Journal of Statistical Physics, Vol. 69, Nos. 1/2, 1992

Ergodic and Statistical Properties of Piecewise Linear Hyperbolic Automorphisms of the 2-Torus

N. I. Chernov[1,2]

Received January 22, 1992; final April 7, 1992

We study generic piecewise linear hyperbolic automorphisms of the 2-torus. We explain why the resulting dynamical system is ergodic and mixing and prove the exponential decay of correlations.

KEY WORDS: Dynamical systems; hyperbolicity; decay of correlations.

1. INTRODUCTION

Two distinct classes of hyperbolic dynamical systems with singularities have attracted attention in the last two to three decades. One of them is the class of piecewise smooth maps of the interval. We only mention the main results in this area: the construction of an absolutely continuous invariant measure, the proof of ergodicity, exactness, and an exponential decay of correlations.

Another interesting class is that of billiard systems with hyperbolic behavior. These systems are also proven to be ergodic, mixing, K- and B-systems.[22,23,13] However, only a subexponential bound for the decay of correlations has been established here.[5,8]

There is a deep analogy between these two classes of dynamical systems. Both have a dual nature: the hyperbolicity leads to exponential instability, but the singularities eventually destroy it. Nonetheless, the hyperbolicity in both cases overcomes the influence of singularities, yielding those ergodic and statistical properties of the systems. The only disturbing inequality between the stochastic properties of these two classes is that the correlations seem to decay more slowly in billiards than in one-dimensional

[1] Joint Institute for Nuclear Research, Dubna, P.O. Box 79, Moscow, Russia.
[2] Present address: Department of Mathematics, UCLA, Los Angeles, California 90024.

822/69/1-2-8

maps. Certain numerical experiments[1,9,25] confirm this conjecture, too. Attempts at qualitative explanations of the phenomenon in question have been made in refs. 12 and 25. Naturally, the slower decay of correlations was ascribed there to the influence of singularities.

However, the actual rate of the decay of correlations in billiards (exponential or subexpotential) is still unknown. This paper will shed more light on this problem. We will show that correlations in systems which are very close to billiards do decay exponentially fast. The hyperbolic properties of our systems are somewhat stronger than those of billiards, but the influence of singularities here seems to be the same. We hope our results can be extended to billiards, as well as to billiard-like Hamiltonian systems considered in refs. 11 and 26 (see also references given there). We also note that a recent numerical experiment[14] has revealed an exponential decay of correlations in a dispersing billiard system. If our conjecture fails, this would mean that the slower decay of correlations in billiards must be understood as a result of nonuniform hyperbolicity combined with specific singularities, and not as an effect of singularities alone.

The systems considered here have never been studied before, except for a particular case treated in a recent work[24] and some illustrative examples in ref. 19. Therefore we start with a short description of their hyperbolic and ergodic properties, including certain elements of Markov partitions. Then we will focus on the proof of the exponential decay of correlations.

2. DEFINITION OF THE SYSTEM

Let M be the standard 2-torus associated with the unit square $K=[0,1)\times[0,1)$. Let T_0 be a linear map of the plane defined by a 2×2 matrix A with $\det A=\pm 1$ and whose eigenvalues are real and not equal to ± 1. For instance,

$$A_1=\begin{pmatrix}2 & 1\\ 1 & 1\end{pmatrix} \quad \text{or} \quad A_2=\begin{pmatrix}1 & k\\ 1 & 1+k\end{pmatrix} \quad \text{with} \quad k>0$$

Denote $\Lambda>1$ and $\lambda=\Lambda^{-1}<1$ the moduli of the eigenvalues of A. The image T_0K is then a parallelogram with a unit area.

Now let us cut the parallelogram T_0K along several compact smooth curves which are either closed or have common endpoints. These curves divide the parallelogram into several pieces (domains). We assume that these pieces can be shifted and put together in such a way that they will make the square K again. As a result we obtain a piecewise linear transformation T of the torus M. We emphasize that only translations of the above parts of T_0K are admitted, while any distortion or rotation is prohibited. Certain generic conditions will be imposed on these cuttings below.

For example, let the matrix A have integer entries, like A_1 above. The projection of T_0K onto K then defines a smooth linear transformation of the torus. Now we can cut the torus into several horizontal strips (Fig. 1) and then shift each strip by the rule $(x, y) \to (x + a, y) \pmod 1$, where a depends on the strip. As another example, consider the above matrix A_2 with noninteger k. The projection of T_0K onto K now specifies a one-to-one piecewise linear *discontinuous* map of the torus M called the *sawtooth map*. In that case we do not need any more, artificial, cuttings. This last example has been studied in detail in ref. 24.

Clearly, the map T of the torus M preserves the Lebesgue measure and is one-to-one almost everywhere. Denote by S_+ and S_- the union of the discontinuity curves for the maps T^{-1} and T, respectively. For any positive integer n set $S_n = T^{n-1}S_+$ and $S_{-n} = T^{-n+1}S_-$; also set $S_0 = \varnothing$. For $m \leqslant n$ set $S_{m,n} = S_{n,m} = S_m \cup \cdots \cup S_n$. Clearly, T^n (T^{-n}) is undefined and discontinuous on S_{-n} (S_n). All the iterates of T are well defined on the subset $M_0 = M \setminus S_{-\infty,\infty}$ of the full measure. We call the union $S = S_+ \cup S_-$ the *singularity set*.

Each point $x \in M_0$ has two Lyapunov exponents $\chi_1 = \ln \Lambda > 0$ and $\chi_2 = \ln \lambda < 0$. The corresponding invariant subspaces E^u_x and (and E^s_x) are all parallel. We denote by E^u (E^s) the eigenspace of A which is parallel to all E^u_x (resp. E^s_x). The rate of expansion of E^u_x (and the rate of contraction of E^s_x) is constant throughout M_0: $\|DTv\| = \Lambda \|v\|$ for $v \in E^u_x$ and $\|DTv\| = \lambda \|v\|$ for $v \in E^s_x$ at any point $x \in M_0$. These are very strong hyperbolic properties which distinguish our systems from billiards.

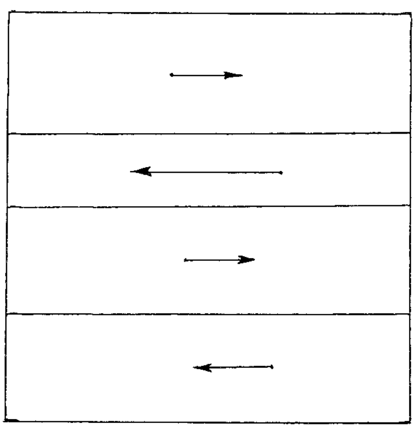

Fig. 1. An example of discontinuous toral automorphism.

According to Pesin's theory[21] (for systems with singularities see ref. 16) almost every point $x \in M$ has a local unstable manifold (LUM) $\gamma^u(x)$ and a local stable manifold (LSM) $\gamma^s(x)$, both passing through x. In our systems all LUMs are parallel rectilinear segments as are all LSMs. Due to the singularities these segments have a finite length, and there are plenty of arbitrary short LUMs and LSMs throughout M. Some points $x \in M_0$ (forming a set of null measure) do not even have an LUM $\gamma^u(x)$ or an LSM $\gamma^s(x)$ (or both). In this respect the situation here is the same as in billiards with hyperbolic behavior.

3. ERGODIC PROPERTIES

Recall that the singularity set S is a finite union of compact smooth curves in M. We now assume additional properties of S:

Property S1 (Double singularity). The intersection $S_m \cap S_n$ for $m \neq n$ is always a finite set of isolated points.

Property S2 (Weak regularity of singularities). The set of points $x \in S$ at with the tangent vector to S is parallel to E^u or to E^s has zero length.

We will also assume a stronger version of the latter property:

Property S3 (Strong regularity of singularities). The tangent vectors to the curves in S, including those at their endpoints, are nowhere parallel to E^s or to E^u.

Property S3 implies that the angles between E^u (E^s) and the tangent vectors to the curves in S are bounded away from zero. In that case we denote by $\alpha_{\min}$ the corresponding lower bound.

Theorem 3.1 (Ergodicity). The mapping T with the properties S1 and S2 are ergodic, mixing, and have the K-property.

Assume for the moment the property S3 instead of S2. Then the proof of Theorem 3.1 is exactly the same as in billiards and similar Hamiltonian systems.[23,17,11,20] This proof has been carried out in full detail in refs. 19 and 24 for particular cases such as the sawtooth map. The extension of this theorem to the case when the property S3 is replaced by S2 has been obtained in ref. 10. We also conjecture that our systems are Bernoulli, as was shown in ref. 24 for the sawtooth map.

Corollary 3.2. The mappings T with the properties S1 and S2 have countable Lebesgue spectrum.

Theorem 3.3 (Entropy). The measure-theoretic entropy of the transformation T with the properties S1 and S2 is given by $h(T) = \ln \Lambda$.

This theorem follows from Pesin's formula[21]; see also ref. 16 for systems with singularities.

In addition, an estimate of the length distribution of LUMs and LSMs can be deduced in the same manner as in refs. 23, 17, and 7. For every $x \in M$ denote by $r^u(x)$ [$r^s(x)$] the distance from x to the nearest endpoint of $\gamma^u(x)$ [resp. $\gamma^s(x)$].

Proposition 3.4 (Distribution of the lengths of LUMs and LSMs). There is a constant $C > 0$ such that $m\{x: r^{u,s}(x) < \varepsilon\} \leqslant C\varepsilon$.

Here and henceforth m denotes the Lebesgue measure in M.

4. ELEMENTS OF MARKOV PARTITIONS

The mixing property of T means that $m(T^n A \cap B) \to m(A)\, m(B)$ as $n \to \infty$ for every two measurable subsets $A, B \subset M$. The rate of the decay of correlations corresponds, roughly speaking, to the rate at which the difference $m(T^n A \cap B) - m(A)\, m(B)$ converges to zero (i.e., the rate of mixing). To facilitate estimation of that rate, one could restrict oneself to special sets A, B whose evolution under T^n is easy to control. One can easily control the evolution of LUMs in the past and that of LSMs in the future. It would thus be reasonable to construct the set A, B out of LUMs and LSMs.

Definition. A *parallelogram* is a subset $A \subset M$ such that for any two points $x, y \in A$ the point $z = \gamma^u(x) \cap \gamma^s(y)$ exists and again belongs to A.

The reason why z may not exist is that there are arbitrary short LUMs and LSMs in M. But if it does exist, it is unique, provided the LUMs and LSMs are not too long. We will always assume that this is the case (otherwise we could introduce some more, artificial, cuttings in M which would reduce the maximal length of the LUMs and LSMs).

Any parallelogram A is a Cantor set with a grid structure. We denote $\gamma_A^{u,s}(x) = \gamma^{u,s}(x) \cap A$ for every x. The sets $\gamma_A^u(x)$ [and $\gamma_A^s(x)$] for all $x \in A$ are congruent; therefore

$$m(A) = c_0 m_1(\gamma_A^u(x))\, m_1(\gamma_A^s(x)) \tag{4.1}$$

for every $x \in A$; here m_1 stands for the one-dimensional Lebesgue measure on the corresponding LUMs and LSMs and c_0 is a normalizing factor.

Evidently, the image $T^n A$ of a parallelogram A is a finite union of parallelograms. Consequently, the intersection $T^n A \cap B$ is again a finite union of parallelograms; here B is another parallelogram. We say that a subparallelogram $C \subset A$ is *u-inscribed* (*s-inscribed*) in A if $\gamma^u_C(x) = \gamma^u_A(x)$ [resp. $\gamma^s_C(x) = \gamma^s_A(x)$] for every $x \in C$. For any pair of parallelograms A, B and $n > 0$ we say that the intersection $T^n A \cap B$ is *regular* if it consists of parallelograms u-inscribed in B and its preimage $A \cap T^{-n} B$ consists of parallelograms s-inscribed in A. Likewise, for $n < 0$ the intersection $T^n A \cap B$ is said to be *regular* if it consists of parallelograms s-inscribed in B and its preimage $A \cap T^{-n} B$ consists of parallelograms u-inscribed in A.

Definition. The *Markov partition* for T is a countable partition (mod 0) of the torus M into parallelograms $\{A_1, A_2, \ldots\}$ such that the intersection $T^n A_i \cap A_j$ is regular for any pair A_i, A_j and any $n \neq 0$.

Note that there are no *finite* partitions of M into parallelograms, because of the presence of arbitrary short LUMs and LSMs.

The construction of the Markov partitions elaborated in refs. 4 and 7 requires the properties S1 and S3 and one extra property of the set S:

Property S4 (Bounds for multiple singularities). For every $m \geqslant 1$ the number of smooth components of $S_{-m,m}$ meeting at a single point of M cannot exceed $K_0 m$, where $K_0 = \text{const}$.

Remark. In fact, any polynomial bound $K_0 m^p$, $p \geqslant 1$, is sufficient here.

Theorem 4.1 (Markov partitions). If the transformation T has the properties S1, S3, and S4, then for any $\varepsilon > 0$ there is a Markov partition for T whose elements have diameters less then ε.

The construction of the Markov partitions for T goes the same way as for billiards; see refs. 4 and 6 and the improved version in ref. 7.

It is well known that the Markov partition provides us with a symbolic representation of the dynamical system in the form of a topological Markov chain (TMC) which has in our case a countable number of states.[4,7] It is also easy to see that the Lebesgue measure in M induces a probability measure in that TMC which makes the TMC a probabilistic Markov chain. Due to the mixing property of T, that chain is recurrent, noncyclic, and ergodic. Unfortunately, this is not enough to ensure an exponential decay of correlations (it *is* enough only for finite-state ergodic Markov chains[2]). One needs to impose more conditions on a countable Markov chain to provide an exponential decay of correlations, like the classic Doeblin condition[4] or the Ibragimov regularity,[15,8] etc. These conditions apparently fail in our systems. Therefore, to estimate the decay of correlations here, we should invoke the *methods* of these classical works rather than the theorems proven there.

If was first noted in ref. 8 that those methods do not require a complete Markov structure of the probabilistic (or symbolic) chain in use. Actually, they need only an *approximation* of this Markov chain by a finite-state chain. It was also noted in ref. 8 that the approximating chain can be constructed directly via a special (*finite*) partition of the phase space which the authors called the *Markov sieve.* That sieve turned out to be much easier to construct than the Markov partition itself. Here we use this idea again, and we actually construct a Markov sieve which is even simpler than in ref. 8. Surprisingly, it gives us a better estimate for the decay of correlations than in ref. 8.

Our Markov sieve is, as in ref. 8, defined through elements of the *pre-Markov partitions* which have been introduced in refs. 4 and 7 as intermediate objects in the construction of the Markov ones. Roughly speaking, elements of pre-Markov partitions serve as frames for parallelograms belonging to the Markov partition.

Definition. Any domain Q in M bounded by two LUMs and two LSMs is called the *quadrilateral.* Its boundary ∂Q consists of two LUMs called the *u-sides of* Q and two LSMs called the *s-sides of* Q. The union of two u-sides is denoted by $\partial^u Q$ and that of two s-sides by $\partial^s Q$.

Remark. Actually, our quadrilaterals look like parallelograms in the usual sense, but we prefer to follow certain traditions is terminology.[4–8]

Fix a sufficiently large $m \geqslant 1$ and let $\varepsilon > 0$ be an arbitrary small real $[\varepsilon < \varepsilon_0(m)]$. A *pre-Markov partition* for the map T^m is a finite partition $\xi_0 = \xi_0(\varepsilon)$ of M into curvilinear polygons $P_1, ..., P_k$ with the following properties:

Property P1. The boundary $\partial\xi_0 = \bigcup \partial P_i$ is the union of $S_{-m,m}$ and a finite collection of LUMs and LSMs. Respectively, we denote $\partial\xi_0 = \partial^0\xi_0 \cup \partial^u\xi_0 \cup \partial^s\xi_0$, where $\partial^0\xi_0 = S_{-m,m}$, and $\partial^u\xi_0$ ($\partial^s\xi_0$) consists of LUMs (LSMs).

Property P2. $T^m(\partial^s\xi_0) \subseteq \partial^s\xi_0$ and $T^{-m}(\partial^u\xi_0) \subseteq \partial^u\xi_0$.

Property P3. All the interior angles of the polygons $P \in \xi_0$ formed by LUMs and LSMs are less than π.

Property P4. The sides of the polygons $P \in \xi_0$ lying on LUMs and LSMs are greater than $c_1\varepsilon$ but less than $c_2\varepsilon$ (here c_1, c_2 are constants determined by the value of m).

The properties P1–P4 immediately imply two others:

Property P5. If a polygon $P_i \in \xi_0$ does not touch the set $S_{-m,m}$, then it is a quadrilateral.

Property P6. The union of all polygons $P_i \in \xi_0$ adjacent to $S_{-m,m}$ has the measure less than $c_3 \varepsilon$ (here c_3 is another constant determined by m).

In other words, ξ_0 is a partition of M into quadrilaterals everywhere except for a vicinity of the singularity set $S_{-m,m}$ whose measure is less than $\text{const} \cdot \varepsilon$.

The pre-Markov partitions for billiard dynamical systems were constructed in ref. 4 and 7 and those methods apply as well to our systems. However, for the sake of completeness we describe briefly the construction of ξ_0 in Appendix C.

Starting with a pre-Markov partition $\xi_0 = \xi_0(\varepsilon)$ for T^m, we obtain a pre-Markov partition for T by $\xi_1 = \xi_1(\varepsilon) = \xi_0 \vee T\xi_0 \vee \cdots \vee T^{m-1}\xi_0$. It also has the properties P1–P6 (maybe with some other values of c_1, c_2, c_3), but now we have $T(\partial^s \xi_1) \subseteq \partial^s \xi_1$ and $T^{-1}(\partial^u \xi_1) \subseteq \partial^u \xi_1$.

The rest of this section is devoted to a preliminary discussion of the evolution of elements of a pre-Markov partition.

Let Q_1 and Q_2 be two arbitrary quadrilaterals. The intersection $T^n Q_1 \cap Q_2$ is a finite union of curvilinear polygons of three types: (i) quadrilaterals whose u-sides lie on the images of u-sides of Q_1 and whose s-sides lie one s-sides of Q_2; (ii) quadrilaterals with either a u-side within $T^n Q_1$ or an s-side within Q_2; (ii) polygons adjacent to $S_{1,n}$. The union of quadrilaterals of type (i) is called the *regular part* of $T^n Q_1 \cap Q_2$ and is denoted by $\mathscr{R}(T^n Q_1 \cap Q_2)$, while the union of polygons of types (ii) and (iii) is called the *irregular part* of $T^n Q_1 \cap Q_2$ and is denoted by $\mathscr{I}(T^n Q_1 \cap Q_2)$. Dual notions are introduced for $T^n Q_1 \cap Q_2$ with $n \leqslant -1$.

Remark. There are at most four quadrilaterals of type (ii). Furthermore, if Q_1 and Q_2 are elements of a pre-Markov partition ξ_1, then due to Property P2 (with $m = 1$) quadrilaterals of type (ii) in $T^n Q_1 \cap Q_2$ never occur.

Similarly, for any two parallelograms A_1, A_2 and $n \geqslant 1$ the intersection $T^n A_1 \cap A_2$ is a finite union of parallelograms. The union of those u-inscribed in A_2 and such that their images under T^{-n} are s-inscribed in A_1 is called the *regular part* of the intersection $T^n A_1 \cap A_2$ and is denoted $\mathscr{R}(T^n A_1 \cap A_2)$, while the union of the others is called the *irregular part* of that intersection and is denoted by $\mathscr{I}(T^n A_1 \cap A_2)$. Again, dual notations are introduced for $T^n A_1 \cap A_2$ with $n \leqslant -1$.

Certain relations between quadrilaterals and parallelograms are established by the following considerations (we note that all of these have been already used in ref. 8).

Definition. For any parallelogram A the minimal closed quadrilateral containing A is called the *support* of A and is denoted by $K(A)$.

Definition. We say that a segment of a LUM (LSM) is *inscribed* in a quadrilateral Q if it lies within Q and terminates on two s-sides (u-sides) of Q.

Definition. A parallelogram A is said to be *maximal* if it intersects all the LUMs and LSMs inscribed in its support $K(A)$.

In other words, to construct a maximal parallelogram one should take a quadrilateral Q, and draw all the LUMs and LSMs inscribed in Q; thus, the maximal parallelogram would consist of the points of intersections of these LUMs and LSMs. The parallelogram so obtained is denoted by $A(Q)$. The following is straightforward:

Lemma 4.2. Let A_1, A_2 be two maximal parallelograms and $Q_1 = K(A_1)$, $Q_2 = K(A_2)$ be their supports. Then $\mathscr{R}(T^n A_1 \cap A_2) \subset \mathscr{R}(T^n Q_1 \cap Q_2)$ and $\mathscr{I}(T^n A_1 \cap A_2) \subset \mathscr{I}(T^n Q_1 \cap Q_2)$.

5. DECAY OF CORRELATIONS

We consider the space of complex-valued Hölder continuous functions $H_\alpha = \{f\colon |f(x) - f(y)| \leqslant C_f |x-y|^\alpha$ for any $x, y \in M\}$. More generally, let H_α^* denote the space of *piecewise* Hölder continuous functions, which are Hölder continuous (with an exponent α) on a finite collection of subdomains in M separated by a finite union of compact smooth curves (for example, the domains where the maps $T^{\pm m}$ are continuous, $m \geqslant 1$ is fixed). Note that the curves and domains here must be fixed for the class H_α^* under considerations.

Theorem 5.1. (Decay of correlations). Let the map T have the properties S1, S3, and S4. Then for any two functions $f, g \in H_\alpha^*$ and any $n \geqslant 1$

$$|\langle f(T^n x)\, g(x)\rangle - \langle f(x)\rangle \langle g(x)\rangle| \leqslant C(f, g)\lambda_0^n$$

where $\lambda_0 \in (0, 1)$ is a constant determined by T and α.

From here on $\langle \cdot \rangle$ denotes the integral over M with respect to the measure m.

Our method for estimating the decay of correlations differs substantially from those developed for one-dimensional maps. (See, e.g., ref. 27.)

While the main idea of those methods consists in analyzing the spectral properties of the associated operator in the function space, our approach is more straightforward. We deduce Theorem 5.1 from certain measure-theoretic properties of our dynamical system, like exponential mixing for a selected collection of parallelograms. These properties are then proven with the help of the pre-Markov partition. Nonetheless, certain analogies between two approaches can also be observed; see below.

For any $n \geqslant 1$ let $G = G_n = \{A_0, A_1, \ldots, A_I\}$ be a finite partition of M with two properties:

Property G1 (Sizes). The diameters of the elements $A_i \in G$ for $i = 1, 2, \ldots, I$ do not exceed $C_1 \lambda_1^n$.

Property G2 (Measure of marginal set). $m(A_0) \leqslant C_1 \lambda_1^n$.

Here $C_1 > 0$ and $\lambda_1 \in (0, 1)$ are constants to be specified below.

For each $A_i \in G$ and $f \in H_\alpha^*$ denote $f_i = (m(A_i))^{-1} \int_{A_i} f(x)\, dm(x)$ and

$$\bar{f}(x) = \sum_{i=0}^{I} f_i \cdot \chi_{A_i}(x)$$

where χ_A stands for the characteristic function of A. In other words, $\bar{f}(x)$ is the conditional expectation of $f(x)$ with resxpect to the partition G.

If an A_i, $i \geqslant 1$, lies wholly within a domain where $f(x)$ is continuous, then the variation of $f(x)$ on A_i does not exceed $C(f)\lambda_1^{\alpha n}$. Hence one can write

$$\langle f(T^n x)\, g(x) \rangle = \langle \bar{f}(T^n x)\, \bar{g}(x) \rangle + \Delta_1 = \sum_{i,j=1}^{I} f_i g_j m(A_i \cap T^n A_j) + \Delta_2$$

where $|\Delta_p| \leqslant C(f, g)\lambda_1^{\alpha n}$ for $p = 1, 2$.

Assume now an additional property of G:

Property G3 (Exponential mixing for parallelograms in G). For any $n \geqslant 1$ and $i, j \in \{1, 2, \ldots, I\}$ one has $m(A_i \cap T^n A_j) \geqslant m(A_i)\, m(A_j)(1 - C_1 \lambda_1^n)$.

For any two real-valued *positive* functions f, g we now obtain from G3

$$\langle f(T^n x)\, g(x) \rangle \geqslant \langle f(x) \rangle \langle g(x) \rangle + \Delta_3 \tag{5.1}$$

with some $|\Delta_3| \leqslant C(f, g)\lambda_1^{\alpha n}$. Substituting $f(x) + C_1$ for $f(x)$ and $g(x) + C_2$ for $g(x)$, C_1 and C_2 being arbitrary real constants, yields the inequality (5.1) for *any* two real-valued functions f, $g \in H_\alpha^*$. Switching the sign of $g(x)$ reverses the inequality of (5.1) and thus completes the proof of Theorem 5.1 for real-valued functions f and g. The extension to complex-valued functions is obvious.

Remark. From the above calculations it follows that λ_0 in Theorem 5.1 can be chosen as λ_1^α, where λ_1 is independent of α. Furthermore, we can set $C(f, g) = (C_f + M_f)(C_g + M_g) \cdot \text{const}$, where C_f is the factor in the Hölder condition and $M_f = \max_M |f(x)|$, and const is independent of f, g. For the class H_α instead of H_α^* one can omit M_f, M_g in the last estimate.

We have thus deduced Theorem 5.1 from the properties G1–G3 of a partition G. Hereafter we call such a partition the *Markov sieve.* It is actually a simpler but more efficient version of a Markov sieve constructed in ref. 8. From now on we forget about the functions f and g and focus on the construction of the Markov sieve G.

We begin with some preliminary considerations of the evolution of parallelograms. Let A, B be two parallelograms. Define two integer functions $k^{\pm}_{A,B}(x)$ on A as $k^{+}_{A,B}(x) = \min\{k \geqslant 1\colon T^k x \in \mathscr{R}(T^k A \cap B)\}$ and $k^{-}_{A,B}(x) = \min\{k \geqslant 1\colon T^{-k} x \in \mathscr{R}(T^{-k} A \cap B)\}$. Next define two transformations $T^{\pm}_{A,B}\colon A \to B$ as

$$T^{\pm}_{A,B}(x) = T^{k^{\pm}_{A,B}(x)}(x)$$

The following two propositions are the main tools in our proof of Theorem 5.1.

Proposition 5.2 (Bound for irregular parts). For any two maximal parallelograms A, B one has $m(\mathscr{I}(T^k A \cap B)) \leqslant C_2 \lambda_2^{|k|}$, where $C_2 > 0$ and $\lambda_2 \in (0, 1)$ are constants determined by the map T.

Remark. The mixing property of T implies that $m(T^k A \cap B) \to m(A)\, m(B)$ as $k \to \pm\infty$. Proposition 5.2 tells us that the intersection $T^k A \cap B$ consists mostly of parallelograms, u-inscribed in B for $k > 0$ or s-inscribed in B for $k < 0$.

Due to Proposition 5.2, the functions $k^{\pm}_{A,B}(x)$ are almost surely finite. We thus can define two probability distributions $p^{\pm}_{A,B}(k) = m\{x \in A\colon k^{\pm}_{A,B}(x) = k\}/m(A)$.

Proposition 5.3 (Tail bound for time distribution). For any two maximal parallelograms A, B and $k \geqslant 1$ one has $p^{\pm}_{A,B}(k) \leqslant C_3 \lambda_3^k / m(A)$, where $C_3 > 0$ and $\lambda_3 \in (0, 1)$ are determined by the map T and the parallelogram B.

The proofs of Propositions 5.2 and 5.3 are supplied in Appendices A and B, respectively.

Proof of Theorem 5.1 from Propositions 5.2 and 5.3. Let $n \geqslant 1$ be sufficiently large integer. Set $\varepsilon = \lambda_4^n$, where $\lambda_4 \in (0, 1)$ is specified below. Consider a pre-Markov partition $\xi_1 = \xi_1(\varepsilon)$. For every quadrilateral $Q \in \xi_1$

consider the maximal parallelogram $A=A(Q)$ with the support Q. It is said to be *ample* if $m(A)\geqslant(1-\sqrt{\varepsilon})\,m(Q)$. Let $A_1, A_2,..., A_l$ be all the ample parallelograms resulting from ξ_1. Denote also $A_0=M\setminus\bigcup A_i$.

Lemma 5.4 (Bound for the measure of marginal set). $m(A_0)\leqslant \text{const}\cdot\sqrt{\varepsilon}$.

Proof. First, the union of all the elements $Q\in\xi_1$ which are *not* quadrilaterals has the measure $\leqslant\text{const}\cdot\varepsilon$. Next, if a point $x\in Q$ of a quadrilateral $Q\in\xi_1$ does not belong to $A(Q)$, then either $\gamma^u(x)$ or $\gamma^s(x)$ has to be too short: it does not meet one of two s-sides or, resp., u-sides of Q. Applying Proposition 3.4 gives the bound for the measure of the set formed by such points as $\text{const}\cdot\varepsilon$. ∎

Lemma 5.4 ensures that the partition $G=\{A_0, A_1,...,A_l\}$ has the properties G1, G2. Now we have to prove the property G3.

Fix any quadrilateral Q from a pre-Markov partition $\xi_1(\varepsilon_0)$ for some $\varepsilon_0>0$ such that the maximal parallelogram $A=A(Q)$ has a nonzero measure.

Lemma 5.5 (Exponential mixing for a fixed parallelogram). For every integer k one has $m(\mathscr{R}(T^kA\cap A))\geqslant m(A)^2(1-C_5\lambda_5^{|k|})$ for some constants $C_5>0$ and $\lambda_5\in(0, 1)$ determined by A.

Proof. Let $k\geqslant 1$ [in the case $k\leqslant -1$ one should observe that $m(\mathscr{R}(T^kA\cap A))=m(\mathscr{R}(T^{-k}A\cap A))$]. Define a probabilistic model that is similar to a random walk. Set $p_k=m(\mathscr{R}(T^kA\cap A))/m(A)$ ("the probability of a return at the kth step") and $q_k=m\{x\in A\colon k^+_{A,A}(x)=k\}/m(A)$ ("the probability of the *first return* at the kth step"). Next take a point $x\in A$ and denote

$$y=T^{k^+_{A,A}(x)}x$$

Consider $k_1=\min\{k>k^+_{A,A}(x)\colon T^kx\in\mathscr{R}(T^kA\cap A)\}$. It is easy to check that $k_1=k^+_{A,A}(x)+k^+_{A,A}(y)$. This enables us to write the following important relation (the convolution law):

$$p_k=q_k+q_{k-1}p_1+\cdots+q_1p_{k-1} \tag{5.2}$$

The rest of the proof of lemma uses quite standard methods from the theory of recurrent events and random walks. Consider the generating functions $P(z)=1+\sum_1^\infty p_kz^k$ and $Q(z)=\sum_1^\infty q_kz^k$ of a complex variable z. Equation (5.2) is then equivalent to

$$P(z)=\frac{1}{1-Q(z)} \tag{5.3}$$

Proposition 5.3 implies that $\sum_1^\infty q_k = 1$ and $q_k \leqslant C_6 \lambda_6^k$ for some $C_6 > 0$ and $\lambda_6 \in (0, 1)$ determined by A. The function $Q(z)$ is therefore analytic in the open disk $|z| < \lambda_6^{-1}$. Obviously, $Q(1) = 1$ and $|Q(z)| < 1$ for all z, $|z| < 1$. In virtue of the mixing property of T the set of integers $\{k\}$ for which $q_k > 0$ is aperiodic (has no common divisors except for unity), and thus $Q(z) \neq 1$ for all z, $|z| = 1$ except for $z = 1$ (compare these reasonings to the study of the spectral properties of the Perron–Frobenius operator for interval maps in ref. 27: the mixing property was used there to rule out all the eigenvalues lying on the unit circle except for unity itself). Since $Q(z)$ is analytic on the unit circle, the equation $Q(z) = 1$ has a unique root $z = 1$ in an open disk $|z| < r$ for some $r > 1$. Therefore the function $D(z) = (z-1)[Q(z) - 1]^{-1}$ is also analytic in the open disk $|z| < r$. Equation (5.3) can be now rewritten as

$$P(z) = \frac{D(z)}{1-z} \tag{5.4}$$

Let $D(z) = d_0 + d_1 z + d_2 z^2 + \cdots$ denote the Taylor series expansion of $D(z)$; then $p_k = d_0 + d_1 + \cdots + d_k$ in virtue of the formula (5.4). The analyticity of $D(z)$ in the open disk $|z| < r$ implies that $|d_k| < \text{const} \cdot r_1^{-k}$ for some r_1, $1 < r_1 < r$. Therefore, the sequence $\{p_k\}$ is convergent and converges exponentially fast. On the other hand, Proposition 5.2 along with the mixing property of T shows that $p_k \to m(A)$ as $k \to \infty$. Lemma is proven. ∎

We now return to the proof of the property G3. Note that the only essential difference between the property G3 and Lemma 5.5 is that the constant C_1 and λ_1 are independent of A_i, $A_j \in G_n$ and of the value of n.

Fix a parallelogram A satisfying the conditions of Lemma 5.5. Our idea now is to transform the parallelogram A_i (A_j) into A by the map $T^+_{A_iA}$ ($T^-_{A_jA}$) and then deduce the property G3 from Lemma 5.5.

Partition the parallelogram A_i into s-inscribed subparallelograms $A_{i1}, A_{i2},\ldots$ such that on each A_{ip} the function $k^+_{A_iA}(x)$ is constant ($= k^+_{ip}$) and

$$T^{k^+_{ip}} A_{ip} = B_{ip}$$

is a parallelogram u-inscribed in A. Likewise, let $A_j = A_{j1} \cup A_{j2} \cup \ldots$, so that on each A_{jq} the function $k^-_{A_jA}(x)$ is constant ($= k^-_{jq}$) and

$$T^{k^-_{jq}} A_{jq} = B_{jq}$$

is s-inscribed in A.

Obviously,

$$m(T^n A_i \cap A_j) = \sum_{p,q} m(T^n A_{ip} \cap A_{jp}) = \sum_{p,q} m(T^{n_{pq}} B_{ip} \cap B_{jq})$$

where $n_{pq} = n - k_{ip}^+ - k_{jq}^-$. Applying the formula (4.1) and then using Lemma 5.5 gives that for each $p, q \geqslant 1$ such that $n_{pq} > 0$

$$\begin{aligned} m(T^n A_{ip} \cap A_{jq}) &= m(T^{n_{pq}} B_{ip} \cap B_{jq}) \\ &\geqslant \frac{m(\mathscr{R}(T^{n_{pq}} A \cap A))}{m(A)^2} m(B_{ip}) \, m(B_{jq}) \\ &\geqslant m(A_{ip}) \, m(A_{jq})(1 - C_5 \lambda_5^{|n_{pq}|}) \end{aligned} \tag{5.5}$$

For each pair p, q such that $k_{ip}^+ \leqslant n/3$ and $k_{jq}^- \leqslant n/3$ we have $n_{pq} \geqslant n/3$. The union of A_{ip} (A_{jq}) over all p (q) such that $k_{ip}^+ > n/3$ ($k_{jp}^- > n/3$) has the measure less than $C_3 \lambda_3^{n/3}$ due to Proposition 5.3. Summing up the inequality (5.5) over p, q thus gives

$$m(T^n A_i \cap A_j) \geqslant [m(A_i) - C_3 \lambda_3^{n/3}][m(A_j) - C_3 \lambda_3^{n/3}](1 - C_5 \lambda_5^{n/3})$$

Finally, observe that $m(A) \geqslant \text{const} \cdot \lambda_4^{2n}$ for every parallelogram $A \in G$ (to ensure that, we had introduced the notion of ample parallelograms). We now complete the construction of the Markov sieve G by setting $\lambda_4 = (\lambda_3/\lambda_5)^{1/6}$ and $\lambda_1 = \max\{\lambda_4, \lambda_5^{1/3}\}$. It is then easy to check that all the properties G1–G3 hold.

Theorem 5.1 is proven.

6. CENTRAL LIMIT THEOREM

Let H_α^* be a class of piecewise Hölder continuous functions on M introduced in the preceding section.

Theorem 6.1 (Central limit theorem). For any real-valued $f \in H_\alpha^*$ the sum

$$\sigma^2 = \sum_{n=-\infty}^{\infty} [\langle f(T^n x) f(x) \rangle - \langle f \rangle^2] \tag{6.1}$$

is finite. In case $\sigma \neq 0$ the sequence

$$\frac{S_n - n \cdot \langle f \rangle}{(n\sigma^2)^{1/2}}$$

where $S_n = f(x) + f(Tx) + \cdots + f(T^{n-1}x)$, converges in distribution with respect to the measure m, as $n \to \infty$, to the standard normal law.

The convergence of the series in (6.1) immediately results from Theorem 5.1. The proof of Theorem 6.1 relies heavily on a fast decay of correlations. The proof has been carried out in full detail for billiard systems in ref. 8. The arguments of ref. 8 are quite general and do not involve any specific feature of billiards, so that we do not need to reproduce that proof here. Note that in the case of one-dimensional maps the central limit theorem can be deduced directly from the spectral properties of the Perron–Frobenius operator (see, e.g., ref. 27).

Remark. As in the case of billiards, the sum (6.1) is zero if and only if the function $f(x)$ is coboundary, i.e., $f(x) = g(Tx) - g(x)$ a.e. with a function $g \in L_2(M, m)$.

7. OPEN QUESTIONS, DISCUSSION

1. Having proven the exponential decay of correlations for piecewise linear hyperbolic maps of the 2-torus, we should try to extend our result to nonlinear maps, like Anosov systems (with additional singularities) or billiards. Apparently, such an extension needs a good approximation of the nonlinear dynamics by a "locally linear" one. In fact, an approximation of that type has been used in ref. 8.

2. We should also try to extend our results to multidimensional hyperbolic toral automorphisms with singularities. However, an explicit geometrical construction of a pre-Markov partition is no longer possible in higher dimensions. The difficulties with a "nonsmooth boundary" arising there were first noticed and described in ref. 3. Besides, our definition of the pre-Markov partition no longer works in that case. Another approach for the construction of Markov partitions for multidimensional systems with singularities has been developed in ref. 18, based on Bowen's shadowing property.[2]

3. L.-S. Young (private communication) has pointed out that Theorem 5.1 could be proven by an alternative method. This consists in reducing our system to a one-dimensional map with the help of the Markov partition, provided certain metric properties of that partition (like our Lemma B.1) have been proven in advance.

APPENDIX A

The proof of Proposition 5.2 involves simple geometrical constructions, in which we omit routine details.

By Lemma 4.2 it is enough to consider only the irregular part $\mathscr{I}(T^n Q(A) \cap Q(B))$. This part consists of curvilinear polygons adjacent to

$S_{1,n}$ and at most four quadrilaterals (see Section 4). Since the s-sides of the latter are less than const $\cdot \lambda^n$, their union has the measure $<$const $\cdot \lambda^n$.

Next, fix an $l \in [1, n]$ and consider the above curvilinear polygons that are adjacent to S_l. Their images under T^{-l+1} are adjacent to S_+ and disjoint. Since the width of each of those images in the u-direction is smaller than const $\cdot \lambda^l$, their union has the measure less than const $\cdot \lambda^l$, too.

Now discard all polygons $P \subset \mathscr{I}(T^n Q(A) \cap Q(B))$ which do not intersect B. After that, each of remaining polygons contains an LUM inscribed in $Q(B)$. Furthermore, if a polygon P of that type is adjacent to S_l, then it contains an LUM $\gamma^u(P)$ u-inscribed in $Q(B)$ and also adjacent to S_l. Two possibilities arise: (i) $\gamma^u(P)$ terminates at a point of intersection of S_l with an s-side of $Q(B)$ or (ii) two smooth curves of S_l have a common endpoint in $\gamma^u(P)$. It is easy to see that the number of points listed above is less than four times the number of smooth curves in S_l, which is, in turn, less than Λ_0^l, where Λ_0 is a constant determined by the map T. Finally, observe that the width of any polygon of that type in the s-direction is less than const $\cdot \lambda^n$. The above considerations lead to the inequality

$$m(\mathscr{I}(T^n A \cap B)) \leqslant \text{const} \cdot \left(\lambda^n + \sum_{l=1}^{n} \min\{\lambda^l, \Lambda_0^l \lambda^n\} \right)$$

which, in turn, yields Proposition 5.2.

APPENDIX B

The proof of Proposition 5.3 is based on several lemmas. The first one describes "gaps" in a maximal parallelogram.

Let A be a maximal parallelogram. Draw all the line segments inside $Q(A)$ parallel to E^u, terminating at s-sides of $Q(A)$ and intersecting the set $S_{1,\infty}$ (note: these are *not* LUMs, since they cross the singularity set). They form a countable union of strips inside $Q(A)$ called here u-gaps. In a similar way we define s-gaps in $Q(A)$. Note that removing all the u-gaps and s-gaps from the quadrilateral $Q = Q(A)$ yields exactly the original parallelogram A.

Lemma B.1 (Distribution of widths of gaps). Let A be a maximal parallelogram and $\varepsilon > 0$. Then the union of all the gaps in A whose width is less than ε has the measure less than $C\varepsilon^a$, where $a > 0$ is a constant determined by T and $C > 0$ depends on A.

Proof. Let H be a u-gap with a curve belonging to S_l inside, for some $l \geqslant 0$. Recall that the minimum angle between E^u and any curve in S_+ is $\alpha_{\min}$. Hence the minimum angle between E^u and any curve in S_l is not less than $\alpha_{\min}\lambda^{2l-2}$. The width of H is therefore greater than $C_1\lambda^{2l}$ for some $C_1 = C_1(A) > 0$. It is thus enough to consider the values $l \geqslant -C_2 \ln \varepsilon$ only [here $C_2 = C_2(A) > 0$ is another constant]. The set $T^{-l+1}H$ is adjacent to S_+ and its width in the u-direction is less than const $\cdot \lambda^l$. The total measure of all u-gaps (for a fixed l) is then less than const $\cdot \lambda^l$. Summing up these bounds over all $l \geqslant -C_2 \ln \varepsilon$ results in the lemma. ▮

The next three lemmas describe the evolution of an LUM under T^n, $n \geqslant 1$. All the statements below have dual forms for the evolution of an LSM under T^n, $n \leqslant -1$.

Let γ^u be an LUM and ρ be its length. For any $n \geqslant 1$ its image $T^n\gamma^u$ is a finite union of LUMs called here *components*. For every $D > 0$ denote $\gamma^u_n(D) = \{x \in \gamma^u$: for every $l = l(x) \in [0, n]$ the component of $T^l\gamma^u$ containing the point T^lx is less than $D\}$. In other words, $\gamma^u_n(D)$ consists of points whose consecutive images never appear in long components during the first n iterates of T.

Lemma B.2 (From short to long components). There are $D > 0$, $C > 0$ and $\beta \in (0, 1)$ all determined by T such that $m_1(\gamma^u_n(D)) \leqslant C\beta^n$ for any $n \geqslant 1$ and any LUM γ^u of length ρ.

In a sense, Lemma B.2 gives an exact expression of the general fact mentioned in the Introduction that the hyperbolicity prevails over singularities. Indeed, if γ^u is *very* short, then typical components of $T^n\gamma^u$ become long (recover) after about $-$const $\cdot \ln m_1(\gamma^u)$ iterates of T, just as in the case of a smooth, uniformly hyperbolic map T_0 the image $T_0^n\gamma^u$ itself recovers after the same number of iterates of T_0.

Next consider an LUM γ^u of length ρ and a quadrilateral Q such that $m(A(Q)) > 0$. For every $n \geqslant 1$ denote by $\gamma^u_n(Q)$ the union of all subintervals in $\gamma^u_n(Q)$ whose images under T^n are LUMs inscribed in Q.

Lemma B.3 (From long components into a fixed quadrilateral). For every $\rho > 0$ and Q such that $m(A(Q)) > 0$ there are $n_0 > 0$ and $\alpha_0 > 0$ such that $m_1(\gamma^u_n(Q)) \geqslant \alpha_0$ for every LUM γ^u of length ρ and every $n \geqslant n_0$.

Lemma B.3 is, of course, interesting for ρ sufficiently large, in particular, for $\rho = D$. We stress that the values of n_0 and α_0 here depend on $\rho = m_1(\gamma^u)$, but not on γ^u itself; thus the bound in Lemma B.3 is uniform in all sufficiently long LUMs.

The proofs of Lemmas B.2 and B.3 are essentially the same as in the case of billiards.[8,6] Here we outline only the ideas of those proofs.

Proof of Lemma B.2. Due to the property S4, for any $m \geqslant 1$ there is $\varepsilon_m > 0$ such that any smooth curve of length ε_m in M intersects at most $K_0 m$ curves of $S_{-m,m}$. Hence the image $T^m\gamma^u$ contains at most $K_0 m + 1$

components, provided the LUM γ^u is less than ε_m. Fix a $k \geqslant 1$ and then count the components of $T^{mk}\gamma_n^u(Q)$. Their number is less than $(K_0 m+1)^k$, and so their images under T^{-mk} form a union of subintervals in γ^u whose total m_1-measure is less than $(K_0 m+1)^k \lambda^{mk} \varepsilon_m$. If m is large enough, then $(K_0 m+1)^{1/m} < \lambda^{-1/2}$ and the lemma follows. ∎

Proof of Lemma B.3. Choose a maximal parallelogram B, $m(B)>0$, such that γ^u intersects both s-sides of $Q(B)$. The existence of such a parallelogram follows from certain ergodic properties of the system under consideration; see ref. 8 for explanations. The mixing property of T and Proposition 5.2 imply that $m(\mathscr{R}(T^n B \cap A)) > m(B)\, m(A)/2$ for all $n \geqslant n_0(Q, B)$, where $A = A(Q)$. The set $\mathscr{R}(T^n B \cap A)$ consists of sub-parallelograms u-inscribed in A. Each has the measure $<\text{const} \cdot \lambda^n$, and so their number must be greater than $\text{const} \cdot \Lambda^n m(B)\, m(A)$. Hence $m_1(\gamma_n^u(Q)) \geqslant \text{const} \cdot m(B)\, m(A)$ as soon as $n \geqslant n_0(Q, B)$.

In order to make the last bound independent of B, we observe that the set of LUMs with the length $\geqslant \rho$ is compact (in C^0 topology). There is therefore a finite collection of maximal parallelograms $\{B_i\}$ such that *each* of those LUMs intersects at least one of the parallelograms $\{B_i\}$ in the way specified above (cf. also ref. 8 for more detail). ∎

Observe now that if $T^n\gamma^u$ contains an LUM γ_1^u inscribed in Q, then a certain portion of that LUM lies in $A = A(Q)$. However, it is important to describe the further evolution of the remaining portion of $\tilde{\gamma}^u$ which has fallen into s-gaps of A. Denote $\tilde{\gamma}^u = \gamma_1^u \setminus A$.

For every $n \geqslant 0$ the image $T^n\tilde{\gamma}^u$ is a countable union of LUMs which we call here *components of second type* to distinguish them from the components defined above. For every $D>0$ denote $\tilde{\gamma}_n^u(D) = \{x \in \tilde{\gamma}^u$: for every $l \in [1, n]$ the component of second type of $T^l\tilde{\gamma}^u$ containing the point $T^l x$ is less than $D\}$.

Lemma B.4 (From gaps to long components). There are $\tilde{C}>0$ and $\tilde{\beta} \in (0,1)$ determined by T such that $m_1(\tilde{\gamma}_n^u(D)) \leqslant \tilde{C}\tilde{\beta}^n$ for every $n \geqslant 1$. Here $D>0$ is the same as that involved in Lemma B.2.

Proof. Denote by $\rho_1, \rho_2, \ldots$ the lengths of the subintervals in γ_1^u of which the set $\tilde{\gamma}^u$ consists. To each of those subintervals we apply Lemma B.2, and so we obtain

$$m_1(\tilde{\gamma}_n^u(D)) \leqslant \sum_i \min\{\rho_i, C\beta^n\}$$

Using Lemma B.1 yields two bounds:

$$\sum_{i:\, \rho_i < C\beta^n} \rho_i \leqslant C^{1+a}\beta^{an} \tag{B.1}$$

and

$$\sum_{i:\, C\beta^{l} \leqslant \rho_i < C\beta^{l-1}} C\beta^n \leqslant \frac{C^{1+a}\beta^{a(l-1)}C\beta^n}{C\beta^l} = \beta^{-a}C^{1+a}\beta^{n-l+al} \qquad \text{(B.2)}$$

for every $l=0, 1,..., n$. Adding the inequalities (B.1) and (B.2) for $l=0, 1,..., n$ results in

$$\sum_i \min\{\rho_i, C\beta^n\} \leqslant \beta^{-a}C^{1+a}(2+n)\beta^{an}$$

that completes the proof. ∎

Remark. The value D involved in Lemmas B.2 and B.4 can be chosen as small as necessary. We can therefore assume that D is less than the sum of widths of s-gaps in A. Furthermore, we assume that $\tilde{C} \geqslant C$ and $\tilde{\beta} \geqslant \beta$.

Consider a special evolution T^* of an LUM γ^u with "absorbing" property of the parallelogram $A = A(Q)$. As soon as a component γ_1^u of $T^n\gamma^u$ intersects both s-sides of Q, the points of $\gamma_1^u \cap A$ stop, and only the remaining portion of γ_1^u, i.e., $\gamma_1^u \backslash A$, keeps moving under T. After n iterates of T^*, a part of γ^u has been already "stuck" with the parallelogram A, while the remaining part is still moving. We denote the part of γ^u which is moving during the first n iterates of T^* by $\tilde{\gamma}^u(n)$. Its image $T^n\tilde{\gamma}^u(n)$ is a countable union of LUMs which we call *components of third type.*

Lemma B.5 (Absorption). For any LUM γ^u and any $n \geqslant 1$ one has $m_1(\tilde{\gamma}^u(n)) \leqslant C_1\beta_1^n$ where $C_1 > 0$ and $\beta_1 \in (0, 1)$ are constants, both depending on the absorbent A alone.

Proof. By Lemma B.2, $m_1(\gamma_{n/2}^u(D)) \leqslant C\beta^{n/2}$. We can therefore neglect the subset $\gamma_{n/2}^u(D)$ of γ^u. The images of points $x \in \tilde{\gamma}^u(n) \backslash \gamma_{n/2}^u(D)$ appear in components of length $\geqslant D$ at least once during the first $n/2$ iterates of T. It is thus enough to prove Lemma B.5 for the case $m_1(\gamma^u) = D$ only.

Let now $m_1(\gamma^u) = D$. For any $x \in \gamma^u$ denote $r(x) = \#\{l \in [1, n]$: the point T^lx belongs to a component of third type of length $\geqslant D\}$.

Sublemma B.6. There are $C_2 > 0$, $\alpha_2 > 0$, and $\beta_2 \in (0, 1)$, all determined by T and Q, such that $m_1\{x \in \gamma^u\colon r(x) < \alpha_2 n\} \leqslant C_2\beta_2^n$.

Proof. Let $r < n$ and $0 \leqslant n_1 < n_2 < \cdots < n_r \leqslant n$. Denote by $\gamma_n^u(n_1,..., n_r)$ a subset of γ^u consisting of points x for which $r(x) = r$ and $T^{n_1}x,..., T^{n_r}x$ belong to components of third type of length $\geqslant D$. Lemmas B.2 and B.4 enable us to prove that

$$m_1(\gamma_n^u(n_1,..., n_r)) \leqslant m_1(\gamma^u)\frac{\tilde{C}\tilde{\beta}^{n_1-1}}{D}\frac{\tilde{C}\tilde{\beta}^{n_2-n_1-1}}{D}\cdots\frac{\tilde{C}\tilde{\beta}^{n-n_r-1}}{D}$$

$$= m_1(\gamma^u)(\tilde{C}/\tilde{\beta})^r\tilde{\beta}^n$$

Hence

$$m_1\{x\in\gamma^u\colon r(x)=r\}\leqslant m_1(\gamma^u)\frac{n!}{r!\,(n-r)!}\left(\frac{\tilde{C}}{\tilde{\beta}}\right)^r\tilde{\beta}^n$$

If $r<\alpha n$ with some $\alpha<1/2$, then the last bound can be rewritten as

$$m_1\{x\in\gamma^u\colon r(x)=r\}\leqslant \text{const}\cdot\left[\frac{\tilde{\beta}(\tilde{C}/\tilde{\beta})^\alpha}{\alpha^\alpha(1-\alpha)^{1-\alpha}}\right]^n$$

For α small enough the constant within the brackets is less than 1. ▮

Sublemma B.6 allows us to disregard, along with $\gamma^u_n(D)$, also the subset $\{x\in\gamma^u\colon r(x)<\alpha_2 n\}$. All the other points of $\tilde{\gamma}^u(n)$, in evolution under T^*, appear in long components of third type at least $\alpha_2 n$ times during the first n iterates of T^*. Any of those components sends a fixed portion of its measure (to be precise, β_0/D) into the absorbent A within n_0 subsequent iterates of T^* by virtue of Lemma B.3 with $\rho=D$.

In terms of probability theory, a point x in question has a chance of being stuck with A during its evolution $\alpha_2 n$ times in succession. It is now clear that the probability of x not being stuck with A within the first n iterates of T^* is less than $(1-\beta_0/D)^{\alpha_2 n}$, so that we come to the desired bound in Lemma B.5. ▮

Remark. Minor modifications of the above arguments are needed to obtain the same bound as in Lemma B.5 for another special evolution T^{**} of γ^u, where the absorption into A takes place only after the $(n/2)$th iterate of T, while within the first $(n/2)$ iterates the maps T^{**} and T coincide.

We now finish the proof of Proposition 5.3. For each $x\in A$ denote $i(x,k)=\min\{i\geqslant k/2\colon\ T^i x\in B$ and the component of $T^i(\gamma^u(x)\cap Q(A))$ containing x intersects both s-sides of $Q(B)\}$. Let $\mathscr{R}(A)$ be a subset of points $x\in A$ such that $T^{i(x,k)}x\in\mathscr{R}(T^{i(x,k)}A\cap B)$. It is easily seen that

$$m\{x\in A\colon k^+_{A,B}(x)\geqslant k\}\leqslant m\{x\in A\colon i(x,k)\geqslant k\}+m(A\backslash\mathscr{R}(A))$$

By Proposition 5.2,

$$m(A\backslash\mathscr{R}(A))\leqslant\sum_{k/2}^{\infty} m(\mathscr{I}(T^iA\cap B))\leqslant\text{const}\cdot\lambda_2^{k/2}$$

On the other hand, Lemma B.5, along with the remark thereto, gives the bound $m\{x\in A\colon\ i(x,k)\geqslant k\}\leqslant C_1\beta_1^k$. This completes the proof of Proposition 5.3.

APPENDIX C

The construction of the pre-Markov partition ξ_0 is accomplished in three steps.

1. Let $m \geqslant 1$ be a sufficiently large integer to be specified below and $\varepsilon > 0$ be a sufficiently small real. First, recall that by the property S4 for any $m \geqslant 1$ there is an $\varepsilon_0 = \varepsilon_0(m)$ such that any ε_0-disk in M intersects at most $K_0 m$ curves of $S_{-m,m}$.

Choose a finite $(\varepsilon/10)$-net $\{x_i^+\}$, $1 \leqslant i \leqslant I_0^+$, in the set $M \setminus U_\varepsilon(S_{1,m})$ and a finite $(\varepsilon/10)$-net $\{x_i^-\}$, $1 \leqslant i \leqslant I_0^-$, in the set $M \setminus U_\varepsilon(S_{-m,-1})$, where $U_\varepsilon(A)$ stands here for the ε-neighborhood of the subset $A \subset M$. Through every point x_i^+, $1 \leqslant i \leqslant I_0^+$, draw a line segment $\hat{\gamma}_i^+$ of length ε, parallel to E^u, and bisected by x_i^+. Consider two subsegments of $\hat{\gamma}_i^+$ adjacent to its endpoints, both of length $\varepsilon/10$. For ε small enough these two segments intersect at most $K_0 m$ curves of $S_{-m,-1}$. Hence each of them contains a smaller subsegment of length $(200\,K_0 m)^{-1}\varepsilon$ which lie outside $U_{c\varepsilon}(S_{-m,-1})$, where $c = (100\,K_0 m)^{-1}$. For each of these smaller segments construct a rhombus with sides parallel to E^u and E^s and with this subsegment as a side. Having constructed two rhombi, we erase two parts of $\hat{\gamma}_i^+$ beyond the two most distant vertices of the rhombi. The remaining part of $\hat{\gamma}_i^+$ and six other sides of two rhombi form a figure called here a *flag*; see Fig. 2. A similar flag also should be constructed for each $\hat{\gamma}_i^-$, $1 \leqslant i \leqslant I_0^-$.

Denote by γ_i^+, $1 \leqslant i \leqslant I^+$, and γ_i^-, $1 \leqslant i \leqslant I^-$, all the segments resulting from the above construction and parallel to E^u and E^s, respectively. These include the remaining parts of $\hat{\gamma}_i^\pm$ and the sides of all the rhombi not lying on $\hat{\gamma}_i^\pm$. Note that $I^+ = 3I_0^+ + 4I_0^-$ and $I^- = 3I_0^- + 4I_0^+$.

2. Take any γ_i^-, $1 \leqslant i \leqslant I^-$. Its image $T^m\gamma_i^-$ is a segment parallel to E^s with the length $\leqslant \varepsilon\lambda^m$. Draw two segments of length $100\,K_0 m\varepsilon$, parallel to E^u, and bisected by the endpoints of $T^m\gamma_i^-$. For ε small enough those segments intersect at most $K_0 m$ curves of $S_{-m,-1}$, so that they lie mostly

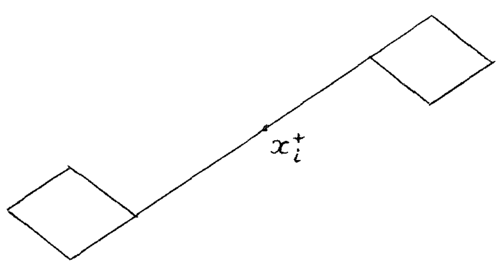

Fig. 2. A flag on a segment $\hat{\gamma}_i^+$.

outside $U_\varepsilon(S_{-m,-1})$. There is, therefore, a point x_j^- in the above net at a distance $\leqslant \varepsilon/10$ to one of those segments, and so the segment $\hat{\gamma}_j^-$ crosses both of them. Denote by $\tilde{\gamma}_j^-$ a part of $\hat{\gamma}_j^-$ confined between those two segments.

Recall that the endpoints of γ_i^- belong to two segments γ_p^+, γ_q^+ for some $1 \leqslant p < q \leqslant I^+$. Suppose now that m is so large that, say, $(1000\, K_0 m)^{-2} > \lambda^m$. The image $T^{-m}\tilde{\gamma}_j^-$ is then a segment lying at a distance $< \varepsilon/100$ to γ_i^- and also terminating at γ_p^+, γ_q^+ or at least at its prolongations beyond the endpoints of γ_i^-. Now replace γ_i^- by $T^{-m}\tilde{\gamma}_j^-$ and adjust the segments γ_p^+, γ_q^+ so that the endpoints of $T^{-m}\tilde{\gamma}_j^-$ will belong to these segments again [this may require either shortening or lengthening of a segments γ_p^+, γ_q^+, changing their lengths by less than $(2000\, K_0 m)^{-1}\varepsilon$].

Next repeat the same procedure with each γ_i^-, $1 \leqslant i \leqslant I^-$, and, likewise, with each γ_i^+, $1 \leqslant i \leqslant I^+$. We thus obtain new segments denoted by $\gamma_{i,1}^-$, $1 \leqslant i \leqslant I^-$, and $\gamma_{i,1}^+$, $1 \leqslant i \leqslant I^+$. They form similar flags to those at step 1. Important relations here are $T^m(\bigcup \gamma_{i,1}^-) \subset \bigcup \gamma_i^-$ and $T^{-m}(\bigcup \gamma_{i,1}^+) \subset \bigcup \gamma_i^+$. These imply, in particular, that none of $\gamma_{i,1}^+$ $(\gamma_{i,1}^-)$ intersects $S_{1,2m}$ (resp., $S_{-2m,-1}$).

We proceed by replacing the system $\gamma_{i,1}^\pm$ with a new one $\gamma_{i,2}^\pm$ in a similar fashion, and so on. The distances between $\gamma_{i,k}^\pm$ and $\gamma_{i,k+1}^\pm$ decrease exponentially in k for each i, so that a limit segment $\gamma_{i,\infty}^\pm = \lim \gamma_{i,k}^\pm$ is well defined for each i. The segments $\gamma_{i,\infty}^\pm$ form similar flags again. Since $\gamma_{i,k}^\pm$ do not intersect $S_{1,\pm(k+1)m}$, the segments $\gamma_{i,\infty}^\pm$ do not intersect $S_{1,\infty}$, and so $\gamma_{i,\infty}^+$ are LUMs and $\gamma_{i,\infty}^-$ are LSMs. Furthermore, $T^m(\bigcup \gamma_{i,\infty}^-) \subset (\bigcup \gamma_{i,\infty}^-)$ and $T^{-m}(\bigcup \gamma_{i,\infty}^+) \subset (\bigcup \gamma_{i,\infty}^+)$.

3. Finally, replace the above segments $\gamma_{i,\infty}^\pm$ by their images $T^\pm \gamma_{i,\infty}^{\pm m}$. They form a collection of LUMs $\tilde{\gamma}_i^+$, $1 \leqslant i \leqslant \tilde{I}^+$, and LSMs $\tilde{\gamma}_i^-$, $1 \leqslant i \leqslant \tilde{I}^-$. These segments together with the curves of $S_{-m,m}$ split M into a finite number of domains, i.e., components of connectedness of the set $M \backslash [(\bigcup \tilde{\gamma}_i^+) \cup (\bigcup \tilde{\gamma}_i^-) \cup S_{-m,m}]$. This gives the required pre-Markov partition $\xi_0 = \xi_0(\varepsilon)$.

The properties P1, P2 obviously hold. To deduce the property P3, observe that each segment $T^m \gamma_{i,\infty}^-$ $(T^{-m}\gamma_{i,\infty}^+)$ lies *strictly inside* another segment $\gamma_{j,\infty}^-$ (resp., $\gamma_{j,\infty}^+$). Consequently, each segment $\tilde{\gamma}_i^+$ $(\tilde{\gamma}_i^-)$ terminates either at $S_{1,m}$ $(S_{-m,-1})$ or at an interior point of another segment $\tilde{\gamma}_p^-$ $(\tilde{\gamma}_p^+)$, so that the property P3 follows.

A part of the property P4, namely, the upper bound for the sides of the polygons $P \in \xi_0$, readily results from our construction. The lower bound can be accomplished by a slightly more careful choice of the $(\varepsilon/10)$-nets and the sides of rhombi at step 1. We omit the details. Note that we could, alternatively, relax from the lower bound in the property P4 and instead remove a certain number of too short quadrilaterals $P \in \xi_1$ in the construction of the Markov sieve in Section 6.

ACKNOWLEDGMENTS

The author is grateful to Ya. G. Sinai, L. A. Bunimovich, D. Szasz, L.-S. Young, and G. Gallavotti for numerous enlightening discussions and to N. Simányi for reading the manuscript and commenting on the text of the paper. This work was done during the author's visit at the University of California at Los Angeles, and the author would like to thank L.-S. Young and P. Fan for their kind hospitality and K. Hirvasaho for her invaluable help in editing the English.

NOTE ADDED IN PROOF

When this paper was in press, L. S. Young pointed out a gap in the proof of Lemma 5.5. The author thanks her for the remark and anticipates submitting an addendum to the proof later.

REFERENCES

1. J.-P. Bouchaud and P. Le Doussal, Numerical study of a *d*-dimensional periodic Lorentz gas with universal properties, *J. Stat. Phys.* **41**:225–248 (1985).
2. R. Bowen, *Equilibrium States and Ergodic Theory of Anosov Diffeomorphisms* (Springer, Berlin, 1975).
3. R. Bowen, Markov partitions are not smooth, *Proc. Am. Math. Soc.* **71**:130–132 (1978).
4. L. A. Bunimovich and Ya. G. Sinai, Markov partitions for dispersed billiards, *Commun. Math. Phys.* **73**:247–280 (1980).
5. L. A. Bunimovich and Ya. G. Sinai, Statistical properties of Lorentz gas with periodic configuration of scatterers, *Commun. Math. Phys.* **78**:479–497 (1981).
6. L. A. Bunimovich and Ya. G. Sinai, Markov partitions for dispersed billiards (Erratum), *Commun. Math. Phys.* **107**:357–358 (1986).
7. L. A. Bunimovich, Ya. G. Sinai, and N. I. Chernov, Markov partitions for two-dimensional hyperbolic billiards, *Russ. Math. Surv.* **45**:105–152 (1990).
8. L. A. Bunimovich, Ya. G. Sinai, and N. I. Chernov, Statistical properties of two-dimensional hyperbolic billiards, *Russ. Math. Surv.* **46** (1991).
9. G. Casati, G. Comparin, and I. Guarneri, Decay of correlations in certain hyperbolic systems, *Phys. Rev. A* **26**:717–719 (1982).
10. N. I. Chernov, On local ergodicity in hyperbolic systems with singularities, Preprint (1991).
11. V. Donnay and C. Liverani, Potentials on the two-torus for which the Hamiltonian flow is ergodic, *Commun. Math. Phys.* **135**:267–302 (1991).
12. Dynamical Systems II, in *Encyclopaedia of Mathematical Sciences*, Vol. 2 (Springer, Berlin, 1989).

13. G. Gallavotti and D. Ornstein, Billiards and Bernoulli schemes, *Commun. Math. Phys.* **38**:83–101 (1974).
14. G. Gallavotti and P. Garrido, private communication.
15. I. A. Ibragimov and Y. V. Linnik, *Independent and Stationary Sequences of Random Variables* (Wolters-Noordhoff, Gröningen, 1971).
16. A. Katok and J.-M. Strelcyn, *Invariant Manifolds, Entropy and Billiards; Smooth Maps with Singularities* (Springer, New York, 1986).
17. A. Krámli, N. Simányi, and D. Szász, A "transversal" fundamental theorem for semi-dispersing billiards, *Commun. Math. Phys.* **129**:535–560 (1990).
18. T. Krüger and S. Troubetzkoy, Markov partitions and shadowing for nonuniformly hyperbolic systems with singularities, Preprint (1991).
19. C. Liverani and M. Wojtkowski, Ergodicity in Hamiltonian systems, Preprint (1991).
20. R. Markarian, The fundamental theorem of Sinai–Chernov for dynamical systems with singularities, Preprint (1991).
21. Ya. B. Pesin, Characteristic Lyapunov exponents and smooth ergodic theory, *Russ. Math. Surv.* **32**:55–114 (1977).
22. Ya. G. Sinai, Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards, *Russ. Math. Surv.* **25**:137–189 (1970).
23. Ya. G. Sinai and N. I. Chernov, Ergodic properties of some systems of 2-dimensional discs and 3-dimensional spheres, *Russ. Math. Surv.* **42**:181–207 (1987).
24. S. Vaienti, Ergodic properties of the discontinuous sawtooth map, *J. Stat. Phys.* **67** (1992).
25. F. Vivaldi, G. Casati, and I. Guarneri, Origin of long-time tails in strongly chaotic systems, *Phys. Rev. Lett.* **51**:727–730 (1983).
26. M. Wojtkowski, A system of one dimensional balls with gravity, *Commun. Math. Phys.* **126**:507–533 (1990).
27. L.-S. Young, Decay of correlations for certain quadratic maps, *Commun. Math. Phys.*, to appear.