MODERN THEORY OF BILLIARDS – AN INTRODUCTION

G. GALLAVOTTI
Istituto Matematico,
Università di Roma,
Rome, Italy

Abstract

MODERN THEORY OF BILLIARDS – AN INTRODUCTION.

This paper presents an elementary introduction to the notions and ideas involved in the proof of the ergodicity of billiards.

1. BILLIARDS, DO YOU RECOGNIZE IT?

Let $N = [0,1] \times [0,1]$ mod 1 be a two-dimensional torus. Let Q_1, Q_2, ..., Q_m be m closed convex regions in N. Assume that Q_i is a C^2-smooth curve with never vanishing curvature and assume also that $Q_i \cap Q_j = \phi$, $i \neq j$.

Let V be the Riemannian manifold (with boundary) obtained by taking out of N the interior of Q_1, ..., Q_m; the metric on V is the one inherited by N (i.e. $ds^2 = dx^2 + dy^2$).

Let T_1V be the unitary tangent bundle of V.

The elements of T_1V can be thought of as applied vectors or as couples (q, θ), $q \in V$ and $0 \leq \theta \leq 2\pi$, where $q \in V$ is the point of application of the vector and θ is its angle with a fixed direction.

Define on T_1V a probability measure $\mu(dq\, d\theta) = (\text{const}) \cdot dq\, d\theta$ and a flow S_t, $-\infty < t < +\infty$, S_t: $T_1V \to T_1V$. This flow is defined almost everywhere and is constructed by means of the billiards rule as is shown in Fig. 1 (where the case $t > 0$ is considered):

One clearly recognizes in the dynamical system (T_1V, S_t, μ) an "ordinary" game of billiards with one ball on a periodic table with m obstacles. The measure μ is preserved by S_t.

Now, the following theorem holds:

Theorem (Sinai): The dynamical system (T_1V, S_t, μ) is ergodic and, more precisely, a K-system.

To attack the problem of the proof of the theorem, first remark that the flow S_t can be more simply described through a "section" of itself. To discuss this point and the following ones, let us choose, from now on, a geometrically simple game of billiards, i.e. let us consider the case in which there is just one circular Q (with radius R).

Let M be the manifold of the applied vectors with point of application on ∂Q and pointing towards the interior of Q. An element $x \in Q$ can be described by two co-ordinates $x = (r, \varphi)$ where $0 \leq r < 2\pi R$ is the clockwise abscissa, over ∂Q, of the point of application of x and φ is the angle which the vector x subtends with the outer normal to ∂Q in r: $\pi/2 \leq \varphi \leq 3\pi/2$. The range of values of φ reflects the fact that M consists of vectors "heading" against ∂Q.

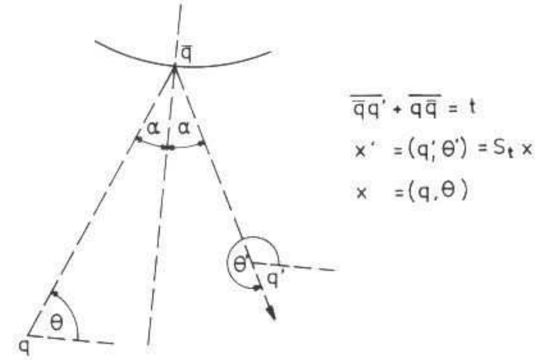

FIG. 1. Construction of flow.

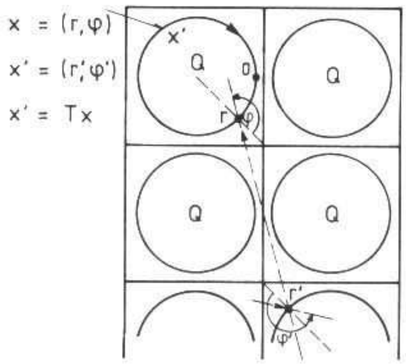

FIG. 2. Transformation T.

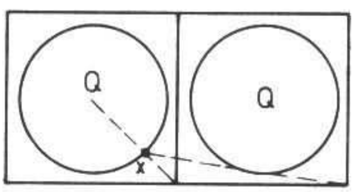

FIG. 3. x as a singular point for T.

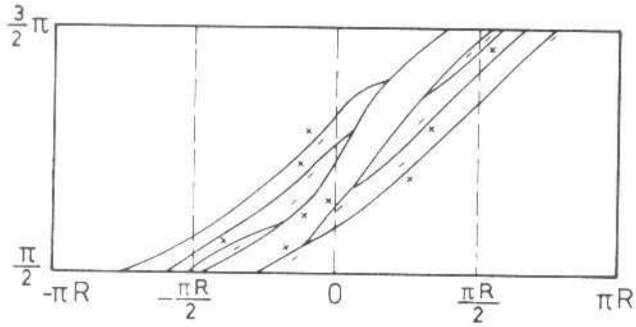

FIG. 4. Some elements of two families of smooth (C^1)-curves.

Let us define a transformation T: $M \to M$ as follows: choose $x \in M$, think of it as the vector describing a billiard ball hitting ∂Q and follow this ball back in time in the past until, at time $\tau(x) < 0$, it hits again ∂Q in a point $x' = Tx$ (Fig. 2).

It is quite easy to check that the measure $\nu(dr\, d\varphi) = -(\text{const}) \cos\varphi\, dr d\varphi$ is preserved by T. The mapping T is only almost everywhere defined.

Therefore, (M, T, ν) is a dynamical system and it can be easily imagined that the properties of (M, T, ν) could be translated into properties of (T_1V, S_t, μ): notice that ν is the natural projection of μ on M. For this reason, we shall concentrate our attention on (M, T, ν) without insisting on how to translate the information on it into information on (T_1V, S_t, μ).

The system (M, T, ν) is still non-smooth since it has a boundary $S = (x/x \in M,\ \varphi(x) = \pi/2,\ (3/2)\pi)$ (there is no boundary in r since this coordinate is periodic) and singularity points which consist in the set $T^{-1}S$. In Fig. 3, x is a singular point for T. One easily finds that the singularity points lie on smooth (C^1)-curves divided into 8 families (if the radius R of the obstacle is not too small compared with the side of the torus). See Fig. 4 where some elements of two such families are drawn.

Clearly, the curves in this figure are also discontinuity curves for $\tau(x)$, which, however, is continuous on them from one side (denoted + in the figure).

A region B which overlaps with the singularity lines is blown into pieces by T (roughly as many pieces as the number of lines intersected by B).

Sinai's idea is to prove that, in spite of its horrifying non-smoothness, (M, T, ν) behaves much in the same way as a C-system (Anosov diffeomorphism).

2. A RELATED PROBLEM

In this section, we shall first discuss the idea behind the proof of the ergodicity of C-systems in a very simple case.

This proof will be used to illustrate the necessity and the meaning of the various mathematical objects that have to be introduced to cope with the problem of billiards (as well as with the theory of the C-systems).

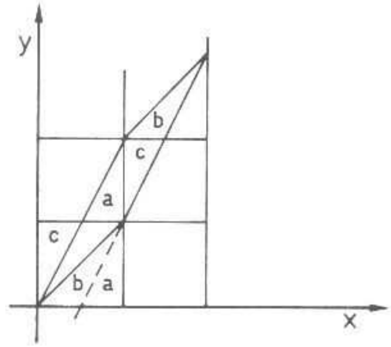

FIG. 5. Comparison system.

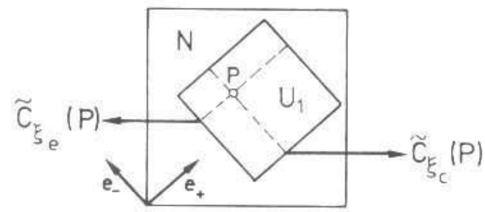

FIG. 6. Definition of $\widetilde{C}_{\xi_e}$ (P) and $\widetilde{C}_{\xi_c}$ (P).

The comparison system is going to be the much publicized – in many other papers of the Proceedings – authomorphism of the torus N = [0,1] × [0,1] mod 1 defined by $\tau(x, y) = (x+y,\ x+2y)$ mod 1; see Fig.5. Let e_+, e_- be the two eigenvectors of the matrix $\binom{1\ 1}{1\ 2}$ and let λ_+, $\lambda_- = \lambda_+^{-1} < 1$ be the two eigenvalues. It is easy to check that the directions e_+, e_- are irrational, i.e. the tangent of the angle of these two directions with the x-axis are irrational numbers. The first claim is that the system (N, T, ν = dxdy) is a C-system and that the contracting and expanding foliations ξ_c and ξ_e consist of parallel straight lines (wrapped on the torus and parallel to e_+ and e_-). Let us note that the irrationality of the directions e_+ and e_- implies that each leaf of the foliation is dense in N. In fact, let $(x, y) \in N$ be a point and let $C_{\xi_c}(x, y)$ be the straight line parallel to e_- and passing through (x, y). Let $(x', y') \in C_{\xi_c}(x, y)$; it is clear that $\tau^n(x, y)$ and $\tau^n(x', y')$ will be at a distance not exceeding their distance counted along the line $\tau^n C_{\xi_c}(x, y)$ which is

$$d_{\tau^n C_{\xi_c}(x,y)}(\tau^n(x, y),\ \tau^n(x', y')) = \lambda_+^{-n}\, d_{C_{\xi_c}(x,y)}((x, y),\ (x', y'))$$

Similarly, one proves that the foliation ξ_e is contracting in the past (i.e. under τ^{-1}) and that the line elements of ξ_e locally expand in the past while the line elements of ξ_e locally expand in the future. The expansion and contraction coefficients are always λ_+ or λ_-. This fact, of course, is responsible for the possibility of an easy treatment and visualition of the above system.

We now show that (N, τ, ν) is ergodic using a proof which, though certainly not the simplest, is extremely instructive since it contains the main idea, due to Hopf, at the base of the proof of the ergodicity of C-systems and billiards.

3. PROOF OF THE ERGODICITY OF $\binom{1\ 1}{1\ 2}$.

Let f be a continuous function $f \in C(N)$. Then, by the Birkhoff theorem, the limits

$$f^+(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} f(\tau^i x)$$

$$f^-(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} f(\tau^{-i} x)$$

exist almost everywhere and are almost everywhere equal: $f^+(x) = f^-(x)$ for $x \in U$, $\nu(U) = 1$.

We have to prove that for all $f \in C(N)$ the functions $f^+(x)$ and $f^-(x)$ are constant almost everywhere. This, of course, implies the ergodicity of (N, τ, ν).

Consider a covering of N with squares $U_1, U_2, \ldots$ with sides not exceeding $1/\sqrt{2}$ and parallel to e_+ and e_-. We shall choose the squares in such a way that they overlap in chain (i.e. if $P, Q \in N$ there is a chain $U_{i_1}, U_{i_2}, \ldots, U_{i_r}$ such that $\nu(U_{i_j} \cap U_{i_{j+1}}) > 0$ and $P \in U_{i_1}$, $Q \in U_{i_r}$).

The family $\{U_i\}$ can be chosen to contain a finite number of elements.

Clearly, it will be sufficient to prove that f^+ and f^- are constant on each U_i (almost everywhere).

So let us fix $f \in C(N)$ and show that its averages in the future and in the past are almost everywhere constant on U_1, say.

If $P \in U_1$ let $C_{\xi_e}(P)$ be the expanding leaf through P and let $\widetilde{C}_{\xi_e}(P)$ be the connected part of $C_{\xi_e}(P) \cap U_1$ containing P; similarly, we define $\widetilde{C}_{\xi_c}(P)$, see Fig. 6.

In U_1, we use a system of orthogonal co-ordinates based on the vectors e_+ and e_-, which are parallel to the sides of U_1. If B is a measurable subset of $\widetilde{C}_{\xi_e}(P)$ or $\widetilde{C}_{\xi_c}(P)$, we denote by $|B|$ its Lebesgue measure with respect to the abscissa or the line; hence, in particular, $|\widetilde{C}_{\xi_c}(P)| = |\widetilde{C}_{\xi_c}(P)| =$ length of the side of U_1.

Let us now consider the set

$$V_e = \left(x/x \in U_1,\ \frac{|\widetilde{C}_{\xi_e}(x) \cap U_1 \cap U|}{|\widetilde{C}_{\xi_e}(x)|} = 1 \right)$$

where $U = (x/x \in N,\ f^+(x)$ and $f^-(x)$ exist, $f^+(x) = f^-(x))$.

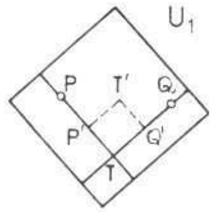

FIG. 7. P and Q with local contracting and expanding leaves.

Since, by the Birkhoff theorem, $\nu(U) = 1$ (see above) it follows by Fubini's theorem that V_e has full measure i.e. $\nu(V_e) = \nu(U_1)$.

Proof: $\nu(U_1) = \nu(U \cap U_1)$, hence if s and s' are the co-ordinates of $P \in U_1$, $P = (s, s')$, we find

$$\nu(U_1) = \int ds' \int ds\, \chi_{U_1 \cap U}(P) = \int ds' \left| \widetilde{C}_{\xi_e}(0, s') \cap U \cap U_1 \right|$$

which implies that for almost all s', we have $\left|\widetilde{C}_{\xi_e}(0, s') \cap U \cap U_1\right| = \left|\widetilde{C}_{\xi_e}(0, s')\right|$ and, again by Fubini's theorem, this implies that $\left|\widetilde{C}_{\xi_e}(x) \cap U \cap U_1\right| = \left|\widetilde{C}'_{\xi_e}(x)\right|$ for almost all $x \in U_1$: in words, we can say that almost all points $x \in U_1$ are such that the line $\widetilde{C}_{\xi_e}(x)$ lies almost entirely in $U \cap U_1$.

Similarly, we can define V_c and show that $\nu(V_c) = \nu(U_1)$.

Let us now consider the set

$$V = U \cap U_1 \cap V_e \cap V_c$$

Clearly, V has full measure in U_1: $\nu(V) = \nu(U_1)$. It is, therefore, enough for our purposes to show that $f^+(x) = f^-(x) = \text{const}$ for $x \in V$.

Let $P, Q \in V$, then draw through P the local contracting leaf $\widetilde{C}_{\xi_c}(P)$ and through Q the local expanding leaf $\widetilde{C}_{\xi_e}(Q)$ (see Fig. 7).

The point T may or may not lie in V. In any case, it is possible to find a point $P' \in \widetilde{C}_{\xi_c}(P)$ and a point $Q' \in \widetilde{C}_{\xi_e}(Q)$ such that the points T', P', Q' are all in V (see Fig. 7). In fact, by construction, almost all points on the two lines $\widetilde{C}_{\xi_c}(P)$ and $\widetilde{C}_{\xi_e}(Q)$ lie in V (remember the choice of P and Q), hence as P' runs over $V \cap \widetilde{C}_{\xi_c}(P)$ and Q' over $V \cap \widetilde{C}_{\xi_e}(Q)$ the point T' spans a set of full measure which, therefore, certainly intersects V.

Now, the game is over; in fact, by construction:

$$d(\tau^n P, \tau^n P') \xrightarrow[n \to \infty]{} 0$$

$$d(\tau^{-n} P', \tau^{-n} T') \xrightarrow[n \to \infty]{} 0$$

$$d(\tau^n T', \tau^n Q') \xrightarrow[n \to \infty]{} 0$$

$$d(\tau^{-n} Q', \tau^{-n} Q) \xrightarrow[n \to \infty]{} 0$$

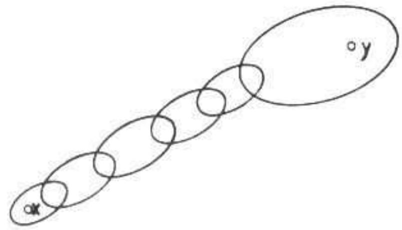

FIG. 8. Sets overlapping in a chain.

(with exponential speed). Hence, by the uniform continuity of f,

$$f(\tau^n P) - f(\tau^n P') \xrightarrow[n\to\infty]{} 0$$

$$f(\tau^{-n} P') - f(\tau^{-n} Q') \xrightarrow[n\to\infty]{} 0 \qquad \text{etc.}$$

Therefore, $f^+(P) = f^+(P')$; $f^-(P') = f^-(T')$; $f^-(T') = f^+(Q')$; $f^-(Q') = f^-(Q)$; but, by construction, it is also true that P, P', Q, Q', T' are in $V \subset U$ hence $f^+(P) = f^-(P)$; $f^+(P') = f^-(P')$; $f^+(T') = f^-(T')$; $f^+(Q') = f^-(Q')$; $f^+(Q) = f^-(Q)$. Hence, all the abovementioned values of f^+ and f^- coincide; in particular,

$$f^+(P) = f^-(Q) = f^-(P) = f^-(Q)$$

which means, by the arbitrarity of P and Q, that f^+ and f^- are constant on V (and, therefore, almost everywhere).

4. HOW TO GENERALIZE THE ABOVE ARGUMENTS

Obviously, in more general situations, things are not so nice and easy; nevertheless, the proof of ergodicity for the case of C-systems or even billiards systems proceeds along the same lines as the above proof of the ergodicity of $\binom{1\ 1}{1\ 2}$.

It is possible, in those cases, to construct a denumerable family of measurable sets $\{U_i\}$, forming a basis for the Borel sets, such that, given any two points x, y in a suitable set of measure 1, one can find a finite number of sets $U_{i_1}, U_{i_2}, \ldots, U_{i_m}$ overlapping in chain (i.e. $\nu(U_{i_j} \cap U_{i_{j+1}}) > 0$) such that $x \in U_{i_1}$ and $y \in U_{i_m}$ (see Fig. 8).

Furthermore, to each of these sets U_i the Hopf idea can be applied: in fact, roughly speaking, the sets U_i can be thought of as unions of pieces of leaves of a contracting foliation and, at the same time, as unions of pieces of leaves of an expanding foliation; furthermore, one can use a "system of co-ordinates" based on the hypersurfaces $C_{\xi_c} \cap U_i = \widetilde{C}_{\xi_c}$ and $C_{\xi_e} \cap U_i = \widetilde{C}_{\xi_e}$ and the measure of a set $B \subset U_i$ can be computed as a double integral on the product of the natural measures $d_{\widetilde{C}_{\xi_c}} \sigma$, $d_{\widetilde{C}_{\xi_e}} \sigma$ induced by the Riemannian metric on $\widetilde{C}_{\xi_c}$ or $\widetilde{C}_{\xi_e}$.

More precisely, it is possible to construct two measurable partitions $\widetilde{\xi}_c$ and $\widetilde{\xi}_e$ for each set U_i which are local contracting or local expanding leaves, and these two partitions are absolutely continuous with respect to each other (see below).

The above discussion should be understood as an intuitive anticipation of the precise definitions to be introduced in the next section.

5. MEASURABLE FOLIATIONS

Here we provide the precise definitions needed to fully understand the sentences of the last section.

Let (M, τ, ν) be a dynamical system.

Definition 1: a measurable set $U \subset M$ is said to be measurably partitioned by a k-dimensional foliation ξ if:

1) ξ is a partition of U.
2) The elements $\widetilde{C}_\xi \in \xi$ are open k-dimensional piecewise smooth manifolds homeomorphic to an open k-sphere[1].
3) If $\widetilde{\nu}$ denotes the restriction of ν to the subalgebra $\sum(\xi)$ of the Borel algebra in U consisting of the measurable sets that are unions of elements of ξ then

$$\nu(B) = \int_{\subset U} \widetilde{\nu}(d\widetilde{C}_\xi) \int_{\widetilde{C}_\xi \cap B} \rho_{\widetilde{C}_\xi}(u)\, d_{\widetilde{C}_\xi}\sigma \quad \forall B \subset U$$

where $\rho_{\widetilde{C}_\xi}(u) > 0$ almost everywhere with respect to the natural measure (surface measure) $d_{\widetilde{C}_\xi}\tau$ on the manifold.

Remark: the above double integral has to be understood in the usual sense, i.e.

$$\int_{\widetilde{C}_\xi \cap B} \rho_{\widetilde{C}_\xi}(u)\, d_{\widetilde{C}_\xi}\sigma$$

must make sense and must be an integrable function with respect to $\widetilde{\nu}$ for all measurable $B \subset U$.

Definition 2: If in U there are two measurable partitions ξ_1, ξ_2, we say that ξ_2 is absolutely continuous with respect to ξ_1 if every element of ξ_2 intersects every element of ξ_1 in just one point and if all $W \in \sum(\xi_2)$ such that $\nu(W) > 0$ are such that $\tau_{\widetilde{C}_{\xi_1}}(W \cap \widetilde{C}_{\xi_1} > 0$ for all $\widetilde{C}_{\xi_1} \in \xi_1$ apart for a set of $\widetilde{C}_\xi$'s which are, however, contained in a null set. In this case, we say $\xi_2 \ll \xi_1$.

---

[1] It is perhaps important to state explicitly what is meant by an open piecewise smooth k-dimensional manifold (see Ref. [1]).

A closed k-dimensional smooth submanifold of a manifold M is a C^2-smooth submanifold N homeomorphic to a closed k-sphere through a mapping which, in the neighbourhood of each point of N, is given in local co-ordinates by C^2 functions having a limit on the boundary of N together with their derivatives; furthermore, if $k > 2$, the boundary ∂N must consist of a finite number of closed smooth k-1 dimensional manifolds.

A closed submanifold is called a closed piecewise smooth submanifold if it consists of the union of a finite number of closed smooth submanifolds and if it is homeomorphic to a closed k-sphere.

An open piecewise smooth submanifold is a submanifold which is homeomorphic to an open k-sphere such that its closure is a closed piecewise smooth submanifold of the same dimensionality k.

Remark: if $\xi_1 \ll \xi_2$ and $\xi_2 \ll \xi_1$, then it is quite clear that, apart for a "bad" set of local leaves $\tilde{C}_{f_1}$ and $\tilde{C}_{f_2}$ contained in a null set, the intersections of any set W of full measure in U with $\tilde{C}_{f_1} \in \xi_1$ and $\tilde{C}_{f_2} \in \xi_2$ have full measure with respect to the natural measures $d_{\tilde{C}_{f_1}}\sigma$ and $d_{\tilde{C}_{f_2}}\sigma$, i.e.

$$\sigma_{\tilde{C}_{f_i}}(W \cap \tilde{C}_{f_i}) \equiv \sigma_{\tilde{C}_{f_i}}(\tilde{C}_{f_i})$$

if the $\tilde{C}_{f_i}$ are not "exceptional".

Definition 3: Let (M, ν, τ) be a dynamical system and let ξ be a measurable decomposition of a measurable set $U \subset M$. Then ξ is said to be "contracting" if for any two points x, y chosen almost everywhere (with respect to the natural measure $d_{\tilde{C}_f}\sigma$) on a leaf $\tilde{C}_f$ are such that $d(\tau^n x, \tau^n y) \xrightarrow[n\to\infty]{} 0$ with the possible exception of a set of $\tilde{C}_f$'s contained in a null set.

Similarly, we define an expanding measurable partition, i.e. in the above definition we replace τ by τ^{-1}.

From the above remarks and definitions we realize that (M, τ, ν) is going to be ergodic if it is possible to construct enough sets U admitting measurable foliations of contracting and dilating type which, furthermore, are absolutely continuous with respect to each other.

6. HEURISTIC CONSTRUCTION OF THE FOLIATIONS FOR BILLIARDS

Let us conclude by discussing how one can attack the problem of finding the contracting and dilating foliations in the case of billiards.

We shall only present heuristic arguments.

Suppose there is a contracting curve γ through $x \in M$. Then the mappings $T^{-1}, T^{-2}, \ldots$ must all be smooth on γ (remember that, by our conventions, T sends back into the past and T^{-1} sends into the future).

From geometric arguments, it is easily shown that if $\varphi = \varphi(r)$ is the equation of a smooth curve Γ and $\varphi' = \varphi'(r')$ is the equation of $T\Gamma$, then

$$\frac{d\varphi'}{dr'} = -\cos\varphi' \left(\frac{1}{R\cos\varphi'} + \cfrac{1}{\tau(r,\varphi) + \cfrac{1}{\cfrac{1}{\cos\varphi}\left(\frac{1}{R} - \frac{d\varphi}{dr}\right)}} \right)$$

Hence, if T^{-i} is smooth on γ for $i = 1, 2, \ldots$ we find, by repeated application of the above formula,

$$\frac{d\varphi}{dr}(x) = -\cos\varphi(x) \left(\frac{1}{R\cos\varphi(x)} + \cfrac{1}{\tau(T^{-1}x) + \cfrac{1}{\cfrac{2}{R\cos\varphi(T^{-1}x)} + \cfrac{1}{\tau(T^{-2}x) + \cfrac{1}{\cfrac{2}{R\cos\varphi(T^{-2}x)} + \ldots}}}} \right)$$

$$= -\kappa^c(x)\cos\varphi(x)$$

where $\kappa^c(x)$ is the function defined by the continued fraction inside the parentheses; it converges since $|\tau(T^{-i}x)| \geq \tau_0 > 0$ because $Q_i \cap Q_j = \phi$, $i \neq j$.

It is also not difficult to see from the above equation that the curve γ (if existing) must be such that the distance between the T^{-i} images of any two points on γ tends to zero as $i \to \infty$ (i.e. γ is actually a contracting curve).

The first real problem is to show that the above differential equation actually has a solution; this seems to be a difficult problem since it is easily realized that $\kappa^c(x)$ is discontinuous over a dense set and, on its set of continuity, it is not at all nicely behaved.

However, it is possible to prove that if the distance of $T^{-i}x$ from the singularity points of T does not tend to zero too fast as $i \to \infty$, then the equation for γ has a solution in a neighbourhood of x.

A similar construction provides the pieces of the expanding leaves.

The set of the x in M for which $T^{-i}x$ does not get "too" close to the singularity lines of T can be shown to have measure 1, so that the differential equation defining the contracting leaves has a solution for a set of initial data having full measure [3].

At this point one has to construct a family $\{U_i\}$ of measurable sets which are measurably decomposable by expanding and contracting foliations.

The constructions and proofs related to this point are a very nice piece of geometry and really at the heart of the theorem. They are rather similar to the analogous constructions encountered in the theory of the C-systems [2, 3]. However, let us note the basic conceptual difference between C-systems and billiards: for C-systems, each point x belongs to contracting and dilating leaves of foliation and, furthermore, the dependence of the leaf upon the point x is smooth. In the billiards case, we have a situation in which the leaves of both the contracting and dilating foliations cover M only almost everywhere; furthermore, they typically end on singularity lines of T or T^{-1} or iterates of them and, if $\gamma(x)$ is a leaf through x, its dependence on x is far from being smooth.

The interested reader is referred to the famous papers [1-3].

REFERENCES

[1] SINAI, Ya. Dynamical systems with countable Lebesgue spectrum, II: Am. Math. Soc. Trans. 2 (1968) 34.

[2] ANOSOV, D., SINAI, Ya., Some smooth ergodic systems, Russ. Math. Surv. 22 (1967) 103.

[3] SINAI, Ya., Systems with elastic reflections, Russ. Math. Surv. 25 (1970) 137.