Current Developments in Mathematics, 1998

Geometric and Ergodic Theory of Hyperbolic Dynamical Systems

LAI-SANG YOUNG*

The topic of this survey lies at the confluence of two branches of dynamical systems, the *geometric theory of ordinary differential equations* and *ergodic theory.* The first goes back to Poincaré, who pioneered the use of geometric methods in dynamical systems in his work on celestial mechanics; the latter goes back to Boltzmann, whose ideas are part of the foundation of modern statistical mechanics.

Hyperobolicity in dynamical systems is a geometric condition describing the exponential divergence of nearby orbits. It leads to irregular, chaotic and unpredictable patterns of behavior. Along with quasi-periodic dynamics or KAM theory, which lies at the opposite end of the ordered–disordered spectrum of dynamical behaviors, hyperbolic theory is one of the better understood areas of dynamical systems today.

This article is about the geometric and ergodic theory of hyperbolic systems. My aim here is not to give a complete list of all the important results, but to focus on a few areas of activity and to describe in as coherent a fashion as I can some of the progress over the last 20-30 years. The topics I have chosen are

I. Billiards and related physical systems
II. Analysis of a class of strange attractors
III. Entropy, Lyapunov exponents and dimension
IV. Correlation decay and related statistical properties

*Courant Institute of Mathematical Sciences, New York University, and Department of Mathematics, University of California, Los Angeles. The author is partially supported by the NSF

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

This article is intended for the broader mathematics community as well as researchers in dynamical systems. Some background material is included at the beginning for readers not in dynamics.

1. BACKGROUND AND DEFINITIONS

1.1. Continuous Time versus Discrete Time.

All the dynamical systems considered in this article take place on finite dimenional manifolds or Euclidean spaces. A **continuous time** system is one defined by an ordinary differential equation; a **discrete time** system is generated by the iteration of a map which is often but not always assumed to be invertible. Many (though not all) of the results discussed here have both continuous and discrete time versions. For simplicity we will often discuss only the discrete time case.

1.2. First Definitions from Ergodic Theory.

The usual setting for abstract ergodic theory consists of the following objects: Let $(X, \mathcal{B}, \mu)$ be a probability space, i.e. $(X, \mathcal{B})$ is a measure space and μ is a measure on $(X, \mathcal{B})$ with $\mu(X) = 1$.

DEFINITION 1.2.1. *$T : (X, \mathcal{B}, \mu) \to (X, \mathcal{B}, \mu)$ is called a* **measure-preserving transformation** (abbrev. **mpt**) *if for every $A \in \mathcal{B}$, $T^{-1}(A) \in \mathcal{B}$ and $\mu(T^{-1}(A)) = \mu(A)$.*

DEFINITION 1.2.2. *A mpt (T, μ) is called* **ergodic** *if for every $A \in \mathcal{B}$, $T^{-1}A = A$ implies that $\mu A = 0$ or 1.*

The most often used theorem in ergodic theory is probably the Birkhoff Ergodic Theorem.

THEOREM 1.2.3 (**Birkhoff Ergodic Theorem** 1932). *Let (T, μ) be a mpt, and let $\varphi \in L^1(\mu)$. Then $\exists\, \varphi^* \in L^1(\mu)$ s.t.*

$$\frac{1}{n}\sum_{0}^{n-1} \varphi \circ T^i \to \varphi^* \quad a.e.$$

Moreover, φ^* satisfies $\varphi^* \circ T = \varphi^*$ a.e. and $\int \varphi^* d\mu = \int \varphi d\mu$. It follows that if (T, μ) is ergodic, then $\varphi^* = \int \varphi d\mu$ a.e.

An often used application of the Birkhoff Ergodic Theorem is the following. Let (T, μ) be ergodic, and let $A \in \mathcal{B}$. Then for μ-a.e. x,

$$\frac{1}{n}\#\{0 \le k < n : T^k x \in A\} \;\to\; \mu(A) \quad \text{as } n \to \infty.$$

DEFINITION 1.2.4. *(T, μ) is called* **mixing** *if $\forall A, B \in \mathcal{B}, \mu(T^{-n}A \cap B) \to \mu(A)\mu(B)$ as $n \to \infty$.*

We remark that mixing is a stronger condition than ergodicity.

1.3. Invariant Measures for Continuous Maps.

Let X be a compact metric space, and let $T : X \to X$ be a continuous transformation.

PROPOSITION 1.3.1. *The set of T-invariant Borel probability measures, denoted $\mathcal{M}_T(X)$, is nonempty.*

PROOF. By the Riesz Representation Theorem we know that there is a one-to-one correspondence between $\mathcal{M}(X)$, the set of all Borel probability measures on X, and $C(X)^*$, where $C(X)$ is the Banach space of continuous real-valued functions on X. Thus $\mathcal{M}(X)$ is a nonempty, compact, convex, metrizable space. Now let $x \in X$, and let δ_x denote the Dirac measure at x, i.e. $\delta_x(E) = 1$ iff $x \in E$ for every Borel set E. Let $\mu_n := \frac{1}{n}\sum_{i=0}^{n-1} \delta_{T^i x}$. Then any accumulation point of μ_n is an element of $\mathcal{M}_T(X)$. ■

Except in the conservative case, differential equations and maps generally do not come equipped with natural invariant measures. Proposition 1.3.1 tells us that invariant measures always exist. The problem is that $\mathcal{M}_T(X)$ is generally too large, and the ergodic properties of (T, μ) depend sensitively on the choice of μ. A question of great importance, then, is:

Given a differential equation or a map, which invariant measures should we consider?

1.4. Physically Relevant Invariant Measures.

In this article, we will adopt the viewpoint that the Lebesgue measure class is of special interest, and that sets of positive Lebesgue measure correspond to the only events that can be observed physically.

Thus for example, in a Hamiltonian system, Liouville measure is our measure of interest, even though most Hamiltonian systems admit many other invariant measures (e.g. every periodic orbit supports one). We will refer to a dynamical system that comes equipped with an invariant measure equivalent to the Riemannian volume as "**a conservative dynamical system**".

For "**dissipative systems**", i.e. systems that are not conservative, the situation is trickier. Consider, for example, the situation of a map $f : U \to U$ where U is an open domain and $f(\bar{U}) \subset U$. Then $\Omega := \cap_{n \geq 0} f^i \bar{U}$ is a compact invariant set that attracts all points in U, meaning that for all $x \in U$, $f^n(x) \to \Omega$ as $n \to \infty$. We call Ω an **attractor** and U its *basin of attraction.* Now for any invariant probability measure μ, since $\mu(f^n U - f^{n+1} U) = 0$ for all n, it follows that μ must be supported on Ω. Furthermore, if f is volume contracting, so that Ω has Lebesgue measure zero, then all the invariant Borel probability measures of f are necessarily singular with respect to Lebesgue measure.

Question: For dissipative systems such as attractors, what does it mean for an invariant measure to be physically relevant, and do such measures always exist?

For a special class of attractors called **Axiom A** or **uniformly hyperbolic attractors**, Sinai [S2] and later Ruelle [R1] and Bowen [BR] discovered certain invariant measures that have the following special property: they govern the behavior of orbits starting from positive Lebesgue measure sets even though they themselves may be singular with respect to Lebesgue measure. The ideas of Sinai, Ruelle and Bowen have since been extended by others to more general dynamical systems, but we will continue to refer to these special invariant measures as **SRB measures**.

We close this section by mentioning the following elementary but very important idea, namely that *expansion is conducive to the existence of invariant measures absolutely continuous with respect to Lebesgue* (abbrev. *acim*). This idea is the backbone of a number of results on invariant densities and SRB measures, some of which are quite sophiscated. The basic principle is contained in the following theorem, a proof of which is given in the Appendix.

DEFINITION 1.4.1. *$f : M \to M$ is* **expanding** *or* **uniformly expanding** *if $\exists \lambda > 1$ s.t. $\forall x \in M$ and $\forall v \in T_x M, ||Df_x v|| \geq \lambda ||v||$.*

THEOREM 1.4.2 [KrS]. *Let $f : M \to M$ be a C^2 uniformly expanding map of a compact Riemannian manifold. Then f admits an acim.*

1.5. Hyperbolic Fixed Points.

Let $f : M \to M$ be a diffeomorphism with a fixed point p.

DEFINITION 1.5.1. *We call p a* **hyperbolic fixed point** *if there is a splitting of the tangent space $T_p M$ at p into two Df-invariant spaces $E^u \bigoplus E^s$ such that all the eigenvalues of $Df_p | E^u$ have modulus > 1 and all the eigenvalues of $Df_p | E^s$ have modulus < 1.*

It is easy to see that via a linear change of coordinates, one may assume that $Df_p|E^u$ is uniformly expanding and $Df_p|E^s$ is uniformly contracting.

Hyperbolic fixed points have **stable** and **unstable manifolds**. We denote them by $W^s(p)$ and $W^u(p)$ respectively. Stable and unstable manifolds are f-invariant immersed submanifolds tangent to E^s and E^u respectively. They are characterized by

$$W^s(p) = \{x \in M : d(f^n x) \to p \text{ as } n \to \infty\};$$

$$W^u(p) = \{x \in M : d(f^{-n} x) \to p \text{ as } n \to \infty\}.$$

The **local stable manifold** of size δ at p, denoted $W^s_\delta(p)$, refers to the disk of radius δ centered at p contained in $W^s(p)$. **Local unstable manifolds** are defined similarly.

1.6. Hyperbolic Invariant Sets.

The concept of hyperbolicity in a global sense was first used by Hedlund and Hopf in their analysis of geodesic flows on manifolds with negative curvature. This notion was axiomatized in the 1960s by Smale [Sm], who initiated a program on the geometric theory of a class of dynamical systems called *Axiom A*.

Let f be a C^2 diffeomorphism of a Riemannian manifold M, and let $\Lambda \subset M$ be a compact f-invariant set.

DEFINITION 1.6.1. *We say that f is* **uniformly hyperbolic** *on Λ if there is a continuous splitting of the tangent bundle over Λ into a direct sum of two Df-invariant subbundles, written*

$$T\Lambda = E^u \oplus E^s,$$

so that for all $x \in \Lambda$ and $n > 0$, the following hold:

$$v \in E^u(x) \Rightarrow |Df_x^{-n} v| \le C\lambda^n |v|$$

and

$$v \in E^s(x) \Rightarrow |Df_x^n v| \le C\lambda^n |v|$$

where $\lambda < 1$ and $C > 0$ are constants independent of x.

When both E^u and E^s are nontrivial, which is the case of primary interest here, the dynamics on Λ can be quite complicated. A prototypical example of a nontrivial hyperbolic invariant set is Smale's horseshoe (see Fig. 1).

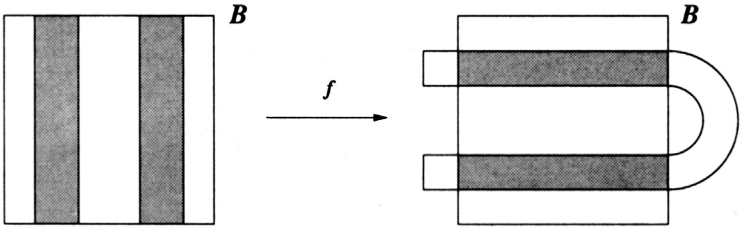


Fig. 1 The horseshoe map: B is a square; f stretches B in the horizontal direction, compresses it in the vertical direction and bends the resulting rectangle into the shape of a horseshoe; the two shaded vertical strips are mapped onto the shaded horizontal strips, and the hyperbolic invariant set $\Lambda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

An important characteristic of hyperbolicity when both E^u and E^s are nontrivial is *dynamic instability*, meaning that the orbits of most pairs of nearby points diverge exponentially fast in both forward and backward times.

A class of dynamical systems introduced in the 1960s that we will encounter later on in this article are **Anosov diffeomorphisms**, which are maps that are hyperbolic on the entire manifold M. **Axiom A** diffeomorphisms are, roughly speaking, maps that are hyperbolic on the nontransient parts of their phase spaces. They are more general than Anosov diffeomorphisms. (We will not need to know the precise definition of Axiom A here.)

PROPOSITION 1.6.2 **Existence of stable and unstable manifolds.** *Let f be uniformly hyperbolic on Λ. Then stable and unstable manifolds as defined in section 1.5 exist at every $x \in \Lambda$.*

The picture is, in fact, identical to that of a hyperbolic fixed point, except that the point x is no longer stationary and the entire local picture is moving with it.

PROPOSITION 1.6.3 *Let Λ be a uniformly hyperbolic attractor, and let U be its basin. Then*

(a) $\Lambda = \cup_{x \in \Lambda} W^u(x)$;

(b) *U is foliated by the stable manifolds of Λ.*

For Anosov diffeomorphisms, we may regard the entire manifold as $M = \Lambda = U$, so that all results about hyperbolic attractors apply.

1.7. Lyapunov Exponents.

In the 1970s, an *almost-everywhere* version of hyperbolicity emerged. The linear part of this theory owes its existence to the following theorem of Oseledec [O].

THEOREM 1.7.1 (**Oseledec's Multiplicative Ergodic Theorem** [O]). *Let $f : M \to M$ be a diffeomorphism of a manifold M, and let μ be an f-invariant Borel probability measure. Then at μ-a.e. x, there exist numbers $\lambda_1(x) < \cdots < \lambda_{r(x)}(x)$ and a decomposition of T_xM into*

$$T_xM = E_1(x) \oplus \cdots \oplus E_{r(x)}(x)$$

s.t.

(1) $\forall v \neq 0 \in E_i(x)$,

$$\lim_{n\to\infty} \frac{1}{n} \log |Df^n(x)v| \quad = \quad - \lim_{n\to\infty} \frac{1}{n} \log |Df^{-n}(x)v| \quad = \lambda_i(x);$$

(2) *for* $j \neq k$, $\lim_{n\to\infty} \frac{1}{n} \log |\sin \sphericalangle(A^{\pm n}E_j(x), A^{\pm n}E_k(x))| = 0$.

The functions $x \mapsto r(x), \lambda_i(x)$ and $E_i(x)$ are measurable.

The numbers $\lambda_i(x)$ are called the **Lyapunov exponents** of f at x; the *multiplicity* of $\lambda_i(x)$ is $dimE_i(x)$. The functions $x \mapsto \lambda_i(x)$ and $dimE_i(x)$ are clearly constant along orbits, so that if (f, μ) is ergodic, then the local properties of f are summed up in the finite set of numbers $\{\lambda_1, \cdots, \lambda_r\}$ counted with multiplicity.

The situation where (f, μ) has no zero Lyapunov exponents almost everywhere is, in some sense, a relaxation of the uniform hyperbolicity condition, with $E^u = \oplus\{E_i : \lambda_i > 0\}$ and $E^s = \oplus\{E_i : \lambda_i < 0\}$. We say that (f, μ) is **nonuniformly hyperbolic**.

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 constructed (sometimes very large) changes of coordinates and used them to prove, among other things, the existence of stable and unstable manifolds at μ-a.e. x. The relation between the uniform and nonuniform settings can be summarized informally as follows:

THEOREM 1.7.2 [P1]. *Let (f, μ) be nonuniformly hyperbolic, i.e. $\lambda_i \neq 0$ μ-a.e. Then there exist closed sets*

$$\Lambda_1 \subset \Lambda_2 \subset \Lambda_3 \subset \cdots \quad \textit{with } \mu(\cup\Lambda_i) = 1$$

such that

(a) orbits starting from each Λ_i are uniformly hyperbolic;
(b) the strength of hyperbolicity decreases as i tends to infinity.

In the same way that measurable functions are approximated by continuous ones defined on sets of measure $1 - \varepsilon$ (Lusin's theorem), Pesin's theorem gives approximations of nonuniformly hyperbolic sets by uniformly hyperbolic ones. The closed sets Λ_i are usually not f-invariant. In the absence of zero Lyapunov exponents, it is also possible to approximate (f, μ) by uniformly hyperbolic invariant sets. These sets in general have zero μ-measure [Ka].

1.8. A Hyperbolic Theory of the Future.

With the aid of computer graphics, dynamicists have become 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 early examples are the Lorenz ("butterfly") attractors and Hénon mappings. The nonuniform theory discussed in section 1.7 provides a more relaxed framework for study, namely that of *almost everywhere, asymptotic hyperbolicity.* This framework, however, requires that one begins with a chosen invariant probability measure. The question of *which invariant measure* aside (see section 1.3), it is also technically very difficult in general to prove that a dynamical system has nonzero Lyapunov exponents.

Thus with all the progress that has been made – and I hope this article will give you a glimpse of it – the challenge for a good hyperbolic theory probably lies ahead. A good hyperbolic theory should give information on **dynamical systems that have a great deal of expansions and contractions on large parts of their phase spaces**. We already know that this information alone is not conclusive, but a good theory should point the way to further relevant characteristics. It should also tell us – based on qualitative or verifiable quantitative properties and not on knowledge of infinitely fine details – what kind of a picture to expect.

APPENDIX: Expanding Maps: Proof of Theorem 1.4.1.

We begin with a few easy facts about f:

(1) $\exists \epsilon_0 > 0$ and $\lambda_0 > 1$ s.t. $d(x,y) < \epsilon_0 \Rightarrow d(fx,fy) \geq \lambda_0 d(x,y)$.

(2) If $\deg(f) = k$, then every $x \in M$ has exactly k inverse images.

(3) If ϵ_1 is sufficiently small, then restricted to any ϵ_1-disk D in M, f^{-n} has exactly k^n well defined branches $\forall n > 0$. (For $n = 1$, this is an immediate consequence of (2). For $n > 1$, use (1) and $\epsilon_1 < \epsilon_0$.)

Let ν_0 be the Riemannian measure on M normalized. For $n = 1, 2, \ldots$, define $\nu_n := f_*^n \nu_0$, i.e. $\nu_n(E) = \nu_0(f^{-n}E)$ for every Borel set E, and let $\varphi_n = \frac{d\nu_n}{d\nu_0}$.

We claim that

$$(*) \qquad \exists\, \alpha, \beta > 0 \text{ s.t. } \alpha \leq \varphi_n \leq \beta \; \forall n\,.$$

To prove this claim, we need the following distortion estimate:

Lemma. *Let f be as above. Then $\exists\, C_0$ (independent of n) s.t. $\forall x, y \in X$, if $d(f^i x, f^i y) < \epsilon_0 \;\; \forall i \leq n$, then*

$$\frac{\det Df^n(x)}{\det Df^n(y)} \leq e^{C_0 d(f^n x, f^n y)}\,.$$

Proof of Lemma.

$$\begin{aligned}\log\frac{\det Df^n(x)}{\det Df^n(y)} &\le \sum_{i=0}^{n-1}|\log\det Df(f^ix)-\log\det Df(f^iy)|\\ &\le \sum_{i=1}^{n-1}C_1d(f^ix,f^iy) \quad \text{for some } C_1\\ &\le \sum_{i=1}^{n-1}C_1\lambda_0^{-(n-i)}d(f^nx,f^ny)\,.\end{aligned}$$

We remark that this estimate relies on the fact that f is C^2.

■

Continuing with our proof of the theorem, consider x, y in some ϵ_1-disk D. Then $\varphi_n = \sum_i \varphi_n^i$, where each φ_n^i is the contribution to the density of ν_n by pushing along the ith branch of $f^{-n}|D$. Assuming that $\epsilon_1 < \epsilon_0$, the lemma above tells us that

$$\frac{\varphi_n^i(x)}{\varphi_n^i(y)} \le e^{C_0 d(x,y)}\,.$$

Summing over i, we obtain

$$\frac{\varphi_n(x)}{\varphi_n(y)} \le e^{C_0 d(x,y)}\,.$$

This together with $\int \varphi_n d\nu_0 = 1$ proves (*).

Let μ be an accumulation point of $\{\frac{1}{n}\sum_{i=0}^{n-1}\nu_i\}_{n=1,2,\dots}$. Then clearly μ is invariant and has a density with the same upper and lower bounds as the φ_n's. This completes our proof.

■

2. BILLIARDS AND RELATED PHYSICAL SYSTEMS

The motivation for this part of dynamical systems is *Boltzmann's Ergodic Hypothesis*, one formulation of which can be stated as follows:

BOLTZMANN'S ERGODIC HYPOTHESIS. *For large systems of interacting particles in equilibrium, time averages are close to the ensemble average.*

In 1963, following earlier observations of Krylov, Sinai [S1] formulated a version of this hypothesis in terms of the *ergodicity of hard balls*, that is, the ergodicity of a system of finitely many balls moving freely in 3-space and interacting completely elastically when they collide. The ergodicity of hard balls remains an open problem today, but we have learned a great deal in the last thirty years about certain related dynamical systems that are simpler, about two-dimensional *billiards* in particular. In this section I will report mostly on results on billiards, returning briefly to hard balls in the last subsection. For a more comprehensive discussion of recent results on billiards and hard balls, see [S4] or [Sz].

2.1. Two-dimensional Billiards: an Introduction.

A billiard flow in 2-dimensions 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 $\mu(M) = 1$.

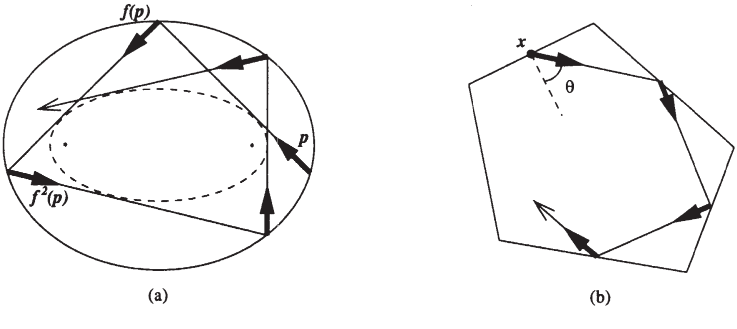


Fig. 2 Examples of nonhyperbolic billiards

As we will see, the dynamical behavior of a billiard flow or billiard map depends greatly on the geometry of the table Ω. Since this article is about hyperbolic systems, most of the billiards discussed here have hyperbolic behavior. Not all billiards are hyperbolic, however. We mention two examples:

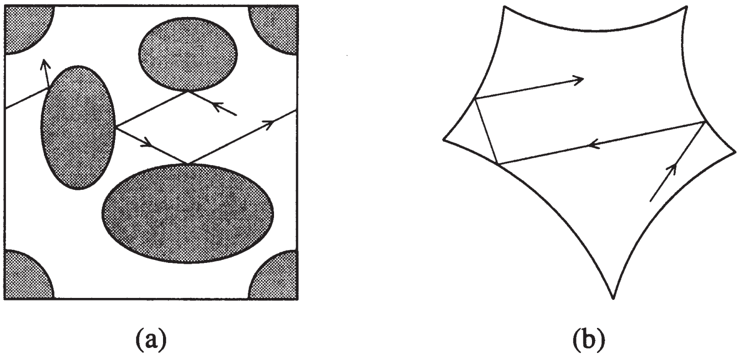
In the case where Ω is an ellipse, it is an exercise to see that the envelope of every (infinite) billiard trajectory is an ellipse or a hyperbola having the same foci as Ω (Fig.2(a)). One could thus picture M, which is diffeomorphic to $\mathbb{S}^1 \times [-\frac{\pi}{2}, \frac{\pi}{2}]$, as being foliated by simple closed curves that wrap around the $\mathbb{S}^1$-direction; these curves are left invariant by the action of f, which "rotates" the points 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.

2.2. Dispersing Billiards and the Lorentz Gas Model.

Sinai was the first to investigate rigorously billiards with hyperbolic properties. He studied in [S3] 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.

The best known example of a dispersing billiard is the **Lorentz gas**, which in 2-dimensions is a model for the free motion of a particle moving in $\mathbb{R}^2$ among a fixed configuration of convex objects called "scatterers". Assuming that the configuration of scatterers is periodic in space, one obtains a billiard flow on $\Omega = \mathbb{T}^2 - \cup_{i=1}^k \Omega_i$ where the Ω_i's are disjoint convex regions (see Fig. 3(a)).


(a)

(b)

Fig. 3 Dispersing billiards

Another example of a domain giving rise to a dispersing billiard is shown in Fig. 3(b). In this example, if a trajectory runs into a "corner" – which happens only for a Lebesgue measure zero set of initial conditions – we will simply not consider that trajectory further.

We explain now some elementary properties of maps f associated with dispersing billiards.

First, we claim that f is essentially uniformly hyperbolic. 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 on the ***projective*** level, at least, Df behaves like a hyperbolic linear map.

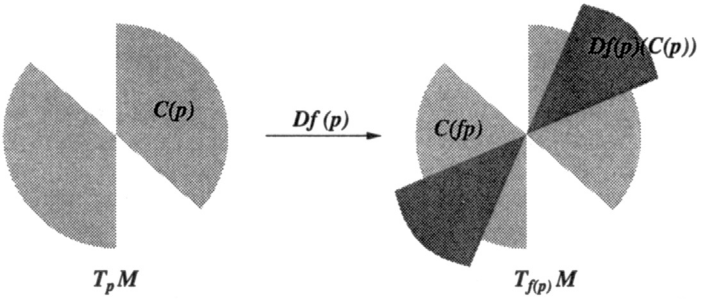


Fig. 4 $Df(p)$ maps $C(p)$, a cone in the tangent space at p, strictly into $C(f(p))$, a cone in the tangent space at $f(p)$. A standard way of proving hyperbolicity is to locate a family of invariant cones.

Next, observe that billiard maps such as those in Fig. 3 are ***discontinuous.*** For the Lorentz gas, for example, consider the trajectory of a 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 $\partial\Omega$. This is illustrated in Fig. 5(b). For the billiards in Fig. 3(b), corners are another source of discontinuity.

We were careful earlier on to claim only that the projectivized action of Df is uniformly hyperbolic. Indeed, in the coordinates introduced in section 2.1, Df does not necessarily expand the vectors in E^u or contract those in E^s (although this is true for Df^n for sufficiently large n). This is because $\det(Df) \neq 1$. In fact, Df becomes unbounded (or arbitrarily close to 0) as one approaches the set of discontinuities. This is illuatrated in Fig. 5(c).

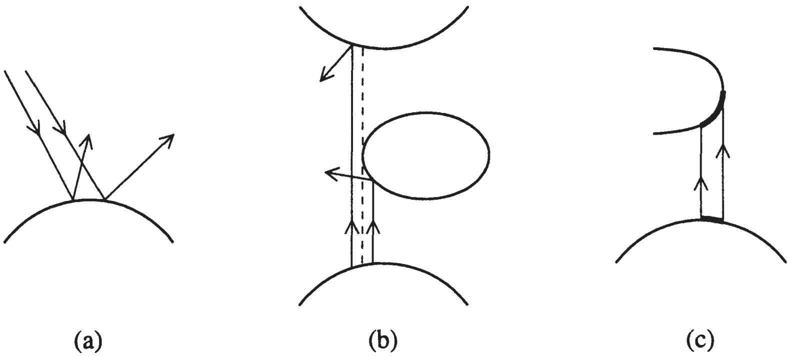


Fig. 5 Properties of dispersing billiards

2.3. Ergodicity of Dispersing Billiards.

An important early breakthrough in the study of billiards is the following theorem of Sinai, which lends support to the hard balls version of the Ergodic Hypothesis (see the beginning of section 2).

THEOREM 2.3.1 [S3]. *Dispersing billiards are ergodic.*

The proof of Sinai's theorem is far too involved to be given here, but I would like to take this opportunity to explain some of the basic issues one has to deal with in proving ergodicity for hyperbolic systems, and also to indicate the difficulties caused by the discontinuities in billiard maps.

When studying questions of ergodicity, particularly for hyperbolic systems, one often distinguishes between two different kinds of issues: *local* and *global.* Local ergodicity is about whether or not ergodic components are localy open modulo sets of Lebesgue measure 0; this is equivalent to asking if the time averages of observables are, as $n \to \infty$, locally constant. Global ergodicity, on the other hand, addresses the issue of transitivity of "larger" regions, such as whether there are "walls" separating the phase space into noninteracting domains.

For Anosov systems, that is, for systems that are uniformly hyperbolic on the entire manifold *without discontinuities*, there is a well established way of proving ergodicity which goes back to Hopf. Let me first give a sketch of the proof in this simpler setting, and then come back to explain why this proof needs to be amended in a nontrivial way for billiard maps on account of the discontinuity curves.

THEOREM 2.3.2 [A]. *Let $f : M \to M$ be a C^2 topologically transitive Anosov diffeomorphism that perserves a Borel probability measure μ equivalent to the Riemannian volume. Then (f, μ) is ergodic.*

Topological transitivity means that for all open sets U and V, there exists n such that $f^n U \cap V \neq \phi$.

SKETCH OF PROOF OF THEOREM 2.3.2. To prove ergodicity, it suffices to show that for every L^1 function φ on M, the trajectory averages $\frac{1}{n}\sum_{i=0}^{n-1}\varphi \circ f^i$ converge μ-a.e. to a constant function; in fact, it suffices to do this for continuous φ. Now f being Anosov, its stable and unstable manifolds form a pair of transversal foliations on M that are invariant under the action of f. For two points x and y with $y \in W^s(x)$, since $d(f^n x, f^n y) \to 0$ 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 is tempting to conclude immediately that the limit function, which is constant on both stable and unstable manifolds, must be locally constant. The validity of this argument actually relies on a rather subtle and very important property of the stable and unstable foliations, namely their *absolute continuity*. The precise definition is as follows:

DEFINITION 2.3.3. *A foliation $\mathcal{F}$ is called* **absolutely continuous** *if for every pair of transversals Σ_1 and Σ_2 connected by a leave L of $\mathcal{F}$, the holonomy map from a neighborhood of $L \cap \Sigma_1$ to Σ_2 carries sets of Lebesgue measure zero to sets of Lebesgue measure zero.*

The absolute continuity of the W^u-foliation tells us that the conditional measures of μ on unstable manifolds are equivalent to the Riemannian volumes induced on these manifolds; and the absolute continuity of the W^s-foliation tells us that if A is a full Lebesgue measure subset of a local unstable leaf $W^u_\delta(x_0)$, then $\cup_{x\in A} W^s_\delta(x)$ occupies, up to a set of Lebesgue measure zero, an open neighborhood of x_0 in M. These technical ideas are needed to make the argument two paragraphs back rigorous.

If a foliation is smooth (meaning its holonomy maps are smooth), then it is, of course, absolutely continuous. Except for maps of algebraic origins, however, stable and unstable foliations of Anosov systems are almost never smooth. They have been shown to be absolutely continuous if f is C^2 [A]. This circle of ideas completes the proof of "local ergodicity". Global ergodicity follows from local ergodicity and the transivity assumption. ■

Returning to the billiard map, we now explain how discontinuities complicate the picture. Intuitively, the closer a point is to the discontinuity set, the shorter is its stable curve, for the discontinuity set breaks up any curve meeting it into two or more components, which may then evolve independently. Indeed, points whose orbits come arbitrarily close to the discontinuity set in forward (respectively backward) time have arbitrarily short stable (respectively unstable) curves. We do not, therefore, have a uniform local product structure, which is relied upon heavily in Hopf's proof of local ergodicity. A great deal of work has to be done to overcome this.

We remark that current understanding of dispersing billiards in 2-dimensions has gone beyond ergodicity to include statistical properties such as decay of correlations, central limit theorem, convergence to Wiener measure etc. (see e.g. [BSC2], [Y3]). Some of these properties will be discussed in section 5 of this article.

2.4. The Stadium and other billiards with Convex Boundaries.

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 [Bu] (see Fig. 6), can also produce hyperbolicity if certain conditions are met. Intuitively, 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.

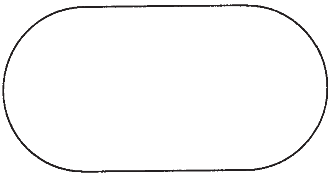


Fig. 6 The stadium

Let us prove rigorously that the billiard map associated with the stadium has nonzero Lyapunov exponents. First we claim that in the (x, θ)-coordinates introduced in section 2.1, the cone corresponding to $x\theta \geq 0$ is invariant under Df. We leave this as a simple exercise in plane geometry.

Note, however, that unlike the case of the Lorentz gas or other dispersing billiards, Df here does not map the cone at certain points $p = (x, \theta)$ *strictly* into the interior of that at $f(p)$. For example, consider billiard trajectories that are nearly perpendicular to the two straight sides. They bounce back and forth for a long time without diverging, and Df has the form $\begin{pmatrix} 1 & c \\ 0 & 1 \end{pmatrix}$. Billiard trajectories that are nearly (though not quite) tangential to the circular parts of the boundary also experience long stretches of *parabolic* behavior.

Clearly, composing matrices of the form $\begin{pmatrix} 1 & c \\ 0 & 1 \end{pmatrix}$ does not lead to hyperbolic behavior; to guarantee hyperbolicity, it is important that Df maps the cones strictly into their interiors. This is where ergodic theory comes to our rescue: having this strict cone invariance happen with a *positive frequency* is good enough, and the Ergodic Theorem says that this is automatic once we know that for μ-a.e. $p \in M$, there exists $n = n(p) \geq 0$ such that one has strict cone invariance at $f^n(p)$. Since the set of initial conditions for which the trajectory bounces back and forth between the two straight sides or skims along the circular pieces forever has measure zero, this last condition is satisfied and the existence of a positive Lyapunov exponent μ-a.e. is proved.

We remark that if $\lambda_1 \geq \lambda_2$ are the Lyapunov exponents of (f, μ), then $\lambda_1 > 0$ implies $\lambda_2 < 0$ since $\lambda_1 + \lambda_2 = \int \log |\det(Df)| d\mu = 0$. We mention also that the invariant cones idea (a simple version of which we have used) is a powerful tool and works in all dimensions; see [W1].

More general geometric conditions on Ω that give rise to nonzero Lyapunov exponents of the billiard maps are formulated in [W2].

2.5. Hard Balls.

First we remark on the relation between hard balls and billiards.

Consider the elastic collision of N balls of unit mass and radii $r > 0$ moving freely in $\mathbb{T}^d$, the d-dimensional torus. Let (q_i, v_i) denote the position and velocity of the ith particle. Then the configuration space of the N particles is

$$Q := \{(q_1, \cdots, q_N) \in \mathbb{T}^{Nd} : |q_i - q_j| \geq 2r \ \forall i \neq j\}.$$

Fixing the energy, total momentum and center of mass, it is not hard to see that this dynamical system is equivalent to that of the uniform motion of a point mass moving in

$$Q \cap \{\sum q_i = 0\} \ \times \ \mathbb{S}^{Nd-d-1}$$

where $\mathbb{S}^k$ denotes the unit sphere in $\mathbb{R}^k$. Thus systems of hard balls can be viewed as special cases of billiards in higher dimensions, where the domains Ω are of a particular shape. In this billiard system, sets corresponding to $\{|q_i - q_j| < 2r\}$ for fixed i, j now play the role of scatterers. These sets are cylinders, which are not strictly convex. The resulting billiard, therefore, is not dispersing or uniformly hyperbolic.

Until recently, most of the results on hard balls have been for small numbers of balls. We are happy to report that Simányi and Szász have now proved that with no restriction on the number of balls, systems of finitely many balls in a torus with typical mass distributions have nonvanishing Lyapunov exponents [SS].

3. ANALYSIS OF A CLASS OF STRANGE ATTRACTORS

Not only are strange attractors familiar objects to dynamicists, they have captured the imagination of scientists in other disciplines. Most of the studies on strange attractors have been via numerical simulations. Few rigorous mathematical analyses have been carried out, in part because they tend to be difficult. The purpose of this section is to report on some recent advances in the understanding of strange attractors that are strongly dissipative with only one direction of instability. The dynamics of these attractors are closely connected to those of 1-dimensional maps, which is where our report begins.

3.1. The quadratic family $x \mapsto 1 - ax^2$.

The last two decades saw an explosion of activity in 1-dimensional dynamics, real and complex. I will limit myself here to the real case, and to issues that are relevant to our discussion of attractors.

Consider $f_a : [-1,1] \to [-1,1]$ defined by $f_a(x) = 1 - ax^2$, where $a \in [0,2]$ is a parameter. We begin with two easy cases.

For $f = f_a$ with a near 0, it is easy to see that $x = -1$ is an attractive fixed point, and every $x \in [-1,1]$ has the property that $f^n(x) \to -1$ as $n \to \infty$.

For $a = 2$, however, we claim that f admits an ergodic invariant probability measure equivalent to Lebesgue, so that for Lebesgue-a.e. x, the orbit of x spends a positive fraction of time on every interval of positive length. One way to see this is to observe that via the change of coordinates $x = h(\theta) = \sin \frac{\pi}{2}\theta$, f is conjugate to the map $g : [-2,2] \to [-2,2]$ with $g(\theta) = 2 + 2\theta$ for $\theta \in [-2,0]$ and $g(\theta) = 2 - 2\theta$ for $\theta \in [0,2]$. It is easy to see that g preserves Lebesgue measure and is ergodic.

The dichotomy suggested by these two sets of behaviors was the subject of much research in the last 20 years.

The term **absolutely continuous invariant probability measures**, meaning invariant measures that have densities with respect to Lebesgue, will be abbreviated as **acim** as is often done in the literature. Recall from section 1.4 that expansions are conducive to the existence of acim's. Away from the critical point 0, the map $f = f_a$ is expanding (assuming a is not too small). When an orbit comes near the critical point, say to a distance δ of 0, it experiences a contraction of order δ. Orbits that come near 0 may or may not recover from this derivative loss. If they do, then the map is essentially (though not uniformly) expanding, and is a candidate for having an acim. If, on the other hand, contraction prevails, then one expects the dynamics to be dominated by an attractive periodic orbit. The question is: **for $f = f_a$, is expansion or contraction more likely to prevail?**

The answer to this very innocent question turns out to be less than simple. The first major theorem that gives insight into this is due to Jakobson, proved around 1980:

THEOREM 3.1.1 [J]. *There exists a positive Lebesgue measure set of parameters a with the property that f_a has an acim with a positive Lyapunov exponent.*

The latest result on this topic is due to Lyubich; part (i) of the theorem below was first announced by Graczyk and Swiatek [GS].

THEOREM 3.1.2 [Ly]. *For the family $f_a(x) = 1 - ax^2$, $a \in [0, 2]$, there are sets A, B in parameter space with $A \cup B = [0, 2]$ up to a set of Lebesgue measure 0 such that*

(i) A is open and dense in $[0, 2]$, and for all $a \in A$, f_a has an attractive periodic orbit which attracts Lebesgue-a.e. $x \in [-1, 1]$.

(ii) B has positive Lebesgue measure, and for every $a \in B$, f_a has an acim.

Thus on a positive measure set of parameters, expansion wins, and on an open and dense set, contraction wins. This intermingling of parameters with diametrically opposite behaviors underscores the complexity of the picture.

Unlike the first theorem, Theorem 3.1.2 uses tools from the complex quadratic family and applies only to the family $f_a(x) = 1 - ax^2$, $a \in [0, 2]$. This theorem is the culmination of ideas from many sources, including previous work on absolutely continuous invariant measures on the interval, remormalization ideas from Feigenbaum, independently Coulette and Tresser, and later Sullivan and McMullen, background material on the Mandelbrot set from various sources, complex analytic techniques such as quadratic-like maps from Douady-Hubbard, puzzle ideas from Yoccoz, etc.

In the remainder of this subsection, we explain a mechanism for "cultivating" expansion for the maps f_a. A natural idea is to take control of the source of nonexpansion, *i.e.* the critical point. Misiurewicz studied in [M] arbitrary multimodal maps whose critical orits are forbidden to get close to the critical set and proved the existence of acims. A more relaxed condition is to require only that the critical orbit has a positive Lyapunov exponent (see [CE]). For purposes of comparing with 2-dimensions, we record below the following result for the family $\{f_a\}$.

THEOREM 3.1.3 [BC1]. *There is a positive measure set of parameters a for which the following hold: for some $\alpha > 0$ and $\lambda > 1$,*

(i) $|f_a^n(0)| \geq e^{-\alpha n}$ for all $n \geq 1$;

(ii) $|(f_a^n)'(f0)| \geq \lambda^n$ for all $n \geq 1$.

Under these conditions, it is easy to see that when an orbit comes near 0, it will follow the critical orbit for some time and copy the derivative of the critical orbit, which by condition (ii) is growing exponentially. This is made precise in the following easy lemma:

LEMMA 3.1.4 [BC1] *For $x \approx e^{-\mu}$, there exists $p \sim \mu$ such that*

- *$f^i x$ stays near $f^i 0$ for all $i \leq p$;*
- *$|(f^p)'x| > \lambda^{\frac{p}{2}}$.*

SKETCH OF PROOF. Since $|f[0,x]| \sim x^2$, the fact that $x^2\lambda^p \sim 1$ implies that $p \sim \frac{2}{-\log\lambda} \cdot \log\frac{1}{x} \sim \mu$; also, $(f^p)'x \sim x\lambda^p \sim \frac{1}{x} \sim \lambda^{\frac{p}{2}}$. ■

3.2. Hénon Maps: Elementary Facts and Basic Questions.

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

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

These equations were first investigated numerically in 1977 by the astronomer Hénon, who observed that in certain parameter regions there are attractors with very complicated dynamics. Many numerical experiments were performed following this intriguing discovery of Hénon, but analytically these maps were to remain intractable for another decade.

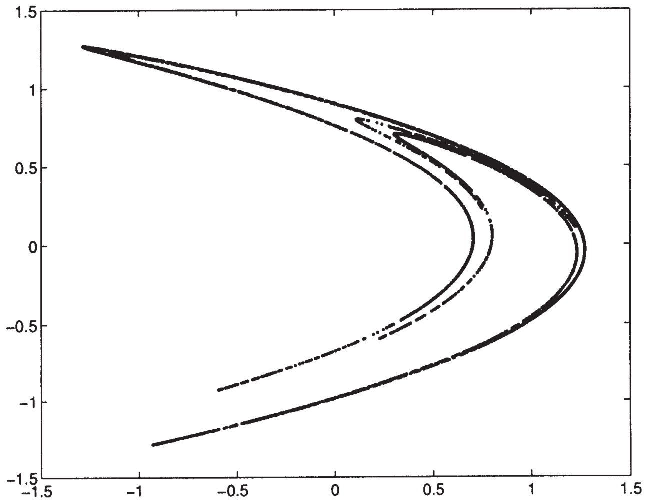


Fig. 7 This is the computer plot of a single orbit of length 5000 for the map $(x,y) \mapsto (1-1.4x^2+0.3y,\; x)$, the original map studied by Hénon. The overall appearance of the picture does not seem to depend on the choice of initial condition provided it is chosen from certain regions of the plane.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

The equations of the map in Fig. 7 differ from those at the beginning of this subsection by a coordinate change. We will continue to use the ones at the beginning of this subsection, and will concentrate on the parameter region $a<2$ and close to 2, and b small.

We begin with some elementary facts. Let $T = T_{a,b}$ with (a, b) fixed. Geometrically, it follows easily from the equations of T that it 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 passes near the y-axis, its directions of expansion and contraction may get confused, and previously established hyperbolicity may be spoiled.

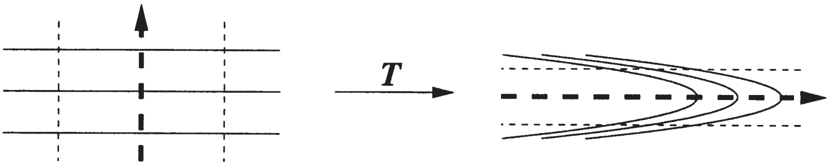


Fig. 8 The geometry of Hénon maps: vertical lines are mapped to horizontal lines; horizontal lines are mapped to parabolas

It is not hard to see that for an open set of parameters in the region of interest, 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 attraction* of Ω. In general, basins are open and relatively large sets, whereas attractors of area decreasing maps are of Lebesgue measure 0. It is not hard to prove that Ω is *not* a uniformly hyperbolic or Axiom A attractor.

Turning to the dynamics of T on Ω, as in 1-dimension 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 $|\det(DT)|$, which is equal to $|b|$, 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 generic conditions not only do periodic sinks exist there are infinitely many of them [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 widely 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 we know that there are no invariant cones, and that the directions of expansion and contraction may get somewhat confused when an orbit passes through the region $\{|x| \leq \sqrt{b}\}$. It is not unreasonable, however, to think that while cancellations can and do occur to some degree, 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. It is known only that periodic sinks are present for an open set of parameters. Numerics as well as intuitive thinking favor the "chaotic" scenario as being fairly "typical", but this thinking was not substantiated until quite recently.

3.3. The Hénon Attractors: First Results.

Recall that for the 1-dimensional maps discussed in section 3.1, one way to gain control of the dynamics is to force the critical orbit to have a positive Lyapunov exponent. Benedicks and Carleson proved that the analogous picture can be arranged for a positive measure set of Hénon maps:

THEOREM 3.3.1 [BC2]. *For all sufficiently small $b > 0$, there exists a positive measure set of a for which the following hold for $T = T_{a,b}$: there is a set $\mathcal{C}$ called the* **critical set** *with the property that for all $z \in \mathcal{C}$, the following hold for all $n \geq 1$:*
(i) $d(T^n z, \mathcal{C}) > e^{-\alpha n}$;
(ii) $||DT_z^n|| > \lambda^n$ for some $\lambda > 1$ independent of z.

Theorem 3.3.1 should be compared to Theorem 3.1.3. The authors of [BC2] showed, in fact, that whenever $d(T^n z, \mathcal{C}) = \delta$ (where $d(\cdot, \cdot)$ here has a special meaning), the orbit suffers a loss of hyperbolicity of order $\sim \delta$, and that this loss is subsequently recovered in a manner similar to that in Lemma 3.1.4.

We remark that the identification of the critical set $\mathcal{C}$ here is not an entirely straightforward matter. In 1-dimension, the critical point is the point of infinite contraction, where all previously accumulated expansion is totally destroyed. T being a diffeomorphism, there obviously is no direct analog for this. In its place, we have the following mechanism for destroying hyperbolicity: the *interchanging of stable and unstable directions.* That is to say, we think of a point z as "non-hyperbolic" if there exists a tangent vector v at z such that for some large n_1, n_2, both $DT^{-n_1}v$ and $DT^{n_2}v$ are strongly contracted. Letting n_1 and n_2 go to infinity, we see that *tangencies of stable and unstable manifolds* are precisely what destroy hyperbolic behavior.

Locating and controlling the tangencies of stable and unstable manifolds in order to establish hyperbolicity is, however, a "chicken-and-egg" proposition: To locate the tangencies of stable and unstable manifolds, it is necessary to first have these manifolds; and to have these manifolds, one needs to first prove hyperbolicity! The solution offered by Benedicks and Carleson is an *inductive* construction. They start with a picture in which hyperbolicity for a finite number of iterates is guaranteed.

Using this hyperbolcity they construct *temporary* stable curves, that is, curves that function as stable curves for a finite period of time, and use them to construct temporary tengencies. Pretending these were the true critical points, they iterate forward, establishing hyperbolicity for a longer period of time. Using this newly created hyperbolicity, they update the location of their tentative critical set, and the bootstrapping process continues.

We mention one other very substantial difference between 1- and 2-dimensions. In 1-dimension, if the orbit from x to $f^{n_1}x$ is expanding, and the orbit from $f^{n_1}x$ to $f^{n_2}x$ is expanding, then the orbit from x to $f^{n_1+n_2}x$ is expanding. This is not the case in dimension 2: *concatenations of hyperbolic orbits do not necessarily give hyperbolic orbits*! This happens when the directions that has been expanded in the first n_1 iterates get contracted in the next n_2 iterates. Careful control of *angles* between stable and unstable directions is required; this is a new element not present in 1-dimension.

In homoclinic bifurcations of surface diffeomorphisms, it is known that for certain parameter ranges one has, for a suitable power of the map, a small attractor whose defining map is a small perturbation of the Hénon map [PT]. Mora and Viana showed that Theorem 3.3.1 can be carried over to this setting [MV].

A couple of years following the work of Benedicks and Carleson, SRB measures were constructed for the Hénon attractors. Recall from section 1.4 that SRB measures are physically relevant invariant measures for attractors, in the sense that they govern the behavior of a positive Lebesgue measure set of initial conditions. SRB measures were first discovered in the context of Anosov diffeomorphisms and Axiom A attractors by Sinai [S2], Ruelle [R1] and Bowen [BR] (see also [Bo]). 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 are constructed. This is done by Benedicks and the author.

THEOREM 3.3.2 [BY1]. *For a positive Lebesgue measure set of parameters (a, b), the Hénon map $T = T_{a,b}$ admits an invariant probability measure μ with supp$(\mu) = \Omega$ such that*

(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 notation δ_z refers to point mass at z. The measure μ in the theorem is the **SRB measure.**

Theorem 3.3.2 can be interpreted as follows. 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.) The resulting plot 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. Theorem 3.3.2 says that as n tends to infinity, this "picture" tends to that of the SRB measure μ for most choices of initial conditions. This explains why μ is physically observable. It also explains why the overall appearance of the computer plot does not seem to depend on the choice of initial conditions (see the caption under Figure 7).

Having seen it in a concrete example, let us now formally define SRB measures in a general context.

DEFINITION 3.3.3. *Let f be a diffeomorphism and μ an f-invariant Borel probability measure with some positive Lyapunov exponents μ-a.e. We call μ an* **SRB measure** *if the conditional measures of μ on unstable manifolds are compatible with the volume elements induced on these submanifolds.*

In the absence of zero Lyapunov exponents, it follows from the absolute continuity of the stable foliation (Definition 2.3.3) that if μ is an SRB measure, then a positive Lebesgue measure set of points in the phase space is connected to the typical points of μ via stable manifolds. The following is a general fact from nonuniform hyperbolic theory:

THEOREM 3.3.4 [PS]. *Let f be a C^2 diffeomorphism of a manifold M preserving an ergodic SRB measure μ with no zero Lyapunov exponents. Then there exists a set $A \subset M$ with positive Lebesgue measure such that for every $x \in A$, $\frac{1}{n}\sum_{i=0}^{n-1}\delta_{f^ix}$ converges weakly to μ.*

To prove Theorem 3.3.2, then, it is sufficient to construct an invariant measure that has absolutely continuous conditional measures on unstable manifolds. This is somewhat analogous to the construction of acim's for 1-dimensional maps (but is more complicated). Once this measure is constructed and proved to be ergodic, the assertion in part (b) follows from Theorem 3.3.4.

3.4. Analysis of a New Class of Attractors.

[BC2] and to some extent the papers that built on it brought new techniques for analyzing concrete nonuniformly hyperbolic systems. There were, however, many unanswered questions. For example, the formulas of the Hénon maps are used explicitly in the estimates in [BC2], which considers only parameters near $a = 2$ and $b = 0$. An immediate question is: *Under what conditions do techniques of this type work?* In a different direction, Theorem 3.3.1 gives information on derivative growth starting from C, an important but very small subset of the attractor constructed in a somewhat *ad hoc* manner. A natural question is: *What can one say about the rest of the attractor, its structure, its geometry and the dynamics on it?*

These and other questions are addressed in a recently completed manuscript by Qiudong Wang and the author [WY]. Behind all the formulas and computations, we believe that the two features of the Hénon family that are truly essential for the type of analysis in the last subsection are (1) $\dim W^u = 1$ and (2) strong contraction. These two properties together imply that locally, the map has some one-dimensional behavior. For simplicity, we consider attractors derived from a single 1-dimensional map: Start with $f : N \to N$ where N is a circle or an interval. Embed N into $M = N \times D_n$ where D_n is an n-dimensional disk, and consider a perturbation of f into a diffeomorphism T that maps M into itself. The attractor of interest is defined by $\Omega = \cap_{i \geq 0} T^i M$.

This construction generalizes that of the *solenoid*, an object familiar to topologists. For the standard solenoid, $f : \mathbb{S}^1 \to \mathbb{S}^1$ is taken to be $f(z) = z^2$ and T maps $\mathbb{S}^1 \times D_2$ diffeomorphically onto a subset of itself winding around two times in the $\mathbb{S}^1$-direction. In [WY] we allow f to be an arbitrary 1-dimensional map, including those with an arbitrary number of critical points. The Hénon maps fit naturally into this context, with $f_a : [-1, 1] \to [-1, 1]$ given $f_a(x) = 1 - ax^2$, as do homoclinic bifurcations. Other examples that have been studied are dissipative twist maps, the most standard of which can be realized as a suitable perturbtation of $f(x) = x + \frac{K}{2\pi}\sin(2\pi x)$, $x \in \mathbb{R}/\mathbb{Z}$.

The technical assumptions in [WY] are as follows: We treat only the case where the phase space is 2-dimensional, and we assume that the 1-dimensional map f satisfies the "Misiurewicz conditon" (see section 3.1). Through f, we pass an arbitrary 2-parameter family of maps $T_{a,b}$, using the first parameter to control movements along the circle or interval and the second to control the unfolding of f into 2-dimensional maps. Two transversality conditions are assumed; both are simple and checkable.

Through a parameter selection process more systematic than that in [BC2], we identify a class of maps for which we are able to obtain a great deal of information. Our set of "good" parameters has positive Lebesgue measure; the b-parameter is small, i.e. we assume strong dissipation. The following is a sample of our results for maps T corresponding to good parameters:

(1) *Hyperbolic behavior.* We prove that T has an intrinsically defined source of nonhyperbolicity, which we also call $\mathcal{C}$. Invariant sets Λ bounded away from $\mathcal{C}$ are uniformly hyperbolic, with the strength of hyperbolicity diminishing as $\text{dist}(\Lambda, \mathcal{C}) \to 0$. These attractors provide concrete illustrations for the abstract nonuniform hyperbolic theory discussed in section 1.7. The set $\mathcal{C}$ is a Cantor set with a very specific structure. Knowledge of this structure is the key to understanding the rest of the attractor.

(2) *Global geometry, symbolic dynamics and topological entropy.* We give a combinatorial description of the approximate shape of the attractor, explaining how its geometry differs from that of 1-dimensional maps. In spite of its fractal nature, $\mathcal{C}$ divides the attractor into well defined regions, allowing us to code orbits on the attractor into symbol sequences in a meaningful way. This opens the door to a number of other methods for studying their dynamics.

(3) *Statistical properties.* Not only do we construct SRB measures for these attractors, we bound the number of ergodic components and prove that for Lebesgue-almost every z in the basin, the "picture" $\frac{1}{n}\sum_{i=0}^{n-1}\delta_{T^i z}$ tends to one of the ergodic

SRB measures. Statistical properties of the type discussed in section 5 are also known.

CONCLUDING REMARKS. To my knowledge, this is the first comprehensive analysis of a reasonably general class of nonuniformly hyperbolic attractors. The attractors in this class are strongly dissipative with only one direction of instability. Their dynamics bear some resemblance to those of one-dimensional maps, but new complexities abound, giving rise to new phenomena and rich geometric structures not present in one-dimension.

4. ENTROPY, LYAPUNOV EXPONENTS AND DIMENSION

4.1. Motivating Examples.

To motivate our results in this section, we consider first a simple-minded way of building fractals from a single template.

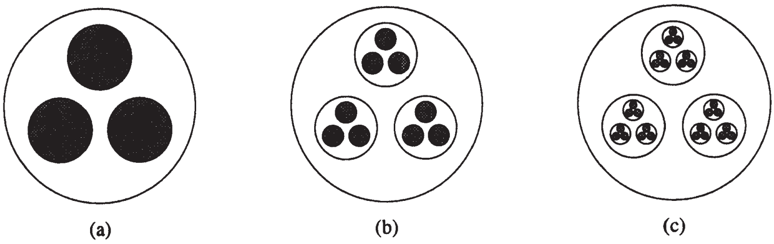


Fig. 9 Construction of fractal from a single template

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 of these balls, 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.

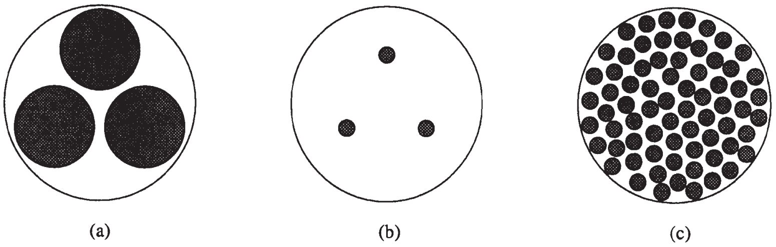


Fig. 10 Three different templates: observe how the dimension of the fractal changes with the number and diameter of the smaller balls

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, which leads naturally to the use of ergodic theory.

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.

4.2. Basic Definitions.

In section 1.7 we introduced the idea of **Lyapunov exponents**, which are the average speeds with which nearby orbits separate. This is a *local, geometric* measure of dynamical complexity.

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 $\alpha^n(x) := \{y \in M : f^i x$ and $f^i y$ belong in the same element of α 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 an invariant measure, they can only be related via a notion of dimension that also reflects properties of this 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.

DEFINITION 4.2.1. *We say the* **dimension of the measure** ν, *written* $\dim(\nu)$, *is well defined and is equal to* α *if for* ν*-a.e.* x,

$$\lim_{\varepsilon \to 0} \frac{\log mB(x, \varepsilon)}{\log \varepsilon} = \alpha.$$

DEFINITION 4.2.1. *The* **Hausdorff dimension** *of* ν *is defined to be*

$$HD(\nu) = \operatorname*{Inf}_{\substack{Y \subset X \\ \nu Y = 1}} HD(Y)$$

where $HD(Y)$ *denotes the Hausdorff dimension of the set* Y.

The notion $\dim(\nu)$ is not always well defined. It is easy to construct examples of measures for which the limits as $\varepsilon \to 0$ do not exist. On the other hand, if ν is an ergodic invariant measure for a locally bi-Lipschitz map, then once these limits exist, they are constant a.e. It is also easy to see that if $\dim(\nu)$ is well defined, then it is equal to $HD(\nu)$. See e.g. [Y1] for relations between $\dim(\nu)$ and other notions of dimension.

4.3. Relation between Entropy and Lyapunov Exponents.

The following are the two most basic results in this direction:

THEOREM 4.3.1 **(Pesin Formula)** [P2]. *Let* f *be a* C^2 *diffeomorphism of a manifold preserving a Borel probability measure* μ. *If* μ *is equivalent to the Riemannian measure on* M, *then*

$$h_\mu(f) = \int \sum \lambda_i^+ m_i d\mu.$$

Here m_i denotes the multiplicities of the Lyapunov exponents λ_i, i.e. $m_i = \dim E_i$ where E_i is the subspace corresponding to λ_i.

THEOREM 4.3.2 **(Ruelle's Inequality)** [R3]. *For C^1 mappings (that are not necessarily invertible) and all invariant Borel probability measures μ, we have*

$$h_\mu(f) \le \int \sum \lambda_i^+ m_i d\mu .$$

Here is one way to interpret these results: clearly, both $h_\mu(f)$ and $\sum \lambda_i^+ m_i$ are measures of dynamical complexity. In a conservative system, all the expansion goes back into the system to make entropy, so these two invariants coincide and Pesin's formula holds. A strict inequality, on the other hand, corresponds to the situation where some of the expansion is "wasted", and that can happen only if there is some "leakage" from the system. Thus the gap in Ruelle's Inequality measures, in some sense, the amount of dissipation in a system.

Since only positive Lyapunov exponents are involved in the relations above, one may suspect that the gap in Ruelle's Inequality measures only dissipativeness *in the unstable direction.* This is indeed the case, as expressed in the following theorem. Recall that an SRB measure is one that has smooth conditional measures on unstable manifolds (see Definition 3.3.3).

THEOREM 4.3.3 [LS], [Le], [LY1]. *For C^2 diffeomorphisms the entropy formula in Theorem 4.3.1 holds if and only if μ is an SRB measure.*

Before the theorems above were proved in their present generality, they had been known for some time in the context of Anosov diffeomorphisms and Axiom A attractors. (See e.g. [S2], [R1] and [Bo]).

A sketch of the proof of Ruelle's Inequality, which is the most elementary of the three results, is given in the Appendix at the end of this section.

4.4. Dimension Enters.

Before giving the full statement of our result, 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

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

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 4.4.1 **(Dimension formula)**. *Assume for simplicity that* (f, μ) *is ergodic. Then corresponding to each* λ_i, *there is a number* δ_i *with* $0 \leq \delta_i \leq \dim E_i$ *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* $\dim(\mu|W^u) + \dim(\mu|W^s)$.

Parts (a), (b), and the "$\leq$" part of (c) of this theorem are proved by Ledrappier and the author [LY1]. The reverse inequality in (c) is proved in a recent preprint by Barreira, Pesin, and Schmeling [BPS].

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.*

Observe that part (a) of this theorem is entirely consistent with the results in the last subsection. First, with $0 \leq \delta_i \leq \dim E_i$, the dimension formula in (a) implies Ruelle's Inequality. When μ is equivalent to the Riemannian measure or is SRB, then all the δ_i's take on their maximum values and Pesin's Formula holds. In this case one also has that $\dim(\mu|W^u)$ is equal to the topological dimension of the unstable manifold.

In the remainder of this subsection we will give an outline of the proof of Theorem 4.4.1(a), and remark on (c).

OUTLINE OF PROOF OF THEOREM 4.4.1 (a). We arrange our (distinct) Lyapunov exponents in decreasing order

$$\lambda_1 > \lambda_2 > \cdots > \lambda_u.$$

For each $i \leq u$, let W^i be the unstable manifolds corresponding to $E_1 \oplus \cdots \oplus E_i$. These manifolds are known to exist a.e. and have the property that $W^{i-1}(x) \subset W^i(x)$.

The strategy here is to work with one exponent at a time and to work our way up the entire hierarchy $W^1 \subset W^2 \subset \cdots W^u$. For each i, we introduce a notion of entropy along W^i, written h_i, measuring the randomness of f along the leaves of W^i and ignoring what happens in the transverse directions. We also prove that the dimensions of the conditional measures are well defined. For brevity write $\hat{\delta}_i = \dim(\mu|W^i)$.

The proof consists of the following steps:

(i) $h_1 = \hat{\delta}_1 \lambda_1$;
(ii) $h_i - h_{i-1} = (\hat{\delta}_i - \hat{\delta}_{i-1})\lambda_i$ for $i = 2, \ldots, u$;
(iii) $h_u = h_\mu(f)$.

The proof of (i) is the same as before since it only involves one exponent, and $\delta_1 = \hat{\delta}_1$.

To give an idea of why (ii) is true, consider the action of f on the leaves of W^i, and pretend somehow that a quotient dynamical system can be defined by collapsing the leaves of W^{i-1} inside W^i. This "quotient" dynamical system has exactly one Lyapunov exponent, namely λ_i. It behaves as though it leaves invariant a measure with dimension $\hat{\delta}_i - \hat{\delta}_{i-1}$ and has entropy $h_i - h_{i-1}$. A fair amount of technical work is needed to turn this into rigorous mathematics, but once properly done, it is again the single-exponent principle at work. Letting $\delta_i = \hat{\delta}_i - \hat{\delta}_{i-1}$ for $i = 2, \ldots, u$, and summing the equations in (ii) over i, we obtain $h_u = \sum\limits_{i=1}^{u} \delta_i \lambda_i$.

Step (iii) says that zero and negative exponents do not contribute to entropy. The influence of negative exponents is easily ruled out, and an argument similar to that in step (ii) tells us that entropy does not increase as we go from the unstable foliation to the "center unstable foliation". This completes the outline to the proof in [LY1].

■

REMARK ON THEOREM 4.4.1(c). Since we know that $\dim(\mu|W^u)$ and $\dim(\mu|W^s)$ are well defined, Theorem 4.4.1(c) asserts only that they add properly. This would have been a straightforward consequence of dimension-theoretic considerations if stable and unstable foliations were Lipschitz; they are, unfortunately, not more than Hölder in general. Another special case that is easy to handle is when μ is the direct product of two measures, one on W^u and the other one on W^s. The main idea in [BPS] is that, in some sense, this product situation is not as special as one might think, meaning that all ergodic measures of diffeomorphisms have, up to slow exponential errors, a kind of local product structure.

4.5. Randomly Perturbed Dynamical Systems.

As an example of the simplified dynamical picture created by the averaging effects of noise, we present the entropy and dimension formulas for random dynamical systems.

Consider as a model of a randomly perturbed dynamical system compositions $\cdots f_2 \circ f_1 \circ f_0$ where the f_i's are an *iid* sequence with repect to a probability measure on the space of C^2 diffeomorphisms of a manifold. (This setup is compatible with that of stochastic differential equations; see e.g. [Ku].) Let μ be an invariant measure for this process, and let $\{\mu_\omega\}$ denote the distintegration of μ on bi-infinite sample paths $\omega = \{f_i\}_{i=-\infty}^{\infty}$.

Dynamical invariants such as Lyapunov exponents, entropy and dimension continue to make sense in this setting; moreover, they are nonrandom. We continue to let $\lambda_1 > \lambda_2 > \cdots > \lambda_r$ denote the distinct Lyapunov exponents. Let σ_i be defined by $\delta_i = \sigma_i \dim E_i$, so that $0 \le \sigma_i \le 1$. Note that these numbers can be defined for all i including those for which $\lambda_i < 0$ (by considering f^{-1} instead of f in the last subsection).

In order to obtain the results below, we need to assume certain conditions that are a little too technical to state here. There are generic conditions, and roughly speaking, they guarantee that the images of points and vectors are sufficiently random that their distributions have densities. For a precise statement see the papers cited.

THEOREM 4.5.1 [LY2], [LY3]. *Assume that the f_i's are sufficiently random as above. Then:*

(a) *if $\lambda_1 > 0$, then a.s. the μ_ω's have the SRB property;*

(b) *if $\lambda_i \neq 0 \ \forall i$, then there is an i^* s.t. $\sigma_i = 1$ for $i < i^*$ and $\sigma_i = 0$ for $i > i^*$.*

Part (b) says, for example, that mass has a tendency to align itself with the more expanding directions when a system is stochastically purturbed.

4.6. APPENDIX: Sketch of Proof of Ruelle's Inequality.

For simplicity we assume that (f, μ) is ergodic. For $\epsilon > 0$, let α_ϵ be a partition of M into approximate ϵ-boxes, and let δ_1, δ_2, and δ_3 be prescribed small numbers.

First we choose N s.t. $\forall x$ in a good set G with $\mu G > 1 - \delta_1, Df^N$ looks like what the Lyapunov exponents say it should.

Next we choose $\epsilon > 0$ small enough that

- in the ϵ-neighborhood of every $x \in G, Df^N$ is a good approximation of f^N; we assume in fact that if $\alpha_\epsilon(x) \cap G \neq \phi$, then $Df^N \alpha_\epsilon(x)$ is contained in an $\epsilon e^{(\lambda_1+\delta_2)N} \times \cdots \times \epsilon e^{(\lambda_r+\delta_2)N}$– box ($\lambda_i$ counted with multiplicity), and
- $h(f^N) \le h(f^N; \alpha_\epsilon) + \delta_3$.

Now

$$h(f) = \frac{1}{N} h(f^N) \quad \textit{and} \quad h(f^N; \alpha_\epsilon) \le H(f^{-N}\alpha_\epsilon | \alpha_\epsilon).$$

We estimate this latter quantity by

$$H(f^{-N}\alpha_\epsilon|\alpha_\epsilon) \leq \sum_{A\in\alpha_\epsilon} \mu A \cdot \log r_{N,\epsilon}(A)$$

where $r_{N,\epsilon}(A)$ is the number of elements of $f^{-N}\alpha_\epsilon$ that meet A, or, equivalently, the number of elements of α_ϵ that intersect $f^N A$. If $A \cap G \neq \phi$, then we have control on the size and shape of $f^N A$, obtaining $r_{N,\epsilon}(A) \lesssim e^{N\Sigma(\lambda_i^+ + \delta_2)m_i}$. If $A \cap G = \phi$, then $r_{N,\epsilon}(A) \leq e^{C_0 N}$ where C_0 is a constant depending only on $||Df||$. We have thus proved

$$\frac{1}{N} H(f^{-N}\alpha_\epsilon|\alpha_\epsilon) \leq (1-\delta_1)\sum(\lambda_i^+ + \delta_2)m_i + \delta_1 C_0\,,$$

which gives the desired result. ■

5. CORRELATION DECAY AND RELATED STATISTICAL PROPERTIES

5.1. Dynamically Generated Processes.

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 almost surely,

$$\frac{1}{n}\sum_0^{n-1} \varphi \circ f^i \to \int \varphi d\mu,$$

holds when (f, μ) is ergodic; this is simply the Birkhoff Ergodic Theorem.

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 **decay of correlations** 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.

There are many other questions along similar lines that one can ask. For example, what about the *law of iterated logarithm*, the *almost sure invariance principle*? Also, do return times to small regions have the *Poisson property*? In principle, at least, every limit law for stochastic processes has a version for observations from dynamical systems.

Of primary interest to us, then, is the following

Question: given (f, μ), do these limit laws hold?

Intuitively, one would guess that the more chaotic a dynamical system is, the more *random* observations of the type above are likely to be, and hence the more likely it is for these laws to be valid.

One must also remember, however, that a dynamical system generated by the iteration of a map (or by a differential equation) is a *deterministic* process: once an initial condition is chosen, all subsequent behavior is totally determined and nothing is left to chance. This would suggest that observations from dynamical systems can never be *completely random*, whatever that means.

5.2. Setting and Previously Known Results.

Since current techniques do not permit us to deal with the question above in complete generality, we will limit ourselves in this section to **maps that geometrically have a great deal of expansion and contraction on large parts of their**

phase spaces. This, of course, is not a rigorous mathematical definition. I prefer to leave the setting vague than to impose artificial boundaries (see section 1.8), and to note only that the class I have in mind includes Axiom A diffeomorphisms as well as all the examples discussed earlier on in this article.

Having settled on a class of maps, we need to decide next on the underlying probability μ. We take the view here that only properties that hold on positive Lebesgue measure sets are observable; that is to say, we are interested primarily in **physically relevant invariant measures**. Accordingly, if a system is "conservative", then the invariant measure of interest will be the one equivalent to Lebesgue measure. If a system is "dissipative", then we will take μ to be an *SRB measure* if one exists. (See sections 1.4 and 3.5 for definitions.) 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.

Assuming that f is either conservative or it admits an SRB measure, called μ in both cases, we preceed next to the question of correlation decay. Observe that $\Phi(n)$ as defined in section 5.1 tending to zero as $n \to \infty$ for all measurable or square-integrable φ is essentially equivalent to the *mixing* property of (f, μ) (see section 1.2 for the definition). For this reason, I will sometimes refer to the speed of correlation decay as **the speed of mixing**.

Certainly not all (f, μ) are ergodic or mixing. There is a theorem due to Pesin and Ledrappier saying, however, 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.[P2]). Thus the question of correlation decay or speed of mixing is always relevant on each *mixing component.*

It is not hard to see mixing can be arbitrarily slow if we allow all measurable or L^2 test functions, and that questions regarding *speeds* of mixing make sense only if we impose some regularity on φ. (This has to do with the remark in the last paragraph of section 5.1. Intuitively, smooth functions are a bit like locally constant ones, and using locally constant test functions is a bit like coarse-graining, which introduces randomness into the system.) From here on φ will always be assumed to be at least Hölder continuous.

This completes the description of the setting in which the questions in the last section will be asked.

In the remainder of this article I would like to report on some recent work that attempts to study systematically these questions, but first I mention some previously known results:

For Anosov diffeomorphisms and attractors of Axiom A maps, SRB measures always exist, correlation decay is exponential, and the central limit theorem always holds (see e.g. [R2]). Correlation decay questions for Axiom A *flows* remain not well understood; for recent progress see [C1], [D]. (Our discussion does not apply to flows.)

Outside of the Axiom A category, much of the progress up until recently has been focused on individual examples or classes of examples, proving for the most part exponential decay. Known techniques for proving exponential decay include spectral gaps of the Perron-Frobenius or transfer operator (see e.g. [R2], [HK], [R4]) and the invariant cones method first used in [FS] and later exploited in [Li] for hyperbolic systems. Some techniques using approximation by Markov chains have also been attempted (e.g. [BSC2]). To my knowledge systematic methods for studying slower decay rates have – up until quite recently – not been developed.

Our reportoire of examples at this point is rather limited, but these examples do suggest that outside of the Axiom A category many distinct behavior types are possible. For instance, there are examples on the boundary of Axiom A that do not admit SRB measures ([HY], [H1]); others do but have polynomial decay ([H2], [Y4]).

5.3. A Generic Scheme: Renewal Times, Growth of Unstable Manifolds, and the Speed of Mixing.

My goals in the research I am reporting on are

(1) to give verifiable conditions for the statistical properties above,

(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. The results contained in this subsection are published in [Y3] and [Y4].

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 $m\{R > n\}$.

For difffeomorphisms with hyperbolic properties, that is, for maps that have contracting as well as expanding directions, the picture is more complicated. To avoid messy estimates I would choose Λ with a product structure (*i.e.* Λ is the intersection of transversal families of W^u and W^s-disks) even though these sets are not open in general. Here m is Lebesgue measure on W^u, and we require that $m(\Lambda \cap W^u) > 0$.

There are also a few technical requirements, the most important of which is a regularity condition for $Df^{R_i}|\Lambda_i$ which puts a uniform bound on the "distortion" or nonlinearity of $f^{R_i}|\Lambda_i$. This is a natural condition for C^2 maps that are sufficiently expanding. This **control of nonlinearities** is essential for ensuring some resemblance to independence for the dynamics between successive returns to the reference set. See [Y3] and [Y4] for the precise formulations.

We now explain how this construction is used to study the questions posed at the beginning of this section. Again we omit details, sketching only the three basic ideas.

First, we relate the statistical properties of f to the asymptotics of $m\{R > n\}$. We call these "abstract" results because they do not depend on the characteristics of the individual dynamical system other than the tail of the return time function R. Let $\Phi(n)$ be as defined in section 5.1.

THEOREM 5.3.1. [Y3], [Y4]. *Let f, Λ, m and R be as above. Then:*

(a) *If $\int R dm < \infty$, then f admits an SRB measure μ.*

(b) *If, additionally,* $\gcd\{R_i\} = 1$, *then (f, μ) is mixing.*

(c) *If $m\{R > n\} < C\theta^n$ for some $\theta < 1$, then $\exists\ \tilde\theta < 1$ s.t. $\forall \varphi$, $\Phi(n) < C\tilde\theta^n$.*

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

(e) *If R is as in* (d) *and $\alpha > 2$, then the CLT holds for all φ.*

Next, we argue that conceptually $m\{R > n\}$ is essentially the speed with which arbitrarily small pieces of unstable manifolds grow to a specified size. (This is *not* the same as Lyapunov exponents, which measure pointwise growth rates.) First we describe the picture:

If f has good hyperbolic properties, then we can cover most of phase space with a finite number of sets $\Gamma_1, \cdots, \Gamma_k$ with product structures (they look like $W^u \times W^s$ trelises). If f is mixing, then in finite time, $f^n\Gamma_i$ crosses over Γ_j in the unstable direction for every i, j. These structures give the dynamics the flavor of a finite Markov chain, but one should not carry the analogy too far, for $\cup\Gamma_i$ is not all of phase space, nor is it an invariant set. The rest of phase space is made up of small bits of stable and unstable manifolds that twist and turn in a rather messy way.

Returning to the problem of estimating $m\{R > n\}$, suppose that Γ_1 is our reference set. Since f is ergodic, it is inevitable that some parts of Γ_1 will get into the messy regions of phase space before they return. It is necessary, therefore, to know how long it takes structures of *arbitrarily small scales* to "straighten out" and grow to the scale of the Γ_i's. This is also sufficient, for once a W^u-leaf reaches a size comparable to the Γ_i's, it will soon cross over one of them, and once it crosses over one Γ_j, it will cross over Γ_1 in a finite number of steps via the Markov-like action on $\cup\Gamma_i$ described earlier on.

Finally, we observe that while in general it is impossible to know the detailed structures of a map to arbitrarily small scales, the type of estimates to which the problem has now been reduced is feasible if we know the *rules* of the game. For example, if there is a recognizable "bad set" – meaning a source of nonhyperbolicity – with known mechanisms, then the messy parts are created by interactions with the "bad set", which also determines how they evolve. The speed in question is therefore related to the speed with which the influence of the "bad set" is overcome.

5.4. Applications.

In section 5.3 we proposed a generic scheme for obtaining statistical information for dynamical systems with some hyperbolic behavior. We now list some examples for which this scheme has been successfully implemented. Most of the results discussed below are not previously known.

EXAMPLE 1 *Expanding maps in 1-d with a neutral fixed point* [Y4]. Here the "bad set" consists exactly of the neutral fixed point, which we call 0. If $f'(0) = 1$ and $f''(0) \approx |x|^{\gamma-1}$ for some $\gamma > 0$, then taking Λ to be a suitable interval, it is an easy exercise to see that $m\{R > n\} = \mathcal{O}(n^{-\alpha})$ where $\alpha = \frac{1}{\gamma}$. Once this is computed, the abstract theorem in 3.2 gives immediately the existence of an invariant probability density with correlation decay rate $\mathcal{O}(n^{-\frac{1}{\gamma}+1})$ for $\gamma < 1$, and the CLT for $\gamma < \frac{1}{2}$. See [H2] for similar results.

EXAMPLE 2 *Logistic, maps, Hénon maps, and the attractors discussed in section 3.4* ([Y2], [KN], [BY2], [WY]). For the "good" parameters, the time that it takes an orbit to regain its hyperbolicity after coming to a distance of δ from the "bad set" is $\sim \log \frac{1}{\delta}$ (for an indication of why this is true, see Lemma 3.1.4). Thus after each visit to the "bad set", it is as though there is unobstructed, uniform growth until the derivative has fully recovered. This translates into the estimate $m\{R > n\} < C\theta^n$ for some $\theta < 1$, from which we conclude exponential decay of correlations and CLT.

EXAMPLE 3 *Billiards.* First we consider the Lorenz gas or billiards on $\mathbb{T}^2$ with convex scatterers (see section 1) and assume in addition a *finite horizon* condition, which says that the times between collisions are uniformly bounded. (This requires that the scatterers be sufficiently dense.) Earlier results [BSC2] have shown that their correlation decay rates are bounded above by $\sim e^{-\sqrt{n}}$. To see if this is the true decay rate, I ran these much studied examples through the analysis in section 5.3. Here is what I found [Y3]:

As observed in section 1, the only obstruction to uniform growth along W^u-curves are a finite number of discontinuity curves transversal to W^u. To get a sense of the worse-case scenario, let γ be a short W^u-curve and imagine that each component of $f^n\gamma$ is expanded by $\frac{3}{2}$ and cut into 2 roughly equal pieces with each iteration – it would be very hard for these components to grow to unit length!

Unlike the situation in Example 2, where parameters are chosen to guarantee full and immediate recovery after each visit to the "bad set", the components of $f^n\gamma$ are not guaranteed to grow long before they get cut again. We rely instead on the geometry of billiards and a *statistical* argument, which goes as follows:

It is observed in [BSC1] that no more than Kn branches of the discontinuity set of f^n can meet in one point, K depending only on the billiard table. Thus in n iterates the image of a sufficiently short W^u-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. This translates, after some work, into the estimate $m\{R > n\} < C\theta^n$, from which we conclude that the speed of correlation decay is actually $\sim e^{-\alpha n}$.

Recently Chernov has used the scheme outlined in the last subsection to obtain further results on the correlation decay rates for other kinds of billiards. He has removed the finite horizon condition for the Lorenz gas model, and has also verified the exponential decay rate for dispersing billiards in domains with corners (see Fig. 3(b)) under certain mild conditions [C2]. We are very hopeful that this scheme of proof will prevail to give further results for similar physical systems.

References

[A] Anosov, D., Geodesic flows on closed Riemann manifolds with negative curvature, AMS Translation (1969) 1-235.

[BPS] Barreira, L., Pesin, Y. and Schmeling, J., Dimension of hyperbolic measures, to appear in Ann. Math.

[BC1] Benedicks, M. and Carleson, L., On iterations of $1 - ax^2$ on $(-1, 1)$, Ann. Math. **122** (1985) 1-25.

[BC2] Benedicks, M. and Carleson, L., The dynamics of the Henon map, Ann. Math. **133** (1991) 73-169.

[BY1] Benedicks, M. and Young, L.-S., SBR measures for certain Henon maps, Inventiones Math. **112** (1993) 541-576.

[BY2] Benedicks, M. and Young, L.-S., Markov extension and decay or correlations for certain Hénon maps, to appear in Asterisque (1999).

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

[BR] Bowen, R. and Ruelle, D., The ergodic theory of Axiom A flows, Invent. Math. **29** (1975) 181-202.

[Bu] Bunimovich, L. A., On the ergodic properties of nowhere dispersing billiards, Commun. Math. Phys. **65** (1979) 295-312.

[BSC1] Bunimovich, L. A., Sinai, Ya. G., and Chernov, N. I., Markov partitions for two-dimensional billiards, Russ. Math. Surv., **45** (1990) 105-152.

[BSC2] Bunimovich, L. A., Sinai, Ya. G., and N. I. Chernov, Statistical properties of 2-dimensional hyperbolic billiards, Russ. Math. Surv., **46** (1991) 47-106.

[C1] Chernov, N. I., Markov approximations and decay of correlations for Anosov flows, Ann. Math., **147** (1998), 269-324

[C2] Chernov, N. I., Decay of correlations and dispersing billiards, 1998 preprint

[CE] Collet, P. and Eckmann, J.-P., Positive Lyapunov exponents and absolute continuity, Ergod. Th. & Dynam. Sys., **3** (1983) 13-46.

[D] Dolgopiat, D., On the decay of correlations in Anosov flows, Ann. Math. **147** (1998)

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

[FS] Ferrero, P. and Schmitt, B., Ruelle-Perron-Frobenius theorems and projective metrics, Colloque Math. Soc. J. Bolyai Random Fields, Estergom (Hungary) (1979)

[GS] Graczyk, G. and Swiatek, G., Generic hyperbolicity in the logistic family, Ann. Math., **146** (1997) 1-52.

[HK] F. Hofbauer and G. Keller, Ergodic properties of invariant measures for piecewise monotonic transformations, Math. Z., **180** (1982), 119-140.

[H1] H. Hu, Conditions for the existence of SRB measures for "almost Anosov" diffeomorphisms, preprint

[H2] Hu, H., polynomial decay for 1-dimensional maps with neutral fixed points, preprint.

[HY] Hu, H. and Young., L.-S., Nonexistence of SBR measures for some systems that are "almost Ansov", Ergod. Th. & Dynam. Sys., **15** (1995) 67-76.

[J] Jakobson, M., Absolutely continuous invariant measures for one-parameter families of one-dimensional maps, Commun. Math. Phys., **81** (1981), 39-88.

[Ka] Katok, A., Lyapunov exponents, entropy and periodic orbits for diffeomorphisms, Publ. Math. IHES, **51** (1980), 137-174.

[KS] A. Katok and J. M. Strelcyn, *Invariant manifolds, entropy and billiards; smooth maps with singularities*, Springer Lecture Notes in Math. **1222** (1986).

[KN] G. Keller and T. Nowicki, Special theory, zeta functions and the distributions of periodic points for Collet-Eckmann maps, Commun. Math. Phys., **149** (1992), 31-69.

[KrS] Krzyzewski, K. and Szlenk, W., On invariant measures for expanding differentiable mappings, Studia Math. **33** (1969), 83-92

[Ku] Kunita, H., *Stochastic flows and stochastic differential equations*, Cambridge Univ. Press (1990).

[Le] Ledrappier, F., Proprietes ergodiques des mesures de Sinai, Publ. Math. IHES **59** (1984) 163-188.

[LS] Ledrappier, F. and Strelcyn, J.-M., A proof of the estimation from below in Pesin entropy formula, Ergod. Th. & Dynam. Sys., **2** (1982) 203-219.

[LY1] Ledrappier, F. and Young, L.-S., The metric entropy of diffeomorphisms, Ann. Math. **122** (1985) 509-574.

[LY2] Ledrappier, F. and Young, L.-S., Entropy formula for random transformations, Prob. Th. Rel. Fields **80** (1988) 217-240.

[LY3] Ledrappier, F. and Young, L.-S., Dimension formula for random transformations, Commun. Math. Phys. **117** (1988) 529-548.

[Li] Liverani, C., Decay of correlations, Ann. Math. **142** (1995) 239-301.

[Ly] Lyubich, M., Almost every real quadratic map is either regular or stochastic, Ann. Math (1999)

[M] Misiurewicz, M., Absolutely continuous invariant measures for certain maps of an interval, Publ. Math. IHES **53** (1981), 17-51

[MV] Mora, L. and Viana, M., Abundance of strange attractors, Acta Math. **171** (1993), 1-71

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

[O] Oseledec, V. I., A multiplicative ergodic theorem: Liapunov characteristic numbers of dynamical systems, Trans. Moscow Math. Soc. **19** (1968) 197-231.

[PT] Palis, J. and Takens, F., Hyperbolicity and sensitive chaotic dynamics at homoclinic bifurcations, Cambridge University Press (1993)

[P1] Pesin, Ya. B., Families if invariant manifolds corresponding to non-zero characteristic exponents, Math. of the USSR, Izvestjia **10** (1978) 1261-1305.

[P2] Pesin, Ya. B., Characteristic Lyapunov exponents and smooth ergoic theory, Russ. Math. Surveys **32** (1977) 55-114.

[PS] Pugh, C. and Shub, M., Ergodic attractors, Trans. AMS **312** (1989) 1-54.

[R1] Ruelle, D., A measure associated with *Axiom A* attractors, Amer. J. Math. **98** (1976) 619-654.

[R2] D. Ruelle, *Thermodynamic formalism*, Addison-Wesley, New York, 1978.

[R3] Ruelle, D., An inequality of the entropy of differentiable maps, Bol. Sc. Bra. Mat. **9** (1978) 83-87.

[R4] Ruelle, D., The thermodynamics formalism for expanding maps, Commun. Math. Phys., Vol. 125 (1989) 239-262.

[SS] Simanyi, N. and Szasz, D., Hard ball systems are completely hyperbolic, Ann. Math (1998)

[S1] Sinai, Ya. G., On the foundations of the ergodic hypothesis for a dynamical system of statistical mechanics, Soviet Math. Dokl. **4** (1963) 1818-1822

[S2] Sinai, Ya. G., Gibbs measures in ergodic theory, Russ. Math. Surveys **27** No. 4 (1972) 21-69.

[S3] Sinai, Ya. G., Dynamical systems with elastic reflections: ergodic properties of dispersing billiards, Russ. Math. Surveys **25**, No. 2 (1970) 137-189.

[S4] Sinai, Ya. G., Dynamicals systems II, Encyclopaedia of Math. Sc. Vol. **2**, Springer-Verlag (1989)

[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.

[WY] Wang, Q. and Young, L.-S., Analysis of a class of strange attractors, 1999 preprint, XXX Mathematics Archive.

[W1] Wojtkowski, M., Invariant faimilies of cones and Lyapunov exponents, Ergod. Th. & Dynam. Sys. (1985) **5**, 145-161.

[W2] Wojtkowski, M., Principles for the design of billiards with nonvanishing Lyapunov exponents, Commun. Math. Phys. **105** (1986) 391-414.

[Y1] Young, L.-S., Dimension, entropy and Lyapunov exponents, Ergod. Th. & Dynam. Sys. **2** (1982) 109-129

[Y2] Young, L.-S., Decay of correlations for certain quadratic maps, Commun. Math. Phys., **146** (1992), 123-138.

[Y3] Young, L.-S., Statistical properties of dynamical systems with some hyperbolicity, Ann. Math. (1998)

[Y4] Young, L.-S., Recurrence times and rates of mixing, to appear in Israel J.