from each other*;

c) (*consistency*) *the end points of each curve $\gamma \in \Gamma_\infty^{\pm}$ lie on two curves in $\Gamma_\infty^{\pm}$.*

Proof. The existence of limit systems $\Gamma_\infty^{\pm}$ and assertions a), b), c) can be derived from Lemma 3.2 by a simple limit transition. By Lemma 2.11 the curves $\gamma \in \Gamma_\infty^{\pm}$ are LUM (LSM).

Having chosen the value of $\bar{c}$ sufficiently small ($\bar{c} \leqslant \lambda_1/200$), we can use the idea of the proof of Proposition 3.1d) for proving the following lemma.

***Lemma* 3.4.** *Any m_0-increasing (m_0-decreasing) curve of p-length $\lambda_1^{-1}\varepsilon$ intersects some segment of a LSM in Γ_∞^- (LUM in Γ_∞^+); moreover, the point of intersection divides this segment into two pieces of p-lengths not less than $\lambda_1\varepsilon/2$.*

Finally, the partition ξ of M_1 is given by the systems of curves Γ_∞^+, Γ_∞^-, and $R_{-m,m}$, which generate $\partial\xi$.

Proposition. *The partition ξ is finite and pre-Markov for T_1.*

Finiteness of ξ follows from the fact that any two smooth components of $R_{-m,m}$, Γ_∞^-, and Γ_∞^+ intersect in at most two points.

3.3. The modified partition ξ_1.

The elements of the above constructed pre-Markov partition ξ are of very complicated shape (in particular those bordering with $R_{-m,m}$). We will construct a pre-Markov partition with elements of a much simpler shape.

We consider the system of curves $\Gamma_{(1)}^{\pm}$ consisting of the curves $\gamma \in \Gamma_\infty^{\pm}$ and the smooth components of their images $T_1^{\pm1}\gamma$. We define the regular partition ξ_1 by giving its boundary using the systems $\Gamma_{(1)}^{\pm}$ and the discontinuity curves $R_{-m,m}$.

***Proposition* 3.5.** *The partition ξ_1 is finite and pre-Markov for T_1. Each curve $\gamma \in \Gamma_{(1)}^{\pm}$ ends on $R_{-m,m}$ or strictly inside certain curves in $\Gamma_{(1)}^{\mp}$.*

Proof. ξ being pre-Markov implies that ξ_1 is pre-Markov. Finiteness is verified as for ξ. The last assertion of Lemma 3.5 follows from relation a) in Lemma 3.3 (recall our convention on the use of the inclusion symbol "$\subset$"!).

The following lemma provides a description of the geometric shapes of the elements of ξ_1.

***Lemma* 3.6.** a) *If the element $\Delta \in \xi_1$ does not border with $R_{-m,m}$, then it is a curvilinear quadrangle, bounded by two LUM and two LSM alternating each other* (Fig. 9a).

b) *If the element $\Delta \in \xi_1$ does border with $R_{-m,m}$, then it is a curvilinear polygon, bounded by segments of LUM, LSM, and discontinuity curves such that all interior angles do not exceed* 180°.

The proof of the lemma involves simple geometrical constructions and the use of Lemma 2.8. It is easily computed that the number of sides of a polygon $\Delta \in \xi_1$ lies between 3 and 6 (Fig. 9b, c).

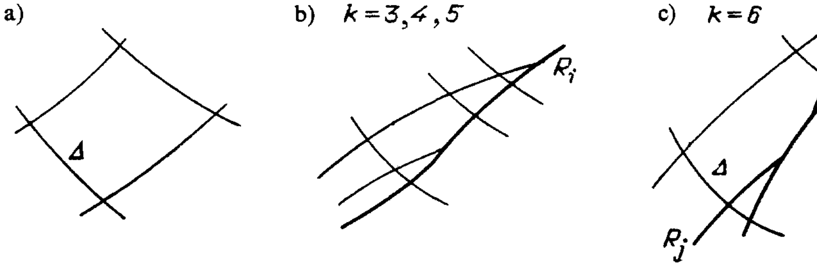


Fig. 9

***Corollary* 3.7.** *The diameters of the elements of the pre-Markov partition ξ_1 do not exceed* $\text{const}(Q)\sqrt{\varepsilon}$.

Proof. We consider an arbitrary monotone curve γ inside an element $\Delta \in \xi_1$. By Lemma 3.4 and relation (2.4), $l(\gamma) \leqslant \text{const}(Q)\sqrt{\varepsilon}$. This and Lemma 3.6 implies the required result.

3.4. In conclusion we consider the partition $\xi_n = T_1^{-n+1}\xi_1 \vee \ldots \vee \xi_1 \vee \vee T_1\xi_1 \vee \ldots \vee T_1^{n-1}\xi_1$. By induction with respect to n it is easily proved that ξ_n is, for any n, a pre-Markov partition. This and Lemma 3.6 imply that all elements of ξ_n are connected and are curvilinear polygons whose interior angles do not exceed 180°. The hyperbolic properties and Lemma 3.4 imply that the diameters of the elements $\Delta \in \xi_n$ tend to zero as n grows. Hence $\lim \xi_n = \varepsilon$ (here ε denotes the partition into individual points).

§4. The case of infinite horizon

At first reading this section may be omitted without loss of understanding the sequel.

4.1. Billiards with infinite horizon are possible only on the torus Tor^2. If the function $\tau^+(x)$ ($\tau^-(x)$) is unbounded in a neighbourhood of a point $z_0 \in M_1$, then:

a) its semitrajectory $\{S^t z_0\}$ for $t > 0$ ($t < 0$) forms a closed periodic winding of the torus;

b) at all points of contact of this semitrajectory and the boundary ∂Q, the corresponding component of ∂Q lies at one side of this semitrajectory (Fig. 10).

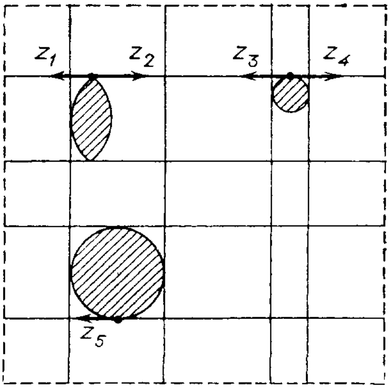


Fig. 10

Such points will be called *u-singular* (*s-singular*). In Fig. 10 the points z_1, z_2 are s-singular but not u-singular, while z_3, z_4 are simultaneously u- and s-singular. The discussion above implies that M_1 contains only finitely many u- or s-singular points, and also that they all lie on R_0. Note that several singular points can have a common semitrajectory (the points z_1 and z_3, and z_2 and z_4 in Fig. 10).

We distinguish three types of singular points z:

1) type S for $z \in S_0 \backslash V_0$;

2) type V for $z \in V_0 \backslash S_0$; and

3) type SV for $z \in S_0 \cap V_0$.

These types will be called *generic*. For each u-singular (s-singular) point z_0 we denote by $Z^u(z_0)$ ($Z^s(z_0)$) the set of s-singular (u-singular) points $\{z\}$ whose supports $\pi(z)$ do not lie on the semitrajectory $\{S^t z_0\}$ for $t < 0$ ($t > 0$) and such that $TV(z) \cap V(z_0) \neq \varnothing$ (respectively, $T^{-1}V(z) \cap V(z_0) \neq \varnothing$), where $V(z)$ is a sufficiently small neighbourhood of the singular point z in M_1. For example, in Fig. 10, $Z^u(z_5) = \{z_1, z_3\}$. For each u-singular (s-singular) point z we introduce the *u-type* (*s-type*). It takes one of the six values: S_{pure}, V_{pure}, SV_{pure}, S_{mix}, V_{mix}, SV_{mix}, where S, V, and SV are the possible generic types of z and the subscripts "pure" and "mix" mean that the set $Z^u(z)$ ($Z^s(z)$) consists of points of the "pure" types (S or V), or contains at least one point of the "mixed" type SV, respectively.

In neighbourhoods of u-singular (s-singular) points there are infinitely (countably) many smooth components of the set R_1 (R_{-1}). The smooth components of R_1 (R_{-1}) partition these neighbourhoods into countably many subdomains, called *u-cells* (*s-cells*) in the sequel. The structure of the cells in a neighbourhood of a singular point is determined by the type of the point and is fairly universal. The positions and dimensions of u-cells of four types are depicted in Fig. 11. Here the notation $O(1/n^\alpha)$ stands for a quantity between const_1/n^α and const_2/n^α; the values of the constants depend on Q (the cells are given a natural enumeration, by the order of approach to the singular point). The corresponding drawings for u-cells of the types SV_{pure} and SV_{mix} can be obtained by cutting the drawings for the types S_{pure} and S_{mix} by a vertical line through the singular point (this line corresponds to the component V_0). The positions and dimensions of s-cells are similar, up to mirror symmetry.

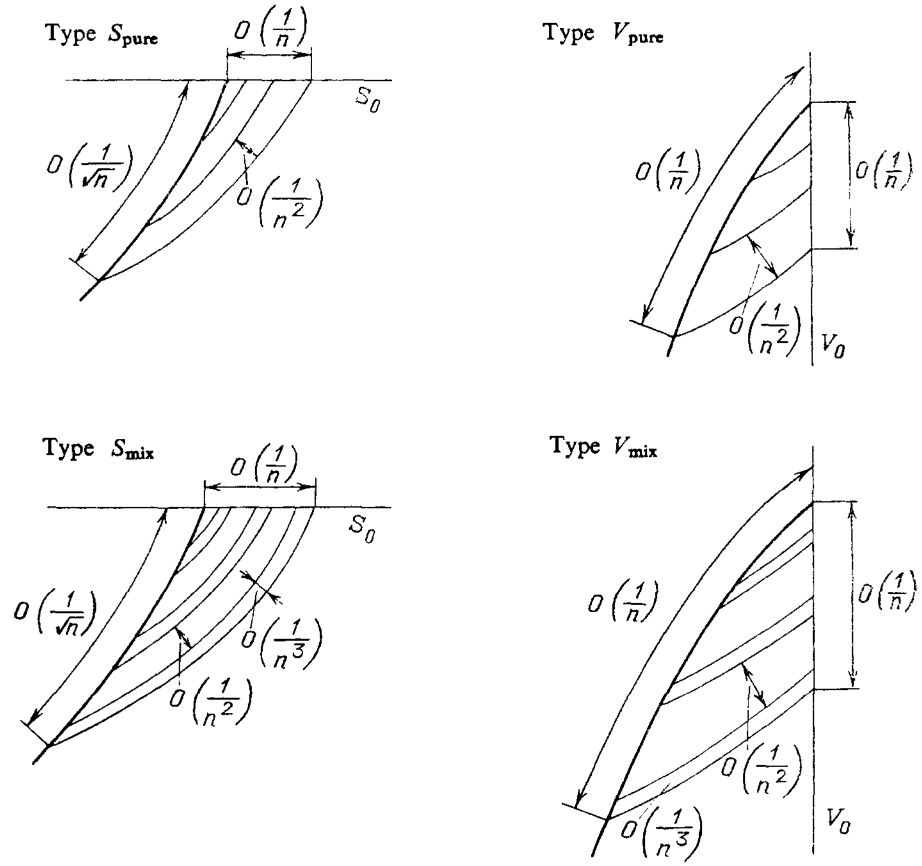

Type S_{pure}
$O\left(\frac{1}{n}\right)$
S_0
$O\left(\frac{1}{\sqrt{n}}\right)$
$O\left(\frac{1}{n^2}\right)$
Type V_{pure}
$O\left(\frac{1}{n}\right)$
$O\left(\frac{1}{n}\right)$
$O\left(\frac{1}{n^2}\right)$
V_0
Type S_{mix}
$O\left(\frac{1}{n}\right)$
S_0
$O\left(\frac{1}{\sqrt{n}}\right)$
$O\left(\frac{1}{n^3}\right)$
$O\left(\frac{1}{n^2}\right)$
Type V_{mix}
$O\left(\frac{1}{n}\right)$
$O\left(\frac{1}{n}\right)$
$O\left(\frac{1}{n^2}\right)$
$O\left(\frac{1}{n^3}\right)$
V_0

Fig. 11

Note that condition B in Theorem 1.1 is not violated in the case of infinite horizon, since through each singular point precisely two components of $R_{0,1}$ $(R_{-1,0})$ pass.

The definition of cells implies the following.

***Lemma* 4.1.** *In a neighbourhood of an s-singular (u-singular) point any u-cell (s-cell) with index n is transformed under the action of* T^{-1} (T) *into an s-cell (u-cell) with index between* $\text{const}_1\, n$ *and* $\text{const}_2\, n$ *(the type of the cell may change).*

4.2. The Lorentz gas.

We consider the special case of a billiard with infinite horizon when all singular points are of type S_{pure} (that is, lie at regular points of ∂Q). As far as is known to us, in all papers on the study of ergodic properties of scattering billiards the authors restrict themselves to this case [6], [21], [25]. The Lorentz model, which is very popular in statistical mechanics, belongs to this class [6].

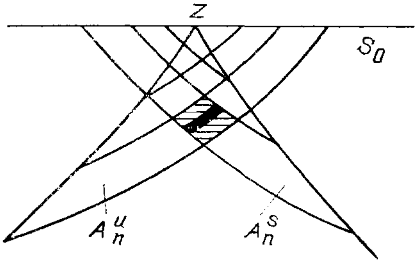


Fig. 12

In the case considered, all u-singular points are s-singular, and conversely. Hence in a neighbourhood of each singular point z there is a sequence of u-cells $A_n^u(z)$ and s-cells $A_u^s(z)$ (Fig. 12). The cell $A_n^u(z)$ $(A_u^s(z))$ intersects the cells $A_i^s(z)$ $(A_i^u(z))$ for all $\text{const}_1\sqrt{n} \leqslant i \leqslant \text{const}_2\, n^2$. This and Lemma 4.1 imply that for all $n \geqslant$ const the s-cell $A_n^s(z)$ intersects the u-cell $TA_n^s(z')$ for some singular point $z' \neq z$, such that their intersection forms a quadrangle, bounded by the long sides of the cells $A_n^s(z)$ and $TA_n^s(z')$ (see the shaded domain in Fig. 12). The number of singular points is finite, hence for any $k \geqslant 1$ there is a (unique) sequence $z_1, \ldots, z_k$ of singular points such that the set $\Delta_{n,k}^s(z) = A_n^s(z) \cap TA_n^s(z_1) \cap \ldots \cap T^k A_n^s(z_k)$ is not empty and forms a small strip joining the two long sides of the cell $A_n^s(z)$ (the blackened part in Fig. 12). The limit of these strips as $k \to \infty$ is a LUM joining the two long sides of $A_n^s(z)$ (see Lemma 2.11). We denote this LUM by $\gamma_n^u(z)$. In a similar manner we can construct the LSM $\gamma_n^s(z)$ joining the two long sides of the u-cell $A_n^u(z)$. This system of LUM (LSM) is semi-invariant:

$$(4.1)\qquad T^{-1}\gamma_n^u(z) \subset \gamma_n^u(z')\,(T\gamma_n^s(z) \subset \gamma_n^s(z'))$$

for some singular point $z' \neq z$. Relation (4.1) follows since the images $T^{-l}\gamma_n^u(z)$ $(T^l\gamma_n^s(z))$ lie, for all $l \geqslant 0$, in the s-cells (u-cells) with fixed index n.

We now turn to the construction of a pre-Markov partition for this case. Let $V(\varepsilon_*)$ be the union of the ε_*-neighbourhoods of all singular points $\{z\}$. The quantity ε_* is chosen so small that under the action of T (T^{-1}) the coefficient of expansion of any increasing (decreasing) curve $\gamma \subset V(\varepsilon_*)$ is at least some Λ_*. By using Lemma 2.2 it is easily proved that such an $\varepsilon_* > 0$ exists for any $\Lambda_* > 1$. It remains to choose Λ_*. This will be done later.

For any $m \geqslant 1$ the set $M_1\backslash V(\varepsilon_*)$ contains only finitely many smooth components of the set $R_{-m,m}$. Hence for sufficiently small $\varepsilon < \varepsilon_0(m, \varepsilon_*)$ we can construct in it a system of curves $\Gamma_0^\pm$ satisfying all the conditions of Proposition 3.1 except d), for the same value of λ_1. Condition d) is fulfilled only for curves lying entirely in $M_1\backslash V(\varepsilon_*)$. At this stage the quantity m, as well as ε_*, is not fixed yet.

We extend the construction of the curves $\Gamma_0^\pm$ to the domains $V(\varepsilon_*)$. We put $N_* = (\tilde{c}\varepsilon)^{1/2}$, where $\tilde{c}$ is a small constant not depending on Λ_* and ε_*; its value is chosen below. We take the above constructed LUM $\gamma_n^u(z)$ and LSM $\gamma_n^s(z)$ for all const $\leqslant n \leqslant N_*$ and all z, and add to them their images $T\gamma_n^u(z)$ and $T^{-1}\gamma_n^s(z)$. The system of LUM (LSM) thus obtained is denoted by Γ_*^+ (Γ_*^-). It has the following properties:

a) semi-invariance: for each $\gamma \in \Gamma_*^\pm$ there is a $\gamma_1 \in \Gamma_*^\pm$ such that $T^{\mp 1}\gamma \subset \gamma_1$;

b) for all z and const $\leqslant n \leqslant$ const N_* there is in the u-cell $A_n^u(z)$ (s-cell $A_n^s(z)$) a $\gamma \in \Gamma_*^+$ (a $\gamma \in \Gamma_*^-$) joining the two short (!) sides of $A_n^u(z)$ $(A_n^s(z))$.

The LUM and LSM constructed are called *supporting*.

In the set $V(\varepsilon_*)\backslash\mathcal{O}_{2\varepsilon,1}^+$ $(V(\varepsilon_*)\backslash\mathcal{O}_{2\varepsilon,1}^-)$ we choose a finite $c_1\varepsilon$-net $\{x_i^+\}$ $(\{x_i^-\})$, where $c_1 = \min\{1, \varkappa_{\min}\}/2$. As in the proof of Proposition 3.1, through each point x_i^+ (x_i^-) we draw a 1-increasing (1-decreasing) curve $\hat{\gamma}_i^+$ $(\hat{\gamma}_i^-)$, divided by the point x_i^+ (x_i^-) into two segments of p-length ε. At the ends of $\hat{\gamma}_i^\pm$ we mark off the points $\hat{a}_i^\pm$, $\hat{b}_i^\pm$, $\hat{a}_{i,1}^\pm$, $\hat{b}_{i,1}^\pm$, as in the proof of Proposition 3.1. If the segment $\hat{\gamma}_i^\pm(\hat{a}_i^\pm, \hat{a}_{i,1}^\pm)$ intersects some supporting LSM (LUM) γ at a point dividing γ into two segments of p-lengths at least $\tilde{c}\,\varepsilon$, then we denote this point of intersection by $a_i^\pm$. If, however, this LSM (LUM) has not been found, then, as is clear from a careful analysis of the structure of the cells, by the choice of N_* and the smallness of $\tilde{c}$ the segment $\hat{\gamma}_i^\pm(\hat{a}_i^\pm, \hat{a}_{i,1}^\pm)$ intersects at most K_1 components of the set R_{-1} (R_1), where $K_1 = K_1(\tilde{c}) < \infty$ is a constant. In this case we construct in a neighbourhood of $\hat{\gamma}_i^\pm(\hat{a}_i^\pm, \hat{a}_{i,1}^\pm)$ a "loop" of three additional curves, in the same manner as in the proof of Proposition 3.1. Similar constructions are carried out for $\hat{\gamma}_i^\pm(\hat{b}_i^\pm, \hat{b}_{i,1}^\pm)$.

We fix the value of $\tilde{c}$ such that $K(\tilde{c})$ is finite. The result of the construction is summarized in the following lemma.

***Lemma* 4.2.** *In the set $V(\varepsilon_*)$ we can find a finite system of* 1*-increasing* (1*-decreasing*) *curves Γ_{00}^{+} (Γ_{00}^{-}) such that*:

a) *the p-lengths of all curves in $\Gamma_{00}^{\pm}$ lie between $\lambda_2\varepsilon$ and $\lambda_2^{-1}\varepsilon$;*

b) *the curves in Γ_{00}^{+} (Γ_{00}^{-}) lie outside $\mathcal{O}_{\lambda_2\varepsilon,1}^{+}$ ($\mathcal{O}_{\lambda_2\varepsilon,1}^{-}$);*

c) *the end-points of each curve in Γ_{00}^{+} (Γ_{00}) lie on two curves in $\Gamma_{00}^{-} \cup \Gamma_*^{-}$ ($\Gamma_{00}^{+} \cup \Gamma_*^{+}$); moreover, the curves in Γ_*^{-} (Γ_*^{+}) divide these ends into two segments of p-lengths at least $\lambda_2\varepsilon$;*

d) *any* 1*-increasing* (1*-decreasing*) *curve $\gamma_0 \subset V(\varepsilon_*)$ of p-length $\lambda_2^{-1}\varepsilon$ intersects at least one of the curves $\gamma \in \Gamma_{00}^{-} \cup \Gamma_*^{-}$ ($\gamma \in \Gamma_{00}^{+} \cup \Gamma_*^{+}$) in such a way that the point of intersection divides γ into two segments of p-lengths at least $\lambda_2\varepsilon$.*

Here $\lambda_2 \in (0, 1)$ is a constant determined by $K_1 = K_1(\tilde{c})$.

Assertion d) is proved as in Proposition 3.1. Note that $\tilde{c}$, K_1, λ_2 do not depend on ε_* and Λ_*. We put $\lambda_3 = \min\{\lambda_1, \lambda_2\}$. Finally we fix m and Λ_* (and hence ε_*) so that $\min\{\Lambda_m, \Lambda_*\} \geqslant (\hat{c}\lambda_3^2)^{-1}$, where $\hat{c} = \lambda_3/200$ (compare with 3.2). Now our construction is completely defined.

We put $\tilde{\Gamma}_0^{\mp} = \Gamma_0^{\mp} \cup \Gamma_{00}^{\pm} \cup \Gamma_*^{\pm}$. We replace each supporting LUM $\gamma \in \Gamma_*^{+}$ (LSM $\gamma \in \Gamma_*^{-}$) by the maximal subsegment $\tilde{\gamma} \subset \gamma$ ending at two curves in $\tilde{\Gamma}_0^{-}$ ($\tilde{\Gamma}_0^{+}$). Then the system $\tilde{\Gamma}_0^{\mp}$ becomes consistent (see 2.5) and together with the component of $R_{-m,m}$ it generates an initial partition ξ_0 for the class of billiards considered.

In constructing the pre-Markov partition ξ we need a differentiated approach to the various components of the boundary $\partial\xi_0$. We apply the transformation $T^{\mp m}$ to the curves $\gamma \in \Gamma_0^{\pm}$, as in 3.2. We apply the transformation $T^{\mp 1}$ to the curves $\gamma \in \Gamma_{00}^{\pm}$. By the choice of Λ_*, all constructions of 3.2 are applicable to these curves. The supporting LSM and LUM lead, without changes, to the boundary of a pre-Markov partition, up to the choice of their end points as described above.

As a result we obtain a consistent system of LUM and LSM, generating together with $R_{-m,m}$ a pre-Markov partition ξ. Moreover, the statement of Lemma 3.4 holds for it, after replacing λ_1 by λ_3.

The construction of the pre-Markov partition ξ_1 of 3.3 and of the partitions ξ_n of 3.4 can be transferred without changes to the case under consideration. We only note that the partition ξ_1 is countable, but for any $\delta > 0$ the set $M_1 \backslash V(\delta)$ contains finitely many elements of it. Moreover, finitely many LUM and LSM are included in $\partial\xi$.

4.3. In a number of cases, similar to the one described above, the u-cells and s-cells cover each other in a neighbourhood of a singular point. This is possible for points of u- and s-types S_{mix}, SV_{mix}, SV_{pure}. In all these cases a pre-Markov partition can be constructed by means of the scheme given in 4.2, with minor changes. The only difference can lie in a somewhat more accurate construction of supporting LUM and LSM. The detailed analysis of these cases does not involve new ideas, and we omit it. We only stress that our construction is based on two properties of the system: strong hyperbolicity in neighbourhoods of singular points (the coefficient of expansion and contraction during one step can become arbitrarily large!) and the covering of the u-cells on the s-cells, which allows the construction of supporting LUM and LSM.

4.4. "Wandering" cells.

There are cells of altogether different types: V_{pure} and V_{mix}. In neighbourhoods of such points the u- and s-cells are not covered by one another. The images of u-singular (s-singular) points of these types under the action of T^m (T^{-m}) for $m \geqslant 1$ are not singular any more—they "wander" somewhat inside M_1. In neighbourhoods of them infinitely many components of R_{m+1} (R_{-m-1}) accumulate (Fig. 13), which complicates the construction.

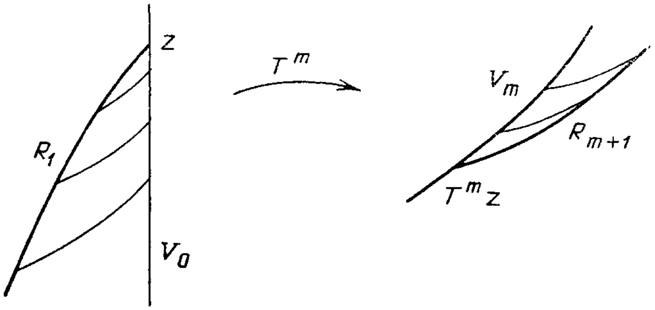


Fig. 13

The construction of supporting LUM and LSM in cells of these types is not successful. Instead we construct for each of them a special chain of increasing and decreasing curves.

First of all we fix m as in §3, and put $m_1 = m+m_0$. Then there is an $n_0 = n_0(m)$ such that in all u-cells $A_n^u(z)$ (s-cells $A_n^s(z)$) of the type considered and with index $n \geqslant n_0$ the transformations $T^{\pm 2m_1}$ are continuous. Moreover, let $\varepsilon > 0$ be a small parameter: $\varepsilon < \varepsilon_0(m)$, as in §3.

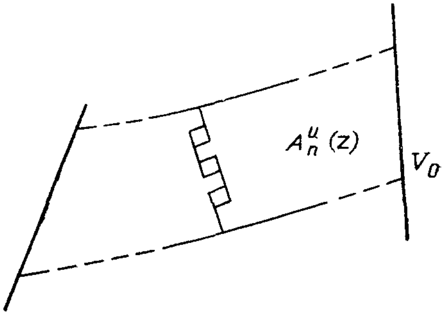


Fig. 14

Let $A_n^u(z)$ be an arbitrary u-cell with index $n_0 \leqslant n \leqslant (\tilde{c}\,\varepsilon)^{-1}$, where $\tilde{c}$ is a sufficiently small constant, chosen below. We construct a chain of $2m_1$-increasing and $2m_1$-decreasing curves in $A_n^u(z)$ as indicated in Fig. 14, such that the following conditions hold:

a) this chain joins the two long sides of the cell $A_n^u(z)$;

b) all end points of increasing (decreasing) curves of this chain (except the two boundary points lying on $\partial A_n^u(z)$) lie on decreasing (increasing) curves of this chain;

c) the p-lengths of all curves of the chain do not exceed $\lambda_1^{-1}\varepsilon$; if the chain consists of more than one decreasing curve, then the p-lengths of all curves must exceed $\lambda_1\varepsilon$;

d) any m_0-increasing curve joining the two short sides of $A_n^u(z)$ intersects at least one decreasing curve γ of the chain such that if γ does not end on $\partial A_n^u(z)$, then the point of intersection divides γ into two segments of p-lengths at least $\lambda_1\varepsilon$.

Here λ_1 is the same as in §3.

Similar chains are constructed in all s-cells $A_n^s(z)$, $n_0 \leqslant n \leqslant (\tilde{c}\,\varepsilon)^{-1}$. (Here, in the statement of conditions a) – d) the monotonicity of each curve must be replaced by the opposite.) The curves of the chains constructed as well as all their images under the action of T^j, $|j| \leqslant m$, are also called *supporting*. A consequence of our constructions is the following.

Lemma 4.3. *Let $\Lambda > 0$, and let $\tilde{c} = \tilde{c}(\Lambda) > 0$ and $\varepsilon < \varepsilon_0(m, \Lambda)$ be sufficiently small. Then any m_0-increasing (m_0-decreasing) curve of p-length less than $\Lambda\varepsilon$ and intersecting more than K_0+3 components of $R_{-m,0}$ ($R_{0,m}$) intersects at least one decreasing (increasing) supporting curve γ in such a way that the point of intersection divides γ into two segments of p-lengths at least $\lambda_1\varepsilon$.*

The subsequent construction of an initial partition is done by analogy with that described in 3.1: one constructs a finite set of m_1-increasing (m_1-decreasing) curves $\Gamma_0^{\pm}$ which together with the supporting curves and the components of $R_{-m,m}$ form a consistent system satisfying conditions a) – d) of Proposition 3.1. In the proof of c) and d) Lemma 4.3 is used, in which $\Lambda = \lambda_1^{-1}$ (and thus $\tilde{c} = \tilde{c}(\lambda_1^{-1})$ is fixed).

The construction of a pre-Markov partion ξ is done as in 3.2, with one modification: the transformations T^m (T^{-m}) are applied not to each supporting decreasing (increasing) curve, but only to those lying in $T^jA_n^u(z)$ and $T^jA_n^s(z)$ for $0 \leqslant j \leqslant m$ ($-m \leqslant j \leqslant 0$). The lengths of such curves do not exceed $\lambda_1^{-1}\varepsilon$. The remaining supporting curves are defined as the images of those indicated above under the action of T^j, $|j| \leqslant m$. Moreover, as in 4.2 it is necessary to replace increasing (decreasing) supporting curves by their maximal subsegments ending at two decreasing (increasing) supporting curves or on curves in $\Gamma_0^{\mp}$. After this the system obtained is consistent.

The constructions in 3.3 and 3.4 can be transferred to the case under consideration without changes.

Note that, as in 4.2, 4.3, the pre-Markov partition ξ_1 is countable, but that $\partial\xi_1$ contains only finitely many LUM and LSM.

§5. Transition from a pre-Markov to a Markov partition

5.1. In §3 it was noted that in the case of smooth two-dimensional systems pre-Markov partitions are simultaneously Markov partitions. In discontinuous systems this is not true, since the elements of a pre-Markov partition are not parallelograms. To obtain a Markov partition an additional construction is necessary, consisting of refining the pre-Markov partition ξ_1 in neighbourhoods of discontinuity curves, since it is there that LUM and LSM of small length are concentrated. Such a construction was proposed for the first time in [18]. In [26] an attempt was made to simplify it (regrettably, it involved a number of inaccuracies). Below we give a reworked and formalized version of this construction.

We construct an increasing sequence of regular partitions $\eta_1 \leqslant \eta_2 \leqslant \dots$ of the space M_1, converging (mod 0) to the required partition η as $n \to \infty$. The partition η_n, $n \geqslant 2$, is constructed recurrently, by refining η_{n-1} in neighbourhoods of the discontinuity curves R_{-k_n,k_n} ($k_n \to \infty$ as $n \to \infty$). The corresponding neighbourhoods will be called *necklaces*. Their thickness and measure decrease sufficiently rapidly as n grows.

5.2. We now turn to strict statements. Let D_n^0, $n \geqslant 1$, be the union of all elements of the partition ξ_n (see 3.4) whose boundary intersects $R_{-m,m}$. Clearly, $D_1^0 \supseteq D_2^0 \supseteq \dots$ and $\bigcap\limits_n D_n^0 = R_{-m,m}$. The set D_n^0, $n \geqslant 1$, is a closed neighbourhood of $R_{-m,m}$.

We define the sequence of integers k_n reccurrently: $k_1 = 1$, $k_2 = 2$, and $k_n = 2k_{n-1}-1$ for $n \geqslant 3$ (its general term is $k_n = 2^{n-2}+1$, $n \geqslant 2$). Let $\xi_{k_n}^0$, $n \geqslant 1$, be a regular partition coinciding with ξ_{k_n} on $D_{k_{n-1}}^0$ and including the complement $M_1\setminus \operatorname{int} D_{k_{n-1}}^0$ as an element. We put $\xi_{k_n}^k = T_1^k \xi_{k_n}^0$ for $n \geqslant 1$, $k \in \mathbb{Z}$.

We recurrently define a sequence of regular partitions η_n : $\eta_1 = \xi_1$, and

$$\eta_n = \eta_{n-1} \vee \Big(\bigvee_{|k| \leqslant k_n - 1} \xi_{k_n}^k \Big) \tag{5.1}$$

for $n \geqslant 2$. It is clearly non-decreasing: $\eta_1 \leqslant \eta_2 \leqslant \dots$.

We put $D_{k_n}^k = T_1^k D_{k_n}^0$ and $E_{k_n}^k = M_1\setminus \operatorname{int} D_{k_n}^k$ for all $n \geqslant 1$, $|k| \leqslant k_{n+1}-1$. Since ξ_{k_n} is a pre-Markov partition, $D_{k_n}^k$ and $E_{k_n}^k$ consist of elements of η_{n+1}. We introduce the following notations:

$$D_{k_n}(k,\ l) = D_{k_n}^k \cup D_{k_n}^{k+1} \cup \dots \cup D_{k_n}^l \quad \text{for } k < l;$$

$$D_{k_n}^+ = D_{k_n}(0, k_{n+1}-1); \quad D_{k_n}^- = D_{k_n}(-k_{n+1}+1, 0); \quad D_{k_n} = D_{k_n}^+ \cup D_{k_n}^-.$$

The sets D_{k_n} will be called *necklaces*.

By (5.1) each element of η_n is the intersection of finitely many elements of ξ_1 and $\xi^k_{k_t}$ for $|k| \leqslant k_t - 1$, $1 \leqslant t \leqslant n$. We divide the elements of the partitions mentioned into two classes:

the 1st class contains all elements of ξ_1 and the elements of the partitions $\xi^k_{k_t}$ lying in the necklace $D_{k_{t-1}}$;

the 2nd class consists of all elements of $E^k_{k_t}$, $|k| \leqslant k_{t+1} - 1$.

It is easily seen that the 1st class contains connected elements of small diameter. On the other hand, the 2nd class includes disconnected elements, each of which fills "almost all" of M_1, except for small neighbourhoods of the discontinuity curves.

Lemma **5.1.** *D_{k_n} is a closed neighbourhood of the set $R_{-k_{n+1}m, k_{n+1}m}$; $D^+_{k_n}$ and $D^-_{k_n}$ are closed neighbourhoods of the sets $R_{0,k_{n+1}m}$ and $R_{-k_{n+1}m,0}$, respectively.*

The lemma can be derived from the construction of the sets $R_{k,l}$ since $D^0_{k_n}$ is a closed neighbourhood of $R_{-m,m}$ for all $n \geqslant 1$.

5.3. We introduce the notion of *rank* of the curves forming the boundary $\partial\xi_n$, $n \geqslant 1$. The ranks of all smooth curves in $\partial\xi_1$ are put equal to 1. For each $n \geqslant 2$ the ranks of the smooth curves in $\partial\xi_n \backslash \partial\xi_{n-1}$ are put equal to n. The rank of a curve γ is denoted by rank γ.

ξ_n being pre-Markov implies the following rules for computing the ranks of increasing curves:

a) if rank $\gamma > 1$, then rank $T_1^{\pm 1}\gamma$ = rank $\gamma \pm 1$;

b) if rank $\gamma = 1$, then rank $T_1^{-1}\gamma = 1$.

Similar rules (with the replacement of T_1 by T_1^{-1} and conversely) are valid for decreasing curves.

Lemma **5.2.** *For all $k_n \leqslant k \leqslant k_{n+1} - 1$, $n \geqslant 1$, the boundary $\partial^n D^{-k}_{k_n}$ $(\partial^s D^k_{k_n})$ consists of curves of rank* 1.

The proof consists of a direct computation of the ranks.

We can prove that the partition η_n is pre-Markov for all $n \geqslant 1$. However, we will not need this fact.

5.4. We introduce some concepts in order to describe the geometrical shapes of the elements of the partitions η_n. A simply-connected closed domain $A \subset M_1$ is called a *polygon* if $A = \mathscr{F}(A)$, if ∂A consists of finitely many LUM, LSM, and discontinuity curves, and if all interior angles formed by intersections of smooth components of ∂A of different monotonicity do not exceed 180°. A *side* of a polygon A is a maximal continuous (not necesarily smooth) monotone or neutral curve $\gamma \subset \partial A$. It is obvious that each side consists either of a chain of curves of the same monotonicity, or of a single neutral segment. Correspondingly we distinguish between *increasing*, *decreasing*, and *neutral* sides of a polygon. Any polygon has at most two increasing and two decreasing sides. This can be proved by calculating the angle of rotation under a full circuit of the boundary ∂A.

The polygons with excactly two increasing and two decreasing sides (and arbitrarily many neutral ones) are called *complete.*

Remark 5.3. For any polygon A and point $x \in A$ the sets $\gamma_A^u(x)$ and $\gamma_A^s(x)$ are connected. The end points of $\gamma_A^u(x)$ $(\gamma_A^s(x))$ either coincide with the "natural" end points of the LUM $\gamma^u(x)$ (LSM $\gamma^s(x)$), or lie on the decreasing (increasing) sides of A.

Remark 5.4. Let A be a complete polygon. We assume that the LUM γ^u intersects both decreasing sides of A, while the LSM γ^s intersects both increasing sides of A. The point $x = \gamma^u \cap \gamma^s$ exists and belongs to A.

In the sequel we will consider only polygons satisfying the following two conditions R_1 and R_2.

Condition R_1. Each smooth component of the boundary of the polygon lies in $\partial\xi_n$ for some $n \geqslant 1$.

Thus the components of the boundaries of polygons have ranks. Let $\operatorname{rank}^+ A$ ($\operatorname{rank}^- A$) be the maximal rank of the increasing (decreasing) components of ∂A.

Condition R_2. The interior of A does not contain increasing (decreasing) curves of ranks not exceeding $\operatorname{rank}^+ A$ ($\operatorname{rank}^- A$).

It is easy to prove that if the intersection of certain polygons has a non-empty interior, then it is also a polygon, and

$$(5.2) \qquad \operatorname{rank}^{\pm}(A_1 \cap \dots \cap A_k) = \max_{1 \leqslant i \leqslant k} \{\operatorname{rank}^{\pm} A_i\}.$$

In 3.4 we have, in fact, proved that all elements of ξ_n, $n \geqslant 1$, are polygons (in particular, they satisfy conditions R_1 and R_2). This implies that all elements of the first class (defined in 5.2) are polygons.

5.5. We show that the necklaces D_{k_n} form (mod 0) a covering of M_1 of at most finite multiplicity. Our version of this assertion is more precise than the corresponding lemma in [18], [26].

***Lemma* 5.5.** *Each point $x \in M_2^+$ ($x \in M_2^-$) belongs to only finitely many sets $D_{k_n}^+$ ($D_{k_n}^-$), $n \geqslant 1$.*

Proof. Let $x \in D_{k_n}^+$. Then $T_1^{-k}x \in D_{k_n}^0$ for some $0 \leqslant k \leqslant k_{n+1}-1$. Let Δ be an element of $\xi_{k_n}^0$ containing $T_1^{-k}x$. It borders $R_{-m,m}$, hence for some $|l| \leqslant m$ the polygon $\Delta_1 = T^l\Delta$ borders $R_{0,1}$ (we stress that the initial automorphism T is involved, and not T_1). We put $x_1 = T^{-km+l}x$, $x_1 \in \Delta_1$.

We show that Δ_1 lies in a sufficiently small neighbourhood of $R_{0,1}$. It is easily seen that for $j_1 = (k_n-1)m-l$ and $j_2 = (k_n-1)m+l$ the transformations T^{j_1} and T^{-j_2} are continuous on int Δ_1. Let γ be an arbitrary increasing (decreasing) curve inside Δ_1. Then $\gamma' = T^{j_1}\gamma$ $(\gamma' = T^{-j_2}\gamma)$ are j_1-increasing (j_2-increasing). By the choice of Δ the curve γ' does not intersect decreasing (increasing) components of $\partial\xi_1$. By Lemma 3.4 we have $p(\gamma') \leqslant \lambda_3^{-1}\varepsilon$ (in the case of finite horizon λ_3 must be replaced by λ_1 (see §3)). Hence $p(\gamma) \leqslant \varepsilon'$, where $\varepsilon' = \lambda_3^{-1}\varepsilon\Lambda_0^{-[j/m_0]}$ and $j = \min\{j_1, j_2\} = (k_n-1)m-|l|$. Similar estimates hold for the p-lengths of all increasing (decreasing) curves inside $T^{-1}\Delta_1$, which borders $R_{-1,0}$. This readily implies that $d^+(x_1) \leqslant \mathrm{const}(Q)\varepsilon'$, or otherwise

$$d^+(T^{-km+l}x) \leqslant \mathrm{const}\,(Q)\,\varepsilon\Lambda_0^{-\left[\frac{(k_n-1)m-|l|}{m_0}\right]} \tag{5.3}$$

(we recall that $0 \leqslant k \leqslant k_{n+1}-1$). If $x \in M_2^+$, the latter inequality can only be valid for finitely many n. A similar reasoning is valid for $x \in M_2^-$. The lemma is proved.

Remark. From the estimates (5.3) and (2.6) we can also derive the following estimate for the measure of $D_{k_n}^0$, which was given in [18], [26] without proof:

$$\nu\,(D_{k_n}^0) \leqslant c\varepsilon\alpha^{m(k_n-2)} \tag{5.4}$$

for all $n \geqslant 2$ (here $c > 0$ and $\alpha < 1$ are certain constants for Q).

For each $x \in M_1$ we put $N_x = \max\{n : x \in D_{k_{n-1}}\}$. If $x \notin D_{k_n}$ for all $n \geqslant 1$, then we put $N_x = 1$. Lemma 5.5 shows that $N_x < \infty$ for all $x \in M_2$.

5.6. We consider in more detail the geometric shape of the elements of η_n, $n \geqslant 2$. We are only interested in the $\Delta \in \eta_n$ that do not lie in the necklace $D_{k_{n-1}}$. These are called *regular elements*. For each regular element $\Delta \in \eta_n$ we put $N_\Delta = \max\{t \leqslant n : \Delta \subset D_{k_{t-1}}\}$. If $\Delta \not\subset D_{k_t}$ for all $t \leqslant n-1$, we put $N_\Delta = 1$. Thus, $1 \leqslant N_\Delta \leqslant n-1$.

We fix an $n \geqslant 2$ and a regular element Δ of η_n. By (5.1) we can write

$$\Delta = \mathcal{F}\,(G_1' \cap G_2' \cap \ldots \cap G_{p_1}' \cap G_1'' \cap G_2'' \cap \ldots \cap G_{p_2}''), \tag{5.5}$$

where G_i' denote the elements of the 1st class and G_i'' the elements of the 2nd class. The results of 5.4 imply that the set $G_\Delta^0 = \mathcal{F}\,(G_1' \cap G_2' \cap \ldots \cap G_{p_1}')$ is a polygon.

***Lemma* 5.6.** $\mathrm{rank}^\pm\, G_\Delta^0 \leqslant k_{N_\Delta+1}$.

For the proof we note that the elements $G_1', G_2', \ldots, G_{p_1}'$ belong to the partitions ξ_1 and $\xi_{k_t}^k$ for certain $|k| \leqslant k_t-1$ and $t \leqslant N_\Delta$. Subsequently we have to estimate the ranks of these elements by the rules a) and b), and apply (5.2).

Among the elements $G_1'', \ldots, G_{p_2}''$ of the 2nd class wę distinguish two groups: to the first group belong those $E_{k_t}^k$ for which $k_t \leqslant k < k_{t+1}$, and to the second group those $E_{k_t}^k$ for which $-k_{t+1} < k \leqslant -k_t$. The intersection of the selected elements from the first group is denoted by G_Δ^+, that from the second by G_Δ^-.

***Lemma* 5.7.** $\Delta = \mathcal{F}\,(G_\Delta^0 \cap G_\Delta^+ \cap G_\Delta^-)$.

For the proof we consider an arbitrary "non-selected" element $G''_{i_0} = E^k_{k_t}$. For it $|k| \leqslant k_t - 1$ and $t \leqslant n-1$. By the construction in 5.2 the set $D^0_{k_{t-1}} \setminus D^0_{k_t}$ is the union of the elements of the 1st class of the partition $\xi^0_{k_t}$. Hence $G''_{i_0} = E^k_{k_t}$ is the union of the element $E^k_{k_{t-1}}$ and certain elements of the 1st class of the partition $\xi^0_{k_t}$. Let $\hat{G}''$ denote the element among those mentioned that contains Δ. Clearly $\hat{G}'' \subseteq G''_{i_0}$ and, moreover, the element $\hat{G}''$ is included in the partition (5.5). This implies the lemma.

We put $G^{+0}_\Delta = \mathscr{F}\,(G^+_\Delta \cap G^0_\Delta)$ and $G^{-0}_\Delta = \mathscr{F}\,(G^-_\Delta \cap G^0_\Delta)$.

***Lemma* 5.8.** *$\partial^s G^{+0}_\Delta \subseteq \partial^s G^0_\Delta$ and $\partial^u G^{-0}_\Delta \subseteq \partial^u G^0_\Delta$.*

Proof. By Lemma 5.2 the boundaries $\partial^s G^+_\Delta$ and $\partial^u G^-_\Delta$ consist of curves of rank 1. By condition R_2 these curves cannot lie strictly in the interior of the polygon G^0_Δ. The lemma is proved.

The typical shape of a regular element Δ is depicted in Fig. 15.

Note that a regular element Δ can contain "splinters" (of the type W_1 and W_2 in Fig. 15), in which there cannot be points of the limit parallelogram $U \in \eta$. We will finally get rid of these "splinters" below, in 5.8.

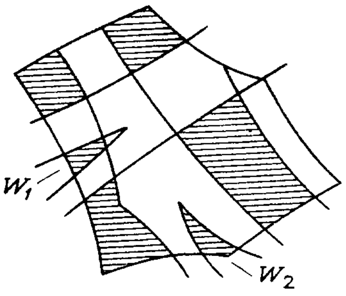


Fig. 15

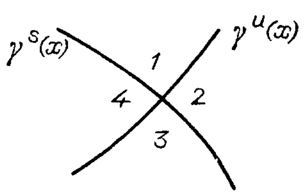


Fig. 16

5.7. We consider the partition of M_1 obtained from η_n as $n \to \infty$.

We fix a point $x \in M_2$. Then for all $n \geqslant N_x + 1$ there is a regular element $\Delta_n(x)$ of η_n containing x. If $x \in \partial\eta_n$, there are several such elements. More precisely, there are at most four such elements; this follows from the next remark.

Remark 5.9. For any n the set $\Delta_n(x)$ contains either a full neighbourhood of x, or a semineighbourhood of the LUM $\gamma^u(x)$ (LSM $\gamma^s(x)$), or the interior of one of the four angles between the LUM $\gamma^u(x)$ and the LSM $\gamma^s(x)$ (Fig. 16).

In any of these cases we can choose a decreasing sequence of regular elements $\Delta_{N_x+1}(x) \supseteq \Delta_{n_x+2}(x) \supseteq \dots$. We put $\Delta_\infty(x) = \bigcap\limits_n \Delta_n(x)$. It is obvious that Δ_∞ is a closed non-empty set (in particular, $x \in \Delta_\infty(x)$). The regularity of $\Delta_{N_x+1}(x)$ readily implies that

$$\Delta_\infty(x) = \lim_{N\to\infty} \mathscr{F}\,\Big(\Delta_{N_x+1}(x) \setminus \bigcup_{n=N_x+1}^{N} D_{k_n}\Big). \tag{5.6}$$

The set of all elements $\Delta_\infty(x)$ of the form (5.6) is denoted by Ω. It is clearly countable.

We put $\tilde{\partial}^u\eta = \bigcup_n \partial^u\eta_n$, $\tilde{\partial}^s\eta = \bigcup_n \partial^s\eta_n$, and $\tilde{\partial}\eta = \tilde{\partial}^u\eta \bigcup \tilde{\partial}^s\eta$. The sign $\sim$ is used here since formally these sets are not boundaries (albeit only because they are dense in M_1). However, $\tilde{\partial}\eta$ serves as an analogue of the boundary of a Markov partition for smooth systems [2], [9]. Namely, if $x \in M_2 \backslash \tilde{\partial}\eta$, then x is covered by precisely one element $\Delta \in \Omega$. If, however, $x \in \tilde{\partial}\eta$, then there can be several such elements, but at most four (this was shown above).

Let $x \in M_2$ and let $\mathring{V}(x)$ be an infinitesimal neighbourhood of x. The regular components of $\tilde{\partial}\eta$ that pass through x or end at x divide $\mathring{V}(x)$ into several parts, called *characteristic neighbourhoods* of x. Any $x \in M_2$ has at most four characteristic neighbourhoods, corresponding to those described in Remark 5.9. The characteristic neighbourhoods of x will be denoted by $V(x)$. We conventionally say that a characteristic neighbourhood $V(x)$ of $x \in M_2$ is *connected* with the element $\Delta \in \Omega$ if $\Delta = \Delta_\infty(x)$ and $V(x)$ belongs to all "prelimit" sets $\Delta_n(x) \in \eta_n$ defining $\Delta_\infty(x)$ by (5.6).

Remark 5.10. Let $x \in M_2$. Every element $\Delta \in \Omega$ containing x is connected with exactly one characteristic neighbourhood of $V(x)$ of x, and conversely.

We recall that $\partial\xi_1$ contains finitely many LUM and LSM (see §§3, 4). ξ_1 pre-Markov implies that these LUM (LSM) contain finitely many periodic (under T_1) points $x_1^u, x_2^u, \dots, x_{n_1}^u$ $(x_1^s, x_2^s, \dots, x_{n_2}^s)$, to which these LUM (LSM) approach under the action of T_1^{-n} (T_1^n) as $n \to \infty$. Let $\Gamma^u(\eta) = \bigcup_{i=1}^{n_1} \Gamma^u(x_i^u)$ and $\Gamma^s(\eta) = \bigcup_{i=1}^{n_2} \Gamma^s(x_i^s)$ be the unions of the LUM of the points x_i^u and of the LSM of x_i^s (see 1.1). The construction of η implies that

$$\tilde{\partial}^u\eta \subset \Gamma^u(\eta) \text{ and } \tilde{\partial}^s\eta \subset \Gamma^s(\eta). \tag{5.7}$$

5.8. Ω is a countable covering of M_2. To obtain a Markov partition from it we must remove certain "non-typical" points from the sets $\Delta \in \Omega$. These sets then become parallelograms.

We introduce some notation. For each $\Delta \in \Omega$ we put $\Delta^0 = \Delta \bigcap M_2$, and let Δ^* be the closure of Δ^0 (clearly $\Delta^* \subset \Delta$). We also put $\Omega^0 = \{\Delta^0\}$ and $\Omega^* = \{\Delta^*\}$. For each $\Delta \in \Omega$ we put $N(\Delta) = \min\{N_x : x \in \Delta\}$, and denote by $C_n(\Delta)$, $n \geqslant 1$, an element of η_n containing Δ. For $n > N(\Delta)$ it is regular. By Lemma 5.7 it has a decomposition, which may be written as

$$C_n(\Delta) = \mathscr{F}(G_n^0(\Delta) \bigcap G_n^+(\Delta) \bigcap G_n^-(\Delta)).$$

We put, as in 5.6, $G_n^{\pm 0}(\Delta) = \mathscr{F}(G_n^{\pm}(\Delta) \bigcap G_n^0(\Delta))$. It is obvious that the sequences of sets $\{C_n(\Delta)\}, \{G_n^{\pm 0}(\Delta)\}$ are non-increasing for $n > N(\Delta)$. We put $G_\infty^{\pm 0}(\Delta) = \bigcap_n G_n^{\pm 0}(\Delta)$. Thus, $\Delta = G_\infty^{+0}(\Delta) \bigcap G_\infty^{-0}(\Delta)$. Note that, by Lemma 5.6, for all $n > N(\Delta)$ the polygons $G_n^0(\Delta)$ coincide. The corresponding polygon is denoted by $G^0(\Delta)$. For $y \in \Delta$ we put

$$\gamma_0^u (y, \Delta) = \gamma^u (y) \cap G^0 (\Delta) \quad \text{and} \quad \gamma_0^s (y, \Delta) = \gamma^s (y) \cap G^0 (\Delta).$$

***Lemma* 5.11.** *Let $\Delta^* \in \Omega^*$. For each $y \in \Delta^*$*

a) *the curve $\gamma^u(y)$ $(\gamma^s(y))$ intersects the two decreasing (increasing) sides of the polygon $G^0(\Delta)$ (in particular, $G^0(\Delta)$ is complete);*

b) *$\gamma_0^u(y, \Delta) \subset G_\infty^{+0}(\Delta)$ and $\gamma_0^s(y, \Delta) \subset G_\infty^{-0}(\Delta)$.*

Proof. By Lemma 2.11 it suffices to consider the case $y \in \Delta^0$. Then $y \in M_2$, and by Lemma 2.13 the end points of the LUM $\gamma^u(y)$ lie on two discontinuity curves $\gamma_1, \gamma_2 \subset R_{0,\infty}$. We put $\gamma_n^u(y) = \gamma^u(y) \cap G_n^{+0}(\Delta)$ for $n > N(\Delta)$. For sufficiently large n we have $k_n > \max\{\text{rank } \gamma_1, \text{rank } \gamma_2\}$, and by Lemma 5.1 the set $G_n^+(\Delta)$ does not intersect γ_1 and γ_2. Then $\gamma_n^u(y)$ is at a positive distance from the end points of $\gamma^u(y)$. Hence all boundary points of $\gamma_n^u(y)$, as a subset of $\gamma^u(y)$, belong to $\partial^s G_n^{+0}(\Delta)$ (if they belonged to $\partial^u G_n^{+0}(\Delta)$, then they would lie on increasing discontinuity curves or on other LUM, which is impossible). By Lemma 5.8 these points belong to $\partial^s G^0(\Delta)$. This implies that $\gamma_n^u(\Delta)$ coincides with $\gamma_0^u(y, \Delta)$. Thus, $\gamma_0^u(y, \Delta) \subset G_\infty^{+0}(\Delta)$. A similar reasoning is valid for the LSM $\gamma^s(y)$. The lemma is proved.

Assertion. *For any $\Delta \in \Omega$ the sets Δ^0 and Δ^* are parallelograms.*

Proof. By Lemma 2.1 it suffices to prove that Δ^0 is a parallelogram. Let $x, y \in \Delta^0$. By Lemma 5.11a) and Remark 5.4 the point $z = [x, y]$ exists and belongs to $G^0(\Delta)$. By Lemma 5.11b), $z \in \gamma_0^u(x, \Delta) \subset G_\infty^{+0}(\Delta)$ and $z \in \gamma_0^s(y, \Delta) \subset G_\infty^{-0}(\Delta)$, therefore $z \in G_\infty^{+0}(\Delta) \cap G_\infty^{-0}(\Delta) = \Delta$. Moreover, $z \in M_2$ since $x, y \in M_2$ (see Lemma 2.12b)). Hence $z \in \Delta^0$. The assertion is proved.

Note that $\nu(\Delta^* \backslash \Delta^0) = 0$ for any Δ, since $\Delta^* \backslash \Delta^0 \subset \Delta \backslash M_2$.

Corollary. *The closed parallelograms $\Delta^* \in \Omega^*$ form a covering* (mod 0) *of M_1. Moreover, they cover all points of M_2. Each $x \in M_2$ is covered by at most four parallelograms.*

We will say that a characteristic neighbourhood $V(x)$ of $x \in M_2$ is *connected with the parallelogram* $\Delta^* \in \Omega^*$ if it is connected with the corresponding $\Delta \in \Omega$. Remark 5.10 naturally generalizes to elements $\Delta^* \in \Omega^*$.

5.9. We now show that the covering Ω^* constructed above is a Markov partition. It suffices to verify the regularity of intersection of parallelograms $U(x)$ and $T_1 U(T_1^{-1}x)$ for ν-almost all x.

***Assertion* 5.12.** *Let $x \in M_2$, and let $\Delta_1^* \in \Omega^*$ be a parallelogram containing x. Then there is a parallelogram $\Delta_2^* \in \Omega^*$ containing $T_1 x$ $(T_1^{-1}x)$ such that the intersection $\Delta_2^* \cap T_1\Delta_1^*$ $(\Delta_1^* \cap T_1\Delta_2^*)$ is regular.*

We give a proof only for $T_1 x$ (for $T_1^{-1}x$ it is completely similar).

***Lemma* 5.13.** *Let $V(x)$ and $V(T_1x)$ be characteristic neighbourhoods of x and T_1x, respectively. We assume that $V(x)$ is connected with the parallelogram $\Delta_1^* \in \Omega^*$, and that $V(T_1x)$ is connected with the parallelogram $\Delta_2^* \in \Omega^*$. If $T_1V(x) \cap V(T_1x) \neq \varnothing$, then the intersection $\Delta_2^* \cap T_1\Delta_1^*$ is regular.*

This lemma immediately implies the statement, since a corresponding characteristic neighbourhood $V(T_1x)$ of T_1x can always be found.

We prove Lemma 5.13. It suffices to show that $T_1\gamma^s_{\Delta_1^*}(x) \subseteq \gamma^s_{\Delta_2^*}(T_1x)$ and $T_1\gamma^u_{\Delta_1^*}(x) \supseteq \gamma^u_{\Delta_2^*}(T_1x)$. We prove only the first relation (the proof of the second is similar). We use the notations introduced in 5.8 and put $n_0 = \max\{N_x, N_{T_1x}\}$. We first prove that

$$(5.8) \qquad T_1\gamma_0^s(x, \Delta_1) \subseteq \gamma_0^s(T_1x, \Delta_2).$$

It is obvious that $T_1\gamma_0^s(x, \Delta_1)$ is a LSM. If (5.8) does not hold, there is a point y in the interior of this LSM lying on $\partial^uG^0(\Delta_2)$. Let γ be the smooth component of $\partial^uG^0(\Delta_2)$ containing y. By the construction of the polygon $G^0(\Delta_2)$, the curve γ lies in the boundary of some element $G' \supseteq G^0(\Delta_2)$ of the 1st class. Two cases are possible:

1st case: G' is an element of $\xi^k_{k_t}$ for some $k \geqslant -k_t+2$ and $t \geqslant 1$. Then $T_1^{-1}G'$ is also an element of the 1st class. It contains $V(x)$, and hence the polygon $G^0(\Delta_1)$. Then $T_1\gamma_0^s(x, \Delta_1) \subset G'$ and we have a contradiction.

2nd case: G' is an element of ξ_1 or $\xi^k_{k_t}$ for $k = -k_t+1$ and $t \geqslant 1$. In this case it is easily verified that rank $\gamma = 1$. Then $T_1^{-1}\gamma \subset \partial\xi_1$. Hence it cannot intersect $\gamma_0^s(x, \Delta_1)$ at the interior point $T_1^{-1}y$. Thus (5.8) is proved.

We show that

$$(5.9) \qquad T_1\gamma^s_{\Delta_1}(x) \subseteq \gamma^s_{\Delta_2}(T_1x).$$

Let y be an arbitrary point in $\gamma^s_{\Delta_1}(x)$. By (5.8) its image $T_1y \in G^0(\Delta_2)$, while by Lemma 5.11b) it lies in $G^-_\infty(\Delta_2)$. Hence it suffices to prove that $T_1y \in G^+_\infty(\Delta_2)$. If this were not true, then $T_1y \notin G_n^+(\Delta_2)$ for some $n > n_0$. Using the definition of $G_n^+(\Delta_2)$ and (5.4) it is easily proved that in this case $y \notin G_n^+(\Delta_1)$. We arrive at a contradiction, proving (5.9).

Finally, let y be an arbitrary point in $\gamma^s_{\Delta_1^*}(x)$. If $y \in M_2$, then $T_1y \in M_2$, and in view of (5.9) we have $T_1y \in \gamma^s_{\Delta_2^*}(T_1x)$. If $y \notin M_2$, then $y_i \to y$ for some sequence $y_i \in \Delta_1^0$. It is easily seen that the points $y_i' = [y_i, x]$ are defined and belong to Δ_1^0. As was shown above, $T_1y_i' \in \gamma^s_{\Delta_2^*}(T_1x)$. Hence $T_1y = \lim T_1y_i' \in$ $\in \gamma^s_{\Delta_2^*}(T_1x)$. Lemma 5.13 is proved.

In this way a Markov partition η for T^m has been constructed. A Markov partition for T can be obtained by Remark 1.2. Note that the following two corollaries follow from our construction:

Corollary. *Any point $x \in M_2$ has a coding $\sigma(x)$ in the symbolic dynamics (Σ_Π, θ) (the notion of coding and all notations were introduced in* 1.2).

Corollary. *If $x \in M_2$ and $T^kx \notin \tilde{\partial}\eta$ for all integers k, then the coding $\sigma(x)$* is unique.

In certain applications it is essential that for each individual point the number of distinct codings $\sigma(x) \in \Sigma_\Pi$ is uniformly bounded. To achieve this we construct in §7 a modified symbolic dynamics.

§6. Non-scattering billiards with hyperbolic behaviour

The results of this section will not be used in §7. Here we construct a Markov partition for certain classes of non-scattering billiards by essentially using the material in §4.

6.1. Semiscattering billiards.

A billiard in a domain Q is called *semiscattering* [14] if ∂Q consists of scattering and neutral components (for terminology see 1.1).

In semiscattering billiards the reflections in neutral components are at most a "disturbing factor", since they do not lead to expansions and contractions (that is, to hyperbolicity). To exclude the influence of this factor we turn to a derived transformation. Let ∂^+Q (∂^0Q) denote the union of all scattering (neutral) components of ∂Q. We put $M_1^+ = \{x : \pi(x) \in \partial^+Q\}$, $M_1^0 = \{x : \pi(x) \in \partial^0Q\}$, and consider the derived automorphism $\tilde{T}$ constructed from T and M_1^+. For a point $x \in M_1$ we denote by $k(x)$ the index of the first reflection of the trajectory of x in ∂^+Q, that is, $Tx = T^{k(x)}x$ for $x \in M_1^+$. The measure $d\nu = \text{const} \cos\varphi dr d\varphi$ is invariant under $\tilde{T}$, as well as under T. It is easily seen that the function $k(x)$ is constant on the continuity domains of $\tilde{T}$ in M_1^+, while the transformations T^i are continuous for $1 \leqslant i \leqslant k(x)$.

Let $x \in M_1^+$ and $k(x) \geqslant 2$. The reflections in the neutral components of ∂Q at the points $Tx, ..., T^{k(x)-1}x$ can be "straightened out" by symmetrically reflecting the domains Q itself with respect to the corresponding component of ∂^0Q (Fig. 17). This shows that locally the properties of $\tilde{T}$ are the same as those of the automorphism T in scattering billiards.

We impose restrictions on Q, as in Theorem 1.1:

A′. All interior angles of ∂Q are strictly positive.

B′. The multiplicity of all points $x \in M_1^+$ is uniformly bounded by a constant $K_0' = K_0(Q) < \infty$ (as in 1.1, the multiplicity is the number of discontinuity curves of the maps T^n, $n \in \mathbb{Z}$, passing through x).

From what we said above it follows that under the conditions A′ and B′ all definitions and assertions in §§1, 2 can, without essential modifications, be transferred to the transformation $\tilde{T}$. On Q we impose the additional restriction:

C′. The function $k(x)$ is uniformly bounded on $M_1^+ : k(x) \leqslant \text{const}(Q) < \infty$.

Under this condition the number of discontinuity curves of $\tilde{T}^n$ is finite for any integer n, and $\tilde{T}$ has all the properties of the automorphism T for scattering billiards with finite horizon. Hence the constructions in §§3, 5 can be transferred to this case. As a result we obtain a Markov partition $\tilde{\eta}$ of M_1^+ for the automorphism $\tilde{T}$.

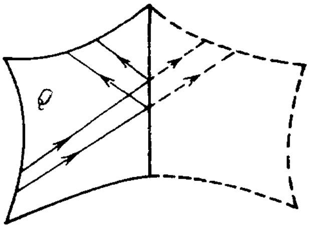


Fig. 17

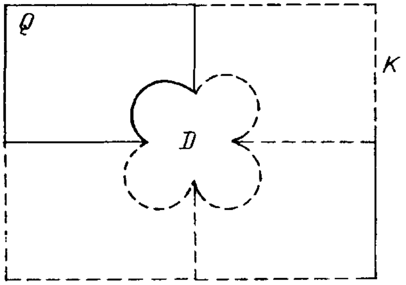


Fig. 18

To construct a Markov partition η in the whole space M_1 we consider an arbitrary element $\tilde{U} \in \tilde{\eta}$ ($\tilde{U} \subset M_1^+$). There is a connected domain $V \supset \tilde{U}$, on which $\tilde{T}$ is continuous. Hence $k(x) \equiv$ const on $\tilde{U}$. The sets $\tilde{U}$, $T\tilde{U}$, ..., $T^{k(x)-1}\tilde{U}$ for all possible $\tilde{U} \in \tilde{\eta}$, where $x \in \tilde{U}$, form a Markov partition of M_1 for T. This can be directly verified. Thus we have proved the following result.

***Theorem* 6.1.** *Let Q be a semiscattering billiard satisfying the conditions* A′, B′, *and* C′. *Then for any $\varepsilon > 0$ there is a Markov partition with elements of diameter at most ε.*

In conclusion we note that condition C′ is very restrictive. If it is not satisfied, then in a number of cases the truth of Theorem 6.1 can be proved by direct reduction to scattering billiards. For instance, in Fig. 18, after three reflections in the neutral components of ∂Q the domain Q fills the square K, from which a domain D with scattering boundary is deleted. Using the procedure of "straightening out" (Fig. 17), the billiard trajectories in Q become billiard trajectories on the torus $\mathrm{Tor}^2 \setminus D$ (the torus Tor^2 is given by a fundamental domain, which coincides with K). The billiard on $\mathrm{Tor}^2 \setminus D$ is scattering, and a Markov partition for it can easily be transformed into a Markov partition for the billiard in Q.

This procedure is applicable if the neutral components of the boundary lie on sides of a rectangle, a regular triangle (or a hexagon). However, in general $\tilde{T}$ has singular points (at which $k(x)$ is unbounded), in neighbourhoods of which the structure of $\tilde{T}$ cannot be given a simple description.

6.2. Regular focussing components (pockets).

Let Q have focussing boundary components. We denote their union by $\partial^- Q$ and put $M_1^- = \{x : \pi(x) \in \partial^- Q\}$. A focussing component $\Gamma \subset \partial Q$ is called *regular* (or a *pocket*) if it is an arc of a circle O_Γ and if the disc K_Γ bounded by O_Γ intersects ∂Q only along Γ ($\Gamma \neq O_\Gamma$).

Billiards with regular focussing components were introduced and studied in [5], [17], and turned out to be very similar to scattering billiards: they are hyperbolic, ergodic, and are K-systems. Subsequently the class of non-scattering hyperbolic billiards was essentially enlarged in [32], [27]. However, we do not have the possibility of encompassing all these cases, and restrict ourselves to regular focussing components only.

We define $k(x)$ as the index of first reflection of the trajectory of the point $x \in M_1$ in $\partial^+ Q \cup \partial^- Q$, and construct the derived automorphism $\widetilde{T}$ from T and $M_1^+ \cup M_1^-$ (that is, $\widetilde{T} x = T^{k(x)}x$). We will assume that Q satisfies the conditions A′, B′, and C′ of 6.1.

The hyperbolic properties of billiards with pockets are described in [5], [17], and we briefly list the necessary results. A curve $\gamma \subset M_1^-$ given by an equation $\varphi = \varphi(r)$ is called *increasing* (*decreasing*) if $d\varphi/dr < 0$ ($d\varphi/dr > 0$), notwithstanding the traditional definition. The property of increase (decrease) is preserved under the action of T^n for $n > 0$ ($n < 0$), and the p-lengths of increasing (decreasing) curves increase under the action of T (T^{-1}). This is related to the so-called *defocussing* property: an increasing curve generates a convergent pencil of trajectories, passing through a defocussing point and becoming divergent already under the next reflection; moreover, the defocussing point lies on the first half of the path between the reflections (Fig. 19). The LUM and LSM in M_1^- are given by differential equations, using the same infinite continued fractions as in §2. Convergence of these continued fractions can be derived from the defocussing condition [6].

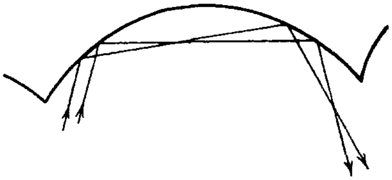


Fig. 19

As distinct from scattering billiards, expansion and contraction in $\hat{M}_1^-$ is not uniform, that is, for any power T^n the coefficients of expansion and contraction are not bounded away from one. This is related to the fact that the sequence of successive reflections in one focussing component does not reduce to expansion and contraction (in particular, billiards in a disc are not hyperbolic!). Hence again we are forced to turn to a derived transformation: we put $\hat{M}_1 = M_1^+ \cup \{x \in M_1^- : \pi(x) \in \Gamma_i \subset \partial^- Q, \pi(T^{-1}x) \in \Gamma_j \subset \partial Q, j \neq i\}$ and consider the derived automorphism $\hat{T} : \hat{M}_1 \to \hat{M}_1$. In the coordinates

(r, φ) the set $\hat{M}_1 \subset M_1^-$ is the union of finitely many parallelograms. Then for some $m_0 \geqslant 1$ the transformation $\hat{T}^{m_0}$ has uniform expansion and contraction (compare with T^{m_0} in §2). Thus, $\hat{T}$ has all the local properties described in §2.

The global properties of $\hat{T}$ are determined by the structure of the discontinuity curves, whose number in the present case is infinite. The discontinuity curves accumulate in neighbourhoods of singular points $z \in \hat{M}_1$, corresponding to the tangent directions to the pockets at their end points (Fig. 20a)). In Fig. 20b) the structure of the discontinuity curves of $\hat{T}$ in a neighbourhood of z_1 is drawn. This structure is similar to the one investigated in 4.4 for scattering billiards with infinite horizon. It is also easily verified that the coefficient of expansion of $\hat{T}$ grows unboundedly in a neighbourhood of z_1. These properties allow us to transfer the methods in 4.3, 4.4 for constructing a Markov partition to the case under consideration without essential modifications. As a result we obtain a Markov partition η for $\hat{T}$.

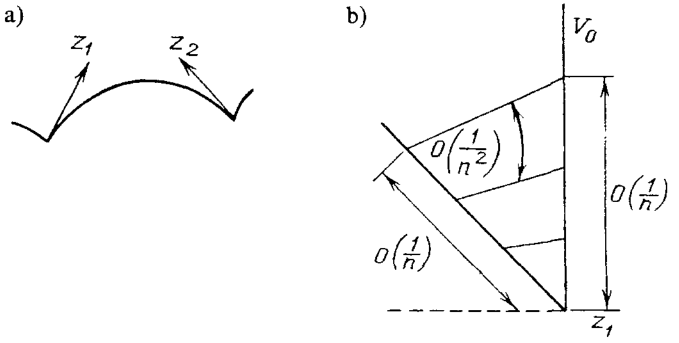

a)
z_1
z_2
b)
V_0
$O\left(\frac{1}{n^2}\right)$
$O\left(\frac{1}{n}\right)$
$O\left(\frac{1}{n}\right)$
z_1

Fig. 20

It remains to pass from $\hat{\eta}$ to a Markov partition η for T. Let $\hat{U} \in \hat{\eta}$ be an arbitrary element $(\hat{U} \subset \hat{M}_1)$. Then $\hat{U}$ belongs to a continuity domain of $\hat{T}$. Hence there is a $k(\hat{U}) \geqslant 1$ such that $T^i\hat{U} \subset M_1 \backslash \hat{M}_1$ for all $1 \leqslant i < k(\hat{U})$, and $T^{k(\hat{U})}\hat{U} \subset \hat{M}_1$. The sets $\hat{U}, T\hat{U}, ..., T^{k(\hat{U})-1}\hat{U}$ for all possible $\hat{U} \in \hat{\eta}$ form a Markov partition for T. This can be verified immediately. Thus we have proved the following result.

Theorem **6.2.** *Let Q be a billiard satisfying the conditions* A′, B′, *and* C′, *and let the focussing part of the boundary $\partial^- Q \neq \varnothing$ consist of pockets only. Then for any $\varepsilon > 0$ there is a Markov partition whose elements have diameter at most ε.*

6.3. The "stadium".

Restriction C′ in Theorem 6.2, as in 6.1, is very restrictive. Here we consider one system not satisfying it—the so-called *stadium*. This is the billiard in the domain bounded by two parallel segments and two arcs of circles (Fig. 21).

This system has applications of its own [14]. The hyperbolicity and ergodicity of the stadium were proved in [17]. Note that if the segments on the boundary of the stadium are not parallel, then the stadium satisfies condition C′, and hence the conditions of Theorem 6.2.

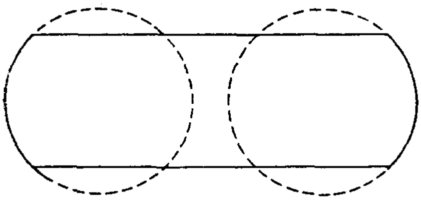
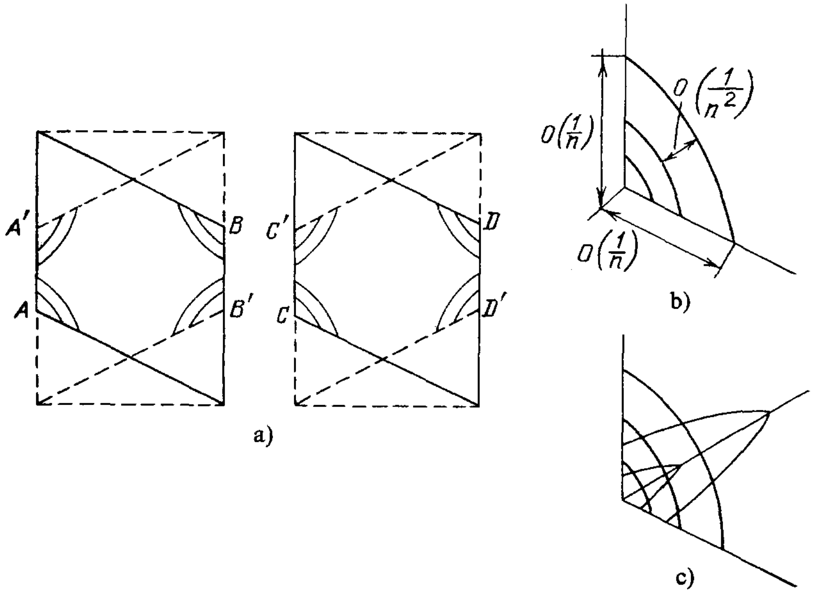


Fig. 21

We preserve for the stadium the notations $\hat{M}_1$, $\widetilde{T}$, and $\hat{T}$ introduced in 6.1 and 6.2. The phase space $\hat{M}_1$ is the union of two parallelograms (Fig. 22a)). In neighbourhoods of the points A, B, C, D infintely many discontinuity curves of $\hat{T}^{-1}$ accumulate (Fig. 22a)). Their structure (Fig. 22b)) is similar to the one investigated in 4.4 for billiards with infinite horizon. It is easily verified that the coefficient of expansion of $\hat{T}^{-1}$ grows unboundedly in neighbourhoods of the points A, B, C, D. If the arcs bounding the stadium are less than a semicircle, then the discontinuity curves of $\hat{T}$ accumulate in neighbourhoods of the four points A', B', C', D' and have a similar structure (Fig. 22a)). If the stadium is bounded by semicircles, then $A = A'$, $B = B'$, $C = C'$, $D = D'$, and the discontinuity curves of $\hat{T}$ and $\hat{T}^{-1}$ overlap (Fig. 22c)).



a)

b)

c)

Fig. 22

This case is similar to that of the singular points of type SV_{mix} in billiards with infinite horizon, which was considered in 4.3. Therefore the methods for constructing a Markov partition developed in §4 can be transferred to the case of the stadium. Finally, the transition from the Markov partiton $\hat{\eta}$ for $\hat{T}$ to a Markov partition η for T does not differ from that described in 6.2.

§7. Estimates for the number of periodic points

The results of this section were obtained by one of the authors— N.I. Chernov.

7.1. Modification of the symbolic dynamics.
As has been noted in §5, for $x \in M_1$ the coding $\sigma(x)$ need not be unique. For computing the number of periodic points it is important that for each individual point $x \in M_1$ the number of distinct codings $\sigma(x)$ is uniformly bounded. To this end we introduce a new, more accurate, definition of the intersection matrix Π.

Take the Markov partition $\eta = \Omega^*$ constructed in §5 for the map T^m. We consider two parallelograms Δ_0^*, $\Delta_1^* \in \Omega^*$ such that $\Delta_0^* \cap T^k\Delta_1^* \neq \varnothing$ for some $|k| \leqslant m$. Clearly, $\Delta_0^* \cap T^k\Delta_1^*$ is a closed parallelogram. We distinguish in it the subset $\Upsilon_0(\Delta_0^* \cap T^k\Delta_1^*)$ of points $x \in M_2$ for which the characteristic neighbourhoods $V(x)$ and $V(T^{-k}x)$ connected with Δ_0^* and Δ_1^*, respectively, satisfy $T^kV(T^{-k}x) \cap V(x) \neq \varnothing$. Further, let $\Upsilon(\Delta_0^* \cap T^k\Delta_1^*)$ be the closure of $\Upsilon_0(\Delta_0^* \cap T^k\Delta_1^*)$ in M_2. Using the construction of Ω^* (in 5.8) we can verify that $\Upsilon(\Delta_0^* \cap T^k\Delta_1^*)$ is a parallelogram. For several $|k_i| \leqslant m$ and $\Delta_i^* \in \Omega^*$ the set $\Upsilon(\Delta_0^* \cap T^{k_1}\Delta_1^* \cap \dots \cap T^{k_p}\Delta_p^*)$ is similarly defined. Remark 5.10 gives the following result.

Lemma **7.1.** *Let $\Delta_0^* \in \Omega^*$ and $|k_i| \leqslant m$ for $1 \leqslant i \leqslant p$, with some $p \geqslant 1$. Then for every point $x \in \Delta_0^* \cap M_2$ there is a parallelogram $\Delta_i^* \in \Omega^*$, $1 \leqslant i \leqslant p$, such that $x \in \Upsilon(\Delta_0^* \cap T^{k_1}\Delta_1^* \cap \dots \cap T^{k_p}\Delta_p^*)$.*

We define the Markov partiton Ω' for T as the collection of parallelograms

$$\Upsilon\,(\Delta_0^* \cap T\Delta_1^* \cap \dots \cap T^{m-1}\Delta_{m-1}^*) \tag{7.1}$$

for all possible $\Delta_0^*, \dots, \Delta_{m-1}^* \in \Omega^*$ for which the set (7.1) is not empty. The elements of Ω' will be denoted by Δ'. By Lemma 7.1, each $x \in M_2$ is covered by at least one parallelogram $\Delta' \in \Omega'$. The same lemma implies that for each $x \in M_2$ there are parallelograms Δ_0^*, Δ_1^*, ..., $\Delta_m^* \in \Omega^*$ such that $x \in \Upsilon(\Delta_0^* \cap T\Delta_1^* \cap \dots \cap T^m\Delta_m^*)$. Moreover, the parallelograms $\Delta_1' = \Upsilon(\Delta_0^* \cap T\Delta_1^* \cap \dots \cap T^{m-1}\Delta_{m-1}^*)$ and $\Delta_2' = \Upsilon(\Delta_1^* \cap T\Delta_2^* \cap \dots \cap T^{m-1}\Delta_m^*)$ are non-empty and hence belong to Ω'. Since $x \in \Delta_1' \cap T\Delta_2'$, Lemma 5.13 implies the following result.

***Lemma* 7.2.** *Under the above described conditions, the intersection* $\Delta_1' \cap T\Delta_2'$ *is regular.*

Thus Ω' is a Markov partition for T.

We redefine the intersection matrix $\Pi = \|\pi_{ij}\|$ as follows. Let

$$\Delta_i' = \Upsilon\,(\Delta_{0,i}^* \cap T\Delta_{1,i}^* \cap \ldots\ T^{m-1}\Delta_{m-1,i}^*)$$

and

$$\Delta_j' = \Upsilon\,(\Delta_{0,j}^* \cap T\Delta_{1,j}^* \cap \ldots \cap T^{m-1}\Delta_{m-1,j}^*)$$

be two parallelogams in Ω'. We put $\pi_{ij} = 1$ if and only if $\Delta_{p,i}^* = \Delta_{p+1,j}^*$ for all $0 \leqslant p \leqslant m-2$ and the set

$$\Upsilon\,(\Delta_{0,i}^* \cap T\Delta_{1,i}^* \cap \ldots \cap T^{m-1}\Delta_{m-1,i}^* \cap T^m\Delta_{m-1,j}^*)$$

is non-empty. Otherwise we put $\pi_{ij} = 0$. Lemma 7.2 implies that for $\pi_{ij} = 1$ the intersection $\Delta_j' \cap T\Delta_i'$ is regular, but the converse need not hold. (It can be shown that for non-degenerate parallelograms the converse is true: if $\nu(\Delta_i') > 0$, $\nu(\Delta_j') > 0$, and the intersection $\Delta_j' \cap T\Delta_i'$ is regular, then $\pi_{ij} = 1$. Thus, our modification of the intersection matrix concerns degenerate parallelograms only.) In the sequel we will consider only the TMC (Σ_Π, θ) constructed using the new matrix Π.

This TMC has the following basic properties:

***Assertion* 7.3.** a) *Each* $x \in M_2$ *has at least one coding* $\sigma(x) \in \Sigma_\Pi$;

b) *for the points* $x \in M_2$ *such that* $T^k x \notin \tilde{\partial}\eta$ *for all integers* k, *the coding* $\sigma(x)$ *is unique*;

c) *each* $x \in M_1$ *has at most four codings* $\sigma(x) \in \Sigma_\Pi$.

For the proof of the last assertion we must consider the four infinitesimal quarter-neighbourhoods of x into which the infinitesimal neighbourhood $\overset{\circ}{V}(x)$ of x is divided by the LUM $\gamma^u(x)$ and the LSM $\gamma^s(x)$, and verify that by the construction of Π each of these generates at most one coding $\sigma(x) \in \Sigma_\Pi$.

7.2. We compare the numbers of periodic points of the automorphism T and the TMC (Σ_Π, θ). For smooth hyperbolic systems the asymptotics of the numbers of periodic points of the system and its corresponding TMC are, as a rule, the same [3], [29]. We show that this coincidence of asymptotics also holds in our case.

Let P_n be the number of periodic points of T of period n (that is, the number of solutions of the equation $T^n x = x$), and let $P_n(\Pi)$ be the number of periodic points of period n of the symbolic system (Σ_Π, θ).

***Theorem* 7.4.** *There is a constant* $C = C(Q) < \infty$ *such that* $|P_n - P_n(\Pi)| < C$ for all $n \geqslant 1$.

The proof consists of a number of lemmas.

***Lemma* 7.5.** *Under condition* B *of Theorem* 1.1, *only finitely many periodic points lie in the set* $R_{-\infty,\infty}$.

Proof. For a billiard with infinite horizon, singular points can be periodic (see §4); there are however only finitely many such points. If $x \in R_{-\infty,\infty}$ is periodic and non-singular (Fig. 23), then it can easily be verified that infinitely many regular components of $R_{-\infty,\infty}$ pass through x, contradicting condition B of Theorem 1.1. The lemma is proved.

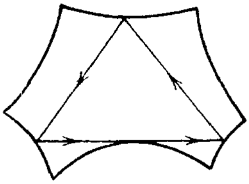


Fig. 23

***Corollary* 7.6.** *All except finitely many periodic points belong to M_2.*

***Lemma* 7.7.** *A point $x \in M_2$ is periodic if and only if every coding $\sigma(x)$ of it is periodic. If $T^k x \notin \tilde{\partial}\eta$ for all integers k, then the period of x equals the period of $\sigma(x)$.*

Proof. Suppose the coding $\sigma(x)$ is periodic. Then from the relation $\Phi \circ T = \theta \circ \Phi$ (see 1.2) it follows that x is periodic. Further, if a periodic point x of period k had a non-periodic coding $\sigma(x)$, it would have infinitely many codings $\sigma(x)$, $\theta^k\sigma(x)$, $\theta^{2k}\sigma(x)$, ..., contradicting Assertion 7.3c). Finally, if $T^k x \notin \tilde{\partial}\eta$ for all integers k, then x has a unique coding $\sigma(x)$ and from the construction of $\sigma(x)$ it follows that x and the sequence $\sigma(x)$ have the same period.

***Lemma* 7.8.** *Only finitely many periodic points lie in $\tilde{\partial}\eta$.*

The proof follows immediately from relation (5.7).

By combining Lemma 7.5, Corollary 7.6, Lemma 7.7, and Lemma 7.8 we obtain Theorem 7.4.

7.3. Estimates for the number of periodic points.

A way of computing the number of periodic points of a TMC is as follows (see, for example, the surveys [1], [2]).

***Proposition* 7.9.** *For any $n \geqslant 1$ we have $P_n(\Pi) = \operatorname{tr}\Pi^n$.*

Here Π^n denotes the nth power of the matrix Π. If Π is a countably-infinite matrix, then in general Π^n can contain infinite elements, and hence it is possible that $P_n(\Pi) = \infty$ for some n. However, in our case this does not happen:

***Proposition* 7.10.** *There is a constant $A_0 = A_0(Q) < \infty$ such that $P_n(\Pi) \leqslant A_0^n$ for all $n \geqslant 1$.*

This proposition follows from Theorem 7.5 and results in [30], in which an exponential upper bound is obtained for the number of periodic points in semiscattering billiards of arbitrary dimensions (without using Markov partitions).

On the other hand we can indicate a finite collection of parallelograms $\Omega'_N = \{\Delta'_{i_1}, \ldots, \Delta'_{i_N}\}$ in Ω' and distinguish in Π the finite $N \times N$ submatrix Π_N of entries $\pi_{i_p i_q}$, $1 \leqslant p, q \leqslant N$, corresponding to the chosen parallelograms. We denote by $P_n(\Pi_N)$ the number of periodic sequences $\sigma \in \Sigma_\Pi$ of period n and consisting of the symbols $i_1, \ldots, i_N$ only. Clearly $P_n(\Pi_N) \leqslant P_n(\Pi)$ for all $n \geqslant 1$. In the sequel we will use the notions of a decomposing, a periodic, and a primitive matrix (see [2]).

***Lemma* 7.11.** *For each $\varepsilon > 0$ there is an $N = N(\varepsilon)$ and a collection of non-degenerate parallelograms $\Omega'_N = \{\Delta'_{i_1}, \ldots, \Delta'_{i_N}\}$ such that*

a) $\mu\ (\Delta'_{i_1} \cup \ldots \cup \Delta'_{i_N}) > 1 - \varepsilon$;

b) *the corresponding matrix Π_N is non-decomposing*;

c) *for sufficiently small ε the matrix Π_N is non-periodic.*

Assertions a), b) readily follow from the ergodicity of T. Periodicity of Π_N for small ε contradicts the mixing property of T. The ergodicity and mixing of T were proved in [10], [4] (see also the simpler proof in [13], [24]). It is well known that a non-decomposing non-periodic matrix of zeros and ones is primitive (that is, some power of it does not contain zeros) [2], which implies the following result.

***Corollary* 7.12.** *There are constants $A_1 = A_1(Q) > 1$ and $n_0 \geqslant 1$ such that for sufficiently small $\varepsilon > 0$ the matrix Π_N, constructed in Lemma* 7.11, *satisfies the estimate $P_n(\Pi_n) > A_1^n$ for all $n \geqslant n_0$.*

Thus the following theorem has been proved.

***Theorem* 7.13.** *There are $A_1 = A_1(Q) > 1$ and $n_0 \geqslant 1$ such that $P_n(\Pi) > A_1^n$ and $P_n > A_1^n$ for all $n \geqslant n_0$.*

This also implies estimates for the number of periodic trajectories of a flow $\{S^t\}$. Let P_T be the number of (closed) periodic trajectories of the flow $\{S^t\}$ whose length does not exceed T. Then we have the following result.

***Theorem* 7.14.** *There are $B_0 = B_0(Q) < \infty$, $B_1 = B_1(Q) > 1$, and $T_0 > 0$ such that $B_1^T \leqslant P_T \leqslant B_0^T$ for all $T > T_0$.*

In the case of a finite horizon this theorem is a direct consequence of Assertions 7.10, 7.13, and 2.3. For an infinite horizon we note that for any $\varepsilon > 0$ the function $\tau^+(x)$ is bounded on the parallelograms $\Delta'_{i_1}, \ldots, \Delta'_{i_N}$ constructed in Lemma 7.11 by a constant $\tau_{\max}(\varepsilon) < \infty$, and subsequently use Corollary 7.12.

Note that a precise asymptotics of P_T has been obtained for certain classes of scattering billiards (see [28]). Also note that for billiards in polygons the rate of growth of the number of periodic points is less than exponential [23].

By Poincaré's recurrence theorem [7], any state i in a TMC (Σ_Π, θ) corresponding to a non-degenerate parallelogram $\Delta_i \in \Omega'$ is recurrent (that is, there are $i_1, \ldots, i_k$ such that $\pi_{ii_1} = \pi_{i_1 i_2} = \ldots = \pi_{i_{k-1} i_k} = \pi_{i_k i} = 1$. Hence any non-degenerate parallelogram contains at least one periodic point. Together with the arbitrariness of ε in the statement of Theorem 1.1 this proves the following result.

***Theorem* 7.15.** *The periodic points are everywhere dense in M_1. The* (*closed*) *periodic trajectories are everywhere dense in M.*

This theorem has a simpler proof, not using Markov partitions (it was communicated to us by Ya.B. Pesin).

§8. Domains with smooth boundary

In this section we consider billiards with hyperbolic behaviour in domains with smooth boundary. In the case of billiards on a torus the boundary ∂Q can be arbitrarily smooth and even analytic (for example, when ∂Q is a circle). When $Q \subset \mathbb{R}^2$ there is no example known of a billiard with hyperbolic behaviour and with boundary of smoothness exceeding C^1. It is possible that such billiards do not exist at all, but up to now this has not been proved. A substantiation of this hypothesis is a result of V.F. Lazutkin about the existence of a caustic for billiards in a convex plane domain bounded by a sufficiently smooth curve.

In the ergodic theory of billiards (and, in particular, in this article), one always considers the case when ∂Q consists of finitely many curves of smoothness class at least C^3 and having sign-definite curvature (see §1). Hence, if Q is simply-connected, its boundary has smoothness at most C^1. The interior boundaries in a plane domain Q can have arbitrarily high smoothness. We show that if ∂Q has smoothness C^1, then in Theorem 1.1 the conditions A, B are superfluous, and hence we need not require their satisfaction in Theorems 6.1, 6.2. For condition A (or A′) this is obvious, since if ∂Q has smoothness C^1, then all interior angles between regular boundary components are equal to π. Condition B (or B′) need not hold even in this case (for example, the boundary of the domain Q depicted in Fig. 23 can be smoothed in such a way that a side of the triangle will touch the scattering component of ∂Q as before). It turns out, however, that for a C^1-smooth boundary ∂Q in the construction of a Markov partition this condition can be circumvented. More precisely, we have the following result.

***Theorem* 8.1.** *Suppose that a domain Q generating a two-dimensional scattering billiard has boundary of smoothness class at least C^1. Then for every $\varepsilon > 0$*

there is a countable Markov partition, the diameters of all elements of which do not exceed ε.

Moreover, analogues of Theorems 6.1 and 6.2 are true.

***Theorem* 8.2.** *Suppose that a domain Q inducing a semiscattering billiard has C^1-smooth boundary and satisfies condition* B′. *Then for every $\varepsilon > 0$ we can construct a Markov partition with elements of diameter at most* ε.

***Theorem* 8.3.** *Suppose that a domain Q has boundary ∂Q of class C^1 whose focussing part consists of rays only (and is not empty) and that Q satisfies Condition* B′. *Then for every $\varepsilon > 0$ there is a Markov partition with elements of diameter at most* ε.

A basic role in the proofs of Theorems 8.1 – 8.3 is played by the following lemma.

***Lemma* 8.4.** *Let the domain Q have boundary of class C^1. Then there is an $a_0 > 0$ such that for any $x \in M_1$ and any integer m the number of curves in $R_{0,m}$ ($R_{0,-m}$) passing through x does not exceed $a_0|m|$.*

Proof. First of all, note that on the set $M_1 \setminus T^{-1}S_0$ ($M_1 \setminus TS_0$) the map T (T^{-1}) is continuous. Hence at most two discontinuity curves belonging to R_1 (R_{-1}) pass through a point $x \in M_1 \setminus T^{-1}S_0$ ($x \in M_1 \setminus TS_0$). Further, consider a rectilinear segment that is part of some billiard trajectory and that touches the boundary ∂Q at least twice (at distinct points of M_1). The set of all such segments will be denoted by D. It is easily seen that $D = D_\mathrm{p} \cup D_\mathrm{n}$, where D_p is the collection of rectilinear segments in Q that are parts of periodic trajectories of the flow generated by the billiard in Q, and $D_\mathrm{n} = D \setminus D_\mathrm{p}$.

Note that D does not include periodic trajectories that touch the boundary ∂Q under every reflection; these correspond to fixed points of T. This situation corresponds to the case of infinite horizon, which was investigated in detail in §4. In particular, such points lie on S_0, and through them there passes one curve belonging to S_1 and one curve belonging to S_{-1}.

The set D consists of finitely many segments, denoted by b_0. Moreover, if some trajectory passes twice through one of these segments, then this trajectory is periodic. Correspondingly, the given segment lies in D_p. Let p_0 be the maximum number of points of tangency that the segments in D have with ∂Q.

***Lemma* 8.5.** *For any point $x \in M_1$ the collection of discontinuity curves belonging to $T^m R_0$, for some fixed m, does not exceed $2b_0(p_0 + 1)$.*

Lemma 8.4 follows immediately from Lemma 8.5 by putting $a_0 = 2b_0(p_0 + 1)$.

We will now formulate a more general assertion concerning the structure of discontinuity curves. For any $x \in M_1 \setminus S_0$ we put $n_+(x) = \min\{n > 0 : x \in S_n\}$ ($n_-(x) = \min\{n > 0 : x \in S_{-n}\}$). Lemma 2.8 and the general properties of T listed in 2.1 – 2.3 imply the following result.

Lemma 8.6. *For any multiple point $x \in M_1 \backslash S_0$ and any $m > 0$ there is a neighbourhood $U(x)$ of x such that:*

a) *the closure $\overline{U(x)}$ does not contain multiple points belonging to $R_{-m,m}$, other than x;*

b) *among the discontinuity curves passing through x there is a unique curve $\Gamma_+(x) \in R_{n_+(x)}$ $(\Gamma_-(x) \in R_{-n_-(x)})$ such that Γ_+ (Γ_-) divides $U(x)$ into two semineighbourhoods $U_1^+(x)$ and $U_2^+(x)$ (respectively, $U_1^-(x)$ and $U_2^-(x)$);*

c) *all decreasing (increasing) discontinuity curves containing x, except Γ_- (Γ_+), intersect only one of $U_1^+(x)$ or $U_2^+(x)$ (respectively, $U_1^-(x)$ or $U_2^-(x)$).*

In the case when ∂Q contains one scattering component, Lemma 8.5 follows immediately from Lemma 8.6, by induction with respect to m.

We now consider the general case, in which ∂Q contains arbitrarily (finitely) many scattering components. First note that if the trajectory of a multiple point x does not intersect D, then the proof of Lemma 8.5 does not differ, for all multiple points on this trajectory, from the case when ∂Q^+ contains a unique component.

So it remains to consider the case when the trajectory of a multiple point x contains a segment that touches two distinct scattering boundary components. Here again two cases are possible, depending on whether $x \in D_{\mathrm{p}}$ or $x \in D_{\mathrm{n}}$. First, let $x \in D_{\mathrm{p}}$, that is, the trajectory of x is periodic. We denote the length of the corresponding period by $p(x)$. Then for arbitrary integers $k_1 > 0$, $k_2 \geqslant 0$ the number of discontinuity curves passing through x and belonging to $R_{k_1 p(x)+k_2}$ coincides with the number of discontinuity curves passing through x and belonging to $R_{p(x)}$. Further, when passing along any (regular) segment in D there emerges (at the point in the phase space M_1 corresponding to the end point of this segment) a number of discontinuity curves (belonging to R_1); this number does not exceed the number of components of ∂Q that are touched by this segment, plus one (because the initial point of the segment can be a singular point of ∂Q). This implies the estimate in Lemma 8.5 for $x \in D_{\mathrm{p}}$.

Finally, let the trajectory of x be non-periodic, that is, $x \in D_{\mathrm{n}}$. Then this trajectory passes at most once through each segment in D_{n}. When passing along each such segment, at its end point at most $p_0 + 1$ discontiniuty curves (belonging to S_0) are "glued", as was shown above. Thus, when the positive semitrajectory of x passes through all segments in D_{n}, at most $(p_0 + 1)b_0$ new discontinuity curves emerge. The remaining part of this semitrajectory consists of segments that can touch at most one scattering component of ∂Q, and in this case, as was shown above, the number of discontinuity curves passing through a given point and belonging to a fixed set R_m does not increase when m becomes larger. Thus Lemma 8.5 is proved.

Using Lemma 8.4, the construction of a Markov partition in the cases governed by the conditions of Theorems 8.1 – 8.3 is carried out completely similar to Theorems 1.1, 6.1, 6.2, respectively. Here "completely similar" literally means the following: all geometric constructions are unchanged, the distinction lies only in the choice of the constants determining the dimensions of the corresponding geometric objects.

The main difference from Theorems 1.1, 6.1, 6.2 is that instead of $K_0 = K_0(Q)$, under the conditions of Theorems 8.1–8.3 we have a_0m. Hence at the end of 3.1 we must put $\lambda_1 = (200a_0m)^{-1}$. Further, formula (3.2) takes the form

$$\Lambda_m \geqq \bar{c}^{-1} (200a_0m)^2. \tag{8.1}$$

This inequality holds for all sufficiently large m, since Λ_m grows exponentially with m. Finally, the right-hand sides of inequalities (5.3) and (5.4) must be multiplied by m, after which their proof remains the same.

References

[1] V.M. Alekseev, *Simbolicheskaya dinamika. Odinnadtsataya matematicheskaya shkola* (Symbolic dynamics. Eleventh mathematical school), Izdat. Inst. Mat. Akad. Nauk Ukrain. SSR, Kiev 1976. MR **57** # 4249.

[2] ——— and M.V. Yakobson, *Simbolicheskaya dinamika i giperbolicheskie dinamicheskie sistemy*: supplement to [3], pp. 196–240.
Translation: Symbolic dynamics and hyperbolic dynamic systems, Phys. Report **75** (1981), 287–325. MR **82j**:58093.

[3] R. Bowen, *Metody simbolicheskoi dinamiki*: *Sb. statei* (Methods of symbolic dynamics: Collection of articles), Mir, Moscow 1979.

[4] L.A. Bunimovich and Ya.G. Sinai, The fundamental theorem of the theory of scattering billiards, Mat. Sb. **90** (1973), 415–431. MR **51** # 3395.
= Math. USSR-Sb. **19** (1973), 407–423.

[5] ———, On billiards close to dispersing, Mat. Sb. **95** (1974), 49–73. MR **49** # 7422.
= Math. USSR-Sb. **23** (1974), 45–67.

[6] ———, On decrease of correlation in dynamical systems with chaotic behaviour, ZhETF **89** (1985), 1452–1470.

[7] I.P. Kornfel'd, Ya.G. Sinai, and S.V. Fomin, *Ergodicheskaya teoria*, Nauka, Moscow 1980.
Translation: Ergodic theory, Springer-Verlag, New York 1982. MR **83a**:28017.

[8] Ya.G. Sinai, Markov partitions and *Y*-diffeomorphisms, Funktsional. Anal. i Prilozhen. **2**:1 (1968), 64–89. MR **38** # 1361.
= Functional Anal. Appl. **2** (1968), 61–82.

[9] ———, Construction of Markov partitions, Funktsional. Anal. i Prilozhen. **2**:3 (1968), 70–80. MR **40** # 3591.
= Functional Anal. Appl. **2** (1968), 245–253.

[10] ———, Dynamical systems with elastic reflections, Uspekhi Mat. Nauk. **25**:2 (1970), 141–192. MR **43** # 481.
= Russian Math. Surveys **25**:2 (1970), 137–189.

[11] ———, Billiard trajectories in a polyhedral angle, Uspekhi Mat. Nauk **33**:1 (1978), 229–230. MR **58** # 7733.
= Russian Math. Surveys **33**:1 (1978), 219–220.

[12] Ya.G. Sinai, Ergodic properties of a Lorentz gas, Funktsional. Anal. i Prilozhen. **13**:3 (1979), 46–59. MR **81b**:28018.
= Functional Anal. Appl. **13** (1979), 192–202.

[13] ——— and N.I. Chernov, Ergodic properties of certain systems of two-dimensional discs and three-dimensional balls, Uspekhi Mat. Nauk. **42**:3 (1987), 153–174. MR **89c**:58097.
= Russian Math. Surveys **42**:3 (1987), 181–207.

[14] *Sovremennye problemy matematiki: Fundamental'nye napravleniya* (Current problems in mathematics: Fundamental directions), vol. 2, VINITI, Moscow 1985.

[15] A.N. Khovanskii, *Prilozheniya tsennykh drobei i ikh obobshchenii k voprosam priblizhennogo analiza*, Gostekhizdat, Moscow 1956.
Translation: A.N. Hovanskii, The application of continued fractions and their generalizations to problems in approximation theory, Noordhoff, Groningen 1963. MR **27** # 6058.

[16] R. Adler and B. Weiss, Entropy, a complete metric invariant for automorphisms of the torus, Proc. Nat. Acad. Sci. USA **57** (1967), 1573–1576. MR **35** # 3031.

[17] L.A. Bunimovich, On the ergodic properties of nowhere dispersing billiards, Comm. Math. Phys. **65** (1979), 295–312. MR **80h**:58037.

[18] ——— and Ya.G. Sinai, Markov partitions for dispersed billiards, Comm. Math. Phys. **73** (1980), 247–280. MR **82e**:58059.

[19] ——— and ———, Statistical properties of Lorentz gas with periodic configuration of scatters, Comm. Math. Phys. **78** (1981), 479–497. MR **82m**:82007.

[20] ——— and ———, Markov partitions for dispersed billiards (Erratum), Comm. Math. Phys. **107** (1986), 357–358. MR **87m**:58090.

[21] G. Gallavotti and D. Ornstein, Billiards and Bernoulli schemes, Comm. Math. Phys. **38** (1974), 83–101. MR **50** # 7480.

[22] A. Katok, J.-M. Strelcyn, F. Ledrappier, and F. Przytycki, Invariant manifolds, entropy and billiards; smooth maps with singularities, Lecture Notes in Math. **1222** (1986). MR **88k**:58076.

[23] ———, The growth rate for the number of singular and periodic orbits for a polygonal billiard, Comm. Math. Phys. **111** (1987), 151–160. MR **88g**:58162.

[24] A. Krámli, N. Simányi, and D. Szász, A transversal fundamental theorem for semi-dispersing billiards, Comm. Math. Phys. (in press).

[25] I. Kubo, Perturbed billiard systems. I, Nagoya Math. J. **61** (1976), 1–57. MR **55** # 6486.

[26] Y.-E. Levy, A note on Sinai and Bunimovich's Markov partition for the billiard, Preprint, Centre de Phys. Theor., Paris 1986.

[27] R. Markarian, Billiards with Pesin region of measure one, Comm. Math. Phys. **118** (1988), 87–97. MR **89m**:58122.

[28] T. Morita, The symbolic representation of billiards without boundary condition, Preprint Tokyo Inst. Technol. Tokyo 1989.

[29] W. Parry and M. Pollicott, An analogue of the prime number theorem for closed orbits of Axiom A flows, Ann. of Math. (2) **118** (1983), 573–591. MR **85i**:58105.

[30] L. Stojanov, An estimate from above of the number of periodic orbits for semi-dispersed billiards, Comm. Math. Phys. **124** (1989), 217–227.

[31] M. Wojtkowski, Invariant families of cones and Lyapunov exponents, Ergod. Theory Dyn. Syst. **5** (1985), 145–161. MR **86h**:58090.

[32] ———, Principles for the design of billiards with non-vanishing Lyapunov exponents, Comm. Math. Phys. **105** (1986), 391–414. MR **87k**:58165.

[33] V.F. Lazutkin, The existence of caustics for a billiard problem in a convex domain, Izv. AN SSSR Ser. Mat. **37** (1973), 188–223. MR **48** # 6561.
= Math. USSR-Izv. **7** (1973), 185–214.

Translated by R.A.M. Hoksbergen

Institute of Oceanology of the USSR Academy of Sciences
Institute of Theoretical Physics of the USSR Academy of Sciences
United Institute of Nuclear Research

Received by the Editors 3 January 1990