DEVELOPMENTS IN CHAOTIC DYNAMICS

Lai-Sang Young*

University of California, Los Angeles
Email: lsy@math.ucla.edu

Dynamical systems as a mathematical discipline goes back to Poincaré, who developed a qualitative approach to problems that arose from celestial mechanics. The subject expanded considerably in scope and underwent some fundamental changes in the last three decades. Today it stands at the crossroad of several areas of mathematics including analysis, geometry, topology, probability, and mathematical physics. It is generally regarded as a study of iterations of maps, of time evolutions of differential equations, and of group actions on manifolds.

This article is about an area of dynamical systems called *hyperbolic dynamics* or *chaotic dynamics.* The concept of hyperbolicity, which we will define shortly, was used by Hedlund and Hopf in their analysis of geodesic flows on manifolds with negative curvature. A systematic study of hyperbolic systems began in the 1960s, when Smale outlined in his 1967 AMS Bulletin article [Sm] a program for the geometric theory of dynamical systems. Another viewpoint, namely the *ergodic theory* or probability approach to hyperbolic dynamics, was introduced several years later by Sinai and Ruelle. These ideas have over the last 30 years developed into a very rich theory, one that has changed the qualitative theory of ordinary differential equations and helped shape modern ideas about chaos.

In this article I would like to report on some developments since the 1960s. This, however, is very far from a survey. I hope that by focusing on a couple of examples and a small sample of ideas, I can convey to the general mathematics community a sense of some of the progress that has been made.

We begin with the meaning of hyperbolicity. For definiteness let us confine ourselves to *discrete time* dynamical systems, that is, to systems generated by the iteration of self-maps of manifolds (as opposed to continuous time systems or flows). A linear map $T : \mathbb{R}^n \to \mathbb{R}^n$ is called hyperbolic if none of its eigenvalues lies on the unit circle. A nonlinear map f is said to have a *hyperbolic fixed point* at p if $f(p) = p$ and $Df(p)$ is a hyperbolic linear map. Hyperbolic fixed points are therefore either *attracting* (corresponding to when all the eigenvalues of $Df(p)$ are inside the unit circle), *repelling* (when they are all outside of the unit circle), or of *saddle type*

*The author is partially supported by a Guggenheim Fellowship and a grant from the NSF

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

(when some are inside and some are outside). The idea of a hyperbolic invariant set as introduced in [Sm] is a globalization of the idea of a hyperbolic fixed point:

Let $f : M \to M$ be a diffeomorphism of a Riemannian manifold, and let $\Lambda \subset M$ be a compact invariant set, *i.e.* $f^{-1}(\Lambda) = \Lambda$. We say that f *is hyperbolic on* Λ if the tangent space at each point $x \in \Lambda$ is the direct sum of two subspaces, one of which is expanded and the other contracted by Df. One requires also that these subspaces vary continuously with x, and that they be respected by Df. When the expanding (or contracting) subspaces are trivial, it can be shown that Λ is the union of a finite number of periodic orbits. The more interesting situation is when both subspaces are nontrivial. In this case orbits are locally saddle-like, and both the structure of Λ and the dynamics on Λ can be quite complicated. A prototypical example of a nontrivial hyperbolic invariant set is Smale's horseshoe (see Fig. 1).

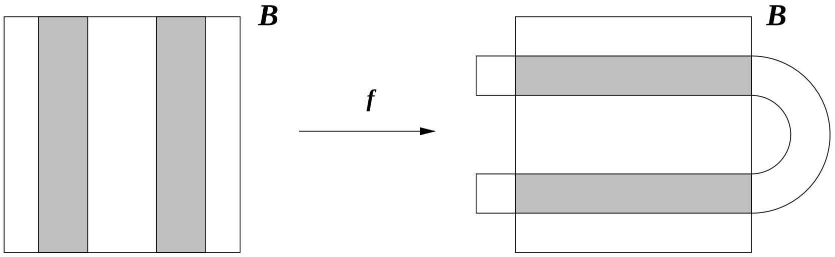

Fig. 1 The horseshoe map: B is a square; f stretches B in the horizontal direction, compresses it in the vertical direction and bends the resulting rectangle into the shape of a horseshoe; the two shaded vertical strips are mapped onto the shaded horizontal strips, and the hyperbolic invariant set $\Lambda = \cap_{i=-\infty}^{\infty} f^i(B)$ is a Cantor set.

An important characteristic of hyperbolicity when both subspaces are nontrivial is *dynamic instability*, meaning that the orbits of most pairs of nearby points diverge exponentially fast in both forward and backward times. Translations, rotations, local isometries, or fixed points with multipliers equal to 1 are among the simplest examples of nonhyperbolic behavior. Two terminologies introduced in the 1960s that we will encounter later on in this article are *Anosov diffeomorphisms*, which refers to maps that are hyperbolic on the entire manifold M, and *Axiom A*, a condition satisfied by maps that are hyperbolic on certain essential parts of M. (We will not need to know its precise definition.)

The 1970s brought new outlooks and new challenges. With the aid of computer graphics, researchers became increasingly aware of the abundance of examples whose dynamics are dominated by expansions and contractions but which do not meet the rather stringent requirements of Axiom A. Two famous examples are the Lorenz ("butterfly") attractors and Hénon mappings. At about the same time, a version of hyperbolicity with considerably weaker assumptions emerged following the works of Oseledec and Pesin. We will discuss this in more detail later, but suffice it to say now that in this weaker version, "expansions and contractions everywhere" on a compact set is replaced by "*asymptotic* expansions and contractions *almost everywhere*". Some old results continue to hold in this more general setting (it is a bit like extending theorems for continuous functions to measurable ones), and new phenomena have been discovered. In light of these developments, we will refer to the definition of hyperbolicity we gave earlier on as *uniform hyperbolicity*, to distinguish it from more general notions which I will refer to loosely as *nonuniform hyperbolicity.*

This article is about nonuniformly hyperbolic systems, with emphasis on their ergodic theory. I would like to focus on the following two directions of progress: the development of a general theory, and the application of hyperbolic techniques to specific examples. I will select two sample results from each one of these directions and discuss some of the ideas behind the theorems. The two applications I have chosen are billiards and Hénon attractors. For the general theory part, my two topics are (1) entropy, Lyapunov exponents, and dimension, and (2) correlation decay and central limit theorem.

Let me try to put things in perspective before continuing. Since the time of its conception, hyperbolic theory has developed in many different directions; the ergodic theory of nonuniformly hyperbolic systems is one of them and this is the topic I have chosen to write about. Other important topics that I will not touch upon include partial hyperbolicity, bifurcation theories, one-dimensional dynamics, real and complex (to the degree that expanding properties are involved), group actions and geometry, etc. Within the topic of ergodic theory of hyperbolic systems, I have also made choices that are clearly biased toward my own interests, although I hope that the results I am presenting are not an unreasonable sample.

BILLIARDS

A billiard flow is the motion of a point mass in a bounded domain $\Omega \subset \mathbb{R}^2$ or $\mathbb{T}^2$ where $\partial\Omega$ is the union of a finite number of smooth curves. The point moves at unit speed, and bounces off $\partial\Omega$ according to the usual laws of reflection, that is, the angle of incidence is equal to the angle of reflection. There is a natural section to this flow given by the surface $M = \partial\Omega \times [-\frac{\pi}{2}, \frac{\pi}{2}]$ which corresponds to collisions with $\partial\Omega$. It is convenient to think of $p = (x, \theta) \in M$ as represented by an arrow with footpoint at $x \in \partial\Omega$ and making an angle θ with the normal pointing into Ω. (See Fig. 2.) We consider the Poincaré map or first return map f from this section to itself and call it the billiard map for the domain Ω. It is straightforward to check that f leaves invariant the probability measure $\mu = c \cos\theta \; dx \; d\theta$ where c is the normalizing constant, *i.e.* $\mu(f^{-1}E) = \mu(E)$ for every Borel measurable set $E \subset M$, and c is chosen so that $\mu(M) = 1$.

Not all billiards have hyperbolic properties. In the case where Ω is an ellipse, for example, it is an exercise to see that the envelope of every (infinite) billiard trajectory is an ellipse or a hyperbola having the same foci as Ω (Fig.2(a)). One could thus picture M as being foliated by simple closed curves left invariant by the action of f, which "rotates" points around within each curve. This kind of dynamics is called *quasi-periodic*; it has a very different flavor from hyperbolic dynamics. In the case of a polygonal domain (Fig.2(b)), it is also easy to see that f does not expand or contract distances.

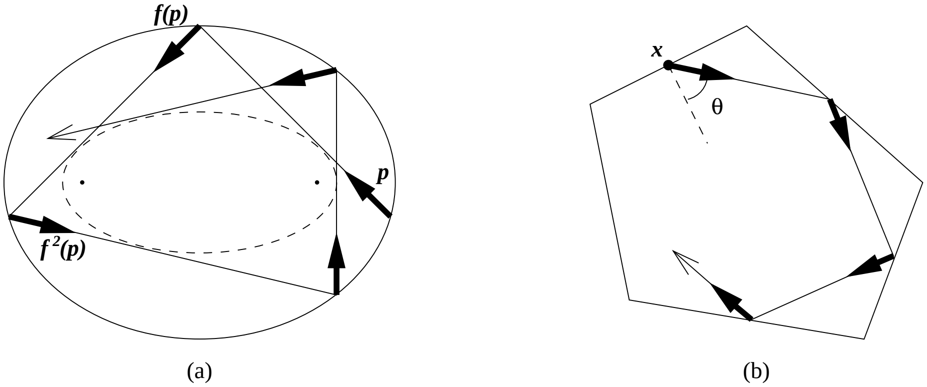

Fig. 2 Examples of nonhyperbolic billiards

Sinai was the first to investigate rigorously billiards with hyperbolic properties. He studied in [S2] billiards of *dispersing* type corresponding to when $\partial\Omega$ is the union of a finite number of "concave" pieces. (Concave boundaries, by convention, refer to boundary curves whose center of curvature at each point lies outside of Ω.) Two standard examples of billiards of this type are billiards on the 2-torus with a finite number of "scatterers" made up of disjoint convex regions (Fig. 3(a)) and those on planar domains as shown in Fig. 3(b).

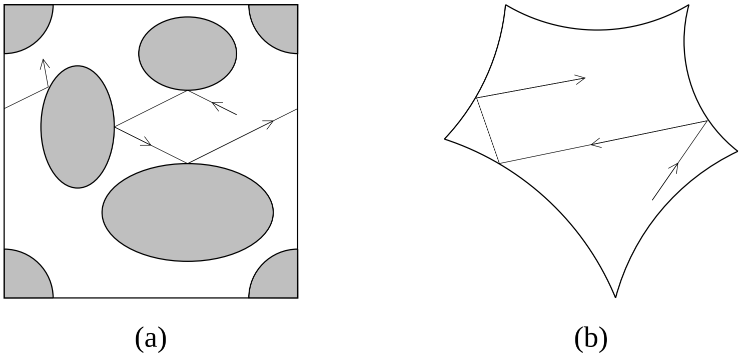

Fig. 3 Dispersing billiards

Let us observe why billiard maps associated with dispersing billiards have hyperbolic properties. A tangent vector at $p \in M$ can be represented by a curve in M, which in turn can be thought of as a parametrized family of arrows containing the one corresponding to p. We distinguish between families of arrows that are *divergent* and those that are *convergent*, and note that divergent families correspond to a sector, or a *cone*, in the tangent space to M at p. Since divergent families of rays become even more divergent upon being reflected off a concave boundary piece (see Fig. 5(a)), we see that Df maps the cone corresponding to divergent rays at p strictly into that at $f(p)$. (See Fig. 4.) Finding a continuous family of cones in tangent spaces that are mapped strictly into themselves by Df is a standard way of proving uniform hyperbolicity – it shows that projectively, at least, Df behaves like a hyperbolic linear map.

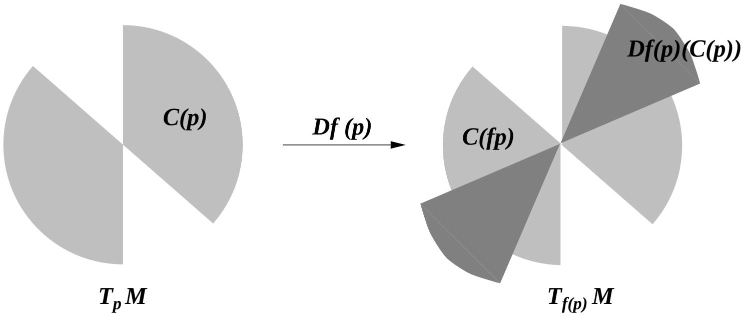

Fig. 4 $Df(p)$ maps $C(p)$, a cone in the tangent space at p, strictly into $C(f(p))$, a cone in the tangent space at $f(p)$. A standard way of proving hyperbolicity is to locate a family of invariant cones.

A few words of caution are in order here. First, billiard maps such as those in Fig. 3(a) are *discontinuous*. Consider a trajectory of the point mass that meets $\partial\Omega$ tangentially. Trajectories slightly to the left and to the right of this one will run into different components of $\partial\Omega$ (see Fig. 5(b)). Second, billiard maps have unbounded derivatives (see Fig. 5(c)). These properties make them considerably more complicated than Anosov diffeomorphisms.

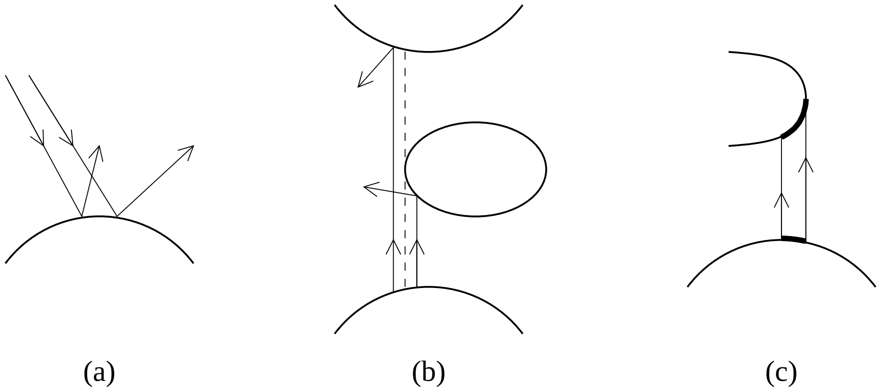

Fig. 5 Properties of dispersing billiards

An important breakthrough in the study of billiards is the following theorem of Sinai, which can be viewed as lending support to Boltzmann's Ergodic Hypothesis, a hypothesis that is part of the foundation of statistical mechanics.

Theorem [S2]. *Dispersing billiards are ergodic.*

For a measure preserving transformation, *ergodicity* means that there are no invariant sets having intermediate measure. The proof of Sinai's theorem is far too involved to be given here, but I would like to try to explain the idea of the proof assuming that the discontinuity curves were not there, and to indicate what problems they cause.

Suppose now that f is Anosov, *i.e.* uniformly hyperbolic (without discontinuities) on the entire manifold, and suppose that f preserves a probability measure μ equivalent to the volume element. By Birkhoff's Ergodic Theorem, we know that for every L^1 function φ, the trajectory averages $\frac{1}{n}\sum_{i=0}^{n-1}\varphi\circ f^i$ converge μ-a.e., and that the limit is equal to $\int\varphi d\mu$ if (f,μ) is ergodic. To prove ergodicity, then, it suffices to check that these limit functions are constant almost everywhere; in fact, it suffices to do this for continuous φ. Our proof follows an idea due to Hopf. It uses strongly the fact that for an Anosov diffeomorphism, the contracting and expanding directions can be integrated to form a pair of invariant foliations. The leaves of these two foliations are called *stable* and *unstable manifolds.* Now for two points x and y on the same stable manifold, since $d(f^nx, f^ny)\to 0$ as $n\to\infty$, it follows that their trajectory averages must tend to the same limit as $n\to\infty$. This argument in backward time gives a similar conclusion for points on the same unstable manifold. Since locally stable and unstable manifolds form a Cartesian coordinate system (topologically, at least), it seems as if it would follow immediately that the limit function, which is constant on both stable and unstable manifolds, would be locally constant. The validity of this argument actually relies on the *absolute continuity* of the foliations, a fairly subtle property that says that the holonomy maps of these foliations carry sets of Lebesgue measure zero on transversals to sets of Lebesgue measure zero. It is a fact that the stable and unstable foliations of Anosov diffeomorphisms are absolutely continuous if f is C^2, and hence "local ergodicity" is proved. To prove that the limits of trajectory averages are globally constant a.e. requires a separate argument that we will omit.

Back with billiard maps, the discontinuity curves "chop up" the stable and unstable curves making some of them arbitrarily short. This destroys the local product structure which is essential in our proof of local ergodicity in the last paragraph. A great deal of work has to be done to overcome this.

We remark that in addition to ergodicity, more refined statistical properties of these and other billiards have been studied by Sinai, Bunimovich, Chernov (see e.g. [BSC]), and others. We will return to some of these properties later on in the article.

We saw from the examples above that the geometry of Ω influences strongly the dynmical properties of the billiard map. It is not the case, however, that hyperbolic behavior is limited to concave boundaries. Convex boundaries, such as those in the *stadium* studied by Bunimovich (see Fig. 6), can also produce hyperbolicity if certain conditions are met. This is because even though nearly parallel rays first become convergent upon reflection, they diverge after focussing, and expansion for the billiard map results if, before the next collision, these rays have diverged more than they have converged.

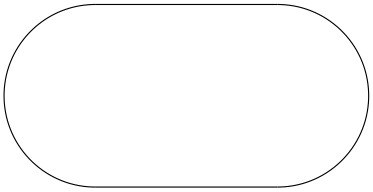

Fig. 6 The stadium

One could in fact prove that the billiard map associated with the stadium has some weak saddle-like behavior almost everywhere. This is called *nonzero Lyapunov exponents.* It means more precisely that almost everywhere on M there is a splitting of the tangent space into two invariant directions E^u and E^s so that for $v \in E^u$, $|Df^n v| \sim e^{n\lambda}$ for some $\lambda > 0$ as $n \to \infty$, and the same holds for $v \in E^s$ with $\lambda < 0$. Note that the hyperbolicity here is very nonuniform: billiard trajectories that are nearly perpendicular to the two straight sides, for example, will bounce back and forth for a long time without diverging, and in the meantime $v \in E^u$ will not expand. Geometric conditions on billiard domains Ω that give rise to nonzero Lyapunov exponents are formulated in [W].

Finally, we remark that billiards are in some sense low-dimensional models of interactions of large numbers of hard balls in, for example, 3-space. We refer the reader to [Sz] for an exposition on what is known about these systems, and close this section with a report on the latest development: it has been announced very recently by Simányi and Szász that with no restriction on the number of balls, systems of finitely many balls in a torus with typical mass distributions have now been proved to have nonvanishing Lyapunov exponents.

HÉNON ATTRACTORS

The Hénon maps are a 2-parameter family of diffeomorphisms of the plane given by

$$T_{a,b}(x,y) \;=\; (1 - ax^2 + y,\; bx).$$

In certain parameter ranges $T_{a,b}$ is known to have an attractor. An *attractor* is an invariant set Ω with the property that it attracts all nearby orbits, that is to say, for any starting point z near Ω, the orbit of z will in time be drawn toward Ω. The equations above were first investigated numerically in 1977 by the astronomer Hénon, who observed that they have attractors with very complicated dynamics. Many numerical studies were carried out in the late 70's and early 80's; analytically these maps remained intractable until quite recently.

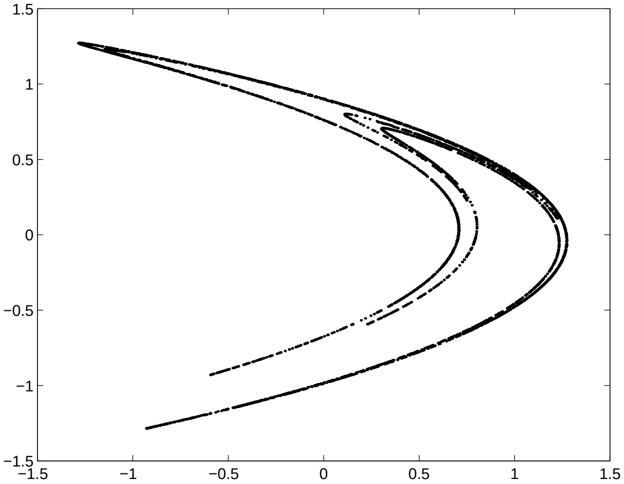

Fig. 7 This is the computer plot of a single orbit of length 5000 for the map $(x, y) \mapsto (1 - 1.4x^2 + 0.3y,\ x)$, the original map studied by Hénon. The overall appearance of the picture does not seem to depend on the choice of initial condition provided it is chosen from certain regions of the plane. This particular picture is generated using the initial condition $(x, y) = (0, 0)$.

We would like to consider here the following two questions: (1) Are the Hénon attractors *chaotic*, and what do we mean by that? (2) What do computer plots such as those in Fig. 7 really represent? (1) and (2) are general questions not at all particular to the Hénon maps, but we will use this family to illustrate some of the underlying issues. For definiteness we consider only parameter values with $a < 2$ and close to 2, and b very small. These are the parameters studied by Benedicks and Carleson; they represent a very small fraction of the parameters for which attractors are known to exist.

We begin with some elementary geometric facts. Let $T = T_{a,b}$ with (a, b) fixed. From the equations of T it follows easily that T maps vertical lines to horizonal lines and sends horizontal lines to parabolas (see Fig. 8). Observe also that T contracts area strongly, with $|\det(DT)| = b$. It is not hard to show that away from the y-axis, say outside of the region $\{|x| > \sqrt{b}\}$, the dynamics is essentially uniformly hyperbolic of saddle type: nearly horizontal tangent vectors are mapped by DT to nearly horizontal vectors, and they grow exponentially after a while. Horizontal segments near the y-axis, however, are mapped to the turns of parabolas. Thus when an orbit gets near the y-axis, directions of expansion and contraction may get mixed up and hyperbolicity may be spoiled.

Another elementary fact is that T has a compact invariant set Ω located near $[-1, 1] \times \{0\}$; Ω is an *attractor* in the sense that there is an open set $U \subset \mathbb{R}^2$ containing it with the property that for every $z \in U$, $d(T^n(z), \Omega) \to 0$ as $n \to \infty$. The maximal set with this property is called the *basin* of Ω; it is an open and relatively large set, whereas Ω, being a compact invariant set of an area contracting map, has Lebesgue measure 0. It is not hard to prove that Ω is *not* a uniformly hyperbolic or Axiom A attractor.

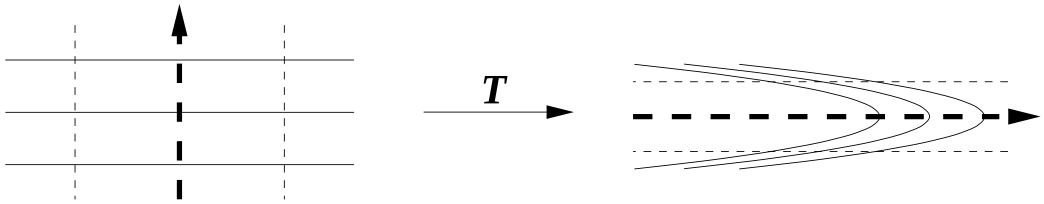

Fig. 8 The geometry of Hénon maps: vertical lines are mapped to horizontal lines; horizontal lines are mapped to parabolas

Consider now the dynamics on Ω. There are two competing scenarios. The first is that most orbits tend eventually to attractive periodic cycles, which are also called *periodic sinks.* To see why this may be the case, recall that $|\det(DT)|$ is very small. If for some z, $T^n z$ comes near z and $DT^n(z)$ is contracting in all directions, then the Contraction Mapping Theorem gives a periodic sink of period n. Newhouse observed some time ago that this happens easily near tangencies of stable and unstable manifolds. He showed in fact that under certain conditions one typically expects to find infinitely many sinks [N].

A counter-scenario is that the dynamics on Ω is predominantly hyperbolic of saddle type, and the resulting dynamic instability gives rise to a rather "chaotic" picture. ("Chaotic" is used as a descriptive word here; to my knowledge it has no accepted mathematical definition.) The reasoning is as follows. If b is small, then the strip $\{|x| \leq \sqrt{b}\}$ is very narrow, and the orbit of an arbitrary point z is likely to spend most of its time outside of this strip where the map is uniformly hyperbolic of saddle type. Now let us not be so naive as to believe that a visit to the region $\{|x| \leq \sqrt{b}\}$ once in a long while cannot do any harm: consider, for instance, the matrix product $A_{2N} \cdots A_1 A_0$ where $A_i = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$ for all $i \neq N$ and $A_N = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. It is not unreasonable, however, to think that while cancellations of this type can and do occur to some degree for the Hénon maps, they are very unlikely to be so severe that no exponential growth in $||DT^n||$ survives.

For most parameters (a, b), it is not known which one of these scenarios occurs, nor is it clear that both cannot co-exist on different parts of Ω; the picture is simply extremely complicated. It is known that sinks exist for an open set of parameters. Numerics as well as intuitive thinking favor the "chaotic" regime in certain parameter ranges, but cancellations of the type in the last paragraph are not easy to deal with. In 1991 Benedicks and Carleson published a 100-page paper [BC] in which they proved that there is a positive measure set of parameters for which the "chaotic" scenario prevails. In [BC] a scheme for tracking the cumulative derivative DT^n is devised and $||DT^n||$ is shown to grow exponentially for many points in the attractor. This work is a breakthrough in hyperbolic dynamics, for even though their analysis is carried out only for the Hénon maps, it has potential applications in other situations. Their scheme is too involved for me to outline here, but I would like to give a much simplified version of it, namely for the case $b = 0$, which corresponds to the 1-dimensional map $f_a(x) = 1 - ax^2,\ x \in [-1, 1]$.

Suppose we wish to prove the exponential growth of $|(f^n)'|$ for arbitrary points in $[-1, 1]$. Clearly, the problem here is the critical point 0. When an orbit gets near 0, its cumulative derivative experiences a sharp drop. A fruitful idea, first used by Collet and Eckmann, is to impose a condition of the type $|(f^n)'(f0)| \geq \lambda^n$ for some $\lambda > 1$. This, together with some control on how fast the critical orbit is allowed to approach 0, gives easily the following estimates: For x near 0, $|f'(x)| \sim |x|$ and $|f(x) - f(0)| \sim x^2$, so the orbit of $f(x)$ will stay near the orbit of $f(0)$ for n times where n is determined by $\lambda^n x^2 \sim 1$. Hence $|(f^{n+1})'x| \sim |x|\lambda^n \sim |x|^{-1} \sim \lambda^{n/2}$ and the small derivative at x is fully compensated for after n iterates. The conditions imposed on the critical orbit have been shown to hold for a positive measure set of a's. The scheme of Benedicks and Carleson has a similar flavor, but with angles as well as lengths to control, the increase in complexity from one to two dimensions is quite substantial.

Let us turn now to the second question. A standard way of making a computer picture of the Hénon attractor is to pick an initial condition in the basin of the attractor and to plot the first few thousand iterates of its orbit. (Initial conditions are typically taken from the basin and not necessarily from the attractor itself because, as we recall, Ω is a measure zero set and it is hard to know exactly which points lie in it.) Since the plotted orbit limits on the attractor, one often *assumes* that the resulting plot is "the picture" of the attractor. (See Fig. 7.) This leads naturally to the following question: We know that orbits of the Hénon map are not all the same; some are periodic, others are not; some come closer to the turns than others. We also know from experience that (for a fixed T) one gets essentially the same picture independent of the choice of initial condition. Is there a mathematical explanation for this?

We do not pretend to have an answer for this intriguing and very important question, but use it instead to motivate the idea of *Sinai-Ruelle-Bowen* or *SRB measures.* Our computer picture can be thought of as the picture of a probability measure which gives mass $\frac{1}{n}$ to each point in an orbit of length n. Let δ_z denote point mass at z. If there is a measure μ with the property that $\frac{1}{n}\sum_{i=0}^{n-1} \delta_{T^i z} \to \mu$ for "most" choices of initial conditions z, that would explain why our pictures tend to look similar. Now this measure μ, if it exists, would have to have the following very special property: like all invariant probability measures, it must be supported on Ω, but somehow μ has the ability to influence orbits starting from various parts of the basin, including points that are far away from the support of μ.

For the parameters studied by Benedicks and Carleson, this very special invariant measure does in fact exist. Recall that a positive Lyapunov exponent at z means that $|DT^n(z)v|$ grows exponentially as $n \to \infty$ for some vector v.

Theorem [BY]. *For a positive Lebesgue measure set of parameters (a, b), the Hénon map $T = T_{a,b}$ admits an invariant probability measure μ on its attractor Ω with the following properties:*

(a) *f has a positive Lyapunov exponent μ-a.e.*;

(b) *for z in a positive Lebesgue measure set in the basin of Ω, the sequence $\frac{1}{n}\sum_{i=0}^{n-1}\delta_{T^i z}$ converges weakly to μ as $n \to \infty$.*

The measure μ in the theorem above is called an *SRB measure*. SRB measures were first discovered in the context of Anosov diffeomorphisms and Axiom A attractors by Sinai [S1], Ruelle, and Bowen (see *e.g.* [B]). Not a great deal is known about their existence outside of the Axiom A category. The Hénon attractors are the first genuinely nonuniformly hyperbolic attractors for which SRB measures were shown to exist; they were constructed by Benedicks and the author. Benedicks and Viana have announced recently that they have extended the result in (b) to almost every point in the basin.

Using the Hénon family as an example, we have introduced in this section several ideas that are state-of-the-art in our understanding of *strange attractors*. We summarize them as follows. First, the existence of an attractor is in itself a stabilizing phenomenon, as a large set of points (namely the basin) is drawn toward what is typically a very small set (namely the attractor). As two nearby orbits from the basin approach the attractor, however, they may diverge quickly. This is due to dynamic instability or chaos within the attractor. Finally, dynamic instability does not imply the absence of coherent behavior, for if an SRB measure exists, then all observable orbits will in the long run have the same statistics, meaning they will visit different parts of the attractor with frequencies governed by this invariant measure.

ENTROPY, LYAPUNOV EXPONENTS and DIMENSION

We now shift to the part on general theory. To motivate our results in this section, we consider first a simple-minded way of building fractals from a single template.

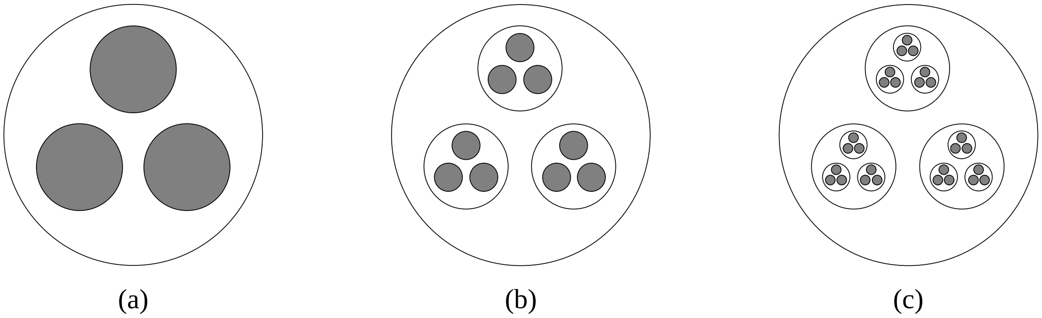

Fig. 9 Construction of fractal from a single template

Fig. 9(a) shows a template consisting of a larger ball with three smaller balls inside. In Fig. 9(b) we put a scaled down copy of this template on each of the 3 smaller balls, constructing 9 balls that are another size smaller. This procedure is repeated in Fig. 9(c) on each of the 9 balls, and so on. Continuing *ad infinitum* and taking the intersection, we obtain a fractal Λ which is nothing other than a standard Cantor set.

All this can be said in the language of dynamical systems. Let us call the large ball in the template B and the smaller balls B_i. Let $f : \cup B_i \to B$ be such that it maps each B_i affinely onto B. Then $\Lambda = \cap_{n=0}^{\infty} f^{-n}(\cup B_i)$. It is natural to try to relate the fractal dimension of Λ to the characteristics of its generating dynamical system. Assume for simplicity that all the B_i's have the same radii. Consider $\lambda := \log(\text{radius } B/\text{radius } B_i)$ and $h := \log \#(B_i)$. To understand the relation among h, λ, and the Hausdorff dimension δ of Λ, we fix one of these numbers, vary a second, and observe the effect on the third. This is illustrated in Fig. 10. From Figs. 10(a) and (b), it seems intuitively clear that if we decrease λ while keeping h fixed, then δ goes down; likewise Figs. 10(b) and (c) should convince us that if we increase h while keeping λ fixed, then δ goes up.

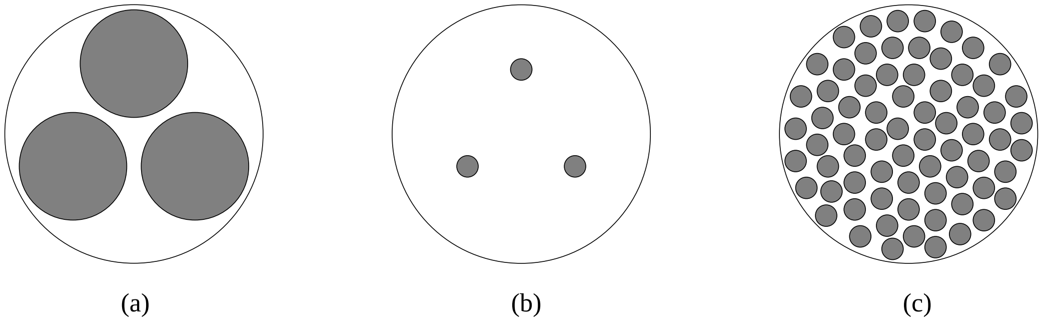

Fig. 10 Three different templates: observe how the dimension of the fractal changes with the number and diameter of the smaller balls

To turn these observations into a theorem that holds for all diffeomorphisms (the derivative of which varies from point to point), one possibility is to consider *averaged* quantities. This leads to the introduction of an invariant measure. We consider for the rest of this section a pair (f, μ), where f is a C^2 diffeomorphism of a compact Riemannian manifold M and μ is an f-invariant Borel probability measure on M.

We have encountered the idea of Lyapunov exponents before. Here is a more systematic discussion. Let v be a tangent vector at x. We call $\lambda = \lambda(x, v)$ the *Lyapunov exponent* at x in the direction of v if $|Df^n(x)v| \sim e^{\lambda n}$ as $n \to \infty$. There is a matrix version of the ergodic theorem due to Oseledec that tells us that $\lambda(x, v)$ is well defined for every v at μ-a.e. x. In fact, at μ-a.e. x, there is a decomposition of the tangent space into a direct sum of subspaces $E_1(x) \oplus \cdots \oplus E_r(x)$ with the property that $\lambda(x, v)$ is constant for all v in $E_i(x)$; this common value of $\lambda(x, v)$ is denoted $\lambda_i(x)$. The decomposition into $\oplus E_i(x)$ is in a sense a generalization of the decomposition into eigenspaces for a single linear map. It is easy to see that $\lambda_i(x) = \lambda_i(fx)$, so that if (f, μ) is ergodic, then these numbers are constant μ-a.e., and the asymptotic properties of Df^n can be summed up in a finite set of numbers $\lambda_1 > \lambda_2 > \cdots > \lambda_r$ with multiplicities $m_1, \cdots, m_r,\ m_i = \dim E_i$.

The translation of this linear theory into a nonlinear one describing the action of f in neighborhoods of typical trajectories was carried out by Pesin, who proved among other things that at μ-a.e. x, there is an *unstable manifold* passing through x. This manifold, denoted $W^u(x)$, is tangent at x to $E^u(x) := \oplus_{\lambda_i > 0} E_i(x)$ and is characterized by $\{y \in M : d(f^{-n}x, f^{-n}y) \to 0 \text{ as } n \to \infty\}$. Analogously there is a *stable manifold* tangent to the stable subspace at almost every point.

While Lyapunov exponents measure *geometrically* how fast orbits diverge, the metric entropy of (f, μ), written $h_\mu(f)$, measures complexity in the sense of *randomness* and *information.* This notion was introduced by Kolmogorov and Sinai around 1959. Roughly speaking, it measures the amount of uncertainty one faces when attempting to predict future behaviors of orbits based on knowledge of their pasts. The formal definition of $h_\mu(f)$ is a little hard to give in this limited space; so let me define it instead via the Shannon-Breiman-McMillan Theorem: Let α be a finite partition of our manifold M. For $n \geq 0$, we let α^n be the partition whose elements are sets of the form $\alpha^n(x) := \{y \in M : f^i x \text{ and } f^i y \text{ belong in the same element of } \alpha \text{ for all } 0 \leq i \leq n\}$. For simplicity let us assume that (f, μ) is ergodic. Then the Shannon-Breiman-McMillan Theorem says that there is a number h (which we will take to be the definition of $h_\mu(f)$) such that if α is a sufficiently fine partition, then for all sufficiently large n, neglecting a set of small μ-measure we may think of M as made up of $\sim e^{nh}$ elements of α^n each having μ-measure $\sim e^{-nh}$. For a more precise statement we refer the reader to a standard ergodic theory text, but for our purposes it suffices to think of e^{nh} as the rate of growth in complexity of f counting only orbits that are "typical" with respect to μ.

Since Lyapunov exponents and metric entropy both reflect properties of the invariant measure, they can only be related via notions that pertain to the measure. Let ν be a Borel probability measure on a compact metric space X, and let $B(x, r)$ denote the ball of radius r about x. We say the *dimension of the measure* ν, $\dim(\nu)$, is well defined and is equal to α if for ν-a.e. x, $\nu B(x, r) \sim r^\alpha$ as $r \to 0$. The relation between $\dim(\cdot)$ and Hausdorff dimension (HD) is that if $\dim(\nu) = \alpha$, then $\text{Inf}\{HD(Y) : Y \subset X, \nu(Y) = 1\} = \alpha$.

Before giving the full statement of our theorem, it is instructive to consider first the special case where f has a single Lyapunov exponent $\lambda > 0$ (such an f is necessarily noninvertible, but that is fine). Let

$$B(x, \epsilon; n) := \{y \in M : d(f^k x, f^k y) < \epsilon \ \forall\ 0 \leq k \leq n\}.$$

Then

$$B(x, \epsilon; n) \sim B(x, \epsilon e^{-\lambda n}),$$

and a small modification of the Shannon-Brieman-McMillan Theorem tells us that

$$\mu B(x, \epsilon; n) \sim e^{-nh}.$$

Putting these two lines together gives

$$\mu B(x, r) \sim r^{\frac{h}{\lambda}},$$

which proves that $\dim(\mu)$ exists and is related to h and λ by $h = \lambda \cdot \dim(\mu)$.

The argument above relies on the fact that we are able to generate dynamically sets that approximate round balls. When there is more than one positive Lyapunov exponent, this is impossible and the proofs become considerably more involved. In the theorem below, f is allowed to be any C^2 diffeomorphism and μ any invariant Borel probability measure. Recall that E_i is the subspace corresponding to the Lyapunov exponent λ_i. The conditional measures of μ on unstable manifolds are denoted $\mu|W^u$, and $a^+ := \max(a, 0)$.

Theorem. *Assume for simplicity that (f, μ) is ergodic. Then corresponding to each λ_i, there is a number δ_i with $0 \le \delta_i \le \dim E_i$ such that*

(a) $h_\mu(f) = \sum_i \lambda_i^+ \delta_i$,

(b) $\dim(\mu|W^u)$ *exists and is equal to* $\sum_{\lambda_i > 0} \delta_i$.

Moreover, if $\lambda_i \neq 0$ for any i, then

(c) $\dim(\mu)$ *exists and is equal to* $\dim(\mu|W^u) + \dim(\mu|W^s)$.

The numbers "δ_i" have geometric interpretations as *partial dimensions* of μ in the directions of E_i. With this in mind, the dimension formula in part (a) can be understood as saying that in general, $h = \vec{\lambda} \cdot \vec{\delta}$ where $\vec{\lambda}$ and $\vec{\delta}$ are the Lyapunov exponent and partial dimension *vectors*. Parts (a), (b), and the "$\le$" part of (c) of this theorem are proved by Ledrappier and the author [LY]. The reverse inequality in (c) is proved in a recent preprint by Barreira, Pesin, and Schmeling.

We remark that the dimension formula above can be viewed as a refinement of two very important results proved earlier on: *Ruelle's Inequality* , which says that $h_\mu(f) \le \int \sum \lambda_i^+ \dim E_i \, d\mu$ [R2], and *Pesin's Formula* , which says that equality holds when μ is equivalent to the Riemannian measure [P]. Since the gap in Ruelle's Inequality can be viewed as an indication of the amount of "dissipation" in a dynamical system, the dimension of an invariant measure has that interpretation as well. For a more detailed survey of this topic see [ER].

CORRELATION DECAY and CENTRAL LIMIT THEOREM

In this last section we consider sequences of observations from dynamical systems and treat them as random variables in probability. More precisely, let $f : M \to M$ be a dynamical system, μ an invariant probability measure, and $\varphi : M \to \mathbb{R}$ a function which we think of as a quantity that can be measured or observed (for example, temperature in an experiment). We regard the sequence of functions

$$\varphi, \quad \varphi \circ f, \quad \varphi \circ f^2, \quad \cdots, \quad \varphi \circ f^n, \quad \cdots$$

as random variables on the underlying probability space (M, μ), and ask how they compare qualitatively with genuinely random stochastic processes (such as outcomes from flipping a coin).

In this context, the *Strong Law of Large Numbers*, which says that $\frac{1}{n}\sum_0^{n-1} \varphi \circ f^i$ converges to $\int \varphi d\mu$ almost surely, holds when (f, μ) is ergodic; this is simply the Birkhoff Ergodic Theorem, which we have encountered many times. One could also ask if the *Central Limit Theorem* holds, that is to say, for φ with $\int \varphi d\mu = 0$ we may ask if

$$\frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} \varphi \circ f^i \quad \overset{\text{distr}}{\longrightarrow} \quad \mathcal{N}(0, \sigma)$$

for some $\sigma > 0$ where $\mathcal{N}(0, \sigma)$ is the normal distribution (or bell-shaped curve) with variance σ^2. Another standard question concerns the correlation between φ and $\varphi \circ f^n$ for large n. More precisely, if

$$\Phi(n) := \left| \int (\varphi \circ f^n) \varphi d\mu \; - \; \left(\int \varphi d\mu \right)^2 \right| ,$$

then one could ask if $\Phi(n)$ tends to zero as n tends to infinity and at what speed. For example, if $\Phi(n) \sim e^{-\alpha n}$ for some $\alpha > 0$ independent of φ, then this is a property of the dynamical system (f, μ) and we say (f, μ) has *exponential decay of correlations*. Similarly, if $\Phi(n) \sim n^{-\alpha}$ for some $\alpha > 0$, then we say (f, μ) has *polynomial decay*, and so on.

These are, I believe, very important questions. Current techniques do not permit us to deal with them for completely general dynamical systems. Thus, following the theme of this article, we will limit ourselves to maps that geometrically have a great deal of expansion and contraction on large parts of their phase spaces. This (not so rigorously defined) class includes Axiom A diffeomorphisms as well as the examples discussed earlier on. We take the view here that only properties that hold on positive Lebesgue measure sets are observable. Accordingly, if a system is "conservative", *i.e.* if f preserves a measure equivalent to Lebesgue measure, then this will be the measure of interest to us. If a system is "dissipative" (meaning not conservative), then we will take μ to be an *SRB measure* if one exists, for as we saw in the section on Hénon attractors, these are the measures that reflect the properties of Lebesgue measure in dissipative systems. Since relatively little is known about the existence of SRB measures in general, this existence question will be our first and foremost challenge with regard to dissipative systems.

Our next remark has to do with the ergodic and mixing properties of (f, μ). *Mixing* is a stronger condition than ergodicity; it says that for every pair of Borel sets A and B, $\mu(f^{-n}A \cap B) \to \mu(A)\mu(B)$ as $n \to \infty$. Some of our questions make sense only if (f,μ) is mixing, which is essentially equivalent to the assertion that the function $\Phi(n)$ above tends to zero as $n \to \infty$. Even though not all (f,μ) are ergodic or mixing, there is a theorem due to Pesin and Ledrappier saying that if μ is smooth or is SRB, and if f has no zero Lyapunov exponents μ-a.e., then (f,μ) is made up of at most a countable number of ergodic components each one of which is mixing up to a finite cycle (see e.g.[P]). Thus our questions are always relevant on each *mixing component.*

Finally, it is not hard to see that in deterministic settings such as ours the speed of mixing can be arbitrarily slow if we do not impose some regularity on our test functions. Thus φ will always be assumed to be Hölder continuous.

For the class of dynamical systems under consideration, the situation can be summarized as follows: For Anosov diffeomorphisms and Axiom A attractors, SRB measures always exist, correlation decay is exponential, and the central limit theorem always holds (see e.g. [R1]). Outside of the Axiom A category, much of the progress up until recently has been focused on individual classes of examples, and these examples suggest that many distinct behaviors are possible. For instance, there are examples on the boundary of Axiom A that do not admit SRB measures; others do but have polynomial decay.

In the remainder of this article I would like to report on some recent work that attempts to study systematically the statistical properties above. My goals are (1) to give verifiable conditions for these properties and (2) to relate them to the geometry of the map. These conditions are formulated in terms of *recurrence times* or *renewal times* and are defined for an object whose construction requires some degree of hyperbolicity. I will begin with a description of this object. For simplicity of exposition, allow me to treat temporarily f as though it were an expanding map, omitting details in connection with collapsing along local stable manifolds for systems with contracting directions. The idea is as follows: Pick an arbitrary set Λ with reasonable properties and with $m(\Lambda) > 0$ where m is Lebesgue measure. Think of Λ as a reference set, and regard $\Lambda' \subset \Lambda$ as having "renewed" itself or "returned" to Λ at time n if f^n maps Λ' diffeomorphically onto Λ. We run the system until almost all points of Λ have returned, decomposing Λ into a disjoint union of subsets $\{\Lambda_i\}$ each returning at a different time. Let R be the return time function. We claim that the statistical properties of f are to a large extent reflected in the asymptotics of the sequence $m\{R > n\}$.

To be sure, the results below are not precisely stated. One important analytic ingredient that we have left out is the *control of nonlinearities*, which is essential for maintaining a certain degree of "independence" for the dynamics between returns to the reference set. Referring the reader to [Y1] and [Y2] for details, we state:

Theorem. *Let f, Λ, m and R be as above.*

(a) *If $\int R dm < \infty$, then f leaves invariant a probability measure μ that is smooth or SRB; μ is unique if we require (f, μ) to be ergodic with $\mu(\Lambda) > 0$.*

(b) *If, in addition,* $\gcd\{R\} = 1$*, then (f, μ) is mixing.*

(c) *If $m\{R > n\} < C\theta^n$ for some $\theta < 1$, then correlation decay is exponential.*

(d) *If $m\{R > n\} = \mathcal{O}(n^{-\alpha})$ for some $\alpha > 1$, then correlation decay is $\mathcal{O}(n^{-\alpha+1})$.*

(e) *If R is as in* (d) *and $\alpha > 2$, then the Central Limit Theorem holds.*

The theorem above relates the statistical properties of f to the tail of its recurrence times. We claim that the latter is closely connected with *the speed with which arbitrarily small pieces of unstable manifolds grow to a fixed size.* This is because in order to return to the reference set Λ in our renewal construction, an unstable disk must grow to the size of Λ; and it is not hard to see that once it has reached a certain size, it will return within a fixed number of iterates. The speed of growth of unstable manifolds varies from system to system; this is where the geometry of the map enters. For example, if f is uniformly hyperbolic, then the diameter of every sufficiently small unstable disk grows by a definite factor with each iteration. For maps that are not uniformly hyperbolic, if the source of its nonhyperbolicity can be identified and its mechanism known, then the degree to which this growth is stunted can be understood in terms of the action of the "bad set" .

What we have proposed is a generic scheme for obtaining statistical information for dynamical systems with some hyperbolic behavior. This scheme has been implemented for a number of well known examples. We finish by illustrating how it works for billiards on $\mathbb{T}^2$ with convex scatterers (see the section on billiards, Fig. 3a in particular). For these billiards, discontinuities in the map are the only hindrance to uniform growth of unstable curves. This suggests that we examine closely the discontinuity set. Under an additional assumption on the billiard called "finite horizon", this set is known to be made up of a finite number of smooth curves some of which meet at certain points. An easy but crucial fact is that no more than Kn branches of the discontinuity set of f^n can meet in one point, K depending only the configuration of scatterers on our billiard table. This observation is due to Bunimovich. In n iterates, then, the image of a sufficiently short unstable curve has at most $Kn + 1$ components while its total length grows by a factor of λ^n for some $\lambda > 1$. On average, therefore, exponential growth prevails. The preceding discussion translates, after some work, into the estimate $m\{R > n\} < C\theta^n$, and we conclude from the theorem above that the central limit theorem holds and the speed of correlation decay is $\sim e^{-\alpha n}$. For more details, see [Y1]. The CLT result is first proved in [BSC].

References

[BC] Benedicks, M. and Carleson, L., The dynamics of the Henon map, Ann. Math. **133** (1991) 73-169.

[BY] Benedicks, M. and Young, L.-S., SBR measures for certain Henon maps, Invent. Math. **112** (1993) 541-576.

[B] Bowen, R., *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Springer Lecture Notes in Math. **470** (1975).

[BSC] Bunimovich, L. A., Sinai, Ya. G., and N. I. Chernov, Statistical properties of 2-dimensional hyperbolic billiards, Russ. Math. Surv., **46** (1991) 47-106.

[ER] Eckmann, J.-P. and Ruelle, D., Ergodic theory of chaos and strange attractors, Rev. Mod. Phys. **57** (1985) 617-656

[LY] Ledrappier, F. and Young, L.-S., The metric entropy of diffeomorphisms, Ann. Math. **122** (1985) 509-574.

[N] Newhouse, S., The abundance of wild hyperblic sets and nonsmooth stable sets for diffeomorphisms, Publ. I.H.E.S. **50** (1979), 101-151.

[P] Pesin, Ya. B., Characteristic Lyapunov exponents and smooth ergodic theory, Russ. Math. Surveys **32** (1977) 55-114.

[R1] Ruelle, D., *Thermodynamic formalism*, Addison-Wesley, New York, 1978.

[R2] Ruelle, D., An inequality of the entropy of differentiable maps, Bol. Sc. Bra. Mat. **9** (1978) 83-87.

[S1] Sinai, Ya. G., Gibbs measures in ergodic theory, Russ. Math. Surveys **27** No. 4 (1972) 21-69.

[S2] Sinai, Ya. G., Dynamical systems with elastic reflections: ergodic properties of dispersing billiards, Russ. Math. Surveys **25**, No. 2 (1970) 137-189.

[Sm] Smale, S., Differentiable dynamical systems, Bull. AMS **73** (1967) 747-817.

[Sz] Szász, D., Boltzmann's Ergodic Hypothesis, a conjecture for centuries?, Studia Sci. Math. Hung. **31** (1996) 299-322.

[W] Wojtkowski, M., Principles for the design of billiards with nonvanishing Lyapunov exponents, Commun. Math. Phys. **105** (1986) 391-414.

[Y1] Young, L.-S., Statistical properties of dynamical systems with some hyperbolicity, Ann. Math. (1998)

[Y2] Young, L.-S., Recurrence times and rates of mixing, to appear in Israel J.