

# Chapter 1

## Introduction to Hyperbolic Dynamics and Ergodic Theory

Boris Hasselblatt

### 1.1 Introduction

#### 1.1.1 Guided Tour

These are notes based on a minicourses given at the Centre International de Rencontres Mathématiques, Marseille in November 2013, at the University of Tokyo in June 2014, and the 2015 Houston Summer School on Dynamical Systems. They owe much to these experiences, and I would like to thank the organizers of these schools for their invitation and hospitality as well as the participants for their engagement and their attentive comments on these notes.

While there are many good introductions to hyperbolic dynamical systems,<sup>1</sup> two aspects of this one are of interest. On one hand, we implement an underappreciated approach due to Bowen, Anosov and Katok [Bo75, Bo78, Ka81] to obtain the basic topological dynamics of uniformly hyperbolic dynamical systems from shadowing and expansivity; this is the content of Sect. 1.3.3. On the other hand, we use the Hopf argument to obtain multiple mixing of hyperbolic dynamical systems—which means that we can do so without any reference to entropy theory or results that use it. Thus, pages 24–44 can be regarded as the principal novelty of these notes.

The lectures gave an introduction to some features of the topological and measurable dynamics of hyperbolic systems, mainly in discrete time, and this is an essentially self-contained account of these basics. The first half is devoted to

---

<sup>1</sup>Notably [Yo95], which inspired Sect. 1.6, where we prove the Stable/Unstable Manifold Theorem using the Perron–Irwin method. Altogether these notes owe much to [Co07, KaHa95, Yo95].

B. Hasselblatt (✉)

Department of Mathematics, Tufts University, Medford, MA 02155, USA

e-mail: [Boris.Hasselblatt@tufts.edu](mailto:Boris.Hasselblatt@tufts.edu)

hyperbolic dynamical systems, and the second half (Sect. 1.7) introduces ergodic theory. While a central point is that these subjects interact deeply, the halves are essentially independent, to the point of having some duplication (mainly between Sects. 1.5 and 1.7). A common feature is that entropy theory is absent, and a novelty is that multiple mixing properties are obtained without it.

The sections on hyperbolic dynamics are modular and can be read largely independently. That is, each part can largely be read on its own or omitted on its own. This is most evidently so for this section and the historical sketch in Sect. 1.2. The remaining “hyperbolic” sections are related as follows.

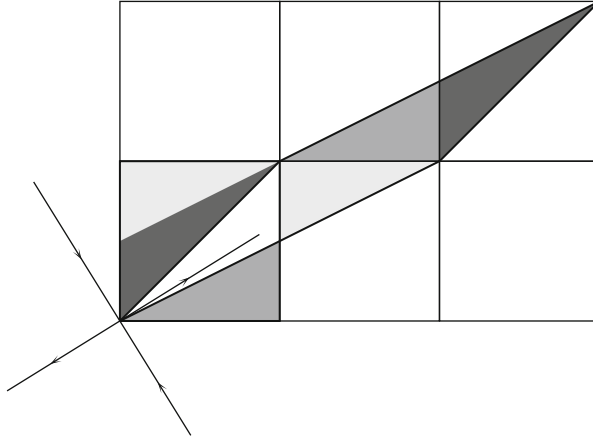
Section 1.3 forms the centerpiece: basic and iconic features of a hyperbolic dynamical system are derived from what thereby appears as the very core features of hyperbolic dynamics: Expansivity and shadowing. The latter is that in a hyperbolic system anything one can imagine approximately happening is, to good approximation, actually happening in the system. Sections 1.3.3–1.3.6 (pages 24–31) show that the Shadowing *Lemma* produces the essential richness and rigidity of the orbit structure of a hyperbolic dynamical system (expansivity, the Anosov Closing Lemma, specification, spectral decomposition, topological stability). The stronger Anosov Shadowing *Theorem* more easily yields structural stability and symbolic descriptions. This whole development uses the Contraction-Mapping Principle (Proposition 1.6.3) but not the technical sections on invariant manifolds (Sects. 1.6.4 and 1.6.5), yet even with complete proofs it only occupies pages 24–37.

Section 1.5 can also be read independently of the preceding material, though the examples in the present section will help. Based on ideas originally due to Babilot and Coudène, it shows the Hopf argument to maximum advantage, which is therefore presented by stating explicitly what is needed for the argument rather than relying on context and background. While the Hopf argument was developed to show ergodicity, we show that it can be used effectively to establish mixing with no added effort, and in the right circumstances multiple mixing for even less effort. A notable ingredient of independent interest is the ergodicity of the stable (or unstable) foliation, which is rarely featured in introductions, and which results from a simple argument.<sup>2</sup>

Section 1.6 provides the Contraction-Mapping Principle and the Hadamard–Perron Stable/Unstable Manifold Theorem. The former is invoked in Sect. 1.3, and we illustrate the major importance of the latter in the subject beyond the arguments in Sect. 1.5 by using these invariant foliations to further the understanding of the topological dynamics of a hyperbolic set. We prove this using the Perron–Irwin method and provide the Hadamard method in a separate chapter in the form of Hadamard’s original presentation [Ha01], translated here into English, presumably for the first time.

---

<sup>2</sup>Self-contained save for invoking absolute continuity, which can be found in [Br02, Chap. 6].



**Fig. 1.1** Example 1.1.1 (©Cambridge University Press, reprinted from [KaHa95] with permission)

The presentations in Sects. 1.3.3–1.3.6 and 1.5 are to our knowledge new to the expository literature.<sup>3</sup> Sections 1.3 and 1.4 implement an approach suggested by Bowen, Anosov and Katok [Bo75, Bo78, Ka81]. The approach in Sect. 1.5 emphasizes the generality of the Hopf argument by using neither compactness nor a smooth structure. In this respect we follow the lead of Babillot and Coudène, and this feature has enabled them to obtain new results.

## 1.1.2 Examples

The first example is an iconic model of hyperbolic dynamics.

*Example 1.1.1 (Toral Automorphism)* Since the matrix  $\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$  has integer entries, it induces a well-defined map  $F_{\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}}$  of the 2-torus  $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$  (Fig. 1.1). Since it has unit determinant, the same goes for the inverse, which means that it defines a diffeomorphism (indeed, algebraic automorphism) of  $\mathbb{T}^2$ —which, furthermore, preserves area. The eigenvalues are

$$\lambda_1 = \frac{3 + \sqrt{5}}{2} > 1 \text{ and } \lambda_1^{-1} = \lambda_2 = \frac{3 - \sqrt{5}}{2} < 1.$$

<sup>3</sup>Other than [Co16], which appeared around the time this minicourse was first given.

The eigenvectors for the first eigenvalue are on the line  $y = \frac{\sqrt{5}-1}{2}x$ . The family of lines parallel to it is invariant, and distances on those lines are expanded by a factor  $\lambda_1$ . Similarly, there is an invariant family of contracting lines  $y = \frac{-\sqrt{5}-1}{2}x + \text{const}$ . This expansion and contraction define hyperbolicity. It is an interesting exercise to show that the collection of periodic points is exactly the set of points with rational coordinates. Thus, periodic orbits are dense. But there are also dense orbits, indeed, almost every orbit is dense.

*Example 1.1.2* More generally, any  $A \in GL(m, \mathbb{Z})$  induces an automorphism  $F_A$  of  $\mathbb{T}^m$  that preserves Lebesgue measure. We say that it is hyperbolic if  $A$  has no eigenvalues on the unit circle.

*Example 1.1.3 (Walters)* In like manner, an area-preserving automorphism of the 4-torus is induced by

$$W := \begin{pmatrix} 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 8 \\ 0 & 1 & 0 & -6 \\ 0 & 0 & 1 & 8 \end{pmatrix}.$$

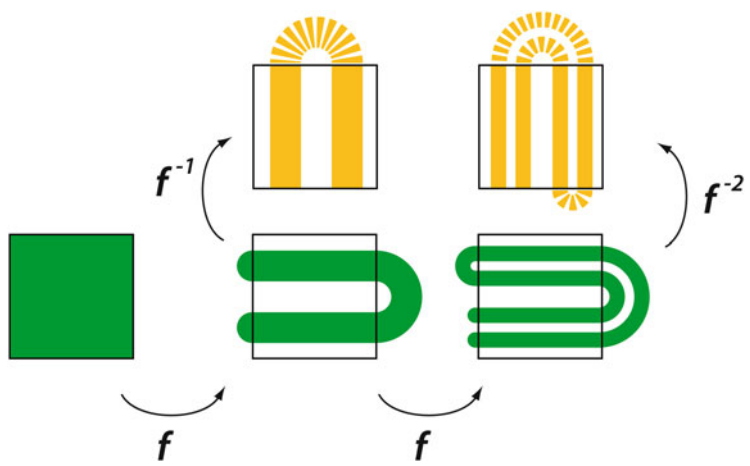
The eigenvalues  $2 - \sqrt{3} \pm i\sqrt{4\sqrt{3}-6}$  lie on the unit circle and the eigenvalues  $\lambda_{\pm} = 2 + \sqrt{3} \pm \sqrt{2(3+2\sqrt{3})}$  are real and satisfy  $0 < \lambda_- < 1 < \lambda_+$ . This automorphism is thus *partially hyperbolic*. The components of the corresponding eigenvectors

$$v^{\pm} := (-2 - \sqrt{3} \pm \sqrt{2(3+2\sqrt{3})}, 3 \mp 2\sqrt{2(-3+2\sqrt{3})}, -6 + \sqrt{3} \pm \sqrt{2(3+2\sqrt{3})}, 1)$$

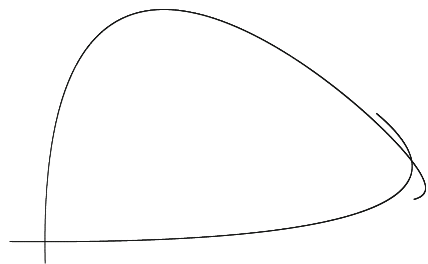
are independent over the rationals, that is, the vector space over  $\mathbb{Q}$  generated by  $-2 - \sqrt{3} + \sqrt{2(3+2\sqrt{3})}$ ,  $3 - 2\sqrt{2(-3+2\sqrt{3})}$ ,  $-6 + \sqrt{3} + \sqrt{2(3+2\sqrt{3})}$ , and 1 is 4-dimensional.

*Example 1.1.4 (Horseshoe)* Hyperbolic Cantor sets are ubiquitous, and an iconic way in which they arise is via *horseshoes*: A map  $f$  of the plane (or sphere, or of any surface) squeezes a rectangle  $\Delta$  vertically, stretches it horizontally and folds it over the original rectangle (Fig. 1.2). The set  $\Lambda := \bigcap_{n \in \mathbb{Z}} f^n(\Delta)$  of points whose orbits are in  $\Delta$  is then a hyperbolic Cantor set with vertical contracting direction and horizontal expanding direction. Nonlinear versions of this have the same qualitative features—this is the content of the Structural Stability Theorem 1.4.6 below.

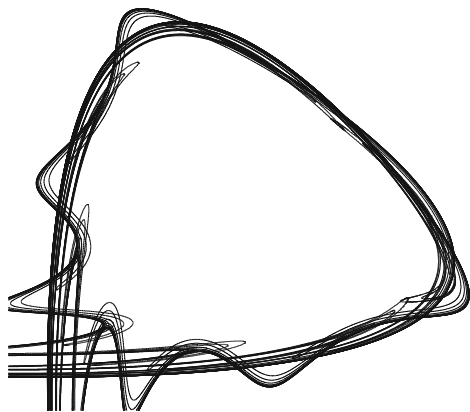
This situation arises whenever there are transverse homoclinic points as in Fig. 1.3. When the stable and unstable curve of a hyperbolic fixed point intersect transversely, they produce tangles (Fig. 1.4), and these in turn produce horseshoes



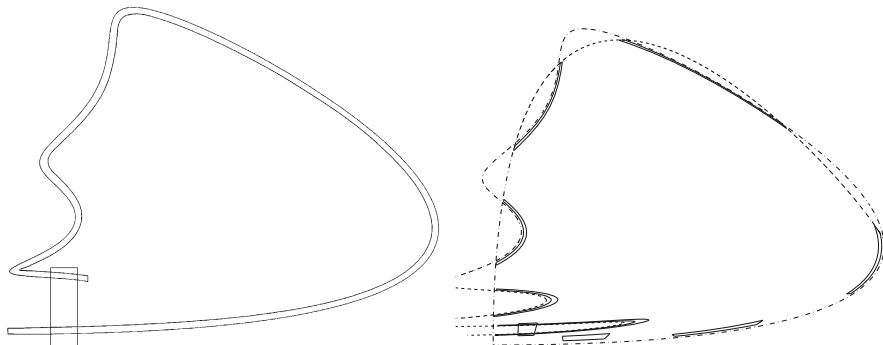
**Fig. 1.2** Horseshoe (Picture from [https://en.wikipedia.org/wiki/Horseshoe\\_map](https://en.wikipedia.org/wiki/Horseshoe_map))



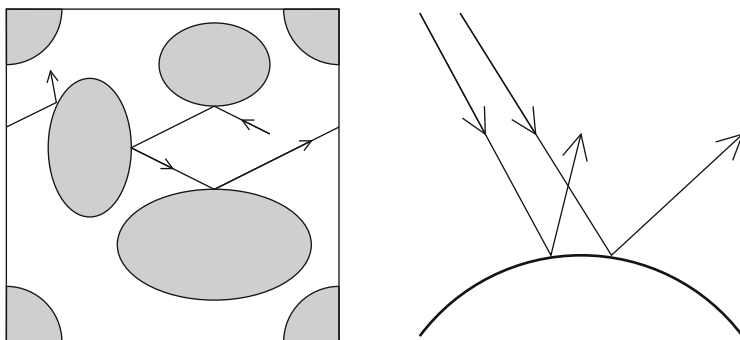
**Fig. 1.3** Transverse homoclinic point (©Cambridge University Press, reprinted from [KaHa95] with permission)



**Fig. 1.4** Homoclinic tangles (©Cambridge University Press, reprinted from [KaHa95] with permission)



**Fig. 1.5** Horseshoes from tangles (©Cambridge University Press, reprinted from [KaHa95, HaKa03] with permission)



**Fig. 1.6** Dispersing billiards (reprinted from [Yo98] with permission)

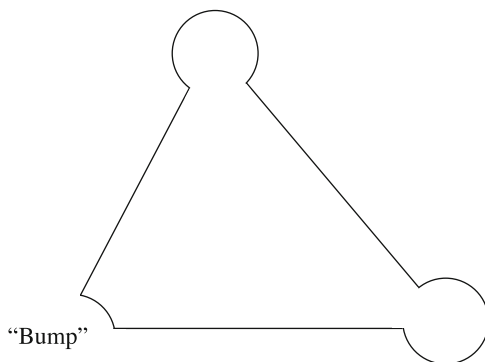
for an iterate of the map; this is the Birkhoff–Smale Theorem, illustrated in Fig. 1.5. Because of their interest with respect to smooth ergodic theory we also present some examples with singularities of various sorts.

*Example 1.1.5* ([ChMa06, p. 67]) A billiard  $\mathcal{D} \subsetneq \mathbb{T}^2$  is said to be *dispersing* if it is defined by reflection in the boundary of smooth strictly convex “scatterers.”<sup>4</sup> If it has no corners or cusps, then Sinai’s Fundamental Theorem of the theory of dispersing billiards [BuSi73, Si70], see also [ChMa06, Theorem 5.70], establishes hyperbolic behavior of the billiard map (Fig. 1.6).

*Example 1.1.6* (Fig. 1.7) Sinai’s Fundamental Theorem also applies to *polygonal billiards with pockets*. These are noncircular billiards obtained from a convex polygon as follows: for each vertex add a disk whose interior contains this vertex and none other [ChTr98, Theorem 4.1]. One can furthermore add “bumps”, i.e., dispersing circle arcs in corners.

<sup>4</sup>One can allow corners at considerable expense of additional effort [ChMa06, p. 69].

**Fig. 1.7** Polygonal billiards with pockets (picture by Serge Troubetzkoy from [ChT98], ©IOP Publishing & London Mathematical Society. Reproduced with permission)



*Example 1.1.7* The *Katok map* is a totally ergodic<sup>5</sup> area-preserving deformation of  $F_{\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}}$  obtained by “damping” the hyperbolicity of  $F_{\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}}$  so as to make the origin a nonhyperbolic fixed point. It is on the boundary of the set of Anosov diffeomorphisms (hence not uniformly hyperbolic) and its stable and unstable partitions are homeomorphic to those of  $F_{\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}}$  [BaPe13, §1.3], [BaPe07, §6.3], [Ka79, §2.2], [PeSeZh].

### 1.1.3 Hyperbolic Dynamics

The primary distinction that sets apart smooth dynamics from general topological dynamics is the availability of the linearization provided by the differential; one can use the linear part of a map to draw conclusions about local behavior of the map itself. Among the elliptic, parabolic, and hyperbolic situations the latter is the one where linearization is most powerful. What makes hyperbolic dynamics distinct from the other two classes is that for the linearization of a map eigenvalues off the unit circle correspond to exponential behavior under iterates, and such behavior is robust enough to produce analogous behavior for the map itself and to engender structural stability.

This local aspect of hyperbolic dynamics combined with the recurrence arising from compactness of the space provides for complex and interesting features of the global structure. In accordance with the main dichotomy between topological and measurable dynamics there are separate but related features of interest.

---

<sup>5</sup>That is, all iterates are ergodic.

In contrast to the individual instability of orbits, the complicated topological dynamics of hyperbolic systems is distinguished by structural stability. Indeed, hyperbolic dynamical systems are *characterized* by structural stability, and to a remarkable degree a classification is possible. Furthermore, even though periodic data give a large number of moduli of differentiable conjugacy, there are interesting results about smooth conjugacy and rigidity.

On the side of measurable dynamics there is the important motivation that Hamiltonian hyperbolic flows, in particular geodesic flows of negatively curved manifolds, are ergodic (with respect to volume); this provides nontrivial classes of examples satisfying Boltzmann's Fundamental Postulate.

Altogether hyperbolic dynamical systems exhibit a remarkable combination of phenomena: Maximally sensitive dependence of an orbit on initial conditions, strong recurrence and mixing properties, many invariant measures, positive entropy, intertwining of periodic and nonperiodic orbits, an abundance of periodic points both in terms of exponential growth of their number as a function of the period and density (topologically as well as in terms of density of their  $\delta$ -measures among all invariant Borel probability measures), structural stability, and the existence of a Markov model both topologically and measure-theoretically.

## 1.2 Historical Sketch

There are several intertwined strands of the history of hyperbolic dynamics: Geodesic flows and statistical mechanics on one hand and hyperbolic phenomena ultimately traceable to some application of dynamical systems. Geodesic flows were studied, e.g., by Hadamard, Hedlund, Hopf (primarily either on surfaces or in the case of constant curvature) and Anosov–Sinai (negatively curved surfaces and higher-dimensional manifolds). Other hyperbolic phenomena appear in the work of Poincaré (homoclinic tangles in celestial mechanics [Po90]), Perron (differential equations [Pe28]), Cartwright, Littlewood (relaxation oscillations in radio circuits [Ca50, CaLi45, Li57]), Levinson (the van der Pol equation, [Le49]) and Smale (horseshoes, [Sm65, Sm63]), as well as countless others in recent history. In looking back, Smale [Sm98] breaks the study of hyperbolic phenomena into three strands: Poincaré–Birkhoff, (Poincaré–) Cartwright–Littlewood–Levinson and Andronov–Pontryagin–Lefschetz–Peixoto (structural stability and topology).

### 1.2.1 Homoclinic Tangles

The advent of complicated dynamics took place in the context of Newtonian mechanics, according to which simple underlying rules governed the evolution of the world in clockwork fashion. The successes of classical and especially celestial mechanics in the eighteenth and nineteenth century were seemingly unlimited and



Pierre Simon de Laplace felt justified in saying (in the opening passage he added to [La95, p. 2]):

Nous devons donc envisager l'état présent de l'univers, comme l'effet de son état antérieur, et comme la cause de celui qui va suivre. Une intelligence qui pour un instant donné, connaîtrait toutes les forces dont la nature est animée, et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule les mouvements des plus grands corps de l'univers et ceux du plus léger atome: rien ne serait incertain pour elle, et l'avenir comme le passé, serait présent à ses yeux.<sup>6</sup>

The enthusiasm in this passage is understandable and its forceful description of (theoretical) determinism is a good anchor for an understanding of one of the basic aspects of dynamical systems. Moreover, the titanic life's work of Laplace in celestial mechanics earned him the right to make such bold pronouncements. Another bold pronouncement of his, that the solar system is stable, came under renewed scrutiny later in the nineteenth century, and Henri Poincaré was expected to win a competition to finally establish this fact. However, Poincaré came upon hyperbolic phenomena in revising his prize memoir [Po90] on the three-body problem before publication. He found that homoclinic tangles (which he had initially overlooked) caused great difficulty and necessitated essentially a reversal of the main thrust of that memoir [Ba97]. He perceived that there is a highly intricate web of invariant curves and that this situation produces dynamics of unprecedented complexity:

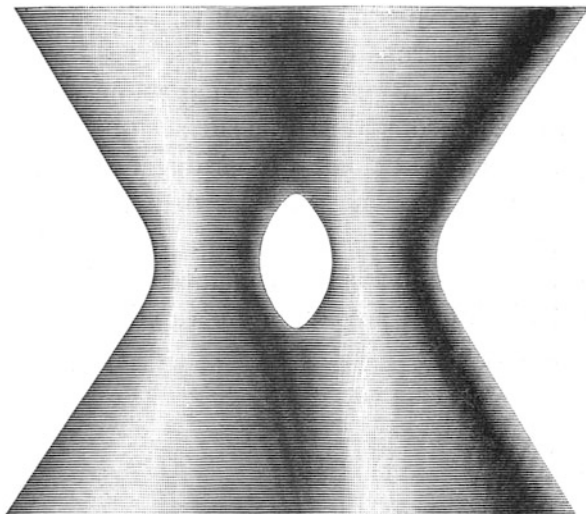
Que l'on cherche à se représenter la figure formée par ces deux courbes et leurs intersections en nombre infini dont chacune correspond à une solution doublement asymptotique, ces intersections forment une sorte de treillis, de tissu, de réseau à mailles infiniment serrées; chacune des deux courbes ne doit jamais se recouper elle-même, mais elle doit se replier sur elle-même d'une manière très complexe pour venir recouper une infinité de fois toutes les mailles du réseau. On sera frappé de la complexité de cette figure, que je ne cherche même pas à tracer.<sup>7</sup>

This is often viewed as the moment chaotic dynamics was first noticed. He concluded that in all likelihood the prize problem could not be solved as posed: To find series expansions for the motions of the bodies in the solar system that

---

<sup>6</sup>We ought then to consider the present state of the universe as the effects of its previous state and as the cause of that which is to follow. An intelligence that, at a given instant, could comprehend all the forces by which nature is animated and the respective situation of the beings that make it up, if moreover it were vast enough to submit these data to analysis, would encompass in the same formula the movements of the greatest bodies of the universe and those of the lightest atoms. For such an intelligence nothing would be uncertain, and the future, like the past, would be open to its eyes.

<sup>7</sup>If one tries to imagine the figure formed by these two curves with an infinite number of intersections, each corresponding to a doubly asymptotic solution, these intersections form a kind of trellis, a fabric, a network of infinitely tight mesh; each of the two curves must not cross itself but it must fold on itself in a very complicated way to intersect all of the meshes of the fabric infinitely many times. One will be struck by the complexity of this picture, which I will not even attempt to draw.



**Fig. 1.8** Negatively curved surface (Reproduced from Hadamard [Ha98] ©1898 Elsevier Masson SAS. All rights reserved)

converge uniformly for all time. Indeed, when Birkhoff picked up the study of this situation in his prize memoir [Bi35] for the Papal Academy of Sciences, he noted that and described how this implies complicated dynamics [Bi35, p. 184] (see also Example 1.1.4):

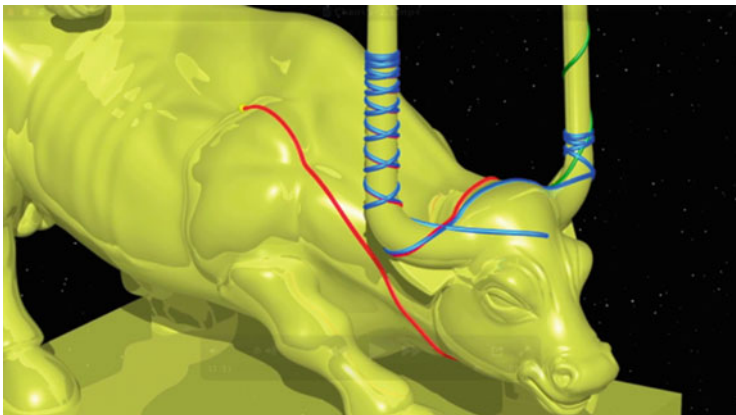
Il paraît donc que tout système dynamique non-intégrable qui admet une seule solution homocline de cette espèce, doit admettre une hiérarchie presque inconcevable de solutions dans le voisinage étendu correspondant.<sup>8</sup>

### 1.2.2 Geodesic Flows

A major class of mathematical examples motivating the development of hyperbolic dynamics is that of geodesic flows of Riemannian manifolds of negative sectional curvature. Hadamard considered (noncompact) surfaces in  $\mathbb{R}^3$  of negative curvature [Ha98] and found, with apparent delight, that if the unbounded parts are “large” (do not pinch to arbitrarily small diameter as you go outward along them) then at any point the initial directions of bounded geodesics form a Cantor set (Figs. 1.8 and 1.9).

---

<sup>8</sup>It thus appears that any nonintegrable dynamical system which admits a single homoclinic solution of this kind must admit an almost inconceivable hierarchy of solutions in the extended neighborhood.



**Fig. 1.9** Duhem's bull (Picture from <http://www.chaos-math.org/fr/chaos-v-billards>)

La génération de l'ensemble  $E$  rappelle évidemment celle de ces ensembles rencontrés par M. Poincaré, introduits plus explicitement dans la Science par M. Bendixson, puis étudiés par M. Cantor et qui, tout en étant parfaits, ne sont *condensés* dans aucun intervalle. Les angles  $\lambda$  jouent ici le rôle des intervalles nommés  $(a_v, b_v)$  par M. Cantor.<sup>9</sup>

Since only countably many directions give geodesics that are periodic or asymptotic to a periodic one, this also proves the existence of more complicated bounded geodesics. Hadamard was fully aware of the connection to Cantor's work and to similar sets discovered by Poincaré, and he appreciated the relation between the complicated dynamics in the two contexts. Hadamard also showed that each homotopy class (except for the “waists” of cusps) contains a unique geodesic. A classic by Duhem [Du91] seized upon this to eloquently describe the dynamics of a geodesic flow in terms of what might now be called deterministic chaos: Duhem used it to illustrate that determinism in classical mechanics does not imply any practical long-term predictability.

Les recherches de M. J. Hadamard nous fournissent un exemple bien saisissant; il est emprunté à l'un des problèmes les plus simples qu'ait à traiter la moins compliquée des théories physiques, la Mécanique.

Une masse matérielle glisse sur une surface; aucune pesanteur, aucune force ne la sollicite; aucun frottement ne gêne son mouvement. Si la surface sur laquelle elle doit demeurer est un plan, elle décrit une ligne droite avec une vitesse uniforme; si la surface est une sphère, elle décrit un arc de grand cercle, également avec une vitesse uniforme. Si notre point matériel se meut sur une surface quelconque, il décrit une ligne que les géomètres nomment une *ligne géodésique* de la surface considérée. Lorsqu'on se donne la position

<sup>9</sup>The manner in which these sets arise clearly recalls that of the sets encountered by Mr. Poincaré, introduced more explicitly to the subject by Mr. Bendixson, then studied by Mr. Cantor and which, while being perfect, are not *dense* in any interval. Here the angles  $\lambda$  play the role of the intervals called  $(a_v, b_v)$  by Mr. Cantor.

initiale de notre point matériel et la direction de sa vitesse initiale, la géodésique qu'il doit décrire est bien déterminée.

Les recherches de M. Hadamard ont porté, en particulier, sur les géodésiques des surfaces à courbures opposées, à connexions multiples, qui présentent des nappes infinies [Ha98]; sans nous attarder ici à définir géométriquement de semblables surfaces, bornons-nous à en donner un exemple.<sup>10</sup>

---

<sup>10</sup>The research of J. Hadamard provides us with a very striking example of such a deduction that can never be useful. It is borrowed from one of the simplest problems that the least complicated of physical theories, mechanics, has to deal with.

A material mass slides on a surface; no weight and no force act on it; no friction interferes with its motion. If the surface on which it is to remain is a plane, it describes a straight line with uniform velocity; if the surface is a sphere, it describes the arc of a great circle, also with uniform velocity. No matter what surface our material point moves on, it describes a line that geometers call a "geodesic line" of the surface considered. When the initial position of our material point and the direction of its initial velocity are given, the geodesic it should describe is well determined.

Hadamard's research has dealt especially with geodesics of surfaces of negative curvature, with multiple connections, and with infinite folds [Ha98]. Without stopping here to define such surfaces geometrically, let us restrict ourselves to giving an illustration of one of them.

Imagine the forehead of a bull, with the protuberances from which the horns and ears start, and with the little mountain passes between these protuberances; but elongate these horns and ears without limit so that they extend to infinity; then you will have one of the surfaces we wish to study.

On such a surface geodesics may show many different aspects.

There are, first of all, geodesics which close on themselves. There are some also which are never infinitely distant from their starting point even though they never exactly pass through it again; some turn continually around the right horn, others around the left horn, or right ear, or left ear; others, more complicated, alternate, in accordance with certain rules, the turns they describe around one horn with the turns they describe around the other horn, or around one of the ears. Finally, on the forehead of our bull with his unlimited horns and ears there will be geodesics going to infinity, some mounting the right horn, others mounting the left horn, and still others following the right or left ear.

Despite this complication, if we know with complete accuracy the initial position of a material point on this bull's forehead and the direction of the initial velocity, the geodesic line that this point will follow in its motion will be determined without any ambiguity. In particular, we shall know whether the moving point will always remain at a finite distance from its starting point or whether it will move away indefinitely so as never to return.

It will be quite a different matter if the initial conditions are not mathematically but practically given: the initial position of our material point will no longer be a determinate point on the surface, but some point taken inside a small spot; the direction of the initial velocity will no longer be a straight line defined without ambiguity, but some one of the lines included in a narrow bundle connected by the contour of the small spot; and our practically determined initial conditions will, for the geometer, correspond to an infinite multiplicity of different initial conditions.

Let us imagine certain of these geometrical data corresponding to a geodesic line that does not go to infinity, for example, a geodesic line that turns continually around the right horn. Geometry permits us to assert the following: Among the innumerable mathematical data corresponding to the same practical data, there are some which determine a geodesic moving indefinitely away from its starting point; after turning a certain number of times around the right horn, this geodesic will go to infinity on the right horn, or on the left horn, or on the right or left ear. More than that: despite the narrow limits which restrict the geometrical data capable of representing the given practical data, we can always take these geometrical data in such a way that the geodesic will go off on that one of the infinite folds which we have chosen in advance.

Imaginons le front d'un taureau, avec les éminences d'où partent les cornes et les oreilles, et les cols qui se creusent entre ces éminences; mais allongeons sans limite ces cornes et ces oreilles, de telle façon qu'elles s'étendent à l'infini; nous aurons une des surfaces que nous voulons étudier.

Sur une telle surface, les géodésiques peuvent présenter bien des aspects différents.

Il est, d'abord, des géodésiques qui se ferment sur elles-mêmes. Il en est aussi qui, sans jamais repasser exactement par leur point de départ, ne s'en éloignent jamais infiniment; les unes tournent sans cesse autour de la corne droite, les autres autour de la corne gauche, ou de l'oreille droite, ou de l'oreille gauche; d'autres, plus compliquées, font alterner suivant certaines règles les tours qu'elles décrivent autour d'une corne avec les tours qu'elles décrivent autour de l'autre corne, ou de l'une des oreilles. Enfin, sur le front de notre taureau aux cornes et aux oreilles illimitées, il y aura des géodésiques qui s'en iront à l'infini, les unes en gravissant la corne droite, les autres en gravissant la corne gauche, d'autres encore en suivant l'oreille droite ou l'oreille gauche.

Malgré cette complication, si l'on connaît avec une entière exactitude la position initiale d'un point matériel sur ce front de taureau et la direction de la vitesse initiale, la ligne géodésique que ce point suivra dans son mouvement sera déterminée sans aucune ambiguïté. On saura très certainement, en particulier, si le mobile doit demeurer toujours à distance finie ou s'il s'éloignera indéfiniment pour ne plus jamais revenir.

Il en sera tout autrement si les conditions initiales ne sont pas données mathématiquement, mais pratiquement; la position initiale de notre point matériel ne sera plus un point déterminé sur la surface, mais un point quelconque pris à l'intérieur d'une petite tache; la direction de la vitesse initiale ne sera plus une droite définie sans ambiguïté, mais une quelconque des droites que comprend un étroit faisceau dont le contour de la petite tache forme le lien; à nos données initiales pratiquement déterminées correspondra, pour le géomètre, une infinie multiplicité de données initiales différentes.

Imaginons que certaines de ces données géométriques correspondent à une ligne géodésique qui ne s'éloigne pas à l'infini, par exemple, à une ligne géodésique qui tourne sans cesse autour de la corne droite. La Géométrie nous permet d'affirmer ceci: Parmi les données mathématiques innombrables qui correspondent aux mêmes données pratiques, il en est qui déterminent une géodésique s'éloignant indéfiniment de son point de départ; après avoir tourné un certain nombre de fois autour de la corne droite, cette géodésique s'en ira à l'infini soit sur la corne droite, soit sur la corne gauche, soit sur l'oreille droite, soit sur l'oreille gauche. Il y a plus; malgré les limites étroites qui resserrent les données géométriques capables de représenter nos données pratiques, on peut toujours prendre ces données géométriques de telle sorte que la géodésique s'éloigne sur celle des nappes infinies qu'on aura choisie d'avance.

---

It will do no good to increase the precision with which the practical data are determined, to diminish the spot where the initial position of the material point is, to tighten the bundle which includes the initial direction of the velocity, for the geodesic which remains at a finite distance while turning continually around the right horn will not be able to get rid of those unfaithful companions who, after turning like itself around the right horn, will go off indefinitely. The only effect of this greater precision in the fixing of the initial data will be to oblige these geodesics to describe a greater number of turns embracing the right horn before producing their infinite branch; but this infinite branch will never be suppressed.

If, therefore, a material point is thrown on the surface studied starting from a geometrically given position with a geometrically given velocity, mathematical deduction can determine the trajectory of this point and tell whether this path goes to infinity or not. But, for the physicist, this deduction is forever unusable. When, indeed, the data are no longer known geometrically, but are determined by physical procedures as precise as we may suppose, the question put remains and will always remain unanswered.

On aura beau augmenter la précision avec laquelle sont déterminées les données pratiques, rendre plus petite la tache où se trouve la position initiale du point matériel, resserrer le faisceau qui comprend la direction initiale de la vitesse, jamais la géodésique qui demeure à distance finie en tournant sans cesse autour de la corne droite ne pourra être débarrassée de ces compagnes infidèles qui, après avoir tournées comme elle autour de la même corne, s'écarteront indéfiniment. Le seul effet de cette plus grande précision dans la fixation des données initiales sera d'obliger ces géodésiques à décrire un plus grand nombre de tours embrassant la corne droite avant de produire leur branche infinie; mais cette branche infinie ne pourra jamais être supprimée.

Si donc un point matériel est lancé sur la surface étudiée à partir d'une position géométriquement donnée, avec une vitesse géométriquement donnée, la déduction mathématique peut déterminer la trajectoire de ce point et dire si cette trajectoire s'éloigne ou non à l'infini. Mais, pour le physicien, cette déduction est à tout jamais inutilisable. Lorsqu'en effet les données ne sont plus connues géométriquement, mais sont déterminées par des procédés physiques, si précis qu'on les suppose, la question posée demeure et demeurera toujours sans réponse.

To today's reader his description amounts to a shrewd translation of symbolic dynamics into everyday language. Indeed, several authors trace back symbolic dynamics to this paper of Hadamard. Birkhoff is among them: In his proof of the Birkhoff–Smale Theorem (see Example 1.1.4) symbolic sequences appear (as well as a picture that resonates with Fig. 1.5), and he remarks [Bi35, p. 184]:

De tels symboles arithmétiques... ressemblent un peu aux symboles effectivement introduits par Hadamard dans son étude remarquable des géodésiques sur certaines surfaces ouvertes de courbure totale négative.<sup>11</sup>

It appears, however, that only in 1944 did symbol spaces begin to be seen as dynamical systems, rather than as a coding device [CoNi08].

### 1.2.3 Boltzmann's Fundamental Postulate

Well before Poincaré's work, James Clerk Maxwell (1831–1879) and Ludwig Boltzmann (1844–1906) had aimed to give a rigorous formulation of the kinetic theory of gases and statistical mechanics. A central ingredient was Boltzmann's Fundamental Postulate, which says that the time and space (phase or ensemble) averages of an observable (a function on the phase space) agree. Apparently without a basis, one often ascribes to him the so-called *Ergodic Hypothesis*:

*The trajectory of the point representing the state of the system in phase space passes through every point on the constant-energy hypersurface of the phase space.*

Poincaré and many physicists doubted its validity since no example satisfying it had been exhibited [Po94]. Accordingly, in 1912 Paul and Tatiana Ehrenfest [EhEh] proposed the alternative *Quasi-Ergodic Hypothesis*:

---

<sup>11</sup>Such arithmetic symbols... resemble a little the symbols effectively introduced by Hadamard in his remarkable study of geodesics on certain open surfaces of negative curvature.

*The trajectory of the point representing the state of the system in phase space is dense on the constant energy hypersurface of the phase space.*

Indeed, within a year proofs (by Rosenthal and Plancherel) appeared that the Ergodic Hypothesis fails [P113, Ro13]. (This is obvious today because a trajectory has measure zero in an energy surface.) These difficulties led to the search for *any* mechanical systems with this property. The motion of a single free particle (also known as the geodesic flow) in a negatively curved space emerged as the first and for a long time sole class of examples with this property. Emil Artin 1924:

Es sei gestattet, auf ein einfaches mechanisches System von zwei Freiheitsgraden mit quasiergodischen Bahnen hinzuweisen, zu dem der Verfasser in einem Briefwechsel mit Herrn G. Herglotz gekommen ist. . .<sup>12</sup>

Daraus geht schon hervor, daß die “quasiergodischen Ketten” die Mächtigkeit des Kontinuums haben. Noch mehr! Nach Resultaten von Herrn Celestyn Burstin haben fast alle Zahlen  $\xi$  eine “quasiergodische” Kettenbruchentwicklung. Von den durch einen Punkt der Fläche gehenden geodätischen Linien sind also fast alle quasiergodisch.

Es möge noch einiges über die physikalische Realisierbarkeit gesagt sein. Man erhält... die Rotationsfläche der Traktrix (Zuglinie) eines Fadens der Länge eins. Bekanntlich hat auch sie das Krümmungsmaß  $K = -1$ , so daß sich unsere Halbebene teilweise auf diese Fläche abwickeln läßt... Damit haben wir aber die physikalische Realisierung... Unser mechanisches System läßt sich dann als die kräftefreie Bewegung eines Massenpunktes... interpretieren (der Punkt sei gezwungen auf der Fläche zu bleiben).

Within a decade, the understanding of the problem led to the pertinent contemporary notion, and this turned out to be probabilistic in nature (Fig. 1.10).<sup>13</sup> The 1931 *Birkhoff Ergodic Theorem* 1.7.20 (“time averages exist a.e.”)<sup>14</sup> laid the foundation for the definition of ergodicity now in use, which is: “No proper invariant set has positive measure.”<sup>15</sup>

---

<sup>12</sup>May it be permitted to point to a simple mechanical system with 2 degrees of freedom and quasiergodic orbits upon which the author came in the course of a correspondence with Mr. G. Herglotz...

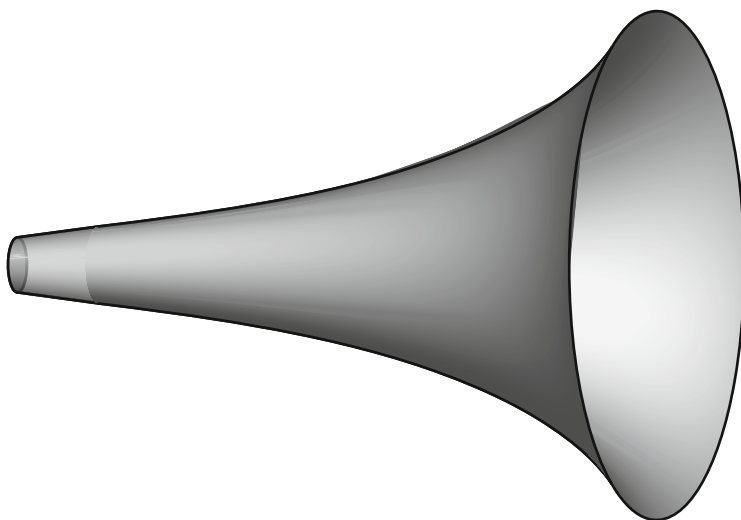
From this one already obtains that the “quasiergodic chains” have the cardinality of the continuum. More! According to results of Mr. Celestyn Burstin almost all numbers  $\xi$  have a “quasiergodic” continued-fraction expansion. Therefore, almost all of the geodesic lines going through a point of the surface are quasiergodic.

Let us remark on the physical realizability. One obtains... the surface of rotation of a tractrix (curve of pursuit) of a string of length 1. It is known to have curvature  $K = -1$ , so our half-plane can be partially developed onto this surface... But with that we have our physical realization... Our mechanical system can be interpreted... as the force-free motion of a point particle (the point being constrained to remain on the surface).

<sup>13</sup>This serves to point out that the earlier quote by Laplace about determinism comes from his *Philosophical essay on probabilities*, where he goes on to say that we often do not have sufficiently detailed initial data, and must hence resort to a probabilistic approach. The motion of a molecule of air was a prominent instant he mentioned in that context.

<sup>14</sup>This was proved after the von Neumann Mean Ergodic Theorem 1.7.33 but published earlier [Zu02].

<sup>15</sup>These two combine to give the Strong Law of Large Numbers.



**Fig. 1.10** The pseudosphere (©Cambridge University Press, reprinted from [KaHa95] with permission)

If this is the case, then time averages agree with space averages—Boltzmann’s Fundamental Postulate. Furthermore, almost every orbit is dense.

The 1930s saw a flurry of work in which Artin’s work was duly extended to other manifolds of *constant* negative curvature. For constant curvature, finite volume and finitely generated fundamental group the geodesic flow was shown to be topologically transitive [Ko29, Lo29], topologically mixing [He36], ergodic [Ho36], and mixing [He39a, Ho39]. (In the case of infinitely generated fundamental group the geodesic flow may be topologically mixing without being ergodic [Se35]). If the curvature is allowed to vary between two negative constants then finite volume implies topological mixing [Gr39] (see also [GrLa09, p. 183]). But as Hedlund noted in an address delivered before the New York meeting of the American Mathematical Society on October 27, 1938 [He39b]:

Outstanding problems remain unsolved, a notable one being the problem of metric transitivity [ergodicity] of the geodesic flow on a closed analytic surface of *variable* negative curvature.

It so happens that Eberhard Hopf was just then working on this problem [Ho39]: He considered compact surfaces of nonconstant (predominantly) negative curvature and was able to show ergodicity of the Liouville measure (phase volume):

Erst die hier entwickelte geometrische Methode der asymptotischen Geodätischen führt wesentlich weiter. Bei ihrer Anwendung auf Flächen mit variablem  $K < 0$  sieht man, daß sie an der wesentlichen Stelle, nämlich der erwähnten Unstabilität der Geodätischen, einsetzt. Was nun Flächen negativer Krümmung  $K$  vor anderen Flächen auszeichnet, ist vor allem die starke Unstabilität ihrer Geodätischen. Der Normalabstand  $n$  einer Geodätischen von einer



infinitesimal benachbarten Geodätischen genügt längs derselben der Variationsgleichung

$$\frac{d^2n}{ds^2} + Kn = 0.$$

Verläuft  $K$  auf  $\mathcal{F}$  zwischen festen negativen Grenzen, so wächst  $n(s)$  in mindestens einer Richtung wie eine Exponentialfunktion.<sup>16</sup>

From Hopf's work there was no progress in the direction of ergodicity of geodesic flows (= free particle motion) for almost 30 years. Hopf's argument had shown roughly that Birkhoff averages of a continuous function must be constant on almost every leaf of the horocycle foliation, and, since these foliations are  $C^1$ , the averages are constant a.e. He realized that much of the argument was independent of the dimension of the manifold (indeed, he carried much of the work out in arbitrary dimension), but could not verify the  $C^1$  condition in higher dimension. Anosov [An69] axiomatized Hopf's instability, defining *Anosov flows*, and he showed that differentiability may indeed fail in higher dimension, but that the Hopf argument can still be used because the invariant laminations have an absolute continuity property [An69, AnSi67, PuSh72, Ba95, Br02, BaPe01]. This extension is interesting because despite the ergodicity paradigm central to statistical mechanics, Boltzmann's Fundamental Postulate, there was a dearth of examples of ergodic Hamiltonian systems. To this day the quintessential model for the Fundamental Postulate, the gas of hard spheres, resists attempts to prove ergodicity.

The Hopf argument remains the main method for establishing ergodicity in hyperbolic dynamical systems without an algebraic structure (the alternative tool being the theory of equilibrium states, see [KaHa95, Theorem 20.4.1]).

### 1.2.4 Picking Up from Poincaré

Like Hadamard, several mathematicians had begun to pick up some of Poincaré's work during his lifetime. Birkhoff did so soon after Poincaré's death. He addressed issues that arose from the mathematical development of mechanics and celestial mechanics such as Poincaré's Last Geometric Theorem and the complex dynamics necessitated by homoclinic tangles [Bi27, Sect. 9]. He was also important in the

---

<sup>16</sup>Only the geometric method of asymptotic geodesics that is developed here takes us significantly further. When applying it to surfaces with variable  $K < 0$  one sees that it homes in on the essential point, namely the aforementioned instability of geodesics.

What distinguishes surfaces of negative curvature  $K$  from other surfaces is primarily the strong instability of its geodesics. The normal distance  $n$  between a geodesic and an infinitesimally close one satisfies the variational equation

$$\frac{d^2n}{ds^2} + Kn = 0$$

along the geodesic. If  $K$  ranges between fixed negative bounds on  $\mathcal{F}$ , then  $n(s)$  grows exponentially in at least one direction.

development of ergodic theory (the Poincaré Recurrence Theorem 1.7.11 is proved in Poincaré’s prize memoir [Po90]), notably by proving the Pointwise Ergodic Theorem.

The work of Cartwright and Littlewood during World War II on relaxation oscillations in radar circuits [CaLi45, Ca50, Li57] consciously built on Poincaré’s work. Further study of the van der Pol equation by Levinson [Le49] contained the first example of a structurally stable diffeomorphism with infinitely many periodic points. (Structural stability originated in 1937 [AnPo37] but began to flourish only 20 years later.) This was brought to the attention of Smale. Inspired by Peixoto’s work, which carried out such a program in dimension two [Pe62], Smale was after a program of studying diffeomorphisms with a view to classification [Sm67]. Until alerted by Levinson, Smale conjectured that only Morse–Smale systems (finitely many periodic points with stable and unstable sets in general position) could be structurally stable [Sm60]. He eventually extracted from Levinson’s work the horseshoe [Sm65, Sm63]. Smale in turn was in contact with the Russian school, where Anosov systems (then C- or U-systems) had been shown to be structurally stable, and their ergodic properties were studied by way of further development of the study of geodesic flows in negative curvature.

## 1.2.5 Modern Hyperbolic Dynamics

It is interesting to note that hyperbolic sets were sometimes said to constitute “a Perron situation”, for example by Alekseev [Al68, Definition 12] (in which the Smale horseshoe makes an appearance as well). Independently, Thom (unpublished) studied hyperbolic toral automorphisms and their structural stability.<sup>17</sup> The initial development of the theory of hyperbolic systems in the 1960s was followed by the founding of the theory of nonuniformly hyperbolic dynamical systems in the 1970s, mostly by Pesin [Os68, Pe76] (during which time the hyperbolic theory continued its development). One of the high points in the development of smooth dynamics is the proof by Robbin, Robinson, Mañé and Hayashi that structural stability indeed characterizes hyperbolic dynamical systems. For diffeomorphisms this was achieved in the 1980s, for flows in the 1990s. Starting in the mid-eighties the field of

---

<sup>17</sup>The automorphism  $\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$  of Example 1.1.1 is often called the “Arnold cat map” by physicists after [ArAv68, Fig. 1.17]. Since there were typewritten notes by Avez, the existence of which was pointed out to me by David Chillingworth, that preceded the joint book and included a similar picture that used a bat, physicists should consider the term “Avez bat map.” There is a possibility that a yet earlier paper by Arnold had the cat picture after all, so the safest way to cut the historical knot is to note that this map also [is the square of that map that] gives the recursion for the Fibonacci numbers, and could hence be called the Fibonacci rabbit map. However, due to Thom’s role (he communicated this example to Smale in response to the short-lived conjecture that only systems with finitely many periodic points are structurally stable), “auThomorphism” has an equal claim, and “Thom cat” (due to Katok) fares fairly finely for faux feline faunetics.

geometric and smooth rigidity came into being. At the same time topological and stochastic properties of attractors began to be better understood with techniques that nowadays blend ideas from hyperbolic and one-dimensional dynamics. Meanwhile, the theory of partially hyperbolic dynamical systems, which goes back to seminal works of Brin and Pesin in the 1970s, has seen explosive development since the last years of the twentieth century, which in turn has entailed renewed interest in the methods of uniformly hyperbolic dynamical systems and their possible extensions to this new realm.

### 1.3 Hyperbolic Sets: Shadowing and Expansiveness

This section and the next develop the core dynamical features of hyperbolicity as a consequence of the shadowing property, deemed by Bowen, Conley [Bo78, p. vii], Anosov, Katok [Bo75, Ka81] and others to be the single most salient feature to characterize hyperbolic dynamical systems. The present section illustrates the array of consequences of the Shadowing Lemma, while the next section builds on a stronger shadowing result that simplifies proofs which use shadowing of whole families of orbits. For explicit comparison, we prove Theorem 1.3.52 both ways.

#### 1.3.1 Definitions

We now describe the class of diffeomorphisms to which the examples in Sect. 1.1.2 belong. First define the *conorm* of a linear map  $A$  by  $\|A\| := \inf\{\|Av\| \mid \|v\| = 1\}$ . This is complementary to the usual norm  $\|A\| := \sup\{\|Av\| \mid \|v\| = 1\}$ .

**Definition 1.3.1 (Hyperbolic, Anosov)** If  $U \subset M$  is open, then an embedding  $f: U \hookrightarrow M$  is said to be *hyperbolic* on a compact set  $\Lambda$  if there exists a Riemannian metric called a *Lyapunov metric* on  $U$  for which there are numbers

$$0 < \lambda < 1 < \mu \quad (1.1)$$

and a pairwise orthogonal invariant splitting into stable and unstable directions

$$T_x M = E^s(x) \oplus E^u(x), \quad D_{xf} E^\tau(x) = E^\tau(f(x)), \quad \tau = s, u$$

such that

$$\|D_{xf} \upharpoonright_{E^s(x)}\| \leq \lambda < 1 < \mu \leq \|D_{xf} \upharpoonright_{E^u(x)}\|. \quad (1.2)$$

If, furthermore,  $\Lambda = M$ , then we say that  $f$  is an *Anosov diffeomorphism*.

*Remark 1.3.2* The usual definition requires that for *any* Riemannian metric there is a constant  $C$  such that instead of (1.2) we have

$$\|D_{xf}^n \upharpoonright_{E^s(x)}\| \leq C\lambda^n \text{ and } C^{-1}\mu \leq \|D_{xf}^n \upharpoonright_{E^u(x)}\|.$$

for all  $n \in \mathbb{N}$  (and omits “orthogonal”—but see Corollary 1.3.8). It is then a theorem [along the lines of (1.15)] that a Lyapunov metric such as in Definition 1.3.1 exists. In this spirit we more generally define *partial hyperbolicity* by requiring that for any Riemannian metric there are a constant  $C$ , numbers  $\lambda < \zeta < \xi < \mu$  with (1.1), and an invariant splitting  $T_x M = E^s(x) \oplus E^c(x) \oplus E^u(x)$  such that  $D_{xf} E^\tau(x) = E^\tau(f(x))$ ,  $\tau = s, c, u$  and

$$\begin{aligned} \|D_{xf}^n \upharpoonright_{E^s(x)}\| &\leq C\lambda^n, \\ \frac{1}{C}\zeta^n &\leq \|D_{xf}^n \upharpoonright_{E^c(x)}\| \leq \|D_{xf}^n \upharpoonright_{E^c(x)}\| \leq C\xi^n, \\ \frac{1}{C}\mu^n &\leq \|D_{xf}^n \upharpoonright_{E^u(x)}\| \end{aligned}$$

for all  $n \in \mathbb{N}$ .

*Example 1.3.3* Example 1.1.3 is partially hyperbolic in this sense.

It is useful to have a characterization of (partial) hyperbolicity in terms of the action of the differential on vector fields.

**Theorem 1.3.4 (Mather)** *Let  $M$  be a smooth manifold,  $U \subset M$  an open subset,  $f: U \rightarrow M$  a  $C^1$  embedding, and  $\Lambda \subset U$  a compact  $f$ -invariant set. Denote by  $\Gamma_b$  the set of bounded vector fields on  $\Lambda$  and by  $\Gamma_c \subset \Gamma_b$  the set of continuous vector fields on  $\Lambda$  (these are sections of the bundle  $T_\Lambda M := TM|_\Lambda$ ), and for a vector field  $X$  on  $\Lambda$  define  $\mathcal{F}(X)$  by*

$$\mathcal{F}(X)(f(x)) := Df_x(X(x)).$$

*Then for  $\ell^- < \ell^+$  the following are equivalent:*

1. *There exist  $\lambda < \ell^-$  and  $\mu > \ell^+$  such that  $\Lambda$  is (partially) hyperbolic with  $\lambda, \mu$  as above.*
2.  $\text{sp}(\mathcal{F}|_{\Gamma_b}) \cap \{z \in \mathbb{C} \mid \ell^- \leq |z| \leq \ell^+\} = \emptyset$ .
3.  $\text{sp}(\mathcal{F}|_{\Gamma_c}) \cap \{z \in \mathbb{C} \mid \ell^- \leq |z| \leq \ell^+\} = \emptyset$ .

*Proof* 1.  $\Rightarrow$  2.: Check that the splitting  $\Gamma_b(T_\Lambda M) = \Gamma_b(E^\lambda) \oplus \Gamma_b(E^\mu)$  has the desired properties.

2.  $\Rightarrow$  3.: Since  $\Gamma_c \subset \Gamma_b$  is an invariant Banach subspace,  $\text{sp}(\mathcal{F}|_{\Gamma_b}) \subset \text{sp}(\mathcal{F}|_{\Gamma_c})$ .

3.  $\Rightarrow$  1.: This involves two simple steps.

**Lemma 1.3.5** *The projections  $\pi^\pm$  that define the splitting  $\Gamma_c = \mathcal{E}^\lambda \oplus \mathcal{E}^\mu$  are  $C^0(\Lambda)$ -linear.*

A map  $L: \Gamma_c \rightarrow \Gamma_c$  is said to be  $C^0(\Lambda)$ -linear if  $L(\varphi X) = \varphi \cdot L(X)$  for all  $\varphi \in C^0(\Lambda)$ . This lets us apply a general fact about continuous maps of bundles.

**Lemma 1.3.6** *A  $C^0(\Lambda)$ -linear map  $L: \Gamma_c \rightarrow \Gamma_c$  is pointwise defined, i.e., there is a continuous family  $(L_x: T_x M \rightarrow T_x M)_{x \in \Lambda}$  of linear maps such that  $L(X)(x) = L_x(X(x))$  for all  $x \in \Lambda$ .*

Now, Lemma 1.3.5 provides the hypotheses for Lemma 1.3.6 applied to  $\pi^\pm$ , so we obtain fiberwise linear maps  $\pi_x^\pm$ , and these are complementary projections since  $\pi^\pm$  are (check that  $(\pi^\pm)^2 = \pi^\pm$  and  $\pi^- + \pi^+ = \text{Id}$  imply the same for  $\pi_x^\pm$ ). This gives continuous subbundles  $E_x^\lambda := \pi_x^+(T_x M)$  and  $E_x^\mu := \pi_x^-(T_x M)$  with the desired properties.  $\square$

*Proof of Lemma 1.3.5* The main point is that the subspaces  $\mathcal{E}^\lambda$  and  $\mathcal{E}^\mu$  are  $C^0(\Lambda)$ -closed: If  $X \in \mathcal{E}^\lambda$  and  $\varphi: \Lambda \rightarrow \mathbb{R}$  is continuous (hence bounded), then  $\varphi X \in \mathcal{E}^\lambda$  because  $\mathcal{F}^n(\varphi X) = \varphi \circ f^{-n} \cdot \mathcal{F}^n(X)$ . Thus  $\Gamma_c = \mathcal{E}^\lambda \oplus \mathcal{E}^\mu$  as  $C^0(\Lambda)$ -modules; since  $\pi^\pm$  is  $C^0(\Lambda)$ -linear on  $\mathcal{E}^\lambda$  and  $\mathcal{E}^\mu$  (it is 0 or Id), the claim follows.  $\square$

*Proof of Lemma 1.3.6* If  $X \equiv 0$  on an open set  $U$  then  $\pi^\pm(X) = 0$  on  $U$ : For  $x \in U$  take  $\varphi \in C^0(\Lambda)$  such that  $\varphi(x) = 1$  and  $\varphi X \equiv 0$  to get

$$\pi^\pm(X)(x) = 1 \cdot \pi^\pm(X)(x) = \varphi(x) \cdot \pi^\pm(X)(x) = \pi^\pm(\varphi X)(x) = \pi^\pm(0)(x) = 0.$$

If  $X \in \Gamma_c$  and  $X(x) = 0$  take  $X_n \rightarrow X$  with  $X_n = 0$  on  $B(x, 1/n)$  and hence  $\pi^\pm(X)(x) = \lim \pi^\pm(X_n)(x) = 0$ .

If  $(x, v) \in T_\Lambda M$ ,  $X \in \Gamma_c$  and  $X(x) = v$ , then  $\pi_x^\pm(v) := \pi^\pm(X)(v)$  is thus independent of such  $X$ .  $\square$

The following useful simple consequence of the definition does not use (1.1).

**Proposition 1.3.7** *Let  $\Lambda$  be a hyperbolic set for  $f: U \rightarrow M$ . Then  $x \mapsto E_x^\tau$  is continuous for  $\tau = u, s$ , and the dimensions of these subspaces are locally constant.*

*Proof* The inequalities  $\|Df_x^n \xi\| \leq \lambda^n \|\xi\|$  characterize  $E_x^\lambda$ , and by continuity of  $Df^n$  the set of  $(x, \xi)$  on which they hold is closed, so  $\lim_{x \rightarrow x_0} E_x^\lambda \subset E_{x_0}^\lambda$ . Similarly,  $\lim_{x \rightarrow x_0} E_x^\mu \subset E_{x_0}^\mu$ . Then  $\dim E_{x_0}^\mu + \dim E_{x_0}^\lambda = \dim M = \dim E_x^\mu + \dim E_x^\lambda$  implies that neither inclusion is proper, so  $E_{x_0}^\lambda = \lim_{x \rightarrow x_0} E_x^\lambda$  and  $E_{x_0}^\mu = \lim_{x \rightarrow x_0} E_x^\mu$ .  $\square$

**Corollary 1.3.8** *The subspaces  $E_x^\lambda$  and  $E_x^\mu$  are uniformly transverse: there is  $\alpha_0 > 0$  such that for any  $x \in \Lambda$ , the angle between  $\xi \in E_x^\lambda$  and  $\eta \in E_x^\mu$  is at least  $\alpha_0$ .*

*Proof* The angle  $\alpha(x)$  between  $\xi \in E_x^\mu$  and  $\eta \in E_x^\lambda$  is continuous by Proposition 1.3.7 and positive since  $E_x^\mu \cap E_x^\lambda = \{0\}$ , so has a positive minimum.  $\square$

### 1.3.2 Invariant Cones

Verifying the conditions in the definition of hyperbolicity requires finding the two invariant subbundles  $E^\pm$ , and it may in a concrete situation not be entirely clear how to do so. Even if one perturbs a hyperbolic system, it is not clear how to find the subbundles for the perturbation from those of the original system. We now present a criterion that addresses both these issues by framing hyperbolicity in terms of invariant cone fields.

We begin by defining cone fields.

**Definition 1.3.9** If a normed vector bundle  $E$  over a metric space  $\Lambda$  decomposes into  $E^1 \oplus E^2$ , then the *standard horizontal  $\gamma$ -cone* field is defined by

$$H_p^\gamma := \{u + v \in E_p^1 \oplus E_p^2 \mid \|v\| \leq \gamma \|u\|\}.$$

The *standard vertical  $\gamma$ -cone* is

$$V_p^\gamma := \{u + v \in E_p^1 \oplus E_p^2 \mid \|u\| \leq \gamma \|v\|\}.$$

Here, a *cone field* is a map that associates to every point  $p \in \mathbb{R}^n$  a cone  $K_p$  in  $T_p \mathbb{R}^n$ . These cone fields are said to be bounded if there is a constant  $c$  such that

$$\|u + v\|/c \leq \|u\| + \|v\| \leq c\|u + v\|$$

for all  $p \in \Lambda$ ,  $u \in E_p^1$ ,  $v \in E_p^2$ . For a given cone  $K$ , the *dual cone*  $K^*$  is the closure of the complement of  $K$ .

If  $\Lambda$  is an invariant set for a diffeomorphism  $f: M \rightarrow M$ , then  $f$  naturally acts on cone fields on  $E := T_\Lambda M$  by

$$(f_*K)_p := Df_{f^{-1}(p)}(K_{f^{-1}(p)}).$$

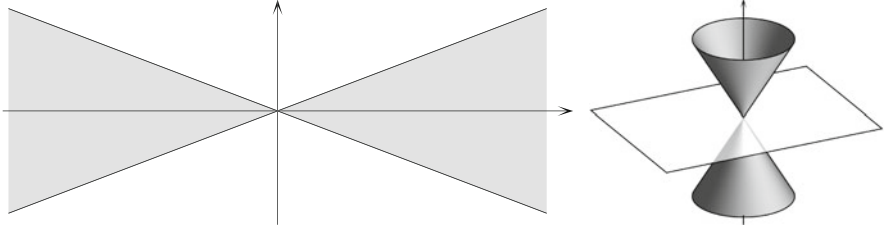
We say that a cone family  $K$  is (strictly) *invariant* if

$$(f_*K)_p \subset \text{Int } K_p \cup \{0\};$$

we write

$$f_*K \Subset K.$$

Let us look at some examples to clarify the picture involved here. In dimension  $n = 2$  a horizontal cone is given by  $|x_2| \leq \gamma|x_1|$ , and its dual cone  $|x_1| \leq |x_2|/\gamma$  is a vertical cone. In dimension  $n = 3$  the following is obviously a cone: Let  $u = x_1$ ,  $v = (x_2, x_3)$ ,  $\sqrt{x_2^2 + x_3^2} \leq \gamma|x_1|$ . So is its dual cone, described by letting  $u =$



**Fig. 1.11** A horizontal cone in  $\mathbb{R}^2$ ; a vertical cone in  $\mathbb{R}^3$  (©Cambridge University Press, reprinted from [KaHa95] with permission)

$(x_2, x_3)$ ,  $v = x_1$  and requiring  $|x_1| \leq \sqrt{x_2^2 + x_3^2}/\gamma$ . This is an example of a cone that does not look like those designed to hold ice cream (Fig. 1.11).

**Theorem 1.3.10 (Aleksyev Cone Field Criterion)** *A compact  $f$ -invariant set  $\Lambda$  is partially hyperbolic (in the broad sense) if and only if there exist  $\lambda < \mu$  such that for every  $x \in \Lambda$  there are*

- a decomposition  $T_x M = S_x \oplus T_x$  (in general, not  $Df$  invariant) and
- a family of horizontal cones  $H_x \supset S_x$  associated with that decomposition

for which

- $\dim S_x = \dim S_{f(x)}$ ,
- $f_* H \subseteq H$ ,
- $\|Df_x v\| \geq \mu \|v\|$  for  $v \in H_x$ , and
- $\|Df_x^{-1} v\| \geq \lambda^{-1} \|v\|$  for  $v \in H_{f(x)}^*$ .

If furthermore  $\lambda < 1 < \mu$ , then  $\Lambda$  is hyperbolic.

*Proof* “Only if” is an easy consequence of the definitions.

Since  $S_x \subset H_x$ ,

$$S_j := Df_{f^{-j}(x)}^j S_{f^{-j}(x)} \subset Df_{f^{-j}(x)}^j H_{f^{-j}(x)} =: H_j.$$

For each  $S_j$  take an ordered orthonormal basis and consider a subsequence such that the sequences of basis elements all converge. Since the intersection of  $H_j$  with the unit sphere is compact it contains the basis consisting of the limits of the basis elements. By the same token any sequence of vectors defined by a fixed set of coefficients converges to a vector in  $H_j$ . Hence the span  $S$  of the limiting basis belongs to all  $H_j$  and thus to the intersection. Indeed,  $S = E_x^\mu$  because we can write  $v \in E_x^\mu$  as  $v = v_S + v_T$  with  $v_S \in S$  and  $v_T \in T_x$  to get

$$\|v_T\| \leq \lambda^n \|Df^{-n}(v_T)\| = \lambda^n \|Df^{-n}(v - v_S)\| \leq \left(\frac{\lambda}{\mu}\right)^n (\|v\| + \|v_S\|) \xrightarrow{n \rightarrow \infty} 0.$$

Likewise one obtains  $E^\lambda$ .

□

While it follows directly from the definitions that every closed invariant subset of a hyperbolic set for  $f$  is also a hyperbolic set, the cone field criterion implies that one can sometimes envelop a given hyperbolic set by a larger one.

**Proposition 1.3.11** *Let  $\Lambda$  be a hyperbolic set for  $f: U \rightarrow M$ . There exists an open neighborhood  $V \supset \Lambda$  such that for any  $g$  sufficiently close to  $f$  in  $C^1$  topology the invariant set*

$$\Lambda_g^V := \bigcap_{n \in \mathbb{Z}} g^n \bar{V}$$

*is hyperbolic.*

**Remark 1.3.12** By construction  $\Lambda \subset \Lambda_f^V$ . Often (e.g., for a hyperbolic periodic orbit or for the “horseshoe”)  $\Lambda_f^V = \Lambda$  if  $V$  is a sufficiently small neighborhood of  $\Lambda$ . For  $g \neq f$  it is not obvious that  $\Lambda_g^V \neq \emptyset$ , and we will prove this soon (Theorem 1.4.6).

*Proof* The inequalities and inclusions in Theorem 1.3.10 persist for continuous extensions of the fields  $S, T, H, V$  from  $\Lambda$  to a neighborhood  $V_1 \supset \Lambda$  and for  $g$  in a  $C^1$ -neighborhood  $U$  of  $f$ . They then hold for  $g$  on  $\Lambda_g^V$ , which is hence a hyperbolic set for  $g$  by Theorem 1.3.10.  $\square$

**Corollary 1.3.13** *The set of Anosov diffeomorphisms is  $C^1$ -open.*

### 1.3.3 Shadowing, Expansiveness, Closing

Hyperbolicity is connected with sensitive dependence on initial conditions and implies a complementary feature called shadowing: behavior that occurs in approximate form does actually occur in the system. This is one of two central underpinnings of the hyperbolic theory (the other being stable and unstable manifolds as described in Sect. 1.6), and it quickly yields the main features of hyperbolic dynamics (by the time we get to page 37): A rich and complex orbit structure with an abundance of periodic orbits, topological and strong structural stability and a symbolic description (Theorems 1.3.15, 1.3.19, 1.3.20, 1.3.33, 1.3.36, 1.3.39, 1.3.45, 1.3.52, 1.4.6, 1.4.11).

Here are formal notions to represent “behavior that occurs in approximate form” and “does actually occur in the system”:

**Definition 1.3.14** Let  $(X, d)$  be a metric space,  $U \subset X$  open and  $f: U \rightarrow X$ . For  $a \in \mathbb{Z} \cup \{-\infty\}$  and  $b \in \mathbb{Z} \cup \{\infty\}$  a sequence  $\{x_n\}_{a < n < b} \subset U$  with  $d(x_{n+1}, f(x_n)) < \epsilon$  for all  $a < n < b$  is called an  $\epsilon$ -orbit or  $\epsilon$ -pseudo-orbit (or just pseudo-orbit) for  $f$ . If  $-\infty < a < b < \infty$ , then it is also referred to as an  $\epsilon$ -chain (or just chain) from  $x_a$  to  $x_b$ . We say that a pseudo-orbit  $(x_n)_{a < n < b}$  is  $\delta$ -shadowed by the orbit  $\mathcal{O}(x)$  of  $x \in U$  if  $d(x_n, f^n(x)) < \delta$  for all  $a < n < b$ .



A point  $x \in X$  is said to be *chain-recurrent* if for every  $\epsilon > 0$  there is an  $\epsilon$ -chain from  $x$  to  $x$ . The set  $\mathcal{R}(f)$  of these points is called the *chain-recurrent set* of  $f$ . A point  $x \in X$  is *nonwandering* with respect to the map  $f: X \rightarrow X$  if for any open set  $U \ni x$  there is an  $N > 0$  such that  $f^N(U) \cap U \neq \emptyset$ . It is said to be *wandering* otherwise. The set of all nonwandering points of  $f$  is denoted by  $NW(f)$ . We say that  $f$  is *regionally recurrent* if  $NW(f) = X$ . (See also Definition 1.7.65.)

We now obtain rather comprehensive information about the topological dynamics of hyperbolic sets from the fact that behavior which occurs in approximate form does actually occur in the system:

**Theorem 1.3.15 (Shadowing Lemma)** *If  $\Lambda$  is a compact hyperbolic set for a diffeomorphism  $f$ , then there is a neighborhood  $U$  of  $\Lambda$  and  $C > 0$  such that any  $\epsilon$ -orbit in  $U$  is  $C\epsilon$ -shadowed by the orbit of some  $x \in \Lambda_V^f$  for a  $C\epsilon$ -neighborhood  $V$  of  $U$ . For sufficiently small  $U$  and  $\epsilon$ ,  $x$  is unique, and  $\Lambda_V^f$  is hyperbolic.*

**Remark 1.3.16** Rufus Bowen observed that much of the topological dynamics of hyperbolic sets arises from this remarkable property. It may be well to note a rather concrete consequence to get a sense of the nature of this statement. If for a given hyperbolic dynamical systems one endeavors to compute a specific orbit numerically, the exponential growth of any errors (due to roundoff and any inaccuracy in representing the dynamical system) ensures that the computed results quickly diverge from the actual orbit to such an extent as to lose any meaningful connection between the computed and the actual orbit. However, the computed orbit is a pseudo-orbit, and therefore the Shadowing Lemma ensures that it reflects, to about the same accuracy as the computation error at every step, *some* actual orbit of the dynamical system (whose initial point is near the intended starting point).<sup>18</sup>

This suggests an exercise. Prove the Shadowing Lemma directly (e.g., by stripping down the proof of the Shadowing Theorem) for the hyperbolic toral automorphism  $F_{\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}}$  in Example 1.1.1 and give, in this case, a plausibly optimal value for the constant  $C$  in the Shadowing Lemma—the point being that this constant is moderate.

Some applications of this property are more easily proved from a slightly strengthened variant, the Shadowing Theorem 1.4.1 below, which asserts the shadowing of entire families of pseudo-orbits in a continuous way. Therefore we do not prove Theorem 1.3.15 here but instead note that it is obtained from the Shadowing Theorem by specializing to shadowing of an individual pseudo-orbit: In Theorem 1.4.1 take  $Y = (\mathbb{Z}, \text{discrete topology})$ ,  $g = f$ ,  $\epsilon_0 = 0$ , and  $\sigma(n) = n + 1$ , i.e., replace  $\alpha \in C^0(Y, V)$  by  $\{x_n\}_{n \in \mathbb{Z}} \subset V$  and “ $\beta \in C^0(Y, V)$  such that  $\beta\sigma = g\beta$ ” by  $\{f^n(x)\}_{n \in \mathbb{Z}} \subset V$ .

We note that the Anosov Shadowing Theorem 1.4.1 is not hard to prove either; the proof fits on one page. We defer that result to focus attention on the fact that

<sup>18</sup>Lest this be taken to be a stronger statement than it actually is, one should note that there is no guarantee that the shadowing orbit is itself in any sense typical. That issue leads to the subject of physical measures and the Sinai–Ruelle–Bowen measure.

the better-known Shadowing *Lemma* suffices for the description of the topological dynamics in the remainder of this section (from here to page 31).

**Corollary 1.3.17** *In this context  $NW(f) \cap \Lambda = \mathcal{B}(f) \cap \Lambda$  (Definition 1.3.14).*

The uniqueness assertion in Theorem 1.3.15 implies 2 more fundamental facts, expansivity and the Anosov Closing Lemma.

**Definition 1.3.18 (Expansivity)** A homeomorphism  $f: X \rightarrow X$  is said to be *expansive* if there exists a constant  $\delta > 0$  such that if  $d(f^n(x), f^n(y)) < \delta$  for all  $n \in \mathbb{Z}$  then  $x = y$ . (For continuous maps replace “ $n \in \mathbb{Z}$ ” by “ $n \in \mathbb{N}$ .”)

**Theorem 1.3.19 (Expansivity)** *The restriction of a diffeomorphism to a hyperbolic set is expansive: if  $\Lambda$  is a compact hyperbolic set for a  $C^1$  diffeomorphism  $f$ , then there are a neighborhood  $U$  of  $\Lambda$  and  $\delta > 0$  such that*

$$\text{if } x, y \in \Lambda_U^f, \text{ then } x = y \text{ or } \sup_{n \in \mathbb{Z}} d(f^n(x), f^n(y)) \geq \delta.$$

**Theorem 1.3.20 (Anosov Closing Lemma)** *Let  $\Lambda$  be a hyperbolic set for  $f: U \rightarrow M$ . Then there are a neighborhood  $V \supset \Lambda$  and  $C, \epsilon_0 > 0$  such that for  $\epsilon < \epsilon_0$  any periodic  $\epsilon$ -orbit  $(x_0, \dots, x_m) \subset V$  is  $C\epsilon$ -shadowed by a point  $y = f^m(y) \in \Lambda_f^U$ . In particular, chain-recurrent points are approximated by periodic ones.*

Here we call a sequence  $x_0, x_1, \dots, x_{m-1}, x_m = x_0$  a *periodic  $\epsilon$ -orbit* or *periodic pseudo-orbit* if  $d(f(x_k), x_{k+1}) < \epsilon$  for  $k = 0, \dots, m-1$ . For almost-closed orbit segments, Corollary 1.6.41 gives additional information.

**Corollary 1.3.21** *Let  $\Lambda$  be a hyperbolic set for  $f: U \rightarrow M$  and  $V$  a neighborhood of  $\Lambda$  such that  $\Lambda_f^V$  is hyperbolic. Then periodic points are dense in  $NW(f|_{\Lambda_f^V})$ .*

*Proof* For  $\epsilon > 0$  sufficiently small denote by  $U_\epsilon$  the  $\epsilon/(2C+1)$ -neighborhood of  $x \in NW(f|_{\Lambda_f^V})$  in  $\Lambda_f^V$ , where  $C$  is as in the Closing Lemma. There exists  $N \in \mathbb{N}$  such that  $f^N(U_\epsilon) \cap U_\epsilon \neq \emptyset$ . If  $y \in f^N(U_\epsilon) \cap U_\epsilon$ , then  $d(f^N(y), y) < 2\epsilon/(2C+1)$ , so the Closing Lemma gives a  $z = f^N(z) \in \Lambda_f^V$  with  $d(f^n(z), f^n(y)) < 2C\epsilon/(2C+1)$  for  $0 \leq n < N$ . Then  $d(x, z) \leq d(x, y) + d(y, z) \leq \frac{(2C+1)\epsilon}{2C+1} = \epsilon$ .  $\square$   
 $V$  and  $\Lambda_f^V$  coincide in our examples (Remark 1.3.12), and this is useful.

**Definition 1.3.22 (Local Maximality, Basic Set)** A hyperbolic set  $\Lambda$  for  $f: U \rightarrow M$  is said to be *locally maximal* or *isolated* if there is a neighborhood  $V$  of  $\Lambda$  (an *isolating neighborhood*) such that  $\Lambda = \Lambda_f^V$ . If furthermore  $f|_{\Lambda}$  has a positive semiorbit that is dense in  $\Lambda$ , then  $\Lambda$  is said to be a *basic set*.

**Remark 1.3.23** If  $\Lambda$  is a basic set, then  $NW(f|_{\Lambda}) = \Lambda$ .

If  $V$  is sufficiently small and  $\Lambda$  is locally maximal then the shadowing orbits in all prior results are in  $\Lambda$ , so  $\Lambda$  has many periodic orbits:

**Corollary 1.3.24** *If  $\Lambda$  is a locally maximal hyperbolic set for  $f: U \rightarrow M$ , then periodic points are dense in  $NW(f|_{\Lambda})$ . In particular, periodic points are dense in basic sets.*

Before pressing on, we take a step back to inventory the dynamical features on which the forthcoming developments are based, or, rather, to disaggregate properties provided by Theorem 1.3.15.

**Definition 1.3.25 (Shadowing Property)** A map  $f: X \rightarrow X$  of a metric space is said to have the *shadowing property* if for all  $\epsilon > 0$  there is a  $\delta > 0$  such that any  $\delta$ -orbit is  $\epsilon$ -shadowed by the orbit of some  $x \in X$ . It is said to have the *Lipschitz-shadowing property* if we can take  $\delta = \epsilon/C$  for some  $C \in \mathbb{R}$ .

*Remark 1.3.26* Note that this notion does not include the uniqueness assertion from Theorem 1.3.15; we noted that this uniqueness assertion implies expansivity, and clearly it also follows from expansivity. (The Lipschitz-shadowing property is not needed here.) Local maximality puts us in this context as follows.

**Theorem 1.3.27** *If  $\Lambda$  is a compact locally maximal hyperbolic set for a diffeomorphism  $f$ , then  $f|_{\Lambda}$  is expansive and has the shadowing property.*

From Theorem 1.3.20 we thus obtain:

**Theorem 1.3.28 (Anosov Closing Lemma)** *For an expansive homeomorphism  $f: X \rightarrow X$  with the shadowing property and any  $\epsilon > 0$  there is a  $\delta > 0$  such that any periodic  $\delta$ -orbit is  $\epsilon$ -shadowed by a periodic point. In particular, periodic points are dense in the chain-recurrent set, which coincides with the nonwandering set. Thus, periodic points are dense in  $X$  if  $f$  is topologically transitive.*

### 1.3.4 Specification

Theorems 1.3.15 and 1.3.20 can be significantly refined: one can prescribe the evolution of an orbit to the extent of specifying a finite collection of arbitrarily long orbit segments and any fixed precision: If one allows for enough time between the specified segments one can find a (periodic) orbit approximating this itinerary. We emphasize that the time between the segments depends only on the quality of the approximation and not on the length of the specified segments. Bowen's Specification Theorem is a useful tool for the study of both the topological structure of hyperbolic sets and statistical properties of orbits within such sets.

**Definition 1.3.29 (Specification)** Let  $f: X \rightarrow X$  be a bijection of a set  $X$ . A *specification*  $S = (\tau, P)$  consists of a finite collection  $\tau = \{I_1, \dots, I_m\}$  of finite intervals  $I_i = [a_i, b_i] \subset \mathbb{Z}$  and a map  $P: T(\tau) := \bigcup_{i=1}^m I_i \rightarrow X$  such that for  $t_1, t_2 \in I \in \tau$  we have  $f^{t_2}(P(t_1)) = f^{t_1}(P(t_2))$ .  $S$  is said to be  *$n$ -spaced* if  $a_{i+1} > b_i + n$  for all  $i \in \{1, \dots, m\}$  and the minimal such  $n$  is called the *spacing* of  $S$ . We say that  $S$  *parameterizes* the collection  $\{P_I \mid I \in \tau\}$  of orbit segments of  $f$ .

We let  $T(S) := T(\tau)$  and  $L(S) := L(\tau) := b_m - a_1$ . If  $(X, d)$  is a metric space we say that  $S$  is  $\epsilon$ -shadowed by  $x \in X$  if  $d(f^n(x), P(n)) < \epsilon$  for all  $n \in T(S)$ .

Thus a specification is a parameterized union of orbit segments  $P \upharpoonright_{I_i}$  of  $f$ .

If  $(X, d)$  is a metric space and  $f: X \rightarrow X$  a homeomorphism then  $f$  is said to have the *weak specification property* if for any  $\epsilon > 0$  there exists an  $M = M_\epsilon \in \mathbb{N}$  such that for any finite collection  $C$  of orbit segments there is an  $M$ -spaced specification  $S$  that parameterizes  $C$  and is  $\epsilon$ -shadowed by some  $x \in X$  as well as by a periodic orbit of period at most  $M + L(S)$ .

If  $(X, d)$  is a metric space and  $f: X \rightarrow X$  a homeomorphism then  $f$  is said to have the *specification property* if for any  $\epsilon > 0$  there exists an  $M = M_\epsilon \in \mathbb{N}$  such that any  $M$ -spaced specification  $S$  is  $\epsilon$ -shadowed by some  $x \in X$  and such that moreover for any  $q \geq M + L(S)$  there is a period- $q$  orbit  $\epsilon$ -shadowing  $S$ .

The difference between the weak and strong specification properties is that in the former we do not have complete freedom in the choice of specification.

**Definition 1.3.30 (Topological Transitivity)** An invertible topological dynamical system  $f: X \rightarrow X$  is said to be *topologically transitive* if there exists a point  $x \in X$  such that its orbit  $\mathcal{O}_f(x) := \{f^n(x)\}_{n \in \mathbb{Z}}$  is dense in  $X$ .

**Proposition 1.3.31 (Topological Transitivity)** If  $X$  is a perfect<sup>19</sup> compact metric space and  $f: X \rightarrow X$  is continuous, then the following are equivalent:

1.  $f$  has a dense positive semiorbit.
2.  $f$  is topologically transitive, i.e., it has a dense orbit.
3. If  $\emptyset \neq U, V \subset X$  are open, then there exists an  $n \in \mathbb{Z}$  such that  $f^n(U) \cap V \neq \emptyset$ .
4. If  $\emptyset \neq U, V \subset X$  are open, then there exists an  $n \in \mathbb{N}$  such that  $f^n(U) \cap V \neq \emptyset$ .

*Remark 1.3.32* Item (3) can be strengthened. Since  $\{B(x, \epsilon/2) \times B(y, \epsilon/2) \mid x, y \in X\}$  has a finite subcover by compactness of  $X$ ,

$$\forall \epsilon > 0 \exists N \in \mathbb{N} \forall x, y \in X \exists n \leq N: f^n(B(x, \epsilon)) \cap B(y, \epsilon) \neq \emptyset. \quad (1.3)$$

**Theorem 1.3.33 (Weak Specification Theorem)** Let  $\Lambda$  be a topologically transitive compact locally maximal hyperbolic set for an embedding  $f$ . Then  $f \upharpoonright_\Lambda$  has the weak specification property. More generally, topologically transitive homeomorphisms with the shadowing property have the weak specification property.

*Proof* Interpolate the specification to a closed pseudo-orbit by orbit segments of length bounded in terms of  $\epsilon$  by (1.3) and apply Theorem 1.3.15.  $\square$

This can be strengthened under the following condition.

**Definition 1.3.34** A continuous map  $f$  of a topological space is *topologically mixing* if for  $V, W$  open there is an  $M \in \mathbb{N}$  such that  $f^m(V) \cap W \neq \emptyset$  when  $m \geq M$ .

---

<sup>19</sup>I.e., without isolated points.

*Remark 1.3.35* Analogously to Remark 1.3.32 this implies a uniform property if  $X$  is compact:  $\forall \epsilon > 0 \exists N \in \mathbb{N} \forall x, y \in X, n \geq N : f^n(B(x, \epsilon)) \cap B(y, \epsilon) \neq \emptyset$ .

Using this in the proof of Theorem 1.3.33 gives

**Theorem 1.3.36 (Bowen Specification Theorem)** *Let  $\Lambda$  be a topologically mixing compact locally maximal hyperbolic set for an embedding  $f$ . Then  $f|_{\Lambda}$  has the specification property. More generally, topologically mixing homeomorphisms with the shadowing property have the specification property.*

### 1.3.5 Spectral Decomposition

It is easy to show that the specification property implies that  $f|_{\Lambda}$  is topologically mixing. Also, Theorem 1.3.33 can be strengthened a little by combining the Spectral Decomposition Theorem 1.3.45 with Theorem 1.3.36; this is Theorem 1.3.50. To this end we introduce the chain decomposition.

**Proposition 1.3.37** *If  $X$  is compact,  $f: X \rightarrow X$  continuous and  $\epsilon > 0$  then the relation defined by  $x \sim y$  if there are  $\epsilon$ -chains from  $x$  to  $y$  and from  $y$  to  $x$  is an equivalence relation on  $\mathcal{R}(f)$ , and each equivalence class is clopen<sup>20</sup> in  $\mathcal{R}(f)$ .*

*Proof* Symmetry of  $\sim$  is obvious, transitivity is easy to check, and reflexivity follows from the definition of  $\mathcal{R}(f)$ . Openness of the equivalence classes then follows from the definition of an  $\epsilon$ -chain. Openness of the complement follows because it is a union of equivalence classes.  $\square$

**Definition 1.3.38 (Chain Components)** If  $f: X \rightarrow X$  continuous, then the equivalence classes of  $\sim$  are called the *chain(-transitive) components* of  $\mathcal{R}(f)$ .  $f$  is said to be *chain transitive* if  $\mathcal{R}(f) = X$  and there is only one chain component. For locally maximal hyperbolic sets Corollary 1.3.17  $\Rightarrow$   $NW(f|_{\Lambda}) = \mathcal{R}(f|_{\Lambda})$ . Together with Theorem 1.3.15 and Proposition 1.3.37 this implies:

**Theorem 1.3.39 (Chain Decomposition)** *Let  $M$  be a Riemannian manifold,  $U \subset M$  open,  $f: U \rightarrow M$  an embedding,  $\Lambda \subset U$  a compact locally maximal hyperbolic set. Then there exist disjoint invariant closed sets  $\Lambda_1, \dots, \Lambda_m$  such that  $NW(f|_{\Lambda}) = \bigcup_{i=1}^m \Lambda_i$  and  $f|_{\Lambda_i}$  is a basic set.*

This can be refined to the spectral decomposition, Theorem 1.3.45. To that end, let  $GCD(f)$  denote the greatest common divisor of the periods of periodic points of  $f: X \rightarrow X$ . We say that  $f$  has *incommensurable periods* if  $GCD(f) = 1$ .

**Theorem 1.3.40** *If  $\Lambda$  is a basic set for an embedding  $f$ ,  $\gamma := GCD(f|_{\Lambda})$ , and  $U, V \subset \Lambda$  are open, then there is an  $N \in \mathbb{N}$  with  $f^{N+k\gamma}(U) \cap V \neq \emptyset$  for all  $k \in \mathbb{N}$ .*

---

<sup>20</sup>Meaning, closed and open.

*Proof* There are finitely many  $x_j = f^{p_j}(x_j) \in \Lambda$  and  $m_j \in \mathbb{Z}$  such that  $\sum_j m_j p_j = \gamma$ . Let  $C, \epsilon$  be as in the Shadowing Lemma (Theorem 1.3.15) and such that  $U$  and  $V$  contain  $C\epsilon$ -balls  $B(x, C\epsilon)$  and  $B(y, C\epsilon)$ , respectively. By chain-transitivity there is an  $\epsilon$ -orbit segment from  $x$  to  $y$  that includes all  $x_j$ . Denoting its length by  $L$ , we find that taking  $N = L + \sum_j |m_j| p_j^2$  and any  $k \in \mathbb{N}$  we can find  $l \in \mathbb{N}$  and  $i < \min\{p_j\}$  with

$$N + k\gamma = L + l \sum_j |m_j| p_j^2 + i \sum m_j p_j = L + \sum_j (l|m_j|p_j + im_j)p_j.$$

By inserting  $l|m_j|p_j + im_j \geq 0$  repeats of the orbits of the  $x_j$ , the Shadowing Lemma gives an orbit of length  $N + k\gamma$  from  $U$  to  $V$ .  $\square$

**Corollary 1.3.41** *If  $\gamma = 1$  in Theorem 1.3.40, then  $f|_{\Lambda}$  is topologically mixing. Here, the contrapositive of the Anosov Closing Lemma (Theorem 1.3.20) implies*

**Proposition 1.3.42** *There is an  $\epsilon > 0$  such that if  $n \not\equiv m \pmod{\text{GCD}(f|_{\Lambda})}$ , then  $d(f^n(x), f^m(x)) \geq \epsilon$  for all  $x \in \Lambda$ .*

In particular,  $\Lambda$  is not mixing if  $\text{GCD}(f) \neq 1$ , i.e.,

**Proposition 1.3.43** *If  $\Lambda$  is a basic set of an embedding  $f$ , then  $f|_{\Lambda}$  is topologically mixing if and only if  $\text{GCD}(f) = 1$  in Theorem 1.3.40.*

A further consequence of Proposition 1.3.42 is

**Corollary 1.3.44** *If  $\Lambda$  is a basic set for an embedding  $f$  and the orbit of  $x \in \Lambda$  is dense, then the  $\gamma := \text{GCD}(f|_{\Lambda})$  sets*

$$\Lambda_i := \overline{\{f^{\gamma k + i}(x) \mid k \in \mathbb{Z}\}}$$

*are pairwise disjoint, invariant and topologically mixing for  $f^\gamma$ .*

Together with Theorem 1.3.39, this implies

**Theorem 1.3.45 (Spectral Decomposition)** *Let  $M$  be a Riemannian manifold,  $U \subset M$  open,  $f: U \rightarrow M$  a diffeomorphism, and  $\Lambda \subset U$  a compact locally maximal hyperbolic set for  $f$ . Then there exist disjoint closed sets  $\Lambda_1, \dots, \Lambda_m$  called homoclinic classes and a permutation  $\sigma$  of  $\{1, \dots, m\}$  such that  $NW(f|_{\Lambda}) = \bigcup_{i=1}^m \Lambda_i$ ,  $f(\Lambda_i) = \Lambda_{\sigma(i)}$ , and when  $\sigma^k(i) = i$  then  $f^k|_{\Lambda_i}$  is topologically mixing.*

The terminology “homoclinic classes” will be explained after Proposition 1.6.48.

**Corollary 1.3.46** *A diffeomorphism  $f$  restricted to a compact locally maximal hyperbolic set is topologically transitive if and only if the permutation  $\sigma$  from Theorem 1.3.45 is cyclic.*

**Corollary 1.3.47** *Let  $\Lambda$  be a connected compact locally maximal hyperbolic set for a diffeomorphism  $f$  such that  $\Lambda = NW(f|_{\Lambda})$  (or equivalently periodic points are dense in  $\Lambda$ ). Then  $f|_{\Lambda}$  is topologically mixing.*

*Proof* The spectral decomposition must be trivial.  $\square$

**Corollary 1.3.48** *A regionally recurrent (Definition 1.3.14) Anosov diffeomorphism  $f: M \rightarrow M$  of a compact connected manifold  $M$  is topologically mixing.*

*Remark 1.3.49* Since the suspension of an Anosov diffeomorphism is an Anosov flow but not topologically mixing, Corollary 1.3.48 fails for flows.

It is not known whether Anosov diffeomorphisms are regionally recurrent. However, Anosov flows need not be.

The spectral decomposition into mixing components (particularly Corollary 1.3.46) plus Theorem 1.3.36 give the following strengthening of Theorem 1.3.33:

**Theorem 1.3.50** *Let  $\Lambda$  be a topologically transitive compact locally maximal hyperbolic set for a diffeomorphism  $f$ . Then there exists  $N \in \mathbb{N}$  such that for  $\epsilon > 0$  every finite collection of  $f$ -orbit segments is parameterized by an  $M$ -spaced specification  $S$  whose spacing depends only on  $\epsilon$  and which is  $\epsilon$ -shadowed by a point of  $\Lambda$  and  $\epsilon$ -shadowed by period- $qN$  orbits for all  $q \geq (M + L(S))/N$ .*

The difference between the conclusion of this result and the specification property is that here we do not have complete freedom in the choice of specification since the periodicity of the permutation of mixing components may only allow transitions at certain times.

### 1.3.6 Stability

Finally, we show that the shadowing property implies a robustness or persistence of the entire orbit structure.

**Definition 1.3.51 (Factor, Stability)** A map  $g: N \rightarrow N$  is a *factor* (or *topological factor*) of  $f: M \rightarrow M$  if there exists a surjective continuous map  $h: M \rightarrow N$  such that  $h \circ f = g \circ h$ . The map  $h$  is called a *factor map*, and a *conjugacy* if  $h$  is a homeomorphism. A  $C^r$  diffeomorphism is said to be *topologically stable* if it is a factor of any homeomorphism sufficiently close to it in the uniform ( $C^0$ ) topology. A  $C^r$  map  $f$  is said to be  *$C^m$  structurally stable* ( $1 \leq m \leq r$ ) if there exists a neighborhood  $U$  of  $f$  in the  $C^m$  topology such that every map  $g \in U$  is topologically conjugate to  $f$ .

**Theorem 1.3.52 (Bowen–Walters)** *Expansive homeomorphisms of compact metric spaces with the shadowing property are topologically stable.*

*Proof* Let  $f: X \rightarrow X$  be an expansive homeomorphism with the shadowing property, and denote by  $E$  an expansivity constant. For  $\epsilon \in (0, E/2)$  and  $\epsilon' < \epsilon$  choose  $\delta < \epsilon'$  such as in the shadowing property. If  $g: X \rightarrow X$  is a homeomorphism with  $d_{C^0}(f, g) < \delta$ . For every  $y \in X$  the  $\delta$ -pseudo-orbit  $n \mapsto g^n(y)$  is  $\epsilon'$ -shadowed by the  $f$ -orbit of a (by expansivity unique) point  $x =: h(y) \in X$ , i.e.,

$$d(f^n(h(y)), g^n(y)) < \epsilon' \text{ for all } n \in \mathbb{Z} \text{ and } y \in X. \quad (1.4)$$

In particular,  $d(h(y), y) < \epsilon'$ , i.e.,  $d_{C^0}(h, \text{Id}) < \epsilon'$ . Furthermore,  $f \circ h = h \circ g$  by expansivity because (1.4) with  $y$  replaced by  $g(y)$  or  $n$  replaced by  $n + 1$  gives

$$d(f^n(h(g(y))), g^n(g(y))) < \epsilon' \quad \text{and} \quad d(f^n(f(h(y))), g^n(g(y))) < \epsilon' \quad \text{for all } n \in \mathbb{Z}.$$

Finally, to show continuity of  $h$  take  $\eta > 0$  and note first that by expansivity there is an  $N \in \mathbb{N}$  such that  $d(f^n(x), f^n(x')) < E$  for  $|n| \leq N \Rightarrow d(x, x') < \eta$ . To apply this with  $x = h(y)$  and  $x' = h(y')$  note that by equicontinuity of  $\{g^n \mid |n| \leq N\}$  there is a  $\gamma > 0$  such that  $d(y, y') < \gamma \Rightarrow d(g^n(y), g^n(y')) < E - 2\epsilon'$  for  $|n| \leq N$  and hence

$$\begin{aligned} d(f^n(h(y)), f^n(h(y'))) &= d(h(g^n(y)), h(g^n(y'))) \\ &\leq d(h(g^n(y)), g^n(y)) + d(g^n(y), g^n(y')) + d(g^n(y'), h(g^n(y'))) \\ &< \epsilon' + (E - 2\epsilon') + \epsilon' = E, \end{aligned}$$

so  $d(h(y), h(y')) < \eta$ , as required.  $\square$

*Remark 1.3.53* We have shown more than stated:  $h$  can be taken close to the identity and is unique when so chosen. Moreover, if  $g$  is assumed expansive with expansivity constant  $E' \geq 2\epsilon'$ , then  $h$  is injective:  $h(x) = h(y) \Rightarrow$

$$\begin{aligned} d(g^n(x), g^n(y)) &\leq \underbrace{d(g^n(x), h(g^n(x)))}_{< \epsilon'} + \underbrace{d(h(g^n(x)), h(g^n(y)))}_{= d(f^n(h(x)), f^n(h(y))) = 0} + \underbrace{d(h(g^n(y)), g^n(y))}_{< \epsilon'} \\ &< 2\epsilon' \leq E' \quad \text{for } n \in \mathbb{Z}, \text{ so } x = y. \end{aligned}$$

This presages structural stability (Theorem 1.4.6), where we obtain a conjugacy rather than just a factor map.

The need to prove continuity at the end of the preceding argument is mirrored by like needs in our next applications, and this motivates strengthening the Shadowing Property to one that produces continuous families of orbits directly, the Shadowing Theorem. We will prove it next, show how much it shortens the proof of the preceding result, and give two further applications that could be derived from the Shadowing Lemma [Wa78], [Bo78, p. 8] but follow more easily from the Shadowing Theorem.

## 1.4 The Shadowing Theorem: Stability, Symbolic Models

### 1.4.1 The Shadowing Theorem

While the preceding section applied the Shadowing *Lemma* and its consequences, we now present the stronger result from which it follows, and then applications of that fact: we complement the result about topological stability by establishing



structural stability, and we provide symbolic models for hyperbolic dynamical systems.

**Theorem 1.4.1 (Anosov Shadowing Theorem [Ka81, p. 57 “Theorem on families of  $\epsilon$ -trajectories”], [KaHa95, Theorem 18.1.3])** *If  $M$  is a Riemannian manifold,  $U \subset M$  open,  $f: U \rightarrow M$  a  $C^1$  embedding, then any compact hyperbolic set  $\Lambda \subset U$  for  $f$  admits a neighborhood  $V$  and  $\epsilon_0, \delta_0, C > 0$  such that if  $g: V \rightarrow M$ ,  $d_{C^1}(f, g) < \epsilon_0$ ,  $Y$  is a topological space,  $\sigma: Y \rightarrow Y$  a homeomorphism,  $\alpha \in C^0(Y, V)$ , and  $d_{C^0}(\alpha\sigma, g\alpha) := \sup_{y \in Y} d(\alpha\sigma(y), g\alpha(y)) < \epsilon < \epsilon_0$ , then there is a  $\beta \in C^0(Y, \Lambda_V^g)$  with  $\beta\sigma = g\beta$  and  $d_{C^0}(\alpha, \beta) < C\epsilon$ .*

*Moreover,  $\beta$  is locally unique: If  $d_{C^0}(\alpha, \bar{\beta}) < \delta_0$  and  $\bar{\beta}\sigma = g\bar{\beta}$ , then  $\bar{\beta} = \beta$ .*

To paraphrase, if  $g|_{\alpha(Y)}$  is  $C^0$ -close to a factor of  $\sigma$  (by  $\alpha$ ), then  $g|_{\beta(Y)}$  is actually a factor of  $\sigma$  by a  $\beta$  that is  $C^0$ -close to  $\alpha$ .

**Remark 1.4.2** As well as the Shadowing Lemma, the Anosov Closing Lemma (Theorem 1.3.20) is a special case: take  $g = f$ ,  $Y = \mathbb{Z}/n\mathbb{Z}$ ,  $\sigma(k) = k + 1 \pmod{n}$ .

*Proof* By the Whitney Embedding Theorem,  $M \subset \mathbb{R}^n$  for suitable  $n$ , so  $M = \mathbb{R}^n$  without loss of generality: If the result is known for  $\mathbb{R}^n$ , embed  $M \hookrightarrow \mathbb{R}^n$ , augment  $U \subset M$  to a tubular neighborhood  $U' \subset \mathbb{R}^n$ , extend  $f$  and a  $C^1$ -close  $g$  to  $U'$  by the same contraction normal to  $M$ , and apply the result. It gives a  $\beta$  consisting of full orbits of the extension of  $g$ , so  $\beta(Y) \subset M$  because  $g$  contracts normally to  $M$  and indeed,  $\beta(Y) \subset V$ , hence  $\beta(Y) \subset \Lambda_V^g$  because  $\beta(Y)$  consists of orbits.

We seek a fixed point of  $F: C^0(Y, V) \rightarrow C^0(Y, \mathbb{R}^n)$ ,  $\beta \mapsto g \circ \beta \circ \sigma^{-1}$ . Represent  $\beta \in C^0(Y, \mathbb{R}^n)$  by the vector field  $v_\beta := \beta - \alpha$  (a section of the bundle  $\{(y, T_{\alpha(y)}\mathbb{R}^n) \mid y \in Y\}$  over  $Y$ ). Then fixed points of  $F$  correspond to fixed points of

$$F^\alpha: v \mapsto g(\alpha \circ \sigma^{-1} + v \circ \sigma^{-1}) - \alpha =: (DF^\alpha|_0 + H)(v)$$

$$\text{or of } T: v \mapsto -((DF^\alpha)_0 - \text{Id})^{-1}H(v).$$

**Lemma 1.4.3** *There are a neighborhood  $V \supset \Lambda$ ,  $\epsilon_0, \epsilon > 0$ , and  $R > 0$  independent of  $Y, g, \alpha$  with  $\|((DF^\alpha)_0 - \text{Id})^{-1}\| < R$  when  $d_{C^1}(f, g) < \epsilon_0$ ,  $d_{C^0}(\alpha\sigma, g\alpha) < \epsilon$ .*

*Proof* For  $\delta > 0$  there are  $\epsilon_0 > 0$ ,  $\mu < 1$  and a neighborhood  $V \supset \Lambda$  to which the splitting  $T_\Lambda M = E^u \oplus E^s$  extends (maybe not invariantly). If  $d_{C^1}(f, g) < \epsilon_0$ , then

$Dg = \begin{pmatrix} a_{uu} & a_{su} \\ a_{us} & a_{ss} \end{pmatrix}$  with respect to  $E^u \oplus E^s$  with  $\|a_{uu}\|^{-1}, \|a_{ss}\| < \mu$ ,  $\|a_{su}\|, \|a_{us}\| < \delta^2\mu$ , With respect to the decomposition into unstable and stable vector fields

$$((DF^\alpha)_0 \xi)(y) = Dg|_{\alpha(\sigma^{-1}(y))} \xi(\sigma^{-1}(y)) \quad \text{splits into} \quad (DF^\alpha)_0 = \begin{pmatrix} A_{uu} & A_{su} \\ A_{us} & A_{ss} \end{pmatrix},$$

where  $d_{C^0}(\alpha, g\alpha\sigma^{-1}) < \epsilon$  and  $d_{C^1}(f, g) < \epsilon_0$  imply

$$\|A_{uu}\|^{-1} < \frac{1 + \mu}{2}, \quad \|A_{su}\| < \delta\mu, \quad \|A_{us}\| < \delta\mu, \quad \|A_{ss}\| < \frac{1 + \mu}{2}.$$

□

To show that  $T$  contracts, we control  $H$ . If  $k_i(t) := H_i(v + th)$  (components with respect to the canonical basis in  $\mathbb{R}^n$ ) then  $k_i(1) - k_i(0) = \int_0^1 k'_i(t) dt$  gives

$$H(v + h) - H(v) = \left( \int_0^1 DH(v + th) dt \right) h = \left( \int_0^1 DF^\alpha|_{v+th} - DF^\alpha|_0 dt \right) h. \quad (1.5)$$

$F^\alpha$  is  $C^1$  since  $g$  is, so there is a  $\delta_0$  such that  $\|DF^\alpha|_{v+th} - DF^\alpha|_0\| \leq \frac{1}{2R}$  for  $\|v\|, \|v + h\| < \delta_0, t \in [0, 1]$ . Thus  $\|v_1\|, \|v_2\| < \delta_0 \Rightarrow \|T(v_1) - T(v_2)\| < \frac{1}{2}\|v_1 - v_2\|$ . With  $\theta = \min(\delta, \delta_0)$  and  $\epsilon < \theta/(2R)$  as in Lemma 1.4.3,  $d_{C^0}(\alpha\sigma, g\alpha) < \epsilon$  gives

$$\|T(0)\| < R\|H(0)\| = R\|g \circ \alpha \circ \sigma^{-1} - \alpha\| = R d_{C^0}(\alpha, g\alpha\sigma^{-1}) = R d_{C^0}(\alpha\sigma, g\alpha) \leq \frac{\theta}{2},$$

so  $\|v\| \leq \delta_0 \Rightarrow \|T(v)\| \leq \|T(0)\| + \|T(v) - T(0)\| < \delta_0/2 + 1/2\|v\| < \delta_0$ , and  $T$  is a  $1/2$ -contraction on the closed ball of vector fields with  $\|v\| \leq \delta_0$ . It has a unique fixed point  $v_\beta$  by Proposition 1.6.3. This yields the desired  $\beta$ . □

*Remark 1.4.4* A slight variation of the argument would be to apply the Hyperbolic Fixed-Point Theorem (Theorem 1.6.5) to  $F^\alpha$  [using Lemma 1.4.3 and (1.5)].

## 1.4.2 Stability

We now turn to Anosov diffeomorphisms to give our first application of the Shadowing Theorem. Here local maximality is automatic. First we note that Theorem 1.3.52 is much easier to prove from the Shadowing Theorem than from the Shadowing Lemma:

**Theorem 1.4.5** *Anosov diffeomorphisms are topologically stable.*

*Proof* An Anosov diffeomorphism  $f: M \rightarrow M$  is a topological factor (via  $h := \beta$ ) of a sufficiently  $C^0$ -close homeomorphism  $g$  by Theorem 1.4.1 with  $\Lambda = U = V = Y = M$ ,  $\sigma = g$  and  $\alpha = \text{Id}$ . Note that  $\beta$  is close to the identity and unique among such maps (i.e.,  $g \mapsto \beta$  is Lipschitz-continuous at  $f$  in the  $C^0$ -topology). □

If we could apply the same reasoning to  $g$  to get a factor map the other way around, we would expect it to be the inverse of the  $h$  in this result, which would then be a homeomorphism. This works if  $g$  is hyperbolic, which requires  $C^1$ -closeness and gives a profound strengthening of Proposition 1.3.11:

**Theorem 1.4.6 (Strong  $C^1$  Structural Stability of Hyperbolic Sets)** *Suppose  $\Lambda$  is a compact hyperbolic set for a  $C^1$  embedding  $f: U \rightarrow M$ . Then there are*

- a  $C^1$ -neighborhood  $U$  of  $f$ ,
- a  $C^0$ -neighborhood  $V$  of the inclusion  $i$  of  $\Lambda$  in  $M$  (viewed as the identity)

- a map  $h: U \rightarrow C(\Lambda, M)$ ,  $g \mapsto h_g$  with  $d_{C^0}(h_g, i) \leq C d_{C^0}(f, g)$ <sup>21</sup>

such that for each  $g \in U$

1.  $h_g$  is a continuous embedding,
2.  $h_g$  is the unique map in  $V$  for which  $g \circ h_g = h_g \circ f|_{\Lambda'}$ ,
3.  $\Lambda_g := h_g(\Lambda)$  is a hyperbolic set for  $g$ .

**Definition 1.4.7** The map  $g \mapsto \Lambda_g$  is called the *continuation* of  $\Lambda$ .

*Proof* We use symmetry and uniqueness by applying the existence part of the Shadowing Theorem 1.4.1 twice and the uniqueness part once.

The Shadowing Theorem with  $0 < \epsilon < \delta_0/2$ ,  $Y = \Lambda$ ,  $\sigma = f$ ,  $\alpha = \text{Id}|_{\Lambda}$ , gives a unique  $h_g := \beta: \Lambda \rightarrow V$   $\delta_0$ -near  $i$  with  $\beta \circ f = g \circ \beta$  which depends Lipschitz-continuously on  $g$ . By Proposition 1.3.11  $\Lambda' := \beta(\Lambda)$  is hyperbolic for  $g$ .

With  $\epsilon$  as before,  $Y = \Lambda'$ ,  $\alpha' = \text{Id}|_{\Lambda'}$ , interchange  $f$  and  $g$  (which we can do if  $\epsilon$  is small enough) to obtain  $\beta'$  with  $\beta' \circ g = f \circ \beta'$  from the Shadowing Theorem.

$\beta: \Lambda \rightarrow \Lambda_g$  is a homeomorphism since  $k := \beta' \circ \beta = \alpha = \text{Id}|_{\Lambda}: k \circ f = f \circ k$ , while

$$d(\alpha, k) = d(\text{Id}, \beta' \circ \beta) \leq d(\text{Id}, \text{Id} \circ \beta) + d(\text{Id} \circ \beta, \beta' \circ \beta) = d(\text{Id}, \beta) + d(\text{Id}, \beta') < \delta_0$$

and  $\alpha \circ f = f \circ \alpha$ , so uniqueness in the Shadowing Theorem implies  $k = \alpha$ .  $\square$

**Corollary 1.4.8** *Anosov diffeomorphisms are structurally stable. The conjugacy is unique when chosen near the identity.*

*Remark 1.4.9* This proof of structural stability rests on the contraction principle because so does that of the Shadowing Theorem. The fixed point of a contraction depends smoothly on the contraction when this is meaningful in a given application, and accordingly, the conjugacy given by structural stability of a hyperbolic  $C^{k+1}$  embedding depends  $C^k$  on the perturbation (in the  $C^0$  topology for conjugacies). We lost one derivative because the composition operator  $\beta \mapsto g \circ \beta \circ \sigma^{-1}$  in the proof of the Shadowing Theorem is  $C^k$  if the maps in question are  $C^{k+1}$ .

### 1.4.3 Markov Models

The final application of the Shadowing Theorem reflects the fact that symbolic descriptions are an effective tool in hyperbolic dynamics. Specifically, we obtain *Markov approximations* (and *Markov partitions* of hyperbolic Cantor sets).

<sup>21</sup>The proof shows that  $g \mapsto (h_g)^{-1}: \Lambda_g \rightarrow \Lambda$  is also Lipschitz-continuous in the  $C^0$ -topology.

**Definition 1.4.10** The standard topology on the space

$$\Omega_N := N^{\mathbb{Z}} := \{0, \dots, N-1\}^{\mathbb{Z}} = \{\omega = (\omega_i)_{i \in \mathbb{Z}} \mid \omega_i \in \{0, \dots, N-1\}\}$$

of two-sided sequences of  $N$  symbols is the product topology arising from the discrete topology on  $\{0, 1, \dots, N-1\}$ , i.e., generated by the *cylinders*

$$C_{\alpha_1, \dots, \alpha_k}^{n_1, \dots, n_k} := \{\omega \in \Omega_N \mid \omega_{n_i} = \alpha_i \text{ for } i = 1, \dots, k\} \quad (1.6)$$

for integers  $n_1 < n_2 < \dots < n_k$  and numbers  $\alpha_1, \dots, \alpha_k \in \{0, 1, \dots, N-1\}$ . ( $k$  is the *rank* of the cylinder.) Thus,  $\Omega_N$  is a Cantor set.

The left shift

$$\sigma_N: \Omega_N \rightarrow \Omega_N, \quad \sigma_N(\omega) = \omega' = (\dots, \omega'_{-1}, \omega'_0, \omega'_1, \dots),$$

where  $\omega'_n = \omega_{n+1}$ , is a one-to-one map and takes cylinders into cylinders. Thus it is a homeomorphism of  $\Omega_N$ .

Let  $A = (a_{ij})_{i,j=0}^{N-1} \in \{0, 1\}^{\{0, \dots, N-1\}^2}$  be an  $N \times N$  matrix with binary entries  $a_{ij}$ . (We call this a 0-1 matrix.) Let

$$\Omega_A := \{\omega \in \Omega_N \mid a_{\omega_n \omega_{n+1}} = 1 \text{ for } n \in \mathbb{Z}\}.$$

The restriction

$$\sigma_N \upharpoonright_{\Omega_A} =: \sigma_A$$

is called the *topological Markov chain* determined by  $A$  or a *subshift of finite type*. *One-sided* shifts  $\sigma_N^R$  and  $\sigma_A^R$  (or again just  $\sigma_N$  and  $\sigma_A$ ) are defined analogously on  $\Omega_N^R := N^{\mathbb{N}} := \{0, \dots, N-1\}^{\mathbb{N}}$  and are not invertible.

**Theorem 1.4.11 (Markov Approximation)** *A compact locally maximal hyperbolic set  $\Lambda$  for a diffeomorphism  $f$  is a factor of a topological Markov chain  $\sigma_A$ . Furthermore for  $\eta > 0$  one can choose  $A$  such that the images of the basic cylinders  $\Omega_A^i := \Omega_A \cap C_i^0$  under the semiconjugacy  $h: \Omega_A \rightarrow M$  have diameter less than  $\eta$ .*

**Remark 1.4.12** One can easily arrange for the symbolic model to represent to any desired accuracy the complexity of the orbit structure in a quantitative sense.<sup>22</sup> The definitive tool of this nature is Markov *partitions*; these provide symbolic representations that represent the dynamical complexity of a hyperbolic dynamical system precisely rather than arbitrarily closely.

*Proof* For  $C, \epsilon > 0$  as in the Shadowing Theorem let  $\mathcal{A} = \{X_0, \dots, X_{N-1}\}$  be an open cover of  $\Lambda$  with  $\text{diam}(X_i) < \epsilon/2$  and  $\text{diam}(f(X_i)) < \epsilon/2$  for all  $i$ . For

<sup>22</sup>That is,  $h_{\text{top}}(\sigma_A) < h_{\text{top}}(f \upharpoonright_{\Lambda}) + \eta$ , see [KaHa95, Theorem 18.2.5].

$i, j \in \{0, \dots, N-1\}$  define  $A_{ij} = 1$  if  $f(X_i) \cap X_j \neq \emptyset$  and  $A_{ij} = 0$  otherwise. Pick  $p_i \in X_i$ . Then  $\alpha: \Omega_A \rightarrow \Lambda$ ,  $\omega \mapsto p_{\omega_0}$  is locally constant, hence continuous.

$d_{C^0}(\alpha\sigma_A, f\alpha) < \epsilon$ : the choice of  $p_i$  and  $X_i$  yields  $x \in f(X_{\omega_0}) \cap X_{\omega_1}$  and hence  $d(\alpha(\sigma_A(\omega)), f(\alpha(\omega))) = d(p_{\omega_1}, f(p_{\omega_0})) \leq d(p_{\omega_1}, x) + d(x, f(p_{\omega_0})) < \epsilon/2 + \epsilon/2$ .

By the Shadowing Theorem the  $\epsilon$ -orbits  $\alpha(\sigma_A^i(\omega)) = p_{\omega_i}$  are  $C\epsilon$ -shadowed by  $\beta(\omega)$  where  $\beta \in C^0(\Omega_A, \Lambda)$  and  $\beta\sigma_A = f\beta$ .

$\beta$  is surjective: If  $x \in \Lambda$  take  $\omega \in \Omega_A$  such that  $f^i(x) \in X_{\omega_i}$ . Then  $x$  and  $\beta(\omega)$  both  $C\epsilon$ -shadow  $(\alpha(\sigma_A^i(\omega)))_{i \in \mathbb{Z}}$  and hence coincide by uniqueness.

Since  $d(\beta, \alpha) < C\epsilon$ , the images of the basic cylinders  $\Omega_A^i = \alpha^{-1}(p_i)$  under the semiconjugacy  $\beta$  have diameter less than  $2C\epsilon$ .  $\square$

*Remark 1.4.13* If  $\beta: \Omega_A \rightarrow \Lambda$  is injective then it is a homeomorphism (invariance of domain), so  $\Lambda$  is a Cantor set. This holds for horseshoes but not for Anosov diffeomorphisms. Indeed, in the context of Example 1.1.4 the preceding argument gives a Markov partition when  $\mathcal{A}$  is chosen to be an open partition. More generally:

**Proposition 1.4.14** *For hyperbolic Cantor sets  $\Lambda$ , the proof of Theorem 1.4.11 using an open partition  $\mathcal{A}$  of  $\Lambda$  gives a Markov partition of  $\Lambda$ .*

## 1.5 Basic Ergodic Theory of Hyperbolic Sets

### 1.5.1 Ergodicity and Related Notions

An  $f$ -invariant probability measure  $\mu$  is said to be ergodic (or  $f$  is said to be *ergodic* with respect to  $\mu$ ) if every  $f$ -invariant measurable set is either a null set or the complement of one (Definition 1.7.36). Equivalently, every bounded measurable  $f$ -invariant function  $\varphi$  is constant a.e. (Proposition 1.7.38):  $\varphi \circ f = \varphi \Rightarrow \varphi \equiv \text{const}$ . If this holds for all iterates  $f^n$ , then  $f$  is said to be *totally ergodic*.

Since the *time-averages* or *Birkhoff averages*  $1/n \sum_{i=0}^{n-1} \varphi \circ f^i$  converge a.e. (Birkhoff Ergodic Theorem 1.7.20) and in  $L^2$  (von Neumann Ergodic Theorem 1.7.33), ergodicity is equivalent to time averages coinciding with space averages ( $\int \varphi$ ), Boltzmann's Fundamental Postulate. The motivation is that such functions  $\varphi$  represent *observables* by associating to each state of the system (each point in the domain of the dynamical system) a number that might be the result of an experimental measurement. We note that in this context we can use all  $L^p$  spaces ( $p \in [1, \infty]$ ) interchangeably. Also, if one takes the probabilistic point of view, then *random variable* is the prevailing term for a real-valued function, since we study probability spaces.

A simple nontrivial example of an ergodic transformation is  $x \mapsto x + \alpha \pmod{1}$  on  $S^1 = \mathbb{R}/\mathbb{Z}$  for irrational  $\alpha$  (Kronecker–Weyl Equidistribution Theorem, Proposition 1.7.68). The preceding examples are also ergodic (with respect to the area measure), but unlike an irrational circle rotation, they have stronger stochastic properties, and we aim to show the mechanisms for this. A colloquial motivation for

these stronger properties is that if  $\varphi$  represents the sugar concentration in a cup with a lump of sugar, then rotation of the cup does little to mix (and dissolve) the sugar.

**Definition 1.5.1** An  $f$ -invariant probability measure is said to be *mixing* (Definition 1.7.118) if two observables become asymptotically independent or uncorrelated when viewed as random variables:

$$\int \varphi \circ f^n \psi \xrightarrow{n \rightarrow \infty} \int \varphi \int \psi \quad \text{for all } \varphi, \psi \in L^2. \quad (1.7)$$

Equivalently (see also Proposition 1.7.140),

$$\varphi \circ f^n \xrightarrow{\text{weakly}} \text{const.} \quad \text{for all } \varphi \in L^2. \quad (1.8)$$

With test function  $\psi \equiv 1$  in (1.7), the left-hand side is independent of  $n$ , so the constant on the right-hand side of (1.8) is  $\int \varphi$ .

**Definition 1.5.2** We say that  $\mu$  is  $N$ -mixing or multiply mixing if (with  $n_0 := 0$ )

$$\prod_{i=1}^N \varphi_i \circ f^{n_i} \xrightarrow[n_i - n_{i-1} \rightarrow \infty]{L^2\text{-weakly}} \prod_{i=1}^N \int \varphi_i d\mu \quad \text{for } \varphi_i \in L^\infty.$$

Made explicit with test function  $\varphi_0$ , this means that  $N + 1$  observables become asymptotically independent. Here, the left-hand side is parametrized by  $\mathbb{Z}^N$ , and the assertion can be checked by considering sequences  $\psi_n = \prod_{i=1}^N \varphi_i \circ f^{n_i(n)}$  with  $n_i(n) - n_{i-1}(n) \xrightarrow{n \rightarrow \infty} \infty$  and  $\psi_n \xrightarrow{\text{weakly}} \psi$ ; then  $\psi$  is an accumulation point, and we describe these as “weak accumulation points  $\psi_n \xrightarrow{\text{weakly}} \psi$  of  $\prod_{i=1}^N \varphi_i \circ f^{n_i}$  with  $n_i - n_{i-1} \xrightarrow{n \rightarrow \infty} \infty$ .”  $N$ -mixing means that for  $\varphi_i \in L^\infty$  there is only one weak accumulation point  $\psi_n \xrightarrow{\text{weakly}} \psi$  of  $\prod_{i=1}^N \varphi_i \circ f^{n_i}$  with  $n_i - n_{i-1} \xrightarrow{n \rightarrow \infty} \infty$ , and it is  $\prod_{i=1}^N \int \varphi_i d\mu$ .

**Proposition 1.5.3** An  $f$ -invariant probability measure  $\mu$  is  $N$ -mixing if and only if given any  $\varphi_i \in L^2(\mu)$ , any weak accumulation point  $\psi_n \xrightarrow[n \rightarrow \infty]{\text{weakly}} \psi$  of  $\prod_{i=1}^N \varphi_i \circ f^{n_i}$  (with  $n_i - n_{i-1} \xrightarrow{n \rightarrow \infty} \infty$ ) is constant.

*Proof* “Only if” is clear. To get “if”, we recursively verify that the constant is correct.

First, take  $\varphi_i \equiv 1$  for  $i \neq 1$ , including taking the test function  $\varphi_0 \equiv 1$ . Then the weak-accumulation statement becomes

$$\int \varphi_1 = \int \varphi_1 \circ f^n \cdot 1 \rightarrow \text{const.} \quad \int 1 = \text{const.},$$

so the constant is  $\int \varphi_1$  for each such subsequence, and thus  $\varphi_1 \circ f^n \xrightarrow{\text{weakly}} \int \varphi_1$ . By symmetry,  $\varphi_i \circ f^n \xrightarrow{\text{weakly}} \int \varphi_i$  for all  $i$ . In particular,  $\varphi_2 \circ f^{n_2-n_1} \xrightarrow[n_2-n_1 \rightarrow \infty]{\text{weakly}} \int \varphi_2$ . Supposing next that  $\varphi_i \equiv 1$  for  $i \notin \{1, 2\}$ , this implies

$$\int \varphi_1 \circ f^{n_1} \varphi_2 \circ f^{n_2} \cdot 1 = \int \varphi_2 \circ f^{n_2-n_1} \varphi_1 \xrightarrow[n_2-n_1 \rightarrow \infty]{} \int \left( \int \varphi_2 \right) \varphi_1 = \int \varphi_1 \int \varphi_2,$$

so  $\varphi_2 \circ f^{n_2} \varphi_1 \circ f^{n_1} \xrightarrow[n_2-n_1 \rightarrow \infty]{\text{weakly}} \int \varphi_1 \int \varphi_2$  with like statements for any pair of the  $\varphi_i$ . This can be continued.  $\square$

Irrational rotations have none of these mixing properties, but the toral automorphisms above do, and the Hopf argument yields them.

Hyperbolic dynamical systems enjoy even stronger stochastic properties, such as the Kolmogorov property and being a Bernoulli system [Ka94, Theorem 3.6], [OrWe98] (Definition 1.7.116). We limit ourselves here to showing how these mixing properties can be established with the Hopf argument.

## 1.5.2 The Hopf Argument

This section presents the Hopf argument for ergodicity, and we will see that it yields mixing. In fairly broad generality it can indeed establish multiple mixing, though that requires additional steps. For example, in the case of volume-preserving Anosov diffeomorphisms, these would be to use the Hopf argument to establish mixing, deduce that the stable partition is ergodic, then apply the one-sided Hopf argument to obtain multiple mixing. While it is useful to keep these steps in mind, we can summarize them in a single theorem (using notations and notions introduced below).

**Theorem 1.5.4** *If  $(X, \mu)$  is a metric Borel probability space,  $\mu$  positive on open sets,  $f: X \rightarrow X$  a continuous invertible  $\mu$ -preserving transformation such that  $W^s$  is absolutely continuous,  $W^s$  and  $W^u$  define a local product structure, and*

$$\varphi \in L^2(\mu) \text{ } f\text{-invariant, } W^s\text{-subordinate and } W^u\text{-subordinate} \Rightarrow \varphi \stackrel{\text{a.e.}}{=} \text{const.}$$

*Then  $f$  is multiply mixing.*

The stable partition of  $f$  is defined by

$$W^s(x) := \{y \in X \mid d(f^n(x), f^n(y)) \xrightarrow[n \rightarrow +\infty]{} 0\}. \quad (1.9)$$

**Definition 1.5.5**  $\varphi: X \rightarrow \mathbb{R}$  is *subordinate* to  $W^s$  or  $W^s$ -*saturated* if there is a set  $G \subset X$  with  $\mu(G) = 1$  such that  $x, y \in G$  and  $y \in W^s(x)$  imply  $\varphi(x) = \varphi(y)$ .

**Theorem 1.5.6** ([Co07, Theorem 2], [Co16, Exercise 4.6]) *If  $X$  is a metric space,  $f: X \rightarrow X$ ,  $\mu$  an  $f$ -invariant Borel probability measure,  $\varphi_i \in L^2(\mu)$ , then any weak accumulation point  $\psi_n \xrightarrow{\text{weakly}} \psi$  of  $\prod_{i=1}^N \varphi_i \circ f^{n_i}$  is  $W^s$ -subordinate.*

Proposition 1.5.3 gives a strong immediate consequence of Theorem 1.5.6:

**Corollary 1.5.7**  *$f$  is multiply mixing if every  $W^s$ -subordinate  $\varphi \in L^2$  is constant a.e.*

*Proof of Theorem 1.5.6* (Coudène) By the Banach–Saks Lemma  $\psi_n \xrightarrow[n \rightarrow \infty]{L^2\text{-weakly}} \psi$  has a subsequence for which  $\frac{1}{n} \sum_{k=0}^{n-1} \psi_{n_k} \xrightarrow[n \rightarrow \infty]{L^2} \psi$ . Note that we gave up a little by passing to a subsequence and to a Birkhoff average rather than a limit but gained  $L^2$ -convergence rather than weak convergence. Furthermore,  $\psi_n \xrightarrow[n \rightarrow \infty]{L^2} \psi$  implies that there is a subsequence with  $\psi_{n_j} \xrightarrow[j \rightarrow \infty]{\text{a.e.}} \psi$ .<sup>23</sup> Again, we give up a little by passing to a subsequence but “upgrade” to pointwise convergence. Thus, we have subsequences  $m_l, n_{ik}$  with

$$\Psi_l := \frac{1}{m_l} \sum_{k=0}^{m_l-1} \psi_{n_{ik}} \xrightarrow[l \rightarrow \infty]{\text{a.e.}} \psi.$$

We passed to pointwise convergence because this is  $W^s$ -subordinate for bounded uniformly continuous  $\varphi_i$ : If  $p_{ij}^l := \varphi_i(f^{(n_i)l}(x_j))$  for  $j = 1, 2$  with  $x_2 \in W^s(x_1)$ , then

$$\prod_{i=1}^N p_{i2}^l - \prod_{i=1}^N p_{i1}^l = \sum_{\ell=1}^N \underbrace{\prod_{i < \ell} p_{i2}^l}_{\text{bounded}} \underbrace{(p_{\ell 2}^l - p_{\ell 1}^l)}_{\xrightarrow[l \rightarrow \infty]{} 0} \underbrace{\prod_{i > \ell} p_{i1}^l}_{\text{bounded}} \xrightarrow[l \rightarrow \infty]{} 0.$$

Finally,  $L^2$ -approximate bounded  $L^2$  functions  $\varphi_i^0$  by bounded uniformly continuous functions  $\varphi_i^k$  within  $1/k$  and this time let  $p_{ij}^l := \varphi_i^k \circ f^{(n_i)l}$  to find that weak limits (of subsequences if necessary) satisfy

$$\begin{aligned} \|\psi - \psi^k\| &\leq \liminf_{l \rightarrow \infty} \left\| \prod_{i=1}^N p_{ik}^l - \prod_{i=1}^N p_{i0}^l \right\| \\ &\leq \sum_{\ell=1}^N \underbrace{\prod_{i < \ell} \|p_{ik}^l\|_\infty}_{\text{bounded}} \underbrace{\|p_{\ell k}^l - p_{\ell 0}^l\|_2}_{\xrightarrow[k \rightarrow \infty]{} 0} \times \underbrace{\prod_{i > \ell} \|p_{i0}^l\|_\infty}_{\text{bounded}} \xrightarrow[k \rightarrow \infty]{} 0 \end{aligned}$$

<sup>23</sup>If  $n_j \geq n_{j-1}$  and  $n > n_j \Rightarrow \|\psi_n - \psi_{n_j}\| < \frac{1}{2^j}$ , then  $\psi_n = \psi_{n_1} + \sum_{i=1}^{j-1} \psi_{n_{i+1}} - \psi_{n_i}$  converges a.e.



so, passing to a subsequence,  $\psi_k \xrightarrow{\text{a.c.}} \psi$ , which is hence  $W^s$ -subordinate.  $\square$   
 If  $f$  is invertible, then we can define

$$W^u(x) := \{y \in X \mid d(f^{-n}(x), f^{-n}(y)) \xrightarrow{n \rightarrow +\infty} 0\},$$

and together with a like conclusion about  $W^u$ -subordination, Theorem 1.5.6 yields mixing. Getting  $W^u$ -subordination requires a slightly subtle argument.

**Theorem 1.5.8 ([Co07, Theorem 3])** *If  $X$  is a metric space,  $f: X \rightarrow X$  invertible,  $\mu$  an  $f$ -invariant Borel probability measure and  $\varphi \in L^2(\mu)$ , then any weak accumulation point of  $U_f^n(\varphi)$  (see Definition 1.7.31) is subordinate to  $W^s$  and to  $W^u$ .*

*Proof (Babillot–Coudène)* If  $\varphi \perp I \subset L^2(\mu)$ , the (closed) subspace of functions subordinate to  $W^u$ , and  $U_f^{n_i} \varphi \xrightarrow{i \rightarrow \infty} \psi$ , then Theorem 1.5.6 applied to  $f^{-1}$  gives a subsequence  $n_{i_k} \rightarrow \infty$  with  $U_f^{-n_{i_k}} \psi \xrightarrow{k \rightarrow \infty} \psi' \in I$ , so  $\langle \psi, \psi \rangle = \lim_{k \rightarrow \infty} \langle U_f^{n_{i_k}} \varphi, \psi \rangle = \lim_{k \rightarrow \infty} \langle \varphi, U_f^{-n_{i_k}} \psi \rangle = \langle \varphi, \psi' \rangle = 0$ , i.e.,  $\psi = 0$ , so  $U_f^n \varphi \xrightarrow{n \rightarrow \infty} 0$ .

For an arbitrary  $\varphi = \varphi_I + \varphi^\perp \in I \oplus I^\perp = L^2$  we then have  $U_f^n \varphi^\perp \xrightarrow{n \rightarrow \infty} 0$ , so the accumulation points of  $U_f^n \varphi$  are accumulation points of  $U_f^n \varphi_I \in I$ .  $\square$

**Corollary 1.5.9** *Suppose  $X$  is a metric space,  $f: X \rightarrow X$  invertible,  $\mu$  an  $f$ -invariant Borel probability measure. If*

$$\varphi \in L^2(\mu) \text{ } f\text{-invariant, } W^s\text{-subordinate and } W^u\text{-subordinate} \Rightarrow \varphi \stackrel{\text{a.c.}}{=} \text{const.},$$

*then  $f$  is mixing.*

### 1.5.3 Mixing from the Hopf Argument

We begin with a “traditional” use of the Hopf argument by applying it to the hyperbolic toral automorphisms of Examples 1.1.1 and 1.1.2 and using both foliations in the process. This reflects the classical use of the Hopf argument to get ergodicity, except that Corollary 1.5.9 yields mixing instead.

**Proposition 1.5.10** *If  $A \in GL(m, \mathbb{Z})$  is hyperbolic, then the induced automorphism  $F_A$  of  $\mathbb{T}^m$  is mixing with respect to Lebesgue measure (cf. Proposition 1.7.95).*

*Proof* For  $q \in \mathbb{T}^m$  the stable subspace  $W^s(q)$  at  $q$  in (1.9) is  $W^s(q) = \pi(E^- + q)$ , where  $E^-$  is the contracting subspace of  $A$  and  $\pi: \mathbb{R}^m \rightarrow \mathbb{T}^m$  is the projection. Likewise,  $W^u(q) = \pi(E^+ + q)$ .

To apply Corollary 1.5.9 consider a  $\varphi \in L^2$  for which there is a set  $G \subset \mathbb{T}^n$  of measure 1 with  $x, y \in G, y \in W^s(x) \Rightarrow \varphi(x) = \varphi(y)$  and  $x, y \in G, y \in W^u(x) \Rightarrow \varphi(x) = \varphi(y)$ . If we can conclude that  $\varphi \stackrel{\text{a.e.}}{=} \text{const.}$ , then Corollary 1.5.9 implies mixing.

Let  $D^\pm \subset E^\pm$  be small disks and  $q \in \mathbb{T}^m$ . Then  $q$  has a neighborhood that is up to rotation and translation of the form  $D^- \times D^+$ , and

$$C := G \cap (D^- \times D^+)$$

has full Lebesgue measure in  $D^- \times D^+$ , i.e., if  $\mu^\pm$  denotes the normalized Lebesgue measure on  $D^\pm$  and  $\mu = \mu^- \times \mu^+$ , then  $\int_{D^- \times D^+} \chi_C d\mu = 1$ . By the Fubini Theorem

$$1 = \int_{D^- \times D^+} \chi_C d\mu = \int_{D^-} \int_{D^+} \chi_C d\mu^+ d\mu^-, \text{ so } \int_{D^+} \chi_C(u, \cdot) d\mu^+ \stackrel{\mu^- \text{ a.e.}}{=} 1.$$

Fix such a  $u_0 \in D^-$ , and note that by construction  $C^- := D^- \times (C \cap (\{u_0\} \times D^+))$  has full Lebesgue measure. If  $(u, v), (u', v') \in C^- \cap C$ , a set of full measure, then

$$\varphi(u, v) = \varphi(u_0, v) = \varphi(u_0, v') = \varphi(u', v').$$

This applies to any such neighborhood of an arbitrary  $q \in \mathbb{T}^n$ , so  $\varphi \stackrel{\text{a.e.}}{=} \text{const.}$   $\square$

This is how Hopf proved ergodicity of geodesic flows of surfaces of negative curvature. The method was extended to geodesic flows of higher-dimensional manifolds by Anosov. The pertinent discrete-time counterpart are Anosov diffeomorphisms, which include the  $F_A$  above. As the preceding argument shows, higher-dimensionality does not directly affect the intrinsic difficulty of the argument. The formidable barrier that Hopf faced and Anosov overcame is related to the use of the Fubini Theorem above—except in Hopf’s context, where local product neighborhoods are indeed diffeomorphic to euclidean patches, one needs to establish the *absolute continuity* of the invariant foliations to apply the Fubini argument [Br02, Chap. 6]. It yields

**Proposition 1.5.11** *Volume-preserving Anosov diffeomorphisms are mixing.*

### 1.5.4 Multiple Mixing from the One-Sided Hopf Argument

The contracting lines in Example 1.1.1 have irrational slope, so each intersects the circle  $S^1 \times \{0\} \subset S^1 \times S^1 = \mathbb{T}^2$  at irrational intervals; the intersections are the orbit of an irrational rotation. Ergodicity of irrational rotations (Proposition 1.7.68) implies that the stable partition  $W^s$  is ergodic, and the “one-sided” Corollary 1.5.7 gives

**Proposition 1.5.12** *The map of  $\mathbb{T}^2$  induced by  $\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$  is multiply mixing.*

*Remark 1.5.13* Simple Fourier analysis also establishes this (see Proposition 1.7.95), but while linearity is helpful for the Hopf argument, it is indispensable for using Fourier analysis. Indeed, the stable partition of volume-preserving Anosov diffeomorphisms is ergodic [An69, Theorem 11], so volume is multiply mixing by Corollary 1.5.7. Likewise, the stable partition for a hyperbolic diffeomorphism is ergodic for the *Margulis measure* of maximal entropy [Ma70, Theorem 2], so this measure is also multiply mixing. These results follow from those in this section.<sup>24</sup> Our final simple application uses that the contracting lines of the partially hyperbolic automorphism in Example 1.1.3 form an ergodic partition because they are generated by a vector whose components are rationally independent. The flow generated by this vector is ergodic analogously to the way an irrational circle rotation is, and its orbits are the stable sets. Therefore, Corollary 1.5.7 applies here as well:

**Proposition 1.5.14** *The automorphism in Example 1.1.3 is multiply mixing.*

The strong conclusion of Corollary 1.5.7 can be obtained in reasonable generality. It turns out that in the original context (of uniformly hyperbolic dynamical systems) in which the Hopf argument applies in the manner shown in Sect. 1.5.3, one can, in fact, apply Corollary 1.5.7. This requires a careful description of the needed properties of  $W^s$  and  $W^u$ .

**Definition 1.5.15** Let  $X$  be a metric space,  $f: X \rightarrow X$  invertible, and  $\mu$  an  $f$ -invariant ergodic Borel probability measure. We say that  $V$  is a *product set* if

- there are  $r_s, r_u > 0$  such that  $x, y \in V \Rightarrow \#(W_{r_s}^s(x) \cap W_{r_u}^u(y)) = 1$ , where

$$W_{r_s}^s(x) := \{y \in W^s(x) \mid d(f^n(x), f^n(y)) \leq r_s \text{ for } r \in \mathbb{N}_0\},$$

$$W_{r_u}^u(x) := \{y \in W^u(x) \mid d(f^{-n}(x), f^{-n}(y)) \leq r_u \text{ for } r \in \mathbb{N}_0\},$$

- $\sup_{x \in V} \text{diam} f^{-n}(W_{r_u}^u(x)) \xrightarrow{n \rightarrow \infty} 0$ , where  $\text{diam}(E) := \sup\{d(x, y) \mid x, y \in E\}$ .

In this case we denote by  $[x, y]$  the unique element of  $W_{r_s}^s(x) \cap W_{r_u}^u(y)$ .

We say that  $W^s$  is *absolutely continuous* on  $V$  (with respect to  $\mu$ ) if for each  $x \in V$  there are measures  $\mu_x^u$  on  $W_{r_u}^u(x)$  and  $\mu_x^s$  on  $W_{r_s}^s(x)$  such that  $\mu_x^u(N) = 0 \Rightarrow \mu_y^u([N, y]) = 0$  and  $\int_V \varphi d\mu = \int_{W_{r_s}^s(z)} \int_{W_{r_u}^u(x)} \varphi d\mu_x^u d\mu_z^s(x)$  for  $\varphi \in L^1(\mu)$ .

**Theorem 1.5.16 ([CoHaTr])** *Let  $X$  be a metric space,  $f: X \rightarrow X$  invertible, and  $\mu$  an  $f$ -invariant ergodic Borel probability measure. If  $W^s$  is absolutely continuous on a product set  $V$  with  $\mu(f^{-1}(V) \cap V) > 0$ , then  $W^s$  is ergodic.*

**Corollary 1.5.17** *Let  $X$  be a metric space,  $\mu$  a Borel probability measure,  $f: X \rightarrow X$   $\mu$ -preserving invertible,  $f^n$  ergodic for all  $n \in \mathbb{N}$ ,  $W^s$  absolutely continuous on a product set  $V$  with  $\mu(V) > 0$ . Then  $f$  is multiply mixing.*

<sup>24</sup>In the case of volume, once one establishes absolute continuity; this is automatic for the Margulis measure.

*Proof* The Poincaré Recurrence Theorem 1.7.11 produces an  $N \in \mathbb{N}$  such that  $\mu(f^{-N}(V) \cap V) > 0$ . Apply Theorem 1.5.16 to  $f^N$ , then Corollary 1.5.7 to  $f$ .  $\square$

*Remark 1.5.18* This applies to volume-preserving Anosov diffeomorphisms [Br02, Chap. 6] but does not use exponential behavior, differentiability or compactness.

**Corollary 1.5.19** *Volume-preserving Anosov diffeomorphisms are multiply mixing.*

**Theorem 1.5.20** *The Liouville measure for dispersing billiards (Example 1.1.5) and for polygonal billiards with pockets (Example 1.1.6) is multiply mixing.*

*Proof* For dispersing billiards, Sinai's Fundamental Theorem of the theory of dispersing billiards [ChMa06, Theorem 5.70] provides product sets [ChMa06, Proposition 7.81] with absolutely continuous holonomies [ChMa06, Theorem 5.42], which implies the absolute continuity property we use. Corollary 1.5.9 then establishes mixing and hence total ergodicity, which by Corollary 1.5.17 implies multiple mixing. This also works for polygonal billiards with pockets [ChTr98, Theorem 4.1].  $\square$

**Theorem 1.5.21** *The Katok map (Example 1.1.7) is multiply mixing.*

*Proof* It is totally ergodic and the stable and unstable partitions are homeomorphic to those of  $F_{\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}}$  (Example 1.1.7), so there is a product neighborhood, which hence has positive measure. Absolute continuity on this neighborhood follows from Pesin theory, so we can apply Corollary 1.5.17.  $\square$

The first step towards proving Theorem 1.5.16 is the following.

**Lemma 1.5.22** *Absolute continuity of  $W^i$  on  $V_f := f^{-1}(V) \cap V$  implies absolute continuity of  $T: V_f \rightarrow X$ ,  $x \mapsto T(x) := [f(x), x]$ , i.e.,  $T_*\mu \ll \mu$ .*

*Proof* If  $N \subset V_f$  and  $\mu(N) = 0$ , then there is a  $W_{\text{loc}}^{su}$ -saturated null set  $N_W$  such that for  $z \notin N_W$  we have  $\int_{W_{\text{loc}}^{su}(z)} \chi_N d\mu_z^{su} = 0$  as well as, by  $f$ -invariance of  $\mu$  and absolute continuity,  $\int_{W_{\text{loc}}^{su}(z)} \chi_{T(N)} d\mu_z^{su} = 0$ . Then

$$\underbrace{(T_*\mu)(N)}_{=\int \chi_{T(N)} d\mu} = \int_{W_{\text{loc}}^i(z) \sim N_W} \underbrace{\int_{W_{\text{loc}}^{su}(x)} \chi_{T(N)} d\mu_x^{su} d\mu_z^i(x)}_{=0} + \underbrace{\int_{N_W} \chi_{T(N)} d\mu}_{=0} = 0.$$

$\square$

Next, we adapt an idea of Thouvenot [Co16, Exercise 4.7]:  $d(f^{-n}(x), f^{-n}(T(x))) \rightarrow 0$  pointwise on  $V_f \cap T^{-1}V_f$ , hence by the Egorov Theorem uniformly on some  $U \subset$

$V_f \cap T^{-1}V_f$  with  $\mu(U) > 0$ . Then  $T_n := \begin{cases} f^{-n} \circ T \circ f^n & \text{on } f^{-n}(U) \\ \text{Id} & \text{elsewhere} \end{cases} \xrightarrow[n \rightarrow \infty]{\text{pointwise}} \text{Id}$ , and

$T_n$  has Radon–Nikodym derivative  $g_n := \left[ \frac{dT_n * \mu}{d\mu} \right] = \left[ \frac{dT_* \mu}{d\mu} \right] \circ f^n$  on  $f^{-n}(U)$  (and

1 elsewhere); this is uniformly integrable, i.e.,  $\sup_{n \in \mathbb{N}} \int_{\{g_n > M\}} g_n d\mu \xrightarrow{M \rightarrow \infty} 0$ .

**Lemma 1.5.23** *Let  $X$  be a metric space with probability measure  $\mu$ ,  $T_n: X \rightarrow X$  such that  $T_n \rightarrow \text{Id}$  a.e.,  $T_{n*}\mu \ll \mu$ , and  $g_n := \left[ \frac{dT_{n*}\mu}{d\mu} \right]$  is uniformly integrable. Then  $\|\varphi \circ T_n - \varphi\|_1 \xrightarrow{n \rightarrow \infty} 0$  for all  $\varphi \in L^\infty$ . ( $\|\cdot\|_p$  denotes the  $L^p$ -norm.)*

*Proof* If  $\psi$  is continuous with  $\|\psi\|_\infty \leq \|\varphi\|_\infty$ , then

$$\|\varphi \circ T_n - \varphi\|_1 \leq \|(\varphi - \psi) \circ T_n\|_1 + \|\psi \circ T_n - \psi\|_1 + \|\psi - \varphi\|_1,$$

and  $\|\psi \circ T_n - \psi\|_1 \rightarrow 0$  by the Bounded Convergence Theorem. Take  $\epsilon > 0$ ,  $M \in \mathbb{R}$  with  $\int_{\{g_n > M\}} g_n d\mu < \epsilon/4\|\varphi\|_\infty$  (uniform integrability),  $\psi$  such that  $\|\psi - \varphi\|_1 < \frac{\epsilon/2}{M+1}$ . Then  $\|(\varphi - \psi) \circ T_n\|_1 = \int |\varphi - \psi|_1 g_n d\mu \leq M\|\psi - \varphi\|_1 + 2\|\varphi\|_\infty \int_{\{g_n > M\}} g_n d\mu$ .  $\square$

*Proof of Theorem 1.5.16* Let  $\varphi \in L^\infty$  be  $W^i$ -saturated. We show  $\varphi$  is  $f$ -invariant. If  $\epsilon' > 0$ , then  $T_n(x) \in W^i(f(x))$  for all  $x \in f^{-n}(U)$  implies that

$$\mu\left(f^{-n}(U) \cap \underbrace{\{|\varphi \circ f - \varphi| > \epsilon'\}}_{=:B}\right) = \mu\left(f^{-n}(U) \cap \{|\varphi \circ T_n - \varphi| > \epsilon'\}\right) \rightarrow 0 \quad (1.10)$$

by Lemma 1.5.23. The Mean Ergodic Theorem and ergodicity of  $f$  imply

$$\frac{1}{n} \sum_{k=0}^{n-1} \chi_U \circ f^k \xrightarrow[n \rightarrow \infty]{L^2} \mu(U), \quad \text{hence} \quad \frac{1}{n} \sum_{k=0}^{n-1} \chi_U \circ f^k \chi_B \xrightarrow[n \rightarrow \infty]{L^2} \mu(U) \chi_B,$$

so  $0 \xleftarrow[n \rightarrow \infty]{(1.10)} \frac{1}{n} \sum_{k=0}^{n-1} \mu\left(f^{-n}(U) \cap \{|\varphi \circ T_n - \varphi| > \epsilon'\}\right) \xrightarrow[n \rightarrow \infty]{} \mu(U) \mu(B)$ , and  $\mu(B) = 0$  because  $\mu(U) > 0$ . Since  $\epsilon$  is arbitrary,  $\varphi$  is  $f$ -invariant, hence constant a.e.  $\square$

## 1.6 Contractions and Invariant Manifolds

### 1.6.1 The Contraction-Mapping Principle

The dynamics of contractions is untypically simple for hyperbolic dynamical systems but useful as a device in auxiliary spaces. Here, a map  $f: X \rightarrow X$  is said to be *contracting* if there exists  $\lambda < 1$  such that for any  $x, y \in X$

$$d(f(x), f(y)) \leq \lambda d(x, y). \quad (1.11)$$

These maps exhibit both stability of equilibria in the sense of ordinary differential equations and in the sense of persistence under perturbation of the dynamical system. All orbits tend to a fixed point, and changing the contracting map slightly does not move the fixed point much. Some pertinent observations are cast in terms of a regularity notion that extends the one of Lipschitz continuity in a natural way.

**Definition 1.6.1 (Lipschitz and Hölder Regularity)** Let  $(X, d)$ ,  $(Y, d)$  be metric spaces. A map  $f: X \rightarrow Y$  is said to be *Lipschitz (continuous)* if there exists  $C > 0$  such that  $d(x, y) < \epsilon$  implies  $d(f(x), f(y)) \leq C(d(x, y))$ , in which case  $f$  is said to be  $C$ -Lipschitz, and the *Lipschitz constant*  $L(f)$  (or  $\text{Lip}(f)$ ) of  $f$  is defined by

$$L(f) := \sup_{x \neq y} \frac{d(f(x), f(y))}{d(x, y)}.$$

We say that  $f$  is *bi-Lipschitz* if it is Lipschitz and has a Lipschitz inverse.

A map  $f: X \rightarrow Y$  is said to be *Hölder-continuous* with exponent  $\alpha$ , or  $\alpha$ -Hölder, if there exist  $C, \epsilon > 0$  such that  $d(x, y) < \epsilon$  implies  $d(f(x), f(y)) \leq C(d(x, y))^\alpha$ . A Hölder-continuous map with Hölder-continuous inverse is said to be bi-Hölder.

*Remark 1.6.2* This notion is both natural and useful in the context of hyperbolic dynamical systems because it corresponds to saying that if  $d(x, y)$  tends to 0 exponentially (as a function of some parameter) then so does  $d(f(x), f(y))$ .

**Proposition 1.6.3 (Contraction-Mapping Principle)** Let  $X$  be a complete metric space and  $f: X \rightarrow X$  a contracting map. Then  $f$  has a unique fixed point  $\varphi$ , and under the action of iterates of  $f$  all points converge exponentially to  $\varphi$ .

Indeed, the error at any step can be estimated in terms of the size of the step:

$$d(x, \varphi) \leq \frac{1}{1 - \lambda} d(x, f(x)). \quad (1.12)$$

Suppose  $X, Y$  are metric spaces,  $X$  complete,  $f: X \times Y \rightarrow X$ ,  $\lambda \in (0, 1)$  such that  $d(f_y(x), f_y(x')) \leq \lambda d(x, x')$  for all  $x, x' \in X, y \in Y$ . Denote the unique fixed point of  $f_y$  by  $\varphi_y$ . Then

1.  $d(\varphi_y, \varphi_{y'}) \leq \frac{1}{1 - \lambda} d(f_{y'}(\varphi_{y'}), f_y(\varphi_{y'}))$ .
2. If  $f$  is continuous then so is  $y \mapsto \varphi_y$ .
3. If  $\alpha \in (0, 1]$  and  $f$  is  $\alpha$ -Hölder-continuous in  $y$ ,<sup>25</sup> then so is  $\varphi$ .
4. If  $X, Y$  are open subsets of Banach spaces and  $f$  is  $C^r$ , then so is  $y \mapsto \varphi_y$ , with derivative

$$(1 - D^Y f|_{(y, \varphi_y)})^{-1} \circ D^X f|_{(y, \varphi_y)},$$

where the superscript denotes the differential in the respective space.

<sup>25</sup>Uniformly in  $x$ , i.e.,  $\exists C \in \mathbb{R}$  such that  $d(f_y(x), f_{y'}(x)) \leq C d(y, y')^\alpha$  for all  $x \in X, y, y' \in Y$ .

5. If  $\lambda \in (0, 1)$  and

$$d(f_y(x), f_{y'}(x')) \leq \lambda \max\{d(x, x'), d(y, y')\}$$

for all  $x, x' \in X, y, y' \in Y$ , then  $d(\varphi_y, \varphi_{y'}) \leq \lambda d(y, y')$ .

*Proof*  $\{f^n(x)\}_{n \in \mathbb{N}}$  is a Cauchy sequence because if  $m \geq n$  then

$$d(f^m(x), f^n(x)) \leq \sum_{k=0}^{m-n-1} \underbrace{d(f^{n+k+1}(x), f^{n+k}(x))}_{\leq \lambda^{n+k} d(f(x), x)} \leq \frac{\lambda^n}{1-\lambda} d(f(x), x) \xrightarrow{n \rightarrow \infty} 0. \quad (1.13)$$

Then  $\varphi := \lim_{n \rightarrow \infty} f^n(x) = \lim_{n \rightarrow \infty} f^{n+1}(x) = \lim_{n \rightarrow \infty} f(f^n(x)) = f(\lim_{n \rightarrow \infty} f^n(x)) = f(\varphi)$  exists since  $X$  is complete. (1.11) implies uniqueness,<sup>26</sup> and letting  $m \rightarrow \infty$  in (1.13) gives

$$d(f^n(x), \varphi) \leq \frac{\lambda^n}{1-\lambda} d(f(x), x).$$

This proves exponential convergence and for  $n = 0$  gives (1.12).

1.: Apply (1.12) with  $x = \varphi_{y'} = f_{y'}(\varphi_{y'})$ .

2. and 3. follow from 1., and 4. from the Implicit-Function Theorem.

To obtain 5., take  $x = \varphi_y = f_y(\varphi_y)$  and  $x' = \varphi_{y'} = f_{y'}(\varphi_{y'})$  in the assumption and note that the maximum on the right-hand side must be  $d(y, y')$ .  $\square$

*Remark 1.6.4* This in particular implies continuous dependence of the fixed point on the contraction when one makes  $C^1$ -perturbations.

The robustness of the asymptotic behavior of contractions in Proposition 1.6.3 has a counterpart for hyperbolic maps, even when they are perturbed so as to be nonlinear.

**Theorem 1.6.5 (Hyperbolic Fixed-Point Theorem)** *If  $A: E \rightarrow E$  is a bounded linear map of a Banach space  $E$  and  $\text{Id} - A$  is invertible, then a continuous map  $F: E \rightarrow E$  has a unique fixed point  $\varphi$  if  $\lambda := L(F - A)\|\text{Id} - A\|^{-1} < 1$ . Furthermore,  $\varphi$  depends continuously on  $F$ , and  $\|\varphi\| \leq \frac{1}{1-\lambda} \|F(0)\|$ .*

*Remark 1.6.6* Boundedness of  $(\text{Id} - A)^{-1}$  follows from the Open-Mapping Theorem.

*Proof*  $\varphi$  is a solution of  $(F - A)(x) = x - A(x) = (\text{Id} - A)x$ , hence a fixed point of the  $\lambda$ -contraction  $(F - A)(\text{Id} - A)^{-1}$ . Apply (1.12) with  $x = 0$ .  $\square$

This is analogous to the persistence of the fixed point of a contraction under perturbations, but a hyperbolic fixed point is harder to find: The fixed point of a contraction is the limit of the forward orbit of any initial condition. Proposition 1.6.25 shows that this fails for hyperbolic maps except with a lucky starting point.

<sup>26</sup> $f(x) = x \Rightarrow y = x$  or  $y \neq f(y)$ .

## 1.6.2 The Spectrum of a Linear Map

If a linear transformation of a finite-dimensional vector space has no eigenvalues on the unit circle, then the space is the direct sum of an expanding subspace (the sum of the generalized eigenspaces for eigenvalues outside of the unit circle) and a contracting subspace (the sum of the generalized eigenspaces for eigenvalues inside of the unit circle). The purpose of this subsection and the next is to prove the same for transformations of Banach spaces (Theorem 1.6.20 below).

While this involves interesting functional analysis that a dynamicist may not otherwise encounter frequently, it is also a good option for the reader to take this conclusion of Theorem 1.6.20 as a definition of hyperbolicity and skip ahead to Sect. 1.6.4.

We now look at a similarly general context that combines contraction and expansion. Here a linear structure helps separate the two, so the natural generality in which this is effective is a Banach space.

It is convenient to consider Banach spaces over the complex numbers. The results we obtain in this context can be used for real Banach spaces  $E$  by passing to the complexification  $E_{\mathbb{C}}$  (i.e., the space  $E \otimes \mathbb{C}$  obtained by allowing complex scalars) and then suitably restricting attention to the real part.

$B(z, r)$  denotes the ball of radius  $r$  around  $z$  in  $\mathbb{C}$ , and  $S(z, r)$  its boundary.

**Definition 1.6.7** Let  $E$  be a Banach space and  $A: E \rightarrow E$  be a bounded linear map, i.e., the norm  $\|A\| := \sup_{\|v\|=1} \|Av\|$  of  $A$  is finite. The *resolvent set*  $R(A)$  of  $A$  is the set of  $\lambda \in \mathbb{C}$  for which  $\lambda \text{Id} - A$  has bounded inverse  $R_A(\lambda)$ , called the *resolvent* of  $A$ . The *spectral radius*  $r(A)$  of  $A$  is defined by  $r(A) := \sup\{|\lambda| \mid \lambda \in \text{sp } A := \mathbb{C} \setminus R(A)\}$ . We call  $\text{sp } A$  the *spectrum* of  $A$ .

- The *point spectrum* consists of the eigenvalues of  $A$  ( $\ker(A - \lambda \text{Id})$  is the corresponding eigenspace).
- The *continuous spectrum* is  $\{\lambda \in \text{sp } A \mid A - \lambda \text{Id} \text{ is injective and } \overline{(A - \lambda \text{Id})(E)} = E\}$ .
- The *residual spectrum* is  $\{\lambda \in \text{sp } A \mid A - \lambda \text{Id} \text{ is injective and } \overline{(A - \lambda \text{Id})(E)} \neq E\}$ .

**Remark 1.6.8** If  $E$  is finite-dimensional, then  $\text{sp } A$  is the set of eigenvalues: these are those  $\lambda$  for which  $A - \lambda \text{Id}$  is not injective, which is also the set of  $\lambda$  for which  $A - \lambda \text{Id}$  is not surjective. Invertibility is the only issue in this context because all linear maps between finite-dimensional spaces are bounded. By the Open Mapping Theorem a bounded linear bijection between Banach spaces has bounded inverse, so

$$\text{sp } A = \{\lambda \in \mathbb{C} \mid \lambda \text{Id} - A \text{ is not injective}\} \cup \{\lambda \in \mathbb{C} \mid \lambda \text{Id} - A \text{ is not surjective}\}.$$

Accordingly, the three items in Definition 1.6.7 are a decomposition of  $\text{sp } A$ .



**Lemma 1.6.9**  $r(A) \leq \|A\|$ :  $|\lambda| > \|A\| \Rightarrow \lambda \notin \text{sp } A, R_A(\lambda) = \sum_{i=0}^{\infty} \frac{A^i}{\lambda^{i+1}}$  (Laurent series).

*Proof*  $(\lambda \text{Id} - A) \sum_{i=0}^{n-1} \frac{A^i}{\lambda^{i+1}} = \sum_{i=0}^{n-1} \frac{A^i}{\lambda^i} - \frac{A^{i+1}}{\lambda^{i+1}} = \text{Id} - \frac{A^n}{\lambda^n} \xrightarrow{n \rightarrow \infty} \text{Id}.$   $\square$

The spectral radius provides an asymptotically sharp bound in a sense made explicit in Proposition 1.6.11. This follows from a widely useful elementary fact.

**Proposition 1.6.10 (Subadditive Sequences)** *If there exist  $k \in \mathbb{N}_0$  and  $L > 0$  such that  $a_{m+n} \leq a_m + a_{n+k} + L$  for all  $m, n \in \mathbb{N}$  then  $\lim_{n \rightarrow \infty} \frac{a_n}{n} \in \mathbb{R} \cup \{-\infty\}$  exists.*

*Proof* Let  $L' := a_{2k} + 2L$ . Then  $a_{m+n} \leq a_m + a_{n+k} + L \leq a_m + a_n + L'$ . In particular,  $\frac{a_n}{n} \leq \frac{a_{n-1} + a_1 + L'}{n} \leq \dots \leq \frac{na_1 + (n-1)L'}{n} \leq a_1 + L'$  for  $n \in \mathbb{N}$ . If  $b > \lim_{n \rightarrow \infty} \frac{a_n}{n}$  and  $\epsilon > 0$  take  $l > 2L'/\epsilon$  such that  $\frac{a_l}{l} < b$ . If  $n = il + r \geq \max(l, 2 \max_{r < l} a_r/\epsilon)$  with  $0 \leq r < l$ , then  $\frac{a_n}{n} \leq \frac{ia_l + a_r + iL'}{n} \leq \frac{a_l}{l} + \frac{a_r}{n} + \frac{L'}{l} < b + \epsilon$ , so  $\overline{\lim}_{n \rightarrow \infty} \frac{a_n}{n} \leq b + \epsilon$ .  $\square$

**Proposition 1.6.11 (Gelfand Spectral Radius Formula)**  $r(A) = \lim_{n \rightarrow \infty} \|A^n\|^{1/n}$ .

*Proof* Since  $a_n := \log \|A^n\|$  is subadditive, the limit exists by Proposition 1.6.10. By Lemma 1.6.9 the domain of convergence of the Laurent series  $\sum_{i=0}^{\infty} A^i/\lambda^{i+1}$  of  $R_A(\cdot)$  is  $\{|\lambda| > r(A)\}$  while by the root test it is  $\{|\lambda| > \lim_{n \rightarrow \infty} \|A^n\|^{1/n}\}$ .  $\square$

**Lemma 1.6.12** *If  $A$  is a bounded linear operator, then  $R(A)$  is the natural domain of analyticity of  $R_A(\cdot)$ . Thus,  $R(A)$  is open, and  $\text{sp } A$  is compact by Lemma 1.6.9.*

*Proof* We show analyticity on  $R(A)$  and that  $d(\lambda, \text{sp } A) \geq \|(R_A(\lambda))^{-1}\|^{-1}$  on  $R(A)$ ; this implies openness and  $\|R_A(\lambda)\| \xrightarrow{d(\lambda, \text{sp } A) \rightarrow 0} \infty$ , hence the claim.

If  $\lambda \in R(A)$  and  $|\mu| < \|(R_A(\lambda))^{-1}\|^{-1}$ , then  $\|\mu R_A(\lambda)\| < 1$ , so  $T(\mu) := \sum_{i=0}^{\infty} \mu^i (R_A(\lambda))^{i+1}$  (Neumann series for the inverse of  $(\lambda - \mu) \text{Id} - A = (\lambda \text{Id} - A) - \mu \text{Id}$ ) converges. Then

$$((\lambda - \mu) \text{Id} - A)T(\mu) = (\lambda \text{Id} - A)T(\mu) - \mu T(\mu) = \sum_{i=0}^{\infty} (\mu R_A(\lambda))^i - (\mu R_A(\lambda))^{i+1} = \text{Id}$$

shows that  $\lambda - \mu \in R(A)$  and  $R_A(\lambda - \mu) = T(\mu)$  is analytic at  $\mu = 0$ .  $\square$

**Remark 1.6.13 (Resolvent Equation)** For  $\mu, \lambda \in R(A)$ , multiplying

$$(\mu \text{Id} - A)(\lambda \text{Id} - A)[R_A(\lambda) - R_A(\mu)] = (\mu \text{Id} - A) - (\lambda \text{Id} - A) = (\mu - \lambda) \text{Id}$$

by  $R_A(\lambda)R_A(\mu)$  gives the *resolvent equation*

$$R_A(\lambda) - R_A(\mu) = (\mu - \lambda)R_A(\lambda)R_A(\mu). \quad (1.14)$$

**Proposition 1.6.14**  $\text{sp} A \neq \emptyset$  unless  $E = \{0\}$ .

*Proof* If  $\text{sp} A = \emptyset$ , then  $R_A$  is entire. It is bounded on  $\overline{B(0, 2\|A\|)}$  by compactness, and  $\|R_A(\lambda)\| \leq \|A\|^{-1}$  for  $|\lambda| > 2\|A\|$  because

$$\|(\lambda \text{Id} - A)v\| = \|\lambda(\text{Id} - \frac{A}{\lambda})v\| \geq 2\|A\| \cdot \frac{1}{2}\|v\|.$$

Being bounded and entire,  $R_A$  is constant by the Liouville Theorem, which implies that  $\text{Id} = 0$ , hence  $E = \{0\}$ .  $\square$

The Liouville Theorem applies to this situation by noticing that for a bounded linear functional  $f \in E^*$ ,  $f \circ R_A$  is an entire bounded scalar function and hence constant.

If  $A$  is diagonal, then clearly  $\|A\| = r(A)$ . The following fact is useful for understanding the dynamics of linear maps even if they cannot be diagonalized.

**Proposition 1.6.15** For every  $\delta > 0$  there exists an equivalent norm on  $E$  with respect to which  $\|A\| < r(A) + \delta$ . This is called an adapted or Lyapunov norm.

*Proof* Take  $n$  such that  $\|A^n\| < (r(A) + \delta)^n$  and  $|v| := \sum_{i=0}^{n-1} \|A^i v\| (r(A) + \delta)^{-i}$ . Then

$$\frac{|Av|}{|v|} = \frac{\sum_{i=1}^n \|A^i v\| (r(A) + \delta)^{1-i}}{\sum_{i=0}^{n-1} \|A^i v\| (r(A) + \delta)^{-i}} = (r(A) + \delta) \underbrace{\left[ 1 + \frac{\|A^n v\| (r(A) + \delta)^{-n} - \|v\|}{\sum_{i=0}^{n-1} \|A^i v\| (r(A) + \delta)^{-i}} \right]}_{< 1}.$$

$\square$

*Remark 1.6.16* One can conclude from this that for any equivalent norm and for every  $\epsilon > 0$  there exists  $C_\epsilon$  such that  $\|A^n v\| \leq C_\epsilon (r(A) + \epsilon)^n \|v\|$  for any  $v \in \mathbb{R}^n$ .

**Corollary 1.6.17** If  $\text{sp}(A) \subset B(0, 1)$ , then there is an equivalent norm on  $E$  such that  $A$  is a contraction with respect to the metric generated by that norm.

*Proof*  $r(A) < 1$  by compactness; apply Proposition 1.6.15 with  $0 < \delta < 1 - r(A)$ .  $\square$   
The concept of exponential convergence does not depend on a particular choice of a norm. Thus Proposition 1.6.3 and Corollary 1.6.17 imply

**Corollary 1.6.18** If  $\text{sp}(A) \subset B(0, 1)$ , then the positive iterates of every point converge exponentially to the origin. If in addition  $A$  is invertible map, then negative iterates of every point go to infinity exponentially.

### 1.6.3 Hyperbolic Linear Maps

Next, we look at fixed point where one sees contraction and expansion.

**Definition 1.6.19** A bounded linear map  $A$  of a Banach space  $E$  is said to be *hyperbolic* if  $\text{sp } A \cap S(0, 1) = \emptyset$ . It is said to be  $(\ell^-, \ell^+)$ -hyperbolic if  $0 < \ell^- < 1 < \ell^+$  and  $\text{sp } A \cap \{z \in \mathbb{C} \mid \ell^- \leq |z| \leq \ell^+\} = \emptyset$ .

**Theorem 1.6.20** If  $E$  is a Banach space,  $A: E \rightarrow E$  continuous linear,  $\gamma := S(0, r) \subset R(A)$ , then there are  $0 < \ell^- < r < \ell^+$  such that  $\{z \in \mathbb{C} \mid \ell^- \leq |z| \leq \ell^+\} \subset R(A)$ ,  $\lambda(A) := r(A^-) < \ell^-$  and  $\mu(A) := 1/r(A^{-1}|_{E^+}) > \ell^+$  (notation as in (2) below), i.e.,  $\|(A^-)^n\| = O(\lambda^n)$  and  $\|(A^+)^{-n}\| = O(\mu^{-n})$ . In particular, if  $A$  is hyperbolic ( $r = 1$ ), then there are  $0 < \ell^- < 1 < \ell^+$  such that  $A$  is  $(\ell^-, \ell^+)$ -hyperbolic.

If  $\gamma \subset \mathbb{C}$  is a smooth curve bounding a topological disk  $D$  and  $\text{sp } A \cap \gamma = \emptyset$ , then there are linear subspaces  $E^-$  and  $E^+$  of  $E$  such that

1.  $E = E^- \oplus E^+$ ,
2.  $AE^- \subset E^-$  (with equality if  $0 \notin \text{sp } A$ ),  $AE^+ = E^+$ ; we write  $A^\pm := A|_{E^\pm}$ ,
3.  $\text{sp } A^- = \text{sp }^- A := \text{sp } A \cap D$ ,  $\text{sp } A^+ = \text{sp }^+ A := \text{sp } A \setminus D$ .

**Remark 1.6.21** We used “big  $O$ ” notation:  $f(n) = O(g(n)) : \Leftrightarrow \frac{f(n)}{g(n)}$  is bounded.

**Remark 1.6.22** If  $\ell^- < 1 < \ell^+$ , then these conditions in turn imply that  $A$  is hyperbolic, so this is a characterization of hyperbolicity.

If  $E^\pm$  are both nontrivial, then the spectrum is contained in 2 annuli. This result readily generalizes to larger numbers of annuli; for instance, if  $0 < r_1 < r_2$  and  $\text{sp } A \cap S(0, r_i) = \emptyset$ , then  $\text{sp } A$  lies in the union of 3 annuli; the corresponding subspaces are  $E_{r_1}^-$ ,  $E_{r_1}^+ \cap E_{r_2}^-$ , and  $E_{r_2}^+$ . Linear maps for which all three subspaces in this decomposition are nontrivial are said to be *partially hyperbolic* if  $r_1 < 1 < r_2$ .

As in Corollary 1.6.17, there is an *adapted norm* (or *Lyapunov norm*) associated with such  $(\ell^-, \ell^+)$ , i.e., a norm  $|\cdot|$  equivalent to the given one and such that

$$\|A^-\| \leq \ell^-, \|(A^+)^{-1}\| \leq 1/\ell^+, \text{ and } |v^- + v^+| = \max(|v^-|, |v^+|) \text{ for } v^\pm \in E^\pm. \quad (1.15)$$

(Take Lyapunov norms  $|\cdot|$  for  $A^\pm$  and  $|v^- + v^+| := \max(|v^-|, |v^+|)$  for  $v^\pm \in E^\pm$ .)

**Definition 1.6.23** If  $\ell^- < 1 < \ell^+$ , then  $E^-$  is called the *contracting* subspace and  $E^+$  the *expanding* subspace.

**Remark 1.6.24** The expanding subspace is not characterized by the fact that vectors in it expand under iterates of the map—all vectors outside the contracting subspace are expanded by a sufficiently large iterate of the map. The characterization of  $E^+$  is given by the description of Remark 1.6.22, namely that preimages contract.

*Proof of Theorem 1.6.20* Compactness of  $\text{sp } A$  implies the first assertions and the existence of a smooth Jordan curve  $\gamma'$  with  $\gamma$  inside it and  $\text{sp } A \setminus D$  outside it.

*Claim*  $\pi^- := \frac{1}{2\pi i} \int_{\gamma} R_A(\lambda) d\lambda = \frac{1}{2\pi i} \int_{\gamma'} R_A(\lambda) d\lambda$  is a projection.

*Proof*  $\frac{1}{2\pi i} \int_c \frac{1}{\mu - \lambda} d\mu = \begin{cases} 1 & \text{if } \lambda \text{ is inside } c \\ 0 & \text{if } \lambda \text{ is outside } c \end{cases}$  for  $c \in \{\gamma, \gamma'\}$ , so

$$\begin{aligned} \pi^- \pi^- &= \frac{1}{2\pi i} \int_{\gamma} R_A(\lambda) d\lambda \cdot \frac{1}{2\pi i} \int_{\gamma'} R_A(\mu) d\mu \\ &= \left( \frac{1}{2\pi i} \right)^2 \int_{\gamma} \int_{\gamma'} \underbrace{R_A(\lambda) R_A(\mu)}_{= \frac{R_A(\lambda) - R_A(\mu)}{\mu - \lambda} \text{ by (1.14)}} d\mu d\lambda \\ &= \left( \frac{1}{2\pi i} \right)^2 \left[ \int_{\gamma} R_A(\lambda) \underbrace{\int_{\gamma'} \frac{1}{\mu - \lambda} d\mu}_{= 2\pi i \text{ since } \lambda \in \gamma \text{ inside } \gamma'} d\lambda - \int_{\gamma'} R_A(\mu) \underbrace{\int_{\gamma} \frac{1}{\mu - \lambda} d\lambda}_{= 0 \text{ since } \mu \in \gamma' \text{ outside } \gamma} d\mu \right] \\ &= \frac{1}{2\pi i} \int_{\gamma} R_A(\lambda) d\lambda = \pi^-. \end{aligned}$$

□

1.  $\pi^+ := \text{Id} - \pi^-$  is then also a projection; take  $E^{\pm} := \pi^{\pm}(E)$ .
2.  $A(E^{\pm}) = A(\pi^{\pm}(E)) = \pi^{\pm}(A(E)) \subset \pi^{\pm}(E) = E^{\pm}$  because  $A$  commutes with  $R_A(\cdot)$  and hence with  $\pi^{\pm}$ .  $AE^+ = E^+$  because below we show that  $0 \notin \text{sp} A^+$ .
3.  $E = E^- \oplus E^+$  and  $A(E^{\pm}) \subset E^{\pm}$  give  $\text{sp} A = \text{sp} A^- \oplus A^+ = \text{sp} A^- \cup \text{sp} A^+$ , so we show  $\text{sp} A^- \subset D$  and  $\text{sp} A^+ \cap D = \emptyset$ .

If  $R_A(\lambda) := \frac{1}{2\pi i} \int_{\gamma} \frac{1}{\lambda - \mu} R_A(\mu) d\mu$  then (1.14) gives

$$R_A(\lambda)(\lambda \text{Id} - A) = \frac{1}{2\pi i} \int_{\gamma} R_A(\mu) - \frac{\text{Id}}{\mu - \lambda} d\mu = \begin{cases} \pi^- & \text{if } \lambda \text{ is outside } \gamma, \\ \pi^- - \text{Id} = -\pi^+ & \text{if } \lambda \text{ is inside } \gamma. \end{cases}$$

If  $\lambda \notin D \cup \gamma$ , then  $R_A(\lambda)(\lambda \text{Id} - A)|_{E^-} = \text{Id}|_{E^-}$ , so  $\lambda \text{Id} - A^-$  is invertible, and  $\lambda \notin \text{sp} A^-$ , hence  $\text{sp} A^- \subset D$ . If  $\lambda \in D$ , restrict to  $E^+$  to get  $\text{sp} A^+ \cap D = \emptyset$ , hence 3. □

We now describe the asymptotic behavior of iterates of a hyperbolic linear map.

**Proposition 1.6.25** *If  $E$  is a Banach space,  $A: E \rightarrow E$  hyperbolic linear, then*

1. *For every  $v \in E^-$ , the positive iterates  $A^n v$  converge to the origin with exponential speed as  $n \rightarrow \infty$  and if  $A$  is invertible then the negative iterates  $A^n v$  go to infinity with exponential speed as  $n \rightarrow -\infty$ .*
2. *For every  $v \in E^+$  the positive iterates of  $v$  go to infinity exponentially and if  $A$  is invertible then the negative iterates converge exponentially to the origin.*

3. For every  $v \in E \setminus (E^- \cup E^+)$  the iterates  $A^n v$  go to infinity exponentially as  $n \rightarrow \infty$  and if  $A$  is invertible also as  $n \rightarrow -\infty$ .

*Proof* This is mainly a restatement of Theorem 1.6.20 and Remark 1.6.22. If  $v \in \mathbb{R}^n \setminus (E^- \cup E^+)$  write  $v = v^- + v^+$  where  $v^- \in E^- \setminus \{0\}$ ,  $v^+ \in E^+ \setminus \{0\}$  to get

$$\|A^n v\| = \|A^n(v^- + v^+)\| \geq \|A^n v^+\| - \|A^n v^-\| \geq \lambda^n c \|v^+\| - \lambda^{-n} c' \|v^-\| \geq \lambda^n c'',$$

for large positive  $n$ , where  $\lambda > 1$  and  $c, c', c'' > 0$  do not depend on  $n$ .

The argument for negative iterates is the same with  $v^+$  and  $v^-$  exchanged.  $\square$

**Remark 1.6.26** With the present notations one can recast Theorem 1.6.5 as follows. Suppose  $A$  is a  $(\lambda, \mu)$ -hyperbolic bounded linear map of a Banach space and  $F: E \rightarrow E$  is such that  $\ell := L(F - A) < \epsilon := \min(1 - \lambda, 1 - \mu^{-1})$  (see Definition 1.6.1). Then  $F$  has a unique fixed point  $\varphi \in E$ , and  $|\varphi| < |F(0)|/(\epsilon - \ell)$ , where  $|\cdot|$  is an adapted norm.  $\varphi$  depends continuously on  $F$ . The advantage of this version is that it is more explicit about the closeness assumption in terms of known parameters. On the other hand, it uses hyperbolicity rather than just  $1 \in R(A)$ .

To prove this, write  $E = E^- \times E^+$ ,  $\pi^\pm: E \rightarrow E^\pm$ ,  $x \mapsto x^\pm$  for the projections,  $F^\pm := \pi^\pm \circ F$  and prove  $\bar{F}(x) := (F^-(x), x^+ + (A^+)^{-1}(x^+ - F^+(x)))$  is a  $(1 + \ell - \epsilon)$ -contraction.

**Remark 1.6.27** The generality of the present context is motivated by its utility when applied in auxiliary spaces. We immediately show one instance of this: Theorem 1.6.5 can be greatly amplified by applying the very same result in a suitable infinite-dimensional space to show that the dynamics of the almost-linear map  $f$  in Theorem 1.6.5 does not only match that of the linear map in that there is a unique fixed point, but that the entire orbit structure of  $f$  is the same as that of  $A$ .

**Theorem 1.6.28** Let  $A$  be a  $(\lambda, \mu)$ -hyperbolic bounded linear map of a Banach space and  $f_1, f_2$  Lipschitz-continuous maps with  $\Delta f_i := f_i - A$  bounded and

$$\ell := \max L(\Delta f_i) < \epsilon := \min(1 - \lambda, 1 - \mu^{-1}, \|A^{-1}\|^{-1}). \quad (1.16)$$

Then there is a unique continuous map  $h = h_{f_1, f_2}: E \rightarrow E$  such that  $f_1 \circ h = h \circ f_2$  and  $\Delta h := h - \text{Id} \in \mathcal{E} := C_b(E, E)$  (bounded continuous maps with the sup norm).

*Proof* The  $f_i$  are invertible:  $f_i(x) = y \Leftrightarrow x = A^{-1}(y - \Delta f_i(x))$ , and the right-hand side is an  $\ell\|A^{-1}\|$ -contraction, so there is a unique such  $x$ .

We can thus rewrite the desired conclusion as  $f_1 \circ h \circ f_2^{-1} = h$  or

$$(A + \Delta f_1) \circ (\text{Id} + \Delta h) \circ f_2^{-1} = \text{Id} + \Delta h \quad \text{or} \\ \mathcal{F}(\Delta h) := \underbrace{A \circ \Delta h \circ f_2^{-1}}_{=: \mathcal{A}(\Delta h) \in \mathcal{E}} + \underbrace{\Delta f_1 \circ (\text{Id} + \Delta h) \circ f_2^{-1} + A \circ f_2^{-1} - \text{Id}}_{=: \Delta \mathcal{F}(\Delta h) \in \mathcal{E}} = \Delta h \in \mathcal{E},$$

a fixed-point problem for  $\mathcal{F} = \mathcal{A} + \Delta \mathcal{F}$ .  $\mathcal{A}$  is hyperbolic:  $\mathcal{E} = \mathcal{E}^- \oplus \mathcal{E}^+$ , where  $\mathcal{E}^\pm := C_b(E, E^\pm) = \mathcal{A}(\mathcal{E}^\pm)$ ,  $\|\mathcal{A}^-\| \leq \lambda$ , and  $\|(\mathcal{A}^+)^{-1}\| \leq 1/\mu$ . Since

$L(\Delta\mathcal{F}) \leq L(\Delta f_1) < \epsilon$ , Theorem 1.6.5 provides the desired unique fixed point  $\Delta h \in \mathcal{E}$ , and  $h := \text{Id} + \Delta h$  is the required continuous map.  $\square$

This does not quite produce what we promised; for the orbit structures of the maps to be the same,  $h$  must be a homeomorphism. This is an easy consequence.

**Corollary 1.6.29 (Hartman–Grobman)** *Let  $A$  be a  $(\lambda, \mu)$ -hyperbolic bounded linear map of a Banach space,  $f: E \rightarrow E$  Lipschitz with  $\Delta f := f - A$  bounded,  $\epsilon$  as in (1.16), and  $\ell := L(\Delta f) < \epsilon$ . Then there is a unique homeomorphism  $h: E \rightarrow E$  depending continuously on  $f$  with  $h - \text{Id}$  bounded and  $h \circ A = f \circ h$ .*

*Proof* We show that the continuous map in Theorem 1.6.28 is a homeomorphism. We have  $f_1 \circ h_{f_1, f_2} = h_{f_1, f_2} \circ f_2$  and (by symmetry)  $f_2 \circ h_{f_2, f_1} = h_{f_2, f_1} \circ f_1$ , hence

$$\begin{aligned} f_2 \circ [h_{f_2, f_1} \circ h_{f_1, f_2}] &= h_{f_2, f_1} \circ f_1 \circ h_{f_1, f_2} = [h_{f_2, f_1} \circ h_{f_1, f_2}] \circ f_2, \\ f_1 \circ [h_{f_1, f_2} \circ h_{f_2, f_1}] &= h_{f_1, f_2} \circ f_2 \circ h_{f_2, f_1} = [h_{f_1, f_2} \circ h_{f_2, f_1}] \circ f_1, \end{aligned}$$

and uniqueness in Theorem 1.6.28 gives  $h_{f_2, f_1} \circ h_{f_1, f_2} = \text{Id} = h_{f_1, f_2} \circ h_{f_2, f_1}$ .  $\square$   
We now describe a localization procedure that connects the global picture in a linear space (such as in Corollary 1.6.29) with local analysis on a manifold.

On a smooth compact manifold  $M$  we can choose a Riemannian metric, and then there is an open set  $B \subset TM$  such that  $0 \in B_x := B \cap T_x M$  and  $\exp_x: B_x \rightarrow M$  is an embedding of  $B_x$  with  $\exp_x(0) = x$ .

**Theorem 1.6.30** *If  $f$  is a  $C^1$ -diffeomorphism of  $M$  with a compact invariant set  $\Lambda$ , take  $\epsilon_0 > 0$  and a  $C^1$ -neighborhood  $U$  of  $f$  such that  $g(\exp_x(v)) \in \exp_{f(x)}(B_{f(x)})$  for  $g \in U$ ,  $x \in \Lambda$ ,  $\|v\| \leq 2\epsilon_0$ . If  $\rho: \mathbb{R} \rightarrow [0, 1]$  is smooth,  $\rho([0, 1]) = \{1\}$ ,  $\rho([2, \infty)) = \{0\}$ , and  $\epsilon < \epsilon_0$  and  $U$  are sufficiently small, then the localization*

$$G_x(v) := D_{x,f}(v) + \rho(\|v\|/\epsilon)(\exp_{f(x)}^{-1} \circ g \circ \exp_x(v) - D_{x,f}(v))$$

of  $g \in U$  by is arbitrarily uniformly  $C^1$ -close to  $D_{x,f}$ .

*Proof* The main observation is that  $\exp_{f(x)}^{-1} \circ g \circ \exp_x$  is  $C^1$ -close to  $D_{x,f}$  when  $\square$

**Remark 1.6.31** The point of this is that the continuous map  $G: T_\Lambda M \rightarrow T_\Lambda M := TM|_\Lambda$  defined by  $G|_{T_x M} = G_x$  fibers over  $f$ , i.e.,  $G_x(T_x M) \subset T_{f(x)} M$ , and satisfies

$$\begin{aligned} G(v) &= D_{x,f}(v) & \text{when } \|v\| \geq 2\epsilon \\ \exp_{f(x)} G(v) &= g(\exp_x(v)) & \text{when } \|v\| \leq \epsilon. \end{aligned}$$

Corollary 1.6.29 immediately translates to the following.

**Theorem 1.6.32 (Hartman–Grobman Theorem)** *Let  $U \subset \mathbb{R}^n$  be open,  $f: U \rightarrow \mathbb{R}^n$  continuously differentiable, and  $O \in U$  a hyperbolic fixed point of  $f$ , i.e.,  $D_O f$  is a hyperbolic linear map. Then there exist neighborhoods  $U_1, U_2, V_1, V_2$  of  $O$  and a homeomorphism  $h: U_1 \cup U_2 \rightarrow V_1 \cup V_2$  such that  $f = h^{-1} \circ Df_O \circ h$  on  $U_1$ , i.e., the*

following diagram commutes:

$$\begin{array}{ccc} U_1 & \xrightarrow{f} & U_2 \\ h \downarrow & & \downarrow h \\ V_1 & \xrightarrow{Df_0} & V_2 \end{array}$$

### 1.6.4 Stable and Unstable Manifolds of a Fixed Point

**Theorem 1.6.33 (Stable Manifold)** *If  $E$  is a Banach space,  $f: E \rightarrow E$ ,  $\eta > 0$ , then*

$$W_\eta^s(f) := \{x \in E \mid \sup_{n \in \mathbb{N}} \eta^{-n} |f^n(x)| < \infty\} \supset f(W_\eta^s(f)). \quad (1.17)$$

Suppose

- $A: E \rightarrow E$  is bounded,
- $\text{sp} A \cap \{z \in \mathbb{C} \mid \lambda \leq |z| \leq \mu\} = \emptyset$ ,
- $|\cdot|$  is an adapted norm (Remark 1.6.22) and
- $\lambda < \eta < \mu$ .

Then  $W_\eta^s(A) = E^-$  (Theorem 1.6.20) and indeed  $\lim_{n \rightarrow \infty} \eta^{-n} |A^n x| = 0$  for all  $x \in E^-$ .

If in this case  $f: E \rightarrow E$  is a Lipschitz-continuous map such that  $f(0) = 0$  and

$$\ell := L(f - A) < \epsilon := \min(\eta - \lambda, \mu - \eta),$$

then  $W_\eta^s(f)$  is the graph of a contraction  $g: E^- \rightarrow E^+$  (Theorem 1.6.20) with  $g(0) = 0$ . Moreover,  $L(f|_{W_\eta^s(f)}) \leq \ell + \lambda < \eta$ , so  $\lim_{n \rightarrow \infty} \eta^{-n} |f^n(x)| = 0$  for all  $x \in W_\eta^s(f)$ .

Finally, if  $\eta < 1$  and  $f$  is  $C^r$  then so is  $g$ .

*Proof* The first two assertions are clear.

$W_\eta^s(f)$  can (when  $\eta \leq 1$ ) be described as the union of bounded forward orbits, i.e., from among all forward orbits it selects the bounded ones. The proof idea is to think instead about selecting from all bounded suitable sequences those that are actually forward orbits. To this end we define a contraction  $\mathcal{F}$  on sequences that depends on a parameter in  $E^-$  and whose fixed points are orbits.

Let  $\pi^\pm: E = E^- \times E^+ \rightarrow E^\pm$ ,  $x \mapsto x^\pm$  (projections),  $A^\pm := A|_{E^\pm}$ ,  $f^\pm := \pi^\pm \circ f$ , and  $\mathcal{F}((x_n)_{n \in \mathbb{N}_0}) = ((y_n^-)_{n \in \mathbb{N}}, (y_n^+)_{n \in \mathbb{N}_0})$ , where

$$\begin{aligned} y_{n+1}^- &= f^-(x_n) \\ y_n^+ &= x_n^+ + (A^+)^{-1}(x_{n+1}^+ - f^+(x_n)). \end{aligned} \quad (1.18)$$

This is in effect Newton's method for finding fixed points, and these are orbits:

$$f(x_n) = x_{n+1} \iff x_{n+1}^- = y_{n+1}^- \text{ and } x_n^+ = y_n^+. \quad (1.19)$$

We next show that  $\mathcal{F}$  is a contraction. If  $x \mapsto y, \tilde{x} \mapsto \tilde{y}$ , then

$$\begin{aligned} |y_{n+1}^- - \tilde{y}_{n+1}^-| &< (\lambda + \ell)|x_n - \tilde{x}_n| \\ |y_n^+ - \tilde{y}_n^+| &\leq (1/\mu)(|x_{n+1} - \tilde{x}_{n+1}| + \ell|x_n - \tilde{x}_n|). \end{aligned} \quad (1.20)$$

If we introduce the Banach spaces

$$\begin{aligned} \mathcal{E}^- &:= \{(x_n^-)_{n \in \mathbb{N}} \mid x_n^- \in E^- \text{ and } \|(x_n^-)_{n \in \mathbb{N}}\|_- := \sup_{n \in \mathbb{N}} \eta^{-n}|x_n^-| < \infty\} \text{ with norm } \|\cdot\|_-, \\ \mathcal{E}^+ &:= \{(x_n^+)_{n \in \mathbb{N}_0} \mid x_n^+ \in E^+ \text{ and } \|(x_n^+)_{n \in \mathbb{N}_0}\|_+ := \sup_{n \in \mathbb{N}_0} \eta^{-n}|x_n^+| < \infty\} \text{ with norm } \|\cdot\|_+, \end{aligned}$$

then (1.18) implies that

$$\mathcal{F}: (E^- \times \mathcal{E}^- \times \mathcal{E}^+, \max\{\|\cdot\|, \|\cdot\|_-, \|\cdot\|_+\}) \rightarrow (\mathcal{E}^- \times \mathcal{E}^+, \max\{\|\cdot\|_-, \|\cdot\|_+\})$$

because  $\mathcal{F}(0) = 0$  and (1.20) give

$$\begin{aligned} \eta^{-n-1}|y_{n+1}^-| &\leq \eta^{-1}(\lambda + \ell)\|x\| \\ \eta^{-n}|y_n^+| &\leq (1/\mu)(\eta + \ell)\|x\|. \end{aligned}$$

By (1.20),  $\mathcal{F}$  is a  $\max\{\eta^{-1}(\lambda + \ell), (1/\mu)(\eta + \ell)\}$ -contraction, so for each  $s \in E^-$ ,

$$\mathcal{F}_s := \mathcal{F}(s, \cdot, \cdot): \mathcal{E}^- \times \mathcal{E}^+ \rightarrow \mathcal{E}^- \times \mathcal{E}^+$$

has a unique fixed point  $p_s = ((p_n^-(s))_{n \in \mathbb{N}}, (p_n^+(s))_{n \in \mathbb{N}_0})$ , and  $s \mapsto p_s$  is a contraction by Proposition 1.6.3. Hence so is  $g := p_0^+: E^- \rightarrow E^+$ , and (1.19) tells us that  $W_\eta^s(f)$  is the graph of  $g: (s, g(s)) \in W_\eta^s(f)$ , and it is the only point  $(s, \cdot)$  for which this is so. Since  $F(0) = 0$  we also have  $p_0 = 0$  and hence  $g(0) = 0$ .

To see that  $L(f|_{W_\eta^s(f)}) \leq \ell + \lambda$ , consider  $x = (s, g(s)), \tilde{x} = (\tilde{s}, g(\tilde{s})) \in W_\eta^s(f)$ . Then  $|f(x) - f(\tilde{x})| = |f^-(x) - f^-(\tilde{x})| \leq (\lambda + \ell)|x - \tilde{x}|$  by (1.20) since  $g$  is a contraction and  $|\cdot|$  takes the maximum of the component norms.

Finally, suppose  $\eta < 1$  and  $f \in C^r$ . We show that

$$F: \mathcal{E} \rightarrow \mathcal{E} := \{x = (x_n)_{n \in \mathbb{N}_0} \mid x_n \in E, \|x\| := \sup_{n \in \mathbb{N}_0} |x_n|/\eta < \infty\},$$

$$(x_n)_{n \in \mathbb{N}_0} \mapsto (f(x_n))_{n \in \mathbb{N}_0}.$$



is  $C^r$ . Then so is  $\mathcal{F}$  and hence  $g$  (Proposition 1.6.3). To see that  $F$  has an  $r$ th-order Taylor expansion reduces to controlling the expansion of  $f$  at  $x_n$  uniformly in  $n$ . For  $x, \Delta x = (\Delta x_n)_{n \in \mathbb{N}_0} \in \mathcal{E}$  and  $n \in \mathbb{N}_0$  denote by  $D_{k,n,t}f$  the  $k$ th total derivative of  $f$  at  $x_n + t\Delta x_n$ . For  $k \leq r$  the Taylor expansion of  $f$  is

$$f(x_n + \Delta x_n) = f(x_n) + \sum_{l=1}^k \frac{1}{l!} D_{l,n,0}f(\underbrace{\Delta x_n, \dots, \Delta x_n}_{l \text{ times}}) + R_k$$

with remainder

$$R_k = \frac{1}{(k-1)!} \int_0^1 (D_{k,n,t}f - D_{k,n,0}f)(\Delta x_n, \dots, \Delta x_n)(1-t)^{k-1} dt.$$

The desired uniform control is the observation that the multilinear map

$$((\Delta x_n^1)_{n \in \mathbb{N}_0}, \dots, (\Delta x_n^l)_{n \in \mathbb{N}_0}) \mapsto (D_{l,n,0}(\Delta x_n^1, \dots, \Delta x_n^l))_{n \in \mathbb{N}_0}$$

maps  $\mathcal{E}^l$  into  $\mathcal{E}$  and has norm at most  $D_l := \sup_{t \in [0,1], n \in \mathbb{N}_0} |D_{l,n,t}f| < \infty$  (since  $\eta < 1$  implies  $\lim_{n \rightarrow \infty} |x_n| = \lim_{n \rightarrow \infty} |\Delta x_n| = 0$ ).  $\square$

*Remark 1.6.34* This proof is attributed to Perron and Irwin. The only other general technique for obtaining this theorem is the Hadamard method [Ha01].

### 1.6.5 Stable and Unstable Foliations

**Theorem 1.6.35 (Stable and Unstable Foliations)** *Let  $\Lambda$  be an invariant set for a  $C^r$  embedding  $f: V \rightarrow M$  (with  $r \geq 1$ ) on which (1.2) holds. Then for each  $x \in \Lambda$  there is an embedded  $C^r$  disk  $W^s(x)$  (resp.  $W^u(x)$ ) called the local stable manifold (resp. local unstable manifold) of  $x$  and depending continuously on  $x$ , such that*

1.  $T_x W^s(x) = E_x^-$  (resp.  $T_x W^u(x) = E_x^+$ );
2.  $f(W^s(x)) \subset W^s(f(x))$  (resp.  $f^{-1}(W^u(x)) \subset W^u(f^{-1}(x))$ );
3. for every  $\delta > 0$  there exists  $C(\delta)$  such that for  $n \in \mathbb{N}$

$$\begin{aligned} d(f^n(x), f^n(y)) &< C(\delta)(\lambda + \delta)^n d(x, y) && \text{for } y \in W^s(x), \\ (\text{resp. } d(f^{-n}(x), f^{-n}(y)) &< C(\delta)(\mu - \delta)^{-n} d(x, y) && \text{for } y \in W^u(x)); \end{aligned}$$

4. there exists  $\beta > 0$  and a family of neighborhoods  $O_x$  containing the ball around  $x \in \Lambda$  of radius  $\beta$  such that

$$W^s(x) = \{y \mid f^n(y) \in O_{f^n(x)} \text{ for } n \in \mathbb{N}\}, \quad W^u(x) = \{y \mid f^{-n}(y) \in O_{f^{-n}(x)} \text{ for } n \in \mathbb{N}\}.$$

Let  $W_\epsilon^s(x) := \{y \mid d(f^n(y), f^n(x)) < \epsilon \text{ for } n \in \mathbb{N}\}$ ,  
 Then complementary  $W_\epsilon^u(x) := \{y \mid d(f^{-n}(y), f^{-n}(x)) < \epsilon \text{ for } n \in \mathbb{N}\}$ .  
 such leaves intersect in exactly one point: Since  $E^+$  and  $E^-$  are continuous on  $\Lambda$  and uniformly transverse (Corollary 1.3.8) the smoothness of  $W^s(x)$  and  $W^u(x)$  and the continuity assertion in Theorem 1.6.35 imply

**Proposition 1.6.36** *There exists an  $\epsilon > 0$  such that for any  $x, y \in \Lambda$  the intersection  $W_\epsilon^s(x) \cap W_\epsilon^u(y)$  consists of at most one point  $[x, y]$ , called the Bowen bracket of  $x$  and  $y$ , and there is a  $\delta > 0$  such that whenever  $d(x, y) < \delta$  for some  $x, y \in \Lambda$  then  $W_\epsilon^s(x) \cap W_\epsilon^u(y) \neq \emptyset$ . Furthermore,  $[\cdot, \cdot]$  is continuous.*

*Proof of Theorem 1.6.35* We reduce this to the Stable Manifold Theorem 1.6.33 for a fixed point via the Localization Theorem 1.6.30; the point is to see how a stable manifold for the action on vector fields gives stable manifolds as asserted here. In particular, we prove the assertions about stable manifolds; those about unstable ones are obtained by considering the inverse.

We assume that the norm is adapted to  $f$ . Take  $\eta \in (\lambda, \min(1, \mu))$  and denote by  $A: \Gamma_b \rightarrow \Gamma_b$  and  $\mathcal{F}: \Gamma_b \rightarrow \Gamma_b$  the actions of  $Df$  and the localization  $F$  of  $f$  (Theorem 1.6.30), respectively, on bounded vector fields. We can ensure that the Lipschitz constant  $\ell := \text{Lip}(\mathcal{F} - A)$  of  $\mathcal{F} - A$  satisfies  $\ell < \epsilon := \min(\eta - \lambda, \mu - \eta)$ . Keeping in mind that the hyperbolic splitting of  $A$  is

$$\Gamma_b = \Gamma_b(E^-) \oplus \Gamma_b(E^+),$$

we can apply the Stable Manifold Theorem 1.6.33 to  $\mathcal{F}$  to conclude that  $W_\eta^s(\mathcal{F})$  [see (1.17)] is the graph of a contraction  $\mathcal{G}: \Gamma_b(E^-) \rightarrow \Gamma_b(E^+)$ . We will show that therefore local stable leaves

$$W_\eta^s(F, x) := W_\eta^s(F) \cap T_x M,$$

where  $W_\eta^s(F) := \{x \in T_\Lambda M \mid \sup_{n \in \mathbb{N}} \eta^{-n} |F^n(x)| < \infty\}$ , are graphs of contractions  $g_x: E_x^- \rightarrow E_x^+$  that depend continuously on  $x$ . The connection is provided by associating to a vector  $v \in T_x M$  the bounded vector field  $\Gamma_v$  given by

$$\Gamma_v(y) := \begin{cases} v & \text{if } y = x \\ 0 & \text{otherwise} \end{cases}$$

because  $\|v\| = \|\Gamma_v\|$  and  $\mathcal{F}^n(\Gamma_v) = \Gamma_{F^n(v)}$  for  $n \in \mathbb{N}$  imply that

$$v \in W_\eta^s(F) \Leftrightarrow \Gamma_v \in W_\eta^s(\mathcal{F}). \quad (1.21)$$

This allows us to define  $g_x(v^-) = v^+ := \mathcal{G}(\Gamma_v^-)(x)$  for  $x \in \Lambda$ ,  $v \in E_x^-$ :

*Claim*  $W_\eta^s(F, x)$  is the graph of  $g_x$ .

*Proof* One one hand,  $v^- + g_x(v^-) \in W_\eta^s(F, x)$  by (1.21) since  $\Gamma_{v^-} + \mathcal{G}(\Gamma_{v^-}) \in W_\eta^s(\mathcal{F})$ . Also, if  $v = v^- + v' \in W_\eta^s(F, x)$  with  $v' \in E_x^+$ , then  $\Gamma_v = \Gamma_{v^-} + \Gamma_{v'} W_\eta^s(\mathcal{F})$ , so

$$v' = \Gamma_{v'}(x) = Gc(\Gamma_{v^-})(x) = g_x(v^-).$$

□

That  $g_x$  is contracting (and  $C^r$ ) follows from the fact that  $\mathcal{G}$  is.

*Claim*  $g_x$  depends continuously on  $x$ .

*Proof* We establish that  $x \mapsto g_x(\Gamma^-(x)) =: \Gamma^+(x)$  is continuous for any continuous stable vector field  $\Gamma^-$ . Applying the Stable Manifold Theorem 1.6.33 on  $\Gamma_c$  rather than  $\Gamma_b$  gives a corresponding map  $\mathcal{G}_c: \Gamma_c(E^-) \rightarrow \Gamma_c(E^+)$  for which

$$\Gamma^- + \mathcal{G}_c(\Gamma^-)(x) \in W_\eta^s(F, x)$$

for all  $x \in \Lambda$ , so necessarily  $\Gamma^+ = \mathcal{G}_c(\Gamma^-) \in \Gamma_c(E^+)$ . □

We produced nonlinear objects for a nonlinear map on  $T_\Lambda M$  by localization, which uses exponential maps. Thus, the desired embeddings of disks are

$$w_x^s: E_x^- \ni B(0, \delta) \rightarrow M, \quad V^- \mapsto \exp_x(v^- + g_x(v^-)).$$

The desired properties follow from the preceding arguments and the Stable Manifold Theorem 1.6.33. □

Global stable and unstable manifolds

$$\widetilde{W}^s(x) = \bigcup_{n=0}^{\infty} f^{-n}(W^s(f^n(x))), \quad \widetilde{W}^u(x) = \bigcup_{n=0}^{\infty} f^n(W^u(f^{-n}(x)))$$

are defined independently of a particular choice of local stable and unstable manifolds and can be characterized topologically:

$$\begin{aligned} \widetilde{W}^s(x) &= \{y \in U \mid d(f^n(x), f^n(y)) \rightarrow 0, \quad n \rightarrow \infty\}, \\ \widetilde{W}^u(x) &= \{y \in U \mid d(f^{-n}(x), f^{-n}(y)) \rightarrow 0, \quad n \rightarrow \infty\}. \end{aligned}$$

*Remark 1.6.37* The literature often denotes by  $W_{\text{loc}}$  the local leaves introduced as  $W$  in Theorem 1.6.35, and then uses  $W$  instead of  $\widetilde{W}$  for the global manifolds. The Shadowing Lemma (Theorem 1.3.15) implies

**Theorem 1.6.38 (In-Phase Theorem)** *If  $\Lambda$  is a compact locally maximal hyperbolic set for  $f: U \rightarrow M$ , then*

$$W^s(\Lambda) := \{y \in U \mid \omega(y) \subset \Lambda\} = \bigcup_{x \in \Lambda} W^s(x),$$

$$W^u(\Lambda) := \{y \in U \mid \alpha(y) \subset \Lambda\} = \bigcup_{x \in \Lambda} W^u(x).$$

*Remark 1.6.39* Here “ $\supset$ ” is obvious from the definition, and “ $\subset$ ” says that a point asymptotic to  $\Lambda$  approaches  $\Lambda$  in a way that is “in phase” with an orbit of  $\Lambda$ .

*Proof* If  $y \in W^s(\Lambda)$  and  $\eta > 0$ , then there is an  $N \in \mathbb{N}$  such that for all  $n \geq N$  we have an  $x_i \in \Lambda$  with  $d(f^i(y), x_i) < \eta$ . If  $\epsilon > 0$  and  $\delta$  is as in the Shadowing Lemma (Theorem 1.3.15), then by uniform continuity of  $f$  we can choose  $\eta$  such that

$$d(f(x_i), x_{i+1}) \leq d(f(x_i), f(f^i(y))) + d(f^{i+1}(y), x_{i+1}) < \delta,$$

and  $(x_i)_{i \geq N}$  is  $\epsilon$ -shadowed by some  $x \in \Lambda$ . For  $i \geq N$  we then have

$$d(f^i(y), f^i(x)) \leq d(f^i(y), x_i) + d(x_i, f^i(x)) \leq \delta + \epsilon,$$

so  $y \in W^s(x)$ . □

## 1.6.6 Applications: Livschitz Theory and Local Product Structure

The presence of stable and unstable manifolds (and our methods for obtaining them) is not only essential for the ergodic theory of hyperbolic sets but also provides the basis for a thorough understanding of the global topological dynamics of hyperbolic sets far beyond the applications of the Shadowing Theorem 1.4.1. We only give a few basic instances of this and in particular omit an important one to which we referred in the context of Markov approximations, the existence of Markov partitions, which establishes an essentially exact correspondence with a symbolic system.

### 1.6.6.1 Exponential Closing

The first among these refinements pertains to the shadowing and closing of orbits. This is the contrapositive of a general property of exponential instability of orbits on and near a hyperbolic set.

**Proposition 1.6.40** *Let  $\Lambda$  be a hyperbolic set for  $f: U \rightarrow M$  and  $\lambda, \mu$  as in Definition 1.3.1. Then for any  $\eta \geq \max(\lambda, \mu^{-1})$  there exist  $\delta > 0$  and  $C > 0$  such that if  $x \in \Lambda$ ,  $y \in U$  and  $d(f^k(y), f^k(x)) < \delta$  for  $k = 0, \dots, n$  then in fact*

$$\begin{aligned} d(f^k(y), f^k(x)) &< C'(\eta^k d(x, y) + \eta^{n-k} d(f^n(x), f^n(y))) \\ &\leq C \eta^{\min(k, n-k)} \cdot (d(x, y) + d(f^n(x), f^n(y))). \end{aligned}$$

*Proof*  $d(f^k(x), f^k(y)) \leq d(f^k(x), f^k([x, y])) + d(f^k([x, y]), f^k(y))$ .  $\square$

The Anosov Closing Lemma implies that near any point in a hyperbolic set whose orbit nearly returns to the point there is a periodic orbit that closely follows the almost-returning segment. This can now be considerably strengthened.

**Corollary 1.6.41** *Let  $\Lambda$  be a hyperbolic set for  $f: U \rightarrow M$  and  $\lambda, \mu$  as in Definition 1.3.1. Then for any  $\eta > \max(\lambda, \mu^{-1})$  there exists a neighborhood  $U \supset \Lambda$ , and  $C_1, \epsilon_0 > 0$  such that if  $f^k(x) \in U$  for  $k = 0, \dots, n$  and  $d(f^k(x), x) < \epsilon_0$  then there exists a periodic point  $y$  such that  $f^n(y) = y$  and*

$$d(f^k(y), f^k(x)) < C_1 \eta^{\min(k, n-k)} d(f^n(x), x).$$

*Proof* Apply Theorem 1.3.20 first to obtain the periodic point  $y$ . By Proposition 1.3.11 one can assume that  $y \in \Lambda$ . Then Proposition 1.6.40 gives the statement.  $\square$

This refinement is crucial in some applications because it implies that the distance or “error” between the orbit segments is summable uniformly in the length of the segments (because a geometric series provides an upper bound). To illustrate this, we next give an important such application.

### 1.6.6.2 The Livschitz Theorem

The Anosov Closing Lemma (Theorem 1.3.20) implies density of periodic orbits, and we now use Proposition 1.6.40 to show that they determine global information in nontrivial ways. One application of this is a sharp criterion for volume-preservation.

**Theorem 1.6.42 (Livschitz Theorem)** *Let  $M$  be a Riemannian manifold,  $U \subset M$  open,  $f: U \rightarrow M$  a smooth embedding,  $\Lambda \subset U$  a topologically transitive compact locally maximal hyperbolic set, and  $\varphi: \Lambda \rightarrow \mathbb{R}$   $\alpha$ -Hölder-continuous with  $\sum_{i=0}^{n-1} \varphi(f^i(x)) = 0$  if  $f^n(x) = x \in \Lambda$ . Then there is a continuous  $\Phi: \Lambda \rightarrow \mathbb{R}$  that solves the cohomological equation  $\varphi = \Phi \circ f - \Phi$ . Moreover  $\Phi$  is unique up to an additive constant and  $\alpha$ -Hölder-continuous.*

*Proof* Since  $f|_{\Lambda}$  is topologically transitive there exists a point  $x_0 \in \Lambda$  such that the orbit  $\mathcal{O}(x_0) = \{f^n(x_0)\}_{n \in \mathbb{Z}}$  is dense in  $\Lambda$ . Once we choose a value  $\Phi(x_0) \in \mathbb{R}$  we

must have  $\Phi(f^n(x_0)) = \Phi(x_0) + \varphi(n, x_0)$ , where  $\varphi(n, x)$  is given by

$$\varphi(n, x) = \sum_{i=0}^{n-1} \varphi(f^i(x)) \text{ for } n \geq 0 \text{ and } \varphi(n, x) = - \sum_{i=n}^{-1} \varphi(f^i(x)) \text{ for } n < 0.$$

**Lemma 1.6.43** *The function  $\Phi$  thus defined on  $\mathcal{O}(x_0)$  is  $\alpha$ -Hölder-continuous.*

*Proof* Suppose  $n, m \in \mathbb{N}$  are such that  $\epsilon := d(f^n(x_0), f^m(x_0))$  is small enough to apply Proposition 1.6.40. Then we obtain  $C > 0$ ,  $\eta \in (0, 1)$ , and  $f^{m-n}(y) = y \in \Lambda$  such that  $d(f^{n+i}(x_0), f^i(y)) \leq C\epsilon\eta^{\min(i, m-n-i)}$ . Since  $\varphi$  is  $\alpha$ -Hölder-continuous there exists  $M > 0$  such that  $|\varphi(x_1) - \varphi(x_2)| \leq M d(x_1, x_2)^\alpha$  whenever  $d(x_1, x_2)$  is small enough. Consequently

$$\begin{aligned} |\Phi(f^n(x_0)) - \Phi(f^m(x_0))| &= \left| \sum_{i=0}^{m-n-1} \varphi(f^{n+i}(x_0)) \right| \\ &= \left| \sum_{i=0}^{m-n-1} (\varphi(f^{n+i}(x_0)) - \varphi(f^i(y))) + \underbrace{\sum_{i=0}^{m-n-1} \varphi(f^i(y))}_{=0} \right| \\ &\leq \sum_{i=0}^{m-n-1} \underbrace{|\varphi(f^{n+i}(x_0)) - \varphi(f^i(y))|}_{\leq MC^\alpha \epsilon^\alpha \eta^{\alpha \min(i, m-n-i)}} \\ &\leq 2MC^\alpha \epsilon^\alpha \sum_{i=0}^{m-n-1} \eta^{\alpha i} < 2MC^\alpha \epsilon^\alpha \frac{1}{1 - \eta^\alpha} \\ &= \frac{2MC^\alpha}{1 - \eta^\alpha} d(f^n(x_0), f^m(x_0))^\alpha. \end{aligned}$$

□

In particular,  $\Phi$  is uniformly continuous on  $\mathcal{O}(x_0)$  and hence extends uniquely to a continuous function  $\Phi$  on  $\Lambda$ , which is uniquely determined by  $\Phi(x_0)$ <sup>27</sup> and has the same Hölder exponent. Since  $\varphi = \Phi \circ f - \Phi$  on a dense set, they coincide by continuity, so  $\Phi$  solves the cohomological equation. □

There is a  $C^1$  version of the Livschitz Theorem 1.6.42 as well:

**Theorem 1.6.44** *Let  $M$  be a Riemannian manifold,  $f: U \rightarrow M$  a smooth embedding with a compact topologically transitive hyperbolic set, and  $\varphi: \Lambda \xrightarrow{C^1} \mathbb{R}$ . Suppose that if  $f^n(x) = x \in M$ , then  $\sum_{i=0}^{n-1} \varphi(f^i(x)) = 0$ . Then there is a  $C^1$  function  $\Phi: \Lambda \rightarrow \mathbb{R}$  such that  $\varphi = \Phi \circ f - \Phi$  and  $\Phi$  is unique up to an additive constant.*

<sup>27</sup>Or:  $\Phi \circ f - \Phi \equiv \Psi \circ f - \Psi \Rightarrow \Phi - \Psi$  is continuous and constant on a dense set.

*Proof* Theorem 1.6.42 gives a Lipschitz-continuous solution  $\Phi$ . To show that it is  $C^1$  we show that the derivatives of  $\Phi$  along stable and unstable leaves exist and are continuous. If  $x$  and  $y$  are nearby points of a stable leaf then

$$\begin{aligned}\Phi(y) - \Phi(x) &= \lim_{n \rightarrow \infty} \left( - \sum_{i=0}^n (\varphi(f^i(y)) - \varphi(f^i(x))) + \Phi(f^n(x)) - \Phi(f^n(y)) \right) \\ &= - \sum_{i=0}^{\infty} (\varphi(f^i(y)) - \varphi(f^i(x))).\end{aligned}$$

Keeping  $x$  fixed and differentiating with respect to  $y = x + tv$  at  $t = 0$  gives by the chain rule  $D_v \Phi(x) = - \sum_{i=0}^{\infty} D_{v_i} \varphi(f^i(x)) D_v(f^i)(x)$ , where  $v_i = Df^i v$ . Since  $v$  is a stable vector,  $D_v(f^i)$  is exponentially small. So is  $D_{v_i} \varphi$  since  $\varphi$  is  $C^1$  and the  $v_i$  are exponentially small. Thus the series on the right converges uniformly and hence to a well-defined and continuous function which is thus the left-hand side. Likewise one obtains differentiability of  $\Phi$  in the unstable direction, so  $\Phi$  has continuous partial derivatives. By Lemma 1.6.45 this implies that  $\Phi$  is  $C^1$ .  $\square$

**Lemma 1.6.45** *Suppose  $\varphi: \mathbb{R}^n \rightarrow \mathbb{R}$  is  $C^1$  along the leaves of two continuous transverse foliations  $W^u$  and  $W^s$  in  $\mathbb{R}^n$ . Then  $\varphi$  is  $C^1$ .*

*Proof* We imitate the argument proving that continuity of partial derivatives implies that a function is  $C^1$ . Given  $x, y$  we show that  $\varphi(y) - \varphi(x) = L(y - x)$  up to higher-order terms in  $|y - x|$  for some linear map  $L$ . Since  $\varphi$  is  $C^1$  along the leaves of  $W^u$  and  $W^s$ , for  $z \in W^u(x) \cap W^s(y)$  we have

$$\varphi(y) - \varphi(x) = \varphi(y) - \varphi(z) + \varphi(z) - \varphi(x) = L_z^s(y - z) + L_x^u(z - x)$$

up to higher order for two linear maps  $L^u$  and  $L^s$  depending continuously on the base-point. But then  $L_z^s \rightarrow L_x^s$  as  $z \rightarrow x$ , hence as  $y \rightarrow x$ , i.e.,  $L_z^s(y - z) = L_x^s(y - z)$  up to higher order, so we can take  $L = L^u \oplus L^s$  on  $TW^u(x) \oplus TW^s(x) = T_x M$ .  $\square$

### 1.6.6.3 Smooth Invariant Measures for Anosov Diffeomorphisms

The obviously necessary condition in Theorem 1.6.46(2) below for existence of a smooth invariant measure is sufficient for topologically transitive Anosov diffeomorphisms. This is an application of the Livschitz Theorem and hence of exponential closing, which was obtained using stable and unstable manifolds. Here,  $Jf(\cdot)$  denotes the Jacobian of a differentiable map with respect to an ambient volume form.

**Theorem 1.6.46** *Let  $M$  be a manifold with volume  $\Omega$  and  $f: M \rightarrow M$  a topologically transitive  $C^2$  Anosov diffeomorphism. Then the following are equivalent:*

1. There is an  $f$ -invariant measure with bounded density that is bounded away from 0.
2.  $Jf^n(x) = 1$  whenever  $f^n(x) = x$ .
3. There is an  $f$ -invariant measure with positive  $C^1$  density.

*Proof* Clearly (3) implies (1). (1) implies (2) because  $e^\Phi \Omega$  is  $f$ -invariant if and only if  $\Phi(x) - \Phi(f^{-1}(x)) = \varphi(x) := -\log Jf(x)$ :

$$\begin{aligned} 0 &= (f^* e^\Phi \Omega)_x - (e^\Phi \Omega)_x = e^{\Phi(f^{-1}(x))} \underbrace{\Omega_{f^{-1}(x)}(Df^{-1}(\cdot), \dots, Df^{-1}(\cdot))}_{=(Jf(x))^{-1} \Omega_x = e^{\varphi(x)} \Omega_x} - e^{\Phi(x)} \Omega_x \\ &= (e^{\Phi(f^{-1}(x))} e^{\varphi(x)} - e^{\Phi(x)}) \Omega_x. \end{aligned}$$

$Jf \in C^1$ , so (2) and Theorem 1.6.44 give a  $C^1$  solution and hence (3).  $\square$

#### 1.6.6.4 Local Product Structure

An important technical property that follows from local maximality is the presence of a *local product structure*:

**Definition 1.6.47** We say that a hyperbolic set  $\Lambda$  has local product structure if for sufficiently small  $\epsilon > 0$  the intersection points provided by Proposition 1.6.36 are always contained in  $\Lambda$ . Equivalently, for such  $\epsilon$  and any  $x \in \Lambda$ , the *Bowen bracket*  $[\cdot, \cdot]$  defines a homeomorphism (a local product parametrization) between  $W_\epsilon^s(x) \times W_\epsilon^u(x)$  and a neighborhood in  $\Lambda$  of  $x$ .

**Proposition 1.6.48** A compact locally maximal hyperbolic set has local product structure.

*Proof* Take  $\epsilon$  such that the  $\epsilon$ -neighborhood  $V$  of  $\Lambda$  satisfies  $\Lambda = \Lambda_V^f$ . Then all points  $[x, y]$  from Proposition 1.6.36 and their orbits are in  $V$ , hence in  $\Lambda$ .  $\square$

*Second Proof of Theorem 1.3.45* Define a relation on  $\text{Per}(f|_\Lambda)$  (which is dense in  $NW(f|_\Lambda)$ ) by Corollary 1.3.21 by  $x \sim y$  if and only if  $W^u(x) \cap W^s(y) \cap \Lambda \neq \emptyset \neq W^s(x) \cap W^u(y) \cap \Lambda$  with both intersections transverse in at least one point. We show that this is an equivalence relation and obtain each  $\Lambda_i$  as the closure of an equivalence class. These closures are called *homoclinic classes*.

Note that  $\sim$  is trivially reflexive and symmetric. To check transitivity suppose  $x, y, z \in \text{Fix}(f^k|_\Lambda)$  and that  $p \in W^u(x) \cap W^s(y) \cap \Lambda$ ,  $q \in W^u(y) \cap W^s(z) \cap \Lambda$  are transverse intersection points. By continuity of unstable leaves the images of a ball around  $p$  in  $W^u(p) = W^u(x) = f^k(W^u(x))$  accumulate on  $W^u(y)$  so  $W^u(x)$  and  $W^s(z)$  have a transverse intersection in  $\Lambda$ .



By Proposition 1.6.48 each equivalence class is open, so by compactness there are finitely many equivalence classes with (pairwise disjoint) closures  $\Lambda_1, \dots, \Lambda_m$ . These are permuted by  $f$  with permutation  $\sigma$ , i.e.,  $f(\Lambda_i) = \Lambda_{\sigma(i)}$ . By Corollary 1.3.21  $NW(f|_{\Lambda}) \subset \overline{\text{Per}(f|_{\Lambda})}$  since  $\Lambda$  is locally maximal, so  $\bigcup_{i=1}^m \Lambda_i = NW(f|_{\Lambda})$ .

To show that  $f^k|_{\Lambda_i}$  is topologically mixing, where  $k$  is the order of  $\sigma$ , note that if  $p \in \Lambda_i$  and  $q \sim p$  are periodic, then there is by definition a heteroclinic point  $z \in W^u(p) \cap W^s(q) \cap \Lambda$ . If  $N$  is the common period then continuity of  $W^u(\cdot)$  shows that  $W^u(p)$  accumulates on  $q$ . So  $W^u(p) \cap \Lambda$  is dense in  $\Lambda_i \cap \text{Per}(f|_{\Lambda})$ , hence its closure  $\Lambda_i$ . To simplify notations assume  $k = 1$ ,  $\Lambda = \Lambda_i$ .

For open  $V$  and  $W$  in  $\Lambda$  we will find  $M \in \mathbb{N}$  with  $f^m(V) \cap W \neq \emptyset$  for all  $m \geq M$  (Definition 1.3.34). Density of periodic points implies the existence of  $f^n(p) = p \in V$ . Since  $V$  is open it contains a neighborhood  $W_\delta^u(p)$  of  $p$  in  $W_{\text{loc}}^u(p) \cap \Lambda$ . Since  $W^u(p) \cap \Lambda = \bigcup_{i=0}^{\infty} f^{in}(W_\delta^u(p))$  is dense there exists  $m_0 \in \mathbb{N}$  such that  $W \cap \bigcup_{i=0}^{m_0} f^{in}(W_\delta^u(p)) \neq \emptyset$ . Since  $f^k(W_\delta^u(p))$  is a neighborhood of  $f^k(p)$  in  $W^u(f^k(p)) \cap \Lambda$  there are  $m_1, \dots, m_{n-1} \in \mathbb{N}$  with  $W \cap \bigcup_{i=0}^{m_k} f^{k+in}(W_\delta^u(p)) \neq \emptyset$ . If  $m \geq M := \max_k(n+1)m_k$ , then  $W \cap \bigcup_{i=0}^m f^i(W_\delta^u(p)) \neq \emptyset$ , so  $W \cap f^m(V) \neq \emptyset$ .  $\square$

This proof gives a new point of view:

**Corollary 1.6.49** *If a compact locally maximal hyperbolic set  $\Lambda$  is topologically mixing then periodic points are dense in  $\Lambda$  and the unstable manifold of every periodic point is dense in  $\Lambda$ .*

*Proof* The spectral decomposition must be trivial.  $\square$

## 1.7 Ergodic Theory

Whereas topological dynamics is about the *possibility* of specific phenomena and evolutions, ergodic theory is about their *probability*. Among the aims of ergodic theory is to sharpen in a quantitative way various recurrence properties such as recurrence of an orbit, topological transitivity, minimality, and topological mixing by considering *asymptotic frequencies* with which corresponding types of recurrence appear. This notion in turn is closely tied in with the notion of time averages discussed in the historical sketch. Indeed, as we saw in Theorem 1.4.11, shifts are a good model for hyperbolic dynamical systems, and since these are mathematical representations of coin tosses or dice, it is natural to seize on probabilistic methods to describe their complexity.

### 1.7.1 Asymptotic Distribution, Invariant Measures

If  $X$  is a metrizable space and  $f: X \rightarrow X$  continuous, the *time average* or *Birkhoff average* of a continuous function  $\varphi$  is

$$E_x(\varphi) := \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi(f^k(x)). \quad (1.22)$$

Whether this exists depends on  $x$  (actually, on the orbit of  $x$ ) and  $\varphi$ . Let  $C(X)$  be the space of continuous functions on  $X$  with the uniform topology. If  $x \in X$  is such that  $E_x(\varphi)$  exists for every  $\varphi \in C(X)$ , then  $E_x: C(X) \rightarrow \mathbb{R}$  has the following properties:

1. Linearity:  $E_x(\alpha\varphi + \beta\psi) = \alpha E_x(\varphi) + \beta E_x(\psi)$ , for  $\alpha, \beta \in \mathbb{R}$ ;
2. Boundedness:  $|E_x(\varphi)| \leq \sup_{y \in X} |\varphi(y)|$ ;
3. Positivity:  $E_x(\varphi) \geq 0$  if  $\varphi \geq 0$  and  $E_x(1) = 1$ ;
4.  $f$ -invariance:  $E_{f(x)}(\varphi) = E_x(\varphi \circ f) = E_x(\varphi)$  because

$$E_x(\varphi \circ f) - E_x(\varphi) = \lim_{n \rightarrow \infty} \frac{1}{n} \left( \underbrace{\sum_{k=0}^{n-1} \varphi(f^{k+1}(x)) - \sum_{k=0}^{n-1} \varphi(f^k(x))}_{=\varphi(f^n(x)) - \varphi(x)} \right) = 0. \quad (1.23)$$

(1)–(3) imply that the Riesz Representation Theorem below gives

**Theorem 1.7.1** *There is a unique probability measure  $\mu_x$  with  $E_x(\varphi) = \int_X \varphi \, d\mu_x$ . Moreover,  $\mu_x$  is  $f$ -invariant: For  $h: (X, \mu) \rightarrow (Y, \nu)$  we write*

$$h_*\mu(A) := \mu(h^{-1}(A)). \quad (1.24)$$

$h^{-1}$  appears because  $\chi_U \circ h = \chi_{h^{-1}(U)}$ . By uniqueness, (4) implies  $f$ -invariance of  $\mu$ , i.e.,  $f_*\mu = \mu$ , and  $\mu \in \mathfrak{M}(f) := \{f\text{-invariant Borel probability measures}\}$ :

**Corollary 1.7.2**  $E_x(\varphi) = \int_X \varphi \, d\mu_x$  for a unique  $\mu_x \in \mathfrak{M}(f)$ .

Two main questions arise:

- (A) Are there  $x \in X$  for which asymptotic distributions  $E_x$  exist?
- (B) When does an invariant measure determine any asymptotic distribution of orbits, i.e., given  $\mu \in \mathfrak{M}(f)$ , is there an  $x \in X$  with  $\int \varphi \, d\mu = E_x(\varphi)$  for all  $\varphi \in C(X)$ ?

We will answer these in due course using a combination of two fundamental theorems from topological dynamics and ergodic theory (see Corollaries 1.7.26 and 1.7.44). But we first pause to introduce the notion of a Borel measure.

**Definition 1.7.3 (Borel and Radon Measures)** Let  $X$  be a separable locally compact Hausdorff space and  $\mathcal{B}$  the  $\sigma$ -algebra of Borel sets, i.e., the  $\sigma$ -algebra generated by closed sets. Then a *Borel measure* is the completion of a measure  $\mu$

defined on  $\mathcal{B}$ . A *Radon measure* is a Borel measure such that  $\mu(B) < \infty$  when  $B$  is compact. (In particular, finite Borel measures are Radon measures.)

**Theorem 1.7.4 (Regularity)** *Radon measures are regular, i.e., for every  $B \in \mathcal{B}$  we have  $\mu(B) = \inf\{\mu(O) \mid B \subset O \text{ open}\} = \sup\{\mu(K) \mid K \subset B \text{ compact}\}$ .*

*Every continuous function  $f: X \rightarrow \mathbb{R}$  is Borel measurable, i.e., preimages of open sets are Borel sets.*

*For every compact set  $K$  there is a decreasing sequence  $\{f_n\}_{n \in \mathbb{N}}$  of nonnegative continuous functions with compact support such that  $f_n \rightarrow \chi_K$  pointwise, where  $\chi_K$  is the characteristic function of  $K$ .*

*Radon measures are separable. (For a dense  $\{x_i\}_{i \in \mathbb{N}}$ , precompact open neighborhoods  $\{B_{ij}\}_{(i,j) \in \mathbb{N}}$  with  $\bigcap_j B_{ij} = \{x_i\}$  define a basis, and every atom is a point by regularity and the Hausdorff assumption.)*

**Theorem 1.7.5** *If  $\mu$  is a Borel probability measure then uniformly continuous functions are dense in  $L^p(\mu)$  for  $1 < p < \infty$ .*

**Theorem 1.7.6 (Riesz Representation Theorem)** *Let  $X$  be a compact Hausdorff space. Then for each bounded linear functional  $F$  on  $C^0(X)$  (i.e., satisfying (1)–(2).) there exists a unique mutually singular pair  $\mu, \nu$  of finite Borel measures (Definition 1.7.3) such that  $F(\varphi) = \int \varphi d\mu - \int \varphi d\nu$  for all  $\varphi \in C^0(X)$ .*

**Remark 1.7.7** It is especially useful that the collection  $\mathfrak{M}(X)$  of Borel probability measures on a compact metrizable space is a convex norm-bounded subset of the dual to  $C(X)$ .  $\mathfrak{M}$  is closed with respect to the *weak\** (or *setwise* or *product*) topology defined by  $\mu_n \rightarrow \mu: \Leftrightarrow \int_X \varphi d\mu_n \rightarrow \int \varphi d\mu \ \forall \varphi \in C(X)$ , hence sequentially compact by the Banach–Alaoglu Theorem.<sup>28</sup>

Coudène [Co16] provides a highly accessible account of these measures and Lebesgue spaces. For some applications these are the appropriate notion of a measure space. They are isomorphic up to a null set to an interval with Lebesgue measure with at most countably many “atoms,” i.e., points of positive measure, added. Surprisingly, this notion is not too restrictive; virtually every probability space in analysis or geometry has this property. In particular, a Borel probability measure on a separable locally compact Hausdorff space defines a Lebesgue space.

## 1.7.2 Existence of Invariant Measures and Recurrence

**Theorem 1.7.8 (Krylov–Bogolubov Theorem)** *A continuous map on a metrizable compact space has an invariant Borel probability measure.*

<sup>28</sup>The (norm-) unit ball in the dual of a normed linear space is weak\*-compact—this implies that norm-bounded weak\*-closed sets are compact; separability gives sequential compactness.

*Proof* If  $f: X \rightarrow X$  continuous,  $\mu \in \mathfrak{M}(X)$ , then by Remark 1.7.7 there is a weak\* accumulation point  $\mu'$  of  $1/n \sum_{k=0}^{n-1} f_*^k \mu \in \mathfrak{M}(X)$ .  $\mu'$  is  $f_*$ -invariant as in (1.23).  $\square$

**Remark 1.7.9** If  $f$  is a homeomorphism,  $\mu$  an  $f$ -invariant measure, and  $A \subset X$  measurable then  $\mu(f(A)) = \mu(A)$ .

**Definition 1.7.10** A surjective  $f: X \rightarrow X$  is said to be  $\mu$ -preserving or measure-preserving if  $A \subset X$  measurable  $\Rightarrow f^{-1}(A)$  measurable and  $\mu(f^{-1}(A)) = \mu(A)$ . In concrete situations an invariant measure may be readily apparent.

**Theorem 1.7.11 (Poincaré Recurrence Theorem)** *Let  $f$  be a probability-preserving transformation of  $(X, \mu)$ ,  $A \subset X$  measurable, and  $N \in \mathbb{N}$ . Then*

$$\mu(\{x \in A \mid \{f^n(x)\}_{n \geq N} \subset X \setminus A\}) = 0.$$

*Proof* Replacing  $f$  by  $f^N$  shows that we may suppose  $N = 1$ . The set

$$\widetilde{A} := \{x \in A \mid \{f^n(x)\}_{n \in \mathbb{N}} \subset X \setminus A\} = A \cap \left( \bigcap_{n=1}^{\infty} f^{-n}(X \setminus A) \right)$$

is measurable.  $f^{-n}(\widetilde{A}) \cap \widetilde{A} = \emptyset$  for every  $n \neq 0$  and hence  $f^{-n}(\widetilde{A}) \cap f^{-m}(\widetilde{A}) = \emptyset$  for all  $m \neq n \in \mathbb{N}$ .  $\mu(f^{-n}(\widetilde{A})) = \mu(\widetilde{A})$  since  $f$  preserves  $\mu$ . Thus  $\mu(\widetilde{A}) = 0$  since  $1 = \mu(X) \geq \mu(\bigcup_{n=0}^{\infty} f^{-n}(\widetilde{A})) = \sum_{n=0}^{\infty} \mu(f^{-n}(\widetilde{A})) = \sum_{n=0}^{\infty} \mu(\widetilde{A})$ .  $\square$

**Remark 1.7.12** The name of this theorem reflects its application to recurrence in Proposition 1.7.66(1).

### 1.7.3 The Birkhoff Ergodic Theorem

The Birkhoff Ergodic Theorem provides the time averages introduced in (1.22). It applies on any probability space, and no topology is involved. Before stating it, we recall a standard result in measure theory in slightly unconventional form.

**Definition 1.7.13 (Absolute Continuity)** If  $(X, \mathcal{S}, \mu)$  and  $(X, \mathcal{T}, \nu)$  are signed measure spaces then  $\nu$  is said to be *absolutely continuous* with respect to  $\mu$ , written  $\nu \ll \mu$ , if every null set for  $\mu$  is a null set for  $\nu$ .

**Theorem 1.7.14 (Radon–Nikodym)** *If  $(X, \mathcal{S}, \mu)$  and  $(X, \mathcal{T}, \nu)$  are  $\sigma$ -finite signed measure spaces and  $\nu \ll \mu$ , then there is a  $\mu$ -a.e. unique density or Radon–Nikodym derivative  $\left[ \frac{d\nu}{d\mu} \right] := \rho: X \rightarrow \mathbb{R}$  of  $\nu$  with respect to  $\mu$  that is measurable with respect to the completion  $\overline{\mathcal{T}}$  of  $\mathcal{T}$  and such that  $\nu(A) = \int_A \rho d\bar{\mu}$ , where  $\bar{\mu}$  is the completion of  $\mu$ , for every  $A$  in the completion of  $\mathcal{T}$ . In particular,  $\overline{\mathcal{T}} \subset \overline{\mathcal{S}}$ .*

**Corollary 1.7.15 (Conditional Expectation)** Suppose  $(X, \mathcal{S}, \lambda)$  is a  $\sigma$ -finite measure space,  $\mathcal{T} \subset \mathcal{S}$  a  $\sigma$ -algebra,  $\varphi \in L^1(X, \mathcal{S}, \lambda)$ . Denote by  $\lambda|_{\mathcal{T}}$  the restriction, that is,  $\lambda|_{\mathcal{T}}(A) = \lambda(A)$  for all  $A \in \mathcal{T} \subset \mathcal{S}$ . Then the conditional expectation

$$E(\varphi | \mathcal{T}) := \varphi_{\mathcal{T}} := \left[ \frac{d(\varphi\lambda)|_{\mathcal{T}}}{d\lambda|_{\mathcal{T}}} \right] \in L^1(X, \mathcal{T}, \lambda|_{\mathcal{T}})$$

of  $\varphi$  on  $\mathcal{T}$  is defined  $\lambda$ -a.e. uniquely by  $\int_A \varphi_{\mathcal{T}} d\lambda = \int_A \varphi d\lambda$  for all  $A \in \mathcal{T}$ .

*Proof* Theorem 1.7.14 with  $\lambda|_{\mathcal{T}} \gg \nu := (\varphi\lambda)|_{\mathcal{T}}, A \mapsto \int \varphi \chi_A d\lambda$  for  $A \in \mathcal{T}$ .  $\square$

**Proposition 1.7.16**

1.  $E(\cdot | \mathcal{T}) =: \pi_{\mathcal{T}}: L^1(\mu) \rightarrow L^1(\mu|_{\mathcal{T}}) \subset L^1(\mu)$  is a projection.
2.  $\pi_{\mathcal{T}}$  is linear and positive, i.e.,  $f \geq 0 \Rightarrow \pi_{\mathcal{T}} f \geq 0$ .
3. If  $g$  is  $\mathcal{T}$ -measurable and bounded, then  $E(gf | \mathcal{T}) = gE(f | \mathcal{T})$ .
4. If  $\mathcal{T}_2 \subset \mathcal{T}_1$  then  $E(\cdot | \mathcal{T}_2) \circ E(\cdot | \mathcal{T}_1) = E(\cdot | \mathcal{T}_2)$ .

The proof is straightforward; we note that (1) follows from (4) but more directly from the obvious fact that  $\pi_{\mathcal{S}} = \text{Id}$ .

We digress briefly to a contemplation of how this plays out in  $L^2$ .

**Definition 1.7.17** If  $H$  is a Hilbert space,  $L \subset H$  a closed subspace, then each  $v \in H$  uniquely<sup>29</sup> decomposes as  $v = v_0 + v_{\perp}$  with  $v_0 \in L$  and  $v_{\perp} \perp L$ , i.e.,  $v_{\perp} \perp w$  for all  $w \in L$ , and the *orthogonal projection to  $L$*  is defined by

$$\pi_L: H \rightarrow L, \quad v_0 + v_{\perp} \mapsto v_0.$$

**Proposition 1.7.18**  $v \in H, w \in L \Rightarrow \|v - \pi(v)\| \leq \|v - w\|$  and  $\langle v, w \rangle = \langle \pi_L(v), w \rangle$ .

*Proof*  $\|v - w\|^2 = \|v_0 + v_{\perp} - w\|^2 = \|v_0 - w\|^2$  is minimal iff  $w = v_0 = \pi(v)$ , and  $\langle v, w \rangle = \langle v_0 + v_{\perp}, w \rangle = \langle v_0, w \rangle = \langle \pi_L(v), w \rangle$ .  $\square$

**Example 1.7.19** Suppose  $(X, \mathcal{T}, \mu)$  is a probability space and  $\mathcal{S} \subset \mathcal{T}$  is a  $\sigma$ -algebra in  $\mathcal{T}$ . Then  $L := L^2(X, \mathcal{S}, \mu) \subset H := L^2(X, \mathcal{T}, \mu)$  is a closed subspace. For  $f \in L^2(X, \mathcal{T}, \mu)$  and  $A \in \mathcal{S}$  we then have  $\chi_A \in L$  and hence by Proposition 1.7.18

$$\int_A f d\mu = \langle f, \chi_A \rangle = \langle \pi_L(f), \chi_A \rangle = \int_A \pi_L(f) d\mu.$$

<sup>29</sup> $v_0 + v_{\perp} = w_0 + w_{\perp} \Rightarrow v_0 - w_0 = w_{\perp} - v_{\perp} \in L \cap L^{\perp}$ .

In light of uniqueness in Corollary 1.7.15 we see that  $\pi_{L^2(X, \mathcal{S}, \mu)} = E(\cdot \mid \mathcal{S}) \upharpoonright_{L^2(X, \mathcal{S}, \mu)}$ , i.e., the orthogonal projection to  $L^2(X, \mathcal{S}, \mu)$  is given by conditional expectation.

If  $f$  is a measure-preserving transformation of a measure space  $(\mathcal{B}, \mu)$  denote by  $\mathcal{I} := \mathcal{I}_f := \{A \in \mathcal{B} \mid f^{-1}(A) = A\}$  the invariant  $\sigma$ -algebra.

**Theorem 1.7.20 (Birkhoff Ergodic Theorem)** *Let  $(X, \mu)$  be a probability space,  $f: X \rightarrow X$   $\mu$ -preserving,  $\varphi \in L^1(X, \mu)$ . Then the time average exists:*

$$\varphi_f := \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi \circ f^k = \varphi_{\mathcal{I}_f} \quad \mu\text{-a.e.}$$

In particular,  $\varphi_f$  is measurable and  $f$ -invariant, and

$$\int \varphi_f d\mu = \int \varphi_{\mathcal{I}} d\mu = \int \varphi d\mu. \quad (1.25)$$

*Proof* If  $\psi \in L^1(\mu)$ , then  $F_n := \max_{k \leq n} \sum_{i=0}^{k-1} \psi \circ f^i \in L^1(\mu)$  is nondecreasing in  $n$ , and

$$\overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \psi \circ f^k \leq \overline{\lim}_{n \rightarrow \infty} \frac{F_n}{n} \leq 0 \quad \text{off } A := \{x \mid F_n(x) \rightarrow \infty\} \in \mathcal{I}. \quad (1.26)$$

$$F_{n+1}(x) = \psi(x) + F_n(f(x)) \Leftrightarrow (F_n \circ f)(x) > 0, \text{ so}$$

$$F_{n+1} - F_n \circ f = \psi - \min(0, F_n \circ f) \searrow \psi \text{ on } A, \quad \text{and}$$

$$0 \leq \int_A (F_{n+1} - F_n) d\mu = \int_A (F_{n+1} - F_n \circ f) d\mu \xrightarrow{n \rightarrow \infty} \int_A \psi d\mu = \int_A \psi_{\mathcal{I}} d\mu \upharpoonright_{\mathcal{I}}$$

by the Monotone-Convergence Theorem. Thus  $\psi_{\mathcal{I}} < 0 \Rightarrow \mu(A) = 0$ , and if  $\psi := \varphi - \varphi_{\mathcal{I}} - \epsilon$ , then  $\psi_{\mathcal{I}} \equiv -\epsilon < 0$ , so (1.26) becomes

$$\overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} (\varphi \circ f^k) - \varphi_{\mathcal{I}} - \epsilon \leq 0 \quad \mu\text{-a.e.}$$

with  $\epsilon > 0$  arbitrary. Replacing here  $\varphi$  by  $-\varphi$  gives

$$\underline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi \circ f^k \geq \varphi_{\mathcal{I}} - \epsilon \quad \mu\text{-a.e.}$$

□

The Birkhoff Ergodic Theorem applies to  $f^{-1}$  when defined, yielding almost-everywhere convergence of negative time averages:

**Proposition 1.7.21** *If  $f$  is invertible, then*

$$\bar{\varphi}_f(x) := \varphi_{f^{-1}}(x) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi(f^{-k}(x)) = \varphi_{\mathcal{I}_{f^{-1}}} = \varphi_{\mathcal{I}_f} = \varphi_f \text{ a.e.,}$$

and the two-sided time averages  $\frac{1}{2n-1} \sum_{|k| \leq n-1} \varphi(f^k(x)) \xrightarrow{\text{a.e.}} \varphi_{\mathcal{I}_f}$  as well.

**Remark 1.7.22** By property (4) of (1.22) i.e., of  $\varphi_f: x \mapsto E_x(\varphi)$  (or Example 1.7.19) and the Birkhoff Ergodic Theorem,  $\varphi \mapsto \varphi_f$  is a projection to the  $f$ -invariant functions.

We note another corollary

**Proposition 1.7.23** *Let  $(X, \mu)$  be a probability space,  $f: X \rightarrow X$   $\mu$ -preserving,  $\varphi \in L^1(X, \mu)$ . Then  $\lim_{n \rightarrow \infty} \frac{1}{n} \varphi \circ f^n \stackrel{\text{a.e.}}{=} 0$ .*

$$\text{Proof } \frac{1}{n} \varphi \circ f^n = \frac{n+1}{n} \left[ \frac{1}{n+1} \sum_{k=0}^n \varphi \circ f^k \right] - \frac{1}{n} \sum_{k=0}^{n-1} \varphi \circ f^k \xrightarrow{n \rightarrow \infty} 1 \cdot \varphi_{\mathcal{I}} - \varphi_{\mathcal{I}} \stackrel{\text{a.e.}}{=} 0.$$

□

### 1.7.4 Existence of Asymptotic Distribution

The exceptional set where the positive or negative time averages do not exist may, of course, depend on the function  $\varphi$ . However, it is negligible for any invariant measure.

**Definition 1.7.24** Given a continuous map  $f$  of a metric space  $X$ , we say that a subset  $A \subset X$  has *total measure* if  $A$  has full measure with respect to *any*  $f$ -invariant Borel probability measure on  $X$ .

**Corollary 1.7.25** *Let  $X$  be compact metrizable,  $f: X \rightarrow X$  continuous. Then*

$$\left\{ x \in X \mid \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi(f^k(x)) \text{ exists for all continuous functions } \varphi \right\}$$

*has total measure, and if  $f$  is a homeomorphism then so does*

$$\left\{ x \in X \mid \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi(f^k(x)) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi(f^{-k}(x)) \text{ for } \varphi \in C(X) \right\}.$$

*Proof* For each  $\varphi_j$  in a countable dense set of functions the averages converge on a set  $E_i$  of total measure. Lipschitz continuity of  $\varphi \mapsto \frac{1}{n} \sum_{k=0}^{n-1} \varphi(f^k(x))$  implies convergence on  $\bigcap_i E_i$  for all continuous  $\varphi$ , and having total measure is stable under countable intersection.  $\square$

Combining this with the Krylov–Bogolubov Theorem 1.7.8 we obtain a positive answer to question (A) on page 66:

**Corollary 1.7.26** *For any continuous map  $f: X \rightarrow X$  of a compact metric space there exists a point  $x \in X$  such that the time average  $\frac{1}{n} \sum_{k=0}^{n-1} \varphi(f^k(x))$  has a limit for every continuous function  $\varphi$  on  $X$  and such that if  $f$  is a homeomorphism, then in addition  $\frac{1}{n} \sum_{k=0}^{n-1} \varphi(f^{-k}(x))$  converges to the same limit.*

### 1.7.5 The Birkhoff Ergodic Theorem for Flows

For flows the statement of the Birkhoff Ergodic Theorem looks as expected and is a straightforward consequence of the Birkhoff Ergodic Theorem 1.7.20 for maps. We state and prove it here mainly to explicitly address the minor subtleties regarding neglected null sets that arise because a flow comprises a continuum of maps. We here use greek letters for flows and roman ones for functions.

**Definition 1.7.27** A 1-parameter group  $\varphi^t$  of probability-preserving transformations of  $(X, \mu)$  is said to be a (measurable) flow if  $(x, t) \mapsto \varphi^t(x)$  is measurable and *measure-preserving* if each  $\varphi^t$  is measure-preserving.

The following is straightforward to verify.

**Theorem 1.7.28** *Consider a flow  $\varphi^t$  on a measure space  $(X, \mu)$  for which there is a  $\lambda > 0$  such that  $\mu(\varphi^t(A)) = \lambda^t \mu(A)$  for each measurable set  $A$  and every  $t \in \mathbb{R}$ . Suppose  $S \subset X$  is measurable and such  $\Phi: [0, a] \times S \rightarrow X$ ,  $(t, x) \mapsto \varphi^t(x)$  is injective for some  $a > 0$ . Then  $S_a := \Phi([0, a] \times S)$  is called a flow box over  $S$ , and*

$$\int f d\mu = \int_0^a \int_S \lambda^t f(\varphi^t(x)) d\mu_S(x) dt,$$

for measurable  $f: S_a \rightarrow \mathbb{R}$ , where

$$\mu_S(A) := \frac{\mu(\varphi^{[t_1, t_2]}(A))}{\int_{t_1}^{t_2} \lambda^t dt}$$

with  $0 \leq t_1 < t_2 \leq a$ .

Note that the case  $\lambda = 1$  corresponds to invariant measures. Now consider a measurable map  $(t, x) \mapsto \varphi^t(x)$  that defines a probability-preserving flow on  $(X, \mu)$ .

**Definition 1.7.29** A function  $f: X \rightarrow \mathbb{R}$  is said to be *almost  $\varphi^t$ -invariant* if there is a null set  $N$  off which  $f \circ \varphi^t = f$  for all  $t$ . The salient point is that  $N$  does not depend



on  $t$ .<sup>30</sup> A set is said to be *almost  $\varphi^t$ -invariant* if its characteristic function is. The  $\sigma$ -algebra of these sets is denoted by  $\mathcal{I}$ .

**Theorem 1.7.30 (Birkhoff Ergodic Theorem for Flows)** *Let  $(X, \mu)$  be a probability space,  $\varphi^t: X \rightarrow X$  a  $\mu$ -preserving flow,  $f \in L^1(X, \mu)$ . Then the time average exists:*

$$f_\varphi(x) := \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t f \circ \varphi^s \, ds = f_{\mathcal{I}} \quad \mu\text{-a.e.}$$

*Proof* We apply the Birkhoff Ergodic Theorem 1.7.20 to establish the existence of the limit and then show that it is  $f_{\mathcal{I}}$ . As a minor convenience we assume  $f \geq 0$ ; the result follows from this by considering positive and negative parts.

First note that by Tonelli's Theorem

$$\infty > n \int f \, d\mu = \int_0^n \int_X f(\varphi^s(x)) \, d\mu \, ds = \int_X \int_0^n f(\varphi^s(x)) \, ds \, d\mu,$$

so  $\int_0^n f(\varphi^s(x)) \, ds$  is defined (and finite) off a null set  $E_n$ , and  $0 \leq f_1 := \int_0^1 f \circ \varphi^s \, ds$  is well-defined a.e. with  $\int f_1 = \int f$ . The Birkhoff Ergodic Theorem 1.7.20 gives

$$\frac{1}{n} \int_0^n f \circ \varphi^s \, ds = \frac{1}{n} \sum_{k=0}^{n-1} f_1 \circ \varphi^k \xrightarrow{n \rightarrow \infty} E(f_1 \mid \mathcal{I}_{\varphi^1}) \text{ off a null set } F. \quad (1.27)$$

To pass from integer times to others, write  $n_t := \lfloor t \rfloor$  and consider  $x$  outside the null set  $N$  defined as the union of the set  $F$  in (1.27), all the  $E_n$  above and the null set implicit in Proposition 1.7.23. Then Proposition 1.7.23 and  $f \geq 0$  imply

$$0 \leq \int_0^{t-n_t} f(\varphi^s(\varphi^{n_t}(x))) \, ds \leq f_1(\varphi^{n_t}(x)) = o(t),$$

so (1.27) gives

$$\begin{aligned} f_\varphi(x) &= \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t f(\varphi^s(x)) \, ds \\ &= \lim_{t \rightarrow \infty} \frac{n_t}{t} \frac{1}{n_t} \sum_{k=0}^{n_t-1} f_1(\varphi^k(x)) + \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^{t-n_t} f(\varphi^s(\varphi^{n_t}(x))) \, ds = (f_1)_{\mathcal{I}_{\varphi^1}} + 0. \end{aligned}$$

<sup>30</sup>It is not required, but in this situation, we can ultimately choose the exceptional set to be invariant, i.e.,  $f$  is measurable with respect to the completion of the  $\sigma$ -algebra of (properly) invariant sets.

Thus  $\int f_\varphi = \int f$ . Now apply what we proved so far to  $g := f\chi_A$  for any  $A \in \mathcal{J}$ :

$$\int_A f_\varphi = \int f_\varphi \chi_A = \int f_\varphi (\chi_A)_\varphi = \int (f\chi_A)_\varphi = \int g_\varphi = \int g = \int f\chi_A = \int_A f,$$

and this, together with  $\varphi$ -invariance, is the very definition of  $f_\varphi = f_{\mathcal{J}}$ .  $\square$

### 1.7.6 The von Neumann Mean Ergodic Theorem

Let us introduce notation that will be much used later, notably in Sect. 1.7.19.

**Definition 1.7.31 (Koopman Operator)** To  $p \geq 1$  and a measure-preserving map  $f: (X, \mu) \rightarrow (Y, \nu)$  associate an isometric operator

$$U_f: L^p(Y, \nu) \rightarrow L^p(X, \mu), \quad \varphi \mapsto \varphi \circ f$$

on complex-valued functions, the *Koopman operator*. Eigenvectors of the Koopman operator  $U_f$  are called eigenfunctions (of  $U_f$  or of  $f$ ), and eigenvalues of  $U_f$  are also referred to as eigenvalues of  $f$ . Constant functions are eigenfunctions of  $U_f$  (for the eigenvalue 1). Therefore, in  $L^2$ , we sometimes explicitly or implicitly restrict attention to  $1^\perp$  when discussing the spectrum of  $U_f$ .

*Remark 1.7.32* The case  $p = 2$  is of particular interest. If  $f: X \rightarrow X$  is invertible then so is  $U_f$  and in this case  $U_f$  defines a unitary operator on  $L^2$ .

As previously mentioned, the Birkhoff Ergodic Theorem 1.7.20 has a counterpart for convergence in  $L^2$ , the *Mean Ergodic Theorem*, which, in fact, predates it [Zu02].

**Theorem 1.7.33 (von Neumann Mean Ergodic Theorem)** Let  $(X, \mu)$  be a measure space,  $f: (X, \mu) \rightarrow (X, \mu)$  a  $\mu$ -preserving transformation,  $\varphi \in L^2(X, \mu)$ . Then

$$\frac{1}{n} \sum_{i=0}^{n-1} \varphi \circ f^i \xrightarrow[n \rightarrow \infty]{L^2} P_f(\varphi),$$

where  $P_f$  is the orthogonal projection to the  $U_f$ -invariant space  $L^2(X, \mathcal{J}, \mu|_{\mathcal{J}})$ .

*Remark 1.7.34* Unsurprisingly,  $P_f = E(\cdot \mid \mathcal{J})$ , the latter being as in Theorem 1.7.20: by definition,  $P_f$  is uniquely defined by  $P_f(\varphi) - \varphi \perp L^2(X, \mathcal{J}, \mu|_{\mathcal{J}})$  for all  $\varphi \in L^2$ , while at the same time clearly  $E(\varphi \mid \mathcal{J}) - \varphi \perp L^2(X, \mathcal{J}, \mu|_{\mathcal{J}})$  for all  $\varphi \in L^2$ .

The measure need not be finite. By passing to  $\bigcup_i (\varphi \circ f^i)^{-1}(\mathbb{R} \setminus \{0\})$ , one may assume  $\sigma$ -finiteness without loss of generality.

The von Neumann Ergodic Theorem follows from a Hilbert-space lemma:

**Proposition 1.7.35** *Suppose  $H$  is a Hilbert space,  $U: H \rightarrow H$  linear such that  $\|U\| \leq 1$ ,  $P: H \rightarrow I := \{x \in H \mid Ux = x\}$  the projection to the  $U$ -invariant subspace. Then  $\frac{1}{n} \sum_{i=0}^{n-1} U^i(x) \xrightarrow{n \rightarrow \infty} P(x)$  for all  $x \in H$ .*

*Proof* Since the sum is telescoping on  $N := \{x - Ux \mid x \in H\} \subset H$  and trivial for invariant elements, the essential step is to show

*Claim*  $I = \bar{N}^\perp$ , i.e.,  $H = \bar{N} \oplus I$ .

*Proof* Let us show that  $x = U^*x \Leftrightarrow x = Ux$ . First,  $\|U^*x\| \leq \|U\|$  because

$$\|U^*x\|^2 = \langle U^*x, U^*x \rangle = \langle UU^*x, x \rangle \leq \|UU^*x\| \|x\| \leq \|U^*x\| \|U\| \|x\|$$

for all  $x \in H$ . Next, if  $\|V\| \leq 1$  and  $x = V^*x$ , then  $x = Vx$  because

$$0 \leq \underbrace{\|x - Vx\|^2}_{=\langle x - Vx, x - Vx \rangle} = \underbrace{\langle x, x \rangle}_{=\langle V^*x, x \rangle = \langle x, x \rangle} - \underbrace{\langle x, Vx \rangle}_{=\langle x, V^*x \rangle = \langle x, x \rangle} + \underbrace{\langle Vx, x \rangle}_{=\langle x, V^*x \rangle = \langle x, x \rangle} = \|Vx\|^2 - \|x\|^2 \leq 0.$$

Applying this to  $U$  and  $U^*$  proves  $x = U^*x \Leftrightarrow x = Ux$ . Then  $x \in I$  iff  $x = U^*x$  iff

$$y \in H \Rightarrow 0 = \langle x - U^*x, y \rangle = \langle (\text{Id} - U)^*x, y \rangle = \langle x, (\text{Id} - U)y \rangle = \langle x, y - Uy \rangle$$

if and only if  $x \in N^\perp = \bar{N}^\perp$ . Therefore  $I = \bar{N}^\perp$ . □

Now write an arbitrary  $x \in H$  as  $x = a + b$  with  $a \in I$  and  $b \in \bar{N}$ . Then

$$\frac{1}{n} \sum_{i=0}^{n-1} U^i x - Px = \frac{1}{n} \sum_{i=0}^{n-1} \underbrace{U^i(a+b)}_{=a+U^i b} - \underbrace{P(a+b)}_{=a} = \frac{1}{n} \sum_{i=0}^{n-1} U^i b \xrightarrow{n \rightarrow \infty} 0$$

because if  $\epsilon > 0$ ,  $b_\epsilon = c_\epsilon - Uc_\epsilon \in N$  such that  $\|b - b_\epsilon\| < \epsilon/2$ , and  $n \geq 4\|c_\epsilon\|$ , then

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=0}^{n-1} U^i b \right\| &= \left\| \frac{1}{n} \sum_{i=0}^{n-1} U^i (b - b_\epsilon) + \frac{1}{n} \sum_{i=0}^{n-1} U^i b_\epsilon \right\| \\ &\leq \|b - b_\epsilon\| + \underbrace{\left\| \frac{1}{n} \sum_{i=0}^{n-1} U^i (c_\epsilon - Uc_\epsilon) \right\|}_{=c_\epsilon - U^n c_\epsilon} \leq \epsilon/2 + 2\|c_\epsilon\|/n \leq \epsilon. \end{aligned}$$

□

### 1.7.7 Ergodicity and Unique Ergodicity

**Definition 1.7.36** A measure  $\mu$  is said to be *ergodic* with respect to  $f$ , or one says that  $f$  is ergodic with respect to  $\mu$ , if for any measurable  $A \subset X$  with  $f^{-1}(A) = A$  either  $\mu(A) = 0$  or  $\mu(X \setminus A) = 0$ . If this holds for all iterates  $f^n$ , then we say that  $f$  is *totally ergodic*.

*Remark 1.7.37* Note that  $f$ -invariance of  $\mu$  is not needed for this definition. Ergodicity can be reformulated in functional language:

**Proposition 1.7.38 (Characterization of Ergodicity)**

- $f: X \rightarrow X$  is ergodic with respect to  $\mu$
- $\Leftrightarrow$  any measurable  $f$ -invariant  $\varphi: X \rightarrow \mathbb{C}$  is constant  $\mu$ -a.e.
- $\Leftrightarrow$  any bounded measurable  $f$ -invariant  $\varphi: X \rightarrow \mathbb{R}$  is constant  $\mu$ -a.e.
- $\Leftrightarrow$  any  $f$ -invariant  $\varphi \in L^p(X, \mu)$  is constant  $\mu$ -a.e.
- $\Leftrightarrow$  any nonnegative measurable  $f$ -invariant  $\varphi: X \rightarrow \mathbb{C}$  is constant  $\mu$ -a.e.

*Proof* These (and other) characterizations arise from the following implications:  $f$  is not ergodic  $\Rightarrow$  there is an invariant characteristic function (namely, of an invariant set of intermediate measure) that is not constant a.e.  $\Rightarrow$  there is a nonnegative bounded invariant measurable function that is not constant a.e.  $\Rightarrow$  there is a nonconstant invariant  $\varphi \in L^p \Rightarrow$  there is an invariant measurable  $\mathbb{C}$ -valued function that is not constant a.e.  $\Rightarrow f$  is not ergodic (because either the real or the imaginary part is an  $f$ -invariant measurable function  $\varphi: X \rightarrow \mathbb{R}$  and not constant almost everywhere, so there exists an  $a \in \mathbb{R}$  such that  $\mu(\varphi^{-1}((a, \infty))) \notin \{0, 1\}$ , and this set is invariant).  $\square$

*Remark 1.7.39* Proposition 1.7.38 simply states in various function spaces that the subspace of  $f$ -invariant functions is the space of constant functions. Remark 1.7.22 lets us determine the space of  $f$ -invariant functions as the range of the projection  $\varphi \mapsto \varphi_f$ , and doing so for a dense set of functions gives the needed information—Theorem 1.7.5 implies the following result.

**Theorem 1.7.40** If  $\varphi_f = \text{const.}$   $\mu$ -a.e. for every  $\varphi \in C(X)$ , then  $\mu$  is ergodic. Considering densities gives:

**Proposition 1.7.41**  $\mu \in \mathfrak{M}(f)$  is ergodic if and only if  $\mu \gg \nu \in \mathfrak{M}(f) \Rightarrow \mu = \nu$ .

*Proof*  $\mu \gg \nu \in \mathfrak{M}(f) \Leftrightarrow \nu = \rho \cdot \mu$ , where  $\rho \in L^1(\mu)$  is the (unique hence  $f$ -invariant) Radon–Nikodym derivative. This is constant ( $\equiv 1$ ) iff  $\nu$  is ergodic.  $\square$

One can strengthen the statement that functions invariant under an ergodic transformation are constant via the following simple observation:

**Proposition 1.7.42** If  $f: X \rightarrow X$  is a transformation preserving a probability measure and  $\varphi: X \rightarrow \mathbb{R}$  satisfies  $\varphi \circ f \leq \varphi$  (“subinvariance”), then  $\varphi$  is  $f$ -invariant.

*Proof* By assumption  $A_r := \{x \in X \mid \varphi(x) \leq r\} \supset \{x \in X \mid \varphi(f(x)) \leq r\} = f^{-1}(A_r)$ , while  $\mu(f^{-1}(A_r)) = \mu(A_r)$ . Thus  $f^{-1}(A_r) \stackrel{\text{a.e.}}{=} A_r$  for all  $r \in \mathbb{R}$ .  $\square$

This and Proposition 1.7.38 yield

**Corollary 1.7.43** *If  $\mu$  is an ergodic invariant probability measure for  $f: X \rightarrow X$ ,  $\varphi: X \rightarrow \mathbb{R}$ , and  $\varphi \circ f \leq \varphi$ , then  $\varphi$  is constant  $\mu$ -a.e.*

An important corollary of the Birkhoff Ergodic Theorem 1.7.20 is that for an ergodic transformation time averages equal space averages almost everywhere.

**Corollary 1.7.44 (Strong Law of Large Numbers)** *If  $\mu(X) = 1$ ,  $f: X \rightarrow X$  is an ergodic  $\mu$ -preserving transformation, and  $\varphi \in L^1(X, \mu)$  then*

$$\varphi_f = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi \circ f^k \stackrel{\text{a.e.}}{=} \int_X \varphi d\mu.$$

Equivalently,  $\sum_{k=0}^{n-1} \varphi(f^k(x)) - n \int_X \varphi d\mu = o(n)$   $\mu$ -a.e.

*Remark 1.7.45* The latter form of the conclusion uses “little  $o$ ” notation:

$$f(n) = o(g(n)) : \Leftrightarrow \frac{f(n)}{g(n)} \xrightarrow{n \rightarrow \infty} 0.$$

*Proof*  $\varphi_f$  is  $f$ -invariant, so constant a.e. By (1.25) the constant is  $\int_X \varphi d\mu$ .  $\square$   
Thus we have answered question (B) after Corollary 1.7.2. An invariant measure determines the asymptotic distribution of  $\mu$ -almost every point if it is ergodic. A nonergodic invariant measure  $\mu$  may also determine the asymptotic distribution of some orbits, but such orbits are always a set of  $\mu$ -measure zero.

**Proposition 1.7.46** *A probability-preserving transformation  $f: X \rightarrow X$  is ergodic iff*

$$\frac{1}{n} \sum_{k=0}^{n-1} \int \varphi \circ f^k \psi \xrightarrow{n \rightarrow \infty} \int \varphi \int \psi \quad (1.28)$$

for all  $\varphi, \psi \in L^2$ , i.e., if and only if  $\frac{1}{n} \sum_{k=0}^{n-1} \varphi \circ f^k \xrightarrow[n \rightarrow \infty]{\text{weakly}} \text{const.}$  for all  $\varphi \in L^2$ .

*Remark 1.7.47* For  $\varphi = \chi_A$  and  $\psi = \chi_B$ , (1.28) becomes

$$\frac{1}{n} \sum_{k=0}^{n-1} \mu(f^{-k}(A) \cap B) \xrightarrow{n \rightarrow \infty} \mu(A)\mu(B). \quad (1.29)$$

*Proof* If  $\varphi = \varphi \circ f$ , then  $\varphi = \frac{1}{n} \sum_{k=0}^{n-1} \varphi \circ f^k \xrightarrow[n \rightarrow \infty]{\text{weakly}} \text{const.}$

If  $f$  is ergodic, then Corollary 1.7.44 and the Dominated-Convergence Theorem give (1.28) for all  $\varphi, \psi \in L^2$ .  $\square$

Corollary 1.7.44 leads to the question of whether every continuous map has an ergodic invariant measure. This is so thanks to a little functional analysis.

**Lemma 1.7.48** *Ergodic measures are the extreme points of  $\mathfrak{M}(f)$ :  $\mu \in \mathfrak{M}(f)$  is not ergodic iff there exist  $\mu_1 \neq \mu_2 \in \mathfrak{M}(f)$  and  $0 < \lambda < 1$  such that  $\mu = \lambda\mu_1 + (1 - \lambda)\mu_2$ .*

*Proof* If  $f^{-1}(A) = A$  and  $0 < \mu(A) < 1$ , then  $\mu = \mu(A)\mu_A + (1 - \mu(A))\mu_{X \setminus A}$ , where

$$\mu_A(B) := \mu(B \mid A) := \frac{\mu(B \cap A)}{\mu(A)}$$

is the density of  $B$  in  $A$ . (Note that  $\mu_A \perp \mu_{X \setminus A}$ .)

If  $i = 1, 2$ , then  $\mu_i \ll \mu$ , so the Radon–Nikodym Theorem gives an  $f$ -invariant  $L^1(\mu)$ -density  $\rho_i$  with  $\int \varphi d\mu_i \equiv \int \rho_i \varphi d\mu$ . By assumption  $\lambda\rho_1 + (1 - \lambda)\rho_2 = 1 = \int \rho_1 d\mu = \int \rho_2 d\mu$ , so  $\mu_1 \neq \mu_2 \Rightarrow \rho_1 \neq \rho_2 \Rightarrow \rho_1 \not\equiv \text{const.}$ , and  $\mu$  is not ergodic.  $\square$

**Theorem 1.7.49** *Every continuous map  $f$  on a metrizable compact space  $X$  has an ergodic invariant Borel probability measure.*

*Proof* By the Krein–Milman Theorem<sup>31</sup>  $\mathfrak{M}(f) \neq \emptyset$  has extreme points.  $\square$

Lemma 1.7.48 connects decomposability of a measure (by convex combination) and decomposability of the space. One can sharpen that connection:

**Proposition 1.7.50** *Different invariant ergodic probability measures for the same transformation are mutually singular.*

*Proof* Call them  $\nu, \mu = \mu_{\text{ac}} + \mu^\perp$  with  $\mu_{\text{ac}} \ll \nu \perp \mu^\perp$  (invariantly by uniqueness of Lebesgue decomposition); since  $\mu$  is ergodic, hence extreme, we have either  $\mu = \mu^\perp$  or  $\mu = \mu_{\text{ac}} = \nu$  by ergodicity of  $\nu$  and Proposition 1.7.41.  $\square$

Proposition 1.7.50 means that any convex combination of finitely many ergodic measures produces a corresponding nontrivial finite partition of the space.

Moreover, every invariant measure for a measure-preserving transformation can be decomposed into *ergodic components*. For continuous maps of compact metrizable spaces the latter fact is a consequence of Lemma 1.7.48 and:

**Theorem 1.7.51 (Choquet Theorem)** *If  $C$  is a compact metrizable convex set in a locally convex<sup>32</sup> topological vector space and  $x \in C$ , then there is a probability measure  $\mu_x$  on  $\text{ex } C$  such that  $x = \int_{\text{ex } C} z d\mu_x(z)$ .*

**Theorem 1.7.52 (Ergodic Decomposition [Co16])** *Every invariant Borel probability measure for a continuous map  $f$  of a metrizable compact space  $X$  decomposes into an integral of ergodic invariant Borel probability measures in the following sense: There is a partition (modulo null sets) of  $X$  into invariant subsets  $X_\alpha$ ,  $\alpha \in A$*

<sup>31</sup>A compact convex set in a locally convex topological vector space is the closed convex hull of its extreme points, i.e.,  $C = \overline{\text{co}} \text{ex}(C)$ .

<sup>32</sup>A topological vector space is *locally convex* if every open set contains a convex open set.

with  $A$  a Lebesgue space, and each  $X_\alpha$  carrying an  $f$ -invariant ergodic measure  $\mu_\alpha$  such that  $\int \varphi d\mu = \iint \varphi d\mu_\alpha d\alpha$  for any function  $\varphi$ .

**Definition 1.7.53** These sets  $X_\alpha$  are called the *ergodic components* of  $(f, \mu)$ .

*Example 1.7.54* The ergodic components of  $T_{(0, \sqrt{2})}$ ,  $(x, y) \mapsto (x, y + \sqrt{2})$  are the vertical circles  $\{x\} \times S^1$  for  $x \in S^1$ .

**Definition 1.7.55** A continuous map  $f: X \rightarrow X$  of a metrizable compact space  $X$  is said to be *uniquely ergodic* if it has only one invariant Borel probability measure.

**Proposition 1.7.56** The invariant probability measure of a uniquely ergodic map  $f$  is ergodic.

*Proof*  $\mathfrak{M}(f) = \{\mu\} = \text{ex}\{\mu\} = \text{ex } \mathfrak{M}(f)$ ; apply Lemma 1.7.48.  $\square$

Unique ergodicity is related to *uniform* convergence of Birkhoff averages:

**Proposition 1.7.57** If  $f: X \rightarrow X$  is uniquely ergodic then for every continuous function  $\varphi$  the time averages  $1/n \sum_{k=0}^{n-1} \varphi(f^k(x))$  converge uniformly.

*Proof* If  $1/n \sum_{k=0}^{n-1} \varphi(f^k(x))$  does not converge uniformly for some continuous function  $\varphi$ , then one can find  $a < b$ ,  $x_k, y_k \in X$ , and  $n_k \rightarrow \infty$  such that

$$\frac{1}{n_k} \sum_{l=0}^{n_k-1} \varphi(f^l(x_k)) < a, \quad \frac{1}{n_k} \sum_{l=0}^{n_k-1} \varphi(f^l(y_k)) > b.$$

A diagonal argument gives a subsequence  $n_{k_j}$  such that for every  $\psi \in C(X)$  both

$$J_1(\psi) = \lim_{j \rightarrow \infty} \frac{1}{n_{k_j}} \sum_{l=0}^{n_{k_j}-1} \psi(f^l(x_{k_j})) \quad \text{and} \quad J_2(\psi) = \lim_{j \rightarrow \infty} \frac{1}{n_{k_j}} \sum_{l=0}^{n_{k_j}-1} \psi(f^l(y_{k_j}))$$

exist.  $J_1$  and  $J_2$  are bounded linear positive  $f$ -invariant functionals; thus  $J_1(\psi) = \int \psi d\mu_1$ ,  $J_2(\psi) = \int \psi d\mu_2$  for  $f$ -invariant probability measures  $\mu_1$  and  $\mu_2$ . Since  $J_1(\varphi) \leq a < b \leq J_2(\varphi)$  we have  $\mu_1 \neq \mu_2$  so  $f$  is not uniquely ergodic.  $\square$

*Remark 1.7.58* The converse fails for  $\text{Id}: X \rightarrow X$  with  $\text{card}(X) > 1$  but holds if  $f$  is transitive or if these uniform limits are always constants (Proposition 1.7.61).

**Corollary 1.7.59** Let  $f: X \rightarrow X$ ,  $\mathfrak{M}(f) = \{\mu\}$ ,  $U \subset X$  open and  $\mu(\partial U) = 0$ . Then  $1/n \sum_{k=0}^{n-1} \chi_U(f^k(x)) \rightarrow \mu(U)$  uniformly.

*Proof* Let  $\varphi_m \leq \chi_U \leq \bar{\varphi}_m$  ( $m \in \mathbb{N}$ ) be sequences of continuous functions such that  $\int \bar{\varphi}_m d\mu \rightarrow \mu(U)$  and  $\int \varphi_m d\mu \rightarrow \mu(U)$ . For each  $n \in \mathbb{N}$  and  $x \in X$  one has

$$\frac{1}{n} \sum_{k=0}^{n-1} \varphi_m(f^k(x)) \leq \frac{1}{n} \sum_{k=0}^{n-1} \chi_U(f^k(x)) \leq \frac{1}{n} \sum_{k=0}^{n-1} \bar{\varphi}_m(f^k(x)). \quad (1.30)$$

Fix  $\delta > 0$  and find  $m$  such that  $\int \underline{\varphi}_m d\mu > \mu(U) - \delta/2$  and  $\int \bar{\varphi}_m d\mu < \mu(U) + \delta/2$ . By Proposition 1.7.57 we have

$$\mu(U) - \delta \leq \frac{1}{n} \sum_{k=0}^{n-1} \chi_U(f^k(x)) \leq \mu(U) + \delta$$

from (1.30) for sufficiently large  $n$ . Since  $\delta$  is arbitrary, the claim follows.  $\square$

**Proposition 1.7.60** *If for every  $\varphi \in C(X)$  the time averages  $1/n \sum_{k=0}^{n-1} \varphi \circ f^k$  converge uniformly to a constant then  $f$  is uniquely ergodic.*

*Proof* Suppose  $\mu$  is an  $f$ -invariant probability measure. For  $\varphi \in C(X)$  we have  $1/n \sum_{k=0}^{n-1} \varphi(f^k(x)) \xrightarrow{\text{uniformly}} \varphi_0 \in \mathbb{R}$ , so  $\int \varphi d\mu = \int \varphi_0 d\mu = \varphi_0$ . Hence  $\mu \in C(X)^*$  is unique.  $\square$

Since any  $\mu \in C(X)^*$  is uniquely determined by its values on a dense set, the preceding argument actually establishes:

**Proposition 1.7.61** *If  $\Phi \subset C(X)$  is dense and for every  $\varphi \in \Phi$  the time averages  $1/n \sum_{k=0}^{n-1} \varphi \circ f^k$  converge uniformly to a constant, then  $f$  is uniquely ergodic.*

## 1.7.8 Isomorphism and Factors

Similarly to the theory of smooth dynamical systems and topological dynamics, ergodic theory has a dual agenda: the classification of various classes of measure-preserving transformations up to natural equivalence relations and the study of various asymptotic properties invariant under those relations. Ergodicity is an example of such an invariant which is a counterpart of topological transitivity; mixing, another recurrence-type invariant, is discussed in Sect. 1.7.16. Right now we define and discuss the most natural equivalence relation in ergodic theory.

**Definition 1.7.62** Let  $f: X \rightarrow X$  and  $g: Y \rightarrow Y$  be measure-preserving transformations of measure spaces  $(X, \mu)$  and  $(Y, \nu)$ , correspondingly.  $f$  and  $g$  are said to be *measure-theoretically isomorphic* if there exists an isomorphism  $h: (X, \mu) \rightarrow (Y, \nu)$ , i.e., an injective (mod 0) transformation such that  $h_*\mu = \nu$  [see (1.24)] and

$$g = h \circ f \circ h^{-1}.$$

$g$  is said to be a (measure-theoretic) *factor* of  $f$  if there is a measure-preserving map  $h: X \rightarrow Y$  (in general noninvertible) such that  $h_*\mu = \nu$  and  $g \circ h = h \circ f$ .

All properties of measure-preserving transformations that we are going to discuss are invariants of measure-theoretic isomorphism<sup>33</sup>; ergodicity quite obviously is.

---

<sup>33</sup>Section 1.7.18 excepted.



Furthermore, a factor of an ergodic transformation is also ergodic: If  $g$  is a factor of  $f$  and  $A \subset Y$  is  $g$ -invariant,  $0 < \nu(A) < 1$ , then  $B := h^{-1}(A)$  is  $f$ -invariant and  $\mu(B) = \nu(A)$ .

In certain cases invariants of measure-theoretic isomorphism provide insights into properties of smooth or topological dynamical systems. For example, the measure-theoretic isomorphism class of a uniquely ergodic map is an important invariant of topological conjugacy.

### 1.7.9 Topological and Probabilistic Recurrence

In spirit, several of the statistical properties we have discussed are close counterparts to topological recurrence properties. We now connect these two realms systematically.

**Proposition 1.7.63** *For a Borel measure  $\mu$  on a separable metrizable space  $X$*

1. *the support  $\text{supp } \mu := \{x \in X \mid \mu(U) > 0 \text{ whenever } x \in U, U \text{ open}\}$  of  $\mu$  is closed,*
2.  *$\mu(X \setminus \text{supp } \mu) = 0$ ,*
3. *any set of full measure is dense in  $\text{supp } \mu$ .*

*Proof*

1. If  $x \notin \text{supp } \mu$  take  $U_x \ni x$  open with  $\mu(U_x) = 0$ . Then  $U_x \cap \text{supp } \mu = \emptyset$ .
2. Since  $X$  is separable,  $X \setminus \text{supp } \mu$  is covered by countably many  $U_x$  as above, so  $\mu(X \setminus \text{supp } \mu) = 0$  by  $\sigma$ -additivity of  $\mu$ .
3. If  $A \subset X$  and  $x \in U := \text{supp } \mu \setminus A$  then  $\mu(X \setminus A) \geq \mu(U) > 0$ .

□

**Remark 1.7.64** If  $\text{supp } \mu = X$  then we say that  $\mu$  has full support or  $\mu$  is positive on open sets.

**Definition 1.7.65** Let  $f: X \rightarrow X$  be a homeomorphism of a topological space. A point  $y \in X$  is said to be an  $\omega$ -limit point resp. an  $\alpha$ -limit point for a point  $x \in X$  if there exists a sequence  $n_i \rightarrow +\infty$  resp.  $n_i \rightarrow -\infty$  such that  $f^{n_i}(x) \rightarrow y$ .

The sets of all  $\omega$ -limit resp.  $\alpha$ -limit points for  $x$  are denoted by

$$\omega(x) = \bigcap_{N \in \mathbb{N}} \overline{\bigcup_{n \geq N} f^n(x)} \quad \text{resp.} \quad \alpha(x) = \bigcap_{n \in \mathbb{N}} \overline{\bigcup_{n \geq N} f^{-n}(x)}$$

and are called its  $\omega$ -limit resp.  $\alpha$ -limit set. The  $\alpha$ - and  $\omega$ -limit set of  $f$  are defined by

$$L_+(f) := \omega(f) := \overline{\bigcup_{x \in X} \omega(x)} \quad \text{and} \quad L_-(f) := \alpha(f) := \overline{\bigcup_{x \in X} \alpha(x)}.$$

The limit set of  $f$  is  $L(f) := L_-(f) \cup L_+(f)$ .

A point  $x \in X$  is *positively recurrent* if  $x \in \omega(x)$ , i.e.,  $x = \lim f^{n_k}(x)$  for some sequence  $n_k \rightarrow \infty$ . If  $f$  is invertible,  $x$  is *negatively recurrent* if  $x \in \alpha(x)$ .

Finally,  $x$  is *recurrent* if it is both positively and negatively recurrent. We denote by  $R^+(f)$ ,  $R^-(f)$ , and  $R(f)$  the closures of the sets of all positively recurrent, negatively recurrent, and recurrent points.

**Proposition 1.7.66** *Let  $f$  be a continuous map of a complete separable metrizable space  $X$ . Then:*

1.  $\text{supp } \mu \subset R^+(f)$  for any  $f$ -invariant Borel probability measure  $\mu$ . If  $f$  is invertible, then  $\text{supp } \mu \subset R(f)$ .
2. If  $\mu$  is ergodic then  $f|_{\text{supp } \mu}$  has a dense positive semiorbit; indeed the set of these has full measure. In particular, ergodicity of a measure with full support implies topological transitivity.
3. If  $\text{supp } \mu$  is compact and  $f|_{\text{supp } \mu}$  is uniquely ergodic, then  $\text{supp } \mu$  is minimal.

*Proof*

1. Take a countable base  $\{U_1, U_2, \dots\}$  of open subsets of  $X$  and let  $R_+$  be the set of all points  $x$  such that if  $x \in U_m$  then infinitely many positive iterates of  $x$  also belong to  $U_m$ . Apply the Poincaré Recurrence Theorem 1.7.11 to each of the  $U_i$  to deduce that  $R_+$  has full measure. If  $f$  is invertible then by the same argument the set  $R_-$  constructed similarly to  $R_+$  but with negative iterates also has full measure. Hence  $R := R_- \cap R_+$  has full measure and is by Proposition 1.7.63(3) dense in  $\text{supp } \mu$ . On the other hand, if  $x \in R$  and  $U \ni x$  is an open set then  $U_m \subset U$  for some  $m$ ; hence infinitely many positive and negative iterates of  $x$  lie in  $U$ , i.e.,  $R$  consists of recurrent points. Hence  $\text{supp } \mu \subset \bar{R} = R(f)$ .
2. Take a countable base  $\{U_1, U_2, \dots\}$  of open sets for the induced topology on  $\text{supp } \mu$ . By definition  $0 < \mu(U_m)$ , and a full-measure set of dense orbits for invertible  $f$  is elementary:  $0 < \mu(U_m) \leq \mu(\bigcup_{i \in \mathbb{Z}} f^{-i}(U_m))$ , so  $\mu(\bigcup_{i \in \mathbb{Z}} f^{-i}(U_m)) = 1$  by ergodicity, hence  $R := \bigcap_{m \in \mathbb{N}} \bigcup_{i \in \mathbb{Z}} f^{-i}(U_m)$  has full measure, and the orbit of any  $x \in R$  intersects all  $U_m$  and hence is dense. To prove the claim as stated, instead apply Corollary 1.7.44 simultaneously to the characteristic functions  $\chi_{U_m}$  to obtain a set  $R$  of full measure such that for  $x \in R$ ,  $m \in \mathbb{N}$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} \chi_{U_m}(f^i(x)) = \mu(U_m) > 0,$$

and which is hence as desired.

3. For closed  $f$ -invariant  $\Lambda \subseteq \text{supp } \mu$  the Krylov–Bogolubov Theorem 1.7.8 gives a  $\nu \in \mathfrak{M}(f|_{\Lambda}) \subset \mathfrak{M}(f)$ . Then  $\mu = \nu \Rightarrow \text{supp } \mu = \text{supp } \nu \subset \Lambda \subset \text{supp } \mu$ .  $\square$

Recurrence implies chain-recurrence (Definition 1.3.14), so Proposition 1.3.37 implies

**Corollary 1.7.67** *Let  $(X, \mu)$  be a complete separable metrizable probability space,  $\text{supp } \mu = X$  (Proposition 1.7.63),  $f: X \rightarrow X$  continuous  $\mu$ -preserving. Then  $\mathcal{R}(f) = R(f) = X$ . If  $X$  is connected, then  $f$  is chain-transitive.*

*Proof* Proposition 1.7.66(1) gives  $X = \text{supp } \mu \subset R(f) \subset \mathcal{R}(f) \subset X$ . Chain-components are clopen (Proposition 1.3.37), so there is only one by connectedness.  $\square$

Thus, having typical behavior with respect to an invariant measure is a probabilistic counterpart for recurrence, ergodicity is such for topological transitivity, and unique ergodicity is likewise for minimality. It is important to note that the converse to any of the statements of Proposition 1.7.66 is not true, even if we assume in addition that  $f$  is a diffeomorphism of a compact manifold.

The following subsections establish ergodicity of various classical examples.

### 1.7.10 Ergodicity of Translations

We begin our study of examples with instances of group translations, the first of which is a circle rotation. We can use multiplicative notation, representing the circle as the unit circle in the complex plane

$$S^1 = \{z \in \mathbb{C} \mid |z| = 1\} = \{e^{2\pi i \varphi} \mid \varphi \in \mathbb{R}\}$$

or additive notation, where  $S^1 = \mathbb{R}/\mathbb{Z}$  is the factor group of the additive group of real numbers modulo the subgroup of integers. The exponential map

$$\mathbb{R}/\mathbb{Z} \ni \varphi \mapsto e^{2\pi i \varphi} \in \{z \in \mathbb{C} \mid |z| = 1\}$$

establishes an isomorphism between these representations. There is a natural projection from  $\mathbb{R}$  to  $S^1$  defined by

$$\pi(x) = x + \mathbb{Z} \quad \text{or} \quad \pi(x) = e^{2\pi i x},$$

respectively. We denote by  $R_\alpha$  the rotation by angle  $2\pi\alpha$ , i.e.,

$$R_\alpha z = z_0 z \text{ with } z_0 = e^{2\pi i \alpha} \quad \text{or} \quad R_\alpha x = x + \alpha \pmod{1},$$

respectively. The iterates of the rotation are correspondingly

$$R_\alpha^n z = R_{n\alpha} z = z_0^n z \quad \text{or} \quad R_\alpha^n x = x + n\alpha \pmod{1}.$$

A crucial distinction appears between the cases of rational and irrational  $\alpha$ .

In the former case, write  $\alpha = p/q$ , where  $p, q$  are relatively prime integers. Then  $R_\alpha^q x = x$  for all  $x$  so  $R_\alpha^q$  is the identity map and after  $q$  iterates the transformation simply repeats itself. The latter case is much more interesting.

### 1.7.10.1 First Proof of Unique Ergodicity

Every rotation preserves Lebesgue measure.

**Proposition 1.7.68 (Kronecker–Weyl Equidistribution Theorem)** *Any irrational rotation is uniquely ergodic.*

*Proof* By Proposition 1.7.61 it is sufficient to check that time averages for every continuous function from a dense set of continuous functions uniformly converge to a constant. By the Weierstrass Theorem, trigonometric polynomials form a dense set among all continuous functions in the uniform topology. Furthermore, uniform convergence to a constant is a linear property; if  $\varphi$  and  $\psi$  have it, then so does  $a\varphi + b\psi$  for  $a, b \in \mathbb{R}$ . Thus, it is enough to check uniform convergence for any complete system of functions, e.g., for the characters (Definition 1.7.162)  $\chi_m(x) = e^{2\pi i m x}$ . For  $m = 0$  this is trivial. If  $m \neq 0$ , then

$$\chi_m(R_\alpha x) = e^{2\pi i m(x+\alpha)} = e^{2\pi i m \alpha} e^{2\pi i m x} = e^{2\pi i m \alpha} \chi_m(x)$$

and

$$\left| \frac{1}{n} \sum_{k=0}^{n-1} \chi_m(R_\alpha^k(x)) \right| \leq \left| \frac{1}{n} \sum_{k=0}^{n-1} e^{2\pi i m k \alpha} \right| = \frac{\overbrace{|1 - e^{2\pi i m n \alpha}|}^{\leq 2}}{n|1 - e^{2\pi i m \alpha}|} \xrightarrow{n \rightarrow \infty} 0.$$

□

This argument extends to any translation  $T_\gamma, x \mapsto x + \gamma$  on the torus where  $\gamma = (\gamma_1, \dots, \gamma_n)$  is such that  $m \in \mathbb{Z}^n \Rightarrow \langle m, \gamma \rangle = m_1 \gamma_1 + \dots + m_n \gamma_n \notin \mathbb{Z}$ , i.e.,  $\{1, \gamma_1, \dots, \gamma_n\}$  is rationally independent. In fact, this is necessary for topological transitivity and hence, since the support of Lebesgue measure is the whole torus, by Proposition 1.7.66(2), also for ergodicity of Lebesgue measure.

**Proposition 1.7.69** *A translation  $T_{(\gamma_1, \dots, \gamma_n)}: \mathbb{T}^n \rightarrow \mathbb{T}^n$  is uniquely ergodic if and only if  $\{1, \gamma_1, \dots, \gamma_n\}$  is rationally independent.*

### 1.7.10.2 Second Proof of Unique Ergodicity

An alternative proof of unique ergodicity for translations on the torus consists of two parts. First, we obtain ergodicity from a Fourier analysis argument.

**Proposition 1.7.70** *If  $\langle k, \gamma \rangle := \sum_{j=1}^n k_j \gamma_j \notin \mathbb{Z}$  for any  $k \in \mathbb{Z}^n \setminus \{0\}$ , then the translation  $T_\gamma$  is ergodic with respect to Lebesgue measure.*

*Proof* If  $\varphi \circ T_\gamma = \varphi: \mathbb{T}^n \rightarrow \mathbb{C}$  is a bounded measurable function, then it is in  $L^2$  and uniqueness of the Fourier expansion

$$\sum_{k \in \mathbb{Z}^n} \varphi_k e^{2\pi i \langle k, x \rangle} = \varphi(x) = \underbrace{\varphi(T_\gamma(x))}_{=\varphi(x+\gamma)} = \sum_{k \in \mathbb{Z}^n} \varphi_k e^{2\pi i \langle k, x+\gamma \rangle} = \sum_{k \in \mathbb{Z}^n} \varphi_k e^{2\pi i \langle k, \gamma \rangle} e^{2\pi i \langle k, x \rangle}$$

implies  $\varphi_k = \varphi_k e^{2\pi i \langle k, \gamma \rangle}$ , i.e., for every  $k$  either  $\varphi_k = 0$  or  $\langle k, \gamma \rangle \in \mathbb{Z}$ .

Unless  $\varphi \stackrel{\text{a.e.}}{=} \text{const.}$ , there is a  $k \neq 0$  with  $\varphi_k \neq 0$ , so  $\langle k, \gamma \rangle \in \mathbb{Z}$ . □

**Remark 1.7.71** Since the exponents are *characters* of the torus considered as a compact abelian group (Definition 1.7.162), the above statements and arguments easily translate to the general case of translations of compact abelian groups. This kind of argument is also useful for other dynamical systems of an algebraic nature, including the expanding maps  $E_m$  (Sect. 1.7.13) and hyperbolic toral automorphisms.

The second step consists of showing that ergodicity with respect to Lebesgue measure implies unique ergodicity. The special property of Lebesgue measure is that it is invariant with respect to *all* translations. The natural context for the argument is thus the multiplication transformation on compact abelian groups.

**Definition 1.7.72 (Haar Measure)** A topological group is said to be locally compact if every point (or equivalently, the identity) has a compact neighborhood. Such a group possesses a locally finite Borel measure invariant with respect to all right translations, which is unique up to a scalar multiple and called the *right Haar measure*. Similarly, the *left Haar measure* is, up to a scalar multiple, the unique measure invariant with respect to all left translations  $L_{g_0}: g \rightarrow g_0 g$ .

These measures are finite if and only if the group is compact. In many interesting cases right-invariant Haar measures are also left-invariant, e.g., when the group is abelian, compact, or, most importantly, a *unimodular linear group*, i.e., a closed subgroup of the group  $SL(n, \mathbb{R})$  of all  $n \times n$  matrices with determinant one. In general, groups for which the left and right Haar measures coincide (and naturally are simply called Haar measures) are said to be *unimodular*.

Let us mention a theorem that yields Haar measure and that can otherwise be useful when looking for invariant measures.

**Theorem 1.7.73 (Kakutani–Markov Fixed-Point Theorem)** *Let  $E$  be a locally convex topological vector space,  $G$  an equicontinuous group of linear maps of  $E$ ,  $K \subset E$  a nonempty  $G$ -invariant compact convex set. Then  $G$  has a fixed point  $p$  in  $K$ , i.e.,  $gp = p$  for all  $g \in G$ .*

Note that since  $\text{supp } \lambda_G \neq \emptyset$  (Proposition 1.7.63) is invariant under all translations,  $\lambda_G$  is positive on open sets. For the torus,  $\lambda_{\mathbb{T}^n}$  is the usual Lebesgue measure.

**Proposition 1.7.74** *If a translation  $L_{g_0}$  on a compact metrizable abelian group  $G$  is ergodic with respect to the Haar measure  $\lambda_G$ , then it is uniquely ergodic.*

*Proof* Let  $\mu$  be any  $L_{g_0}$ -invariant Borel probability measure. Then so is the pullback measure  $\mu_g: A \mapsto \mu(L_g A)$ :

$$\mu_g(L_{g_0} A) = \mu(L_g L_{g_0} A) = \mu(L_{g_0} L_g A) = \mu(L_g A) = \mu_g(A).$$

Since  $\mathfrak{M}(L_{g_0})$  is weak\*-closed and convex, we can average over any measurable set  $E$  of positive Haar measure to get an  $L_{g_0}$ -invariant measure

$$\mu_E: A \mapsto \frac{1}{\lambda_G(E)} \int_E \mu_g(A) d\lambda_G(g). \quad (1.31)$$

If  $E \cap F = \emptyset$  then

$$\lambda_G(E \cup F) \mu_{E \cup F} = \lambda_G(E) \mu_E + \lambda_G(F) \mu_F. \quad (1.32)$$

A change of variables in (1.31) shows that  $\mu_G$  is  $L_g$ -invariant for any  $g \in G$ ; hence  $\mu_G = \lambda_G$  by uniqueness of Haar measure.

If  $\mu \neq \lambda_G$ , then there exists a continuous function  $\varphi$  such that

$$\int_G \varphi d\mu \neq \int_G \varphi d\lambda_G = \int_G \varphi d\mu_G = \int_G \left( \int_G \varphi d\mu_g \right) d\lambda_G = \int_G \left( \int_G (\varphi \circ L_{g^{-1}}) d\mu \right) d\lambda_G.$$

$g \mapsto \bar{\varphi}_g := \int_G \varphi \circ L_{g^{-1}} d\mu$  is continuous and not constant since  $\bar{\varphi}_{\text{Id}} \neq \int_G \bar{\varphi}_g d\lambda_G$ . Thus we can find a number  $a$  such that  $\lambda_G(E) > 0$  and  $\lambda_G(F) > 0$ , where  $E = \{g \mid \bar{\varphi}_g \geq a\}$ ,  $F = G \setminus E$ . Then  $\int_G \varphi d\mu_E \geq a$  and  $\int_G \varphi d\mu_F < a$  so

$$\mu_E \neq \mu_F,$$

while (1.32) implies  $\lambda_G(E) \mu_E + \lambda_G(F) \mu_F = \mu_{E \cup F} = \mu_G = \lambda_G$ . So  $\lambda_G$  is not ergodic by Lemma 1.7.48.  $\square$

### 1.7.10.3 A Third Proof and an Application

We sketch a geometric proof that introduces ideas useful in the study of broad classes of dynamical systems, including those with no apparent algebraic structure.

Every measurable set on a small scale is densely concentrated; it fills some small balls or cubes almost completely and almost misses others because it can be approximated arbitrarily well (in measure) by finite collections of cubes. Fix an invariant set  $A$  and  $\epsilon > 0$  and find a small cube  $\Delta$  such that  $\lambda(A \cap \Delta) > (1 - \epsilon)\lambda(\Delta)$ . Images of  $\Delta$  under the iterates of our map have the same property since both  $\lambda$  and  $A$  are invariant. Since our map is an isometry, any image of  $\Delta$  is again a cube of the same size. By topological transitivity one can find a collection of images that cover the whole phase space almost uniformly, without much overlap. In fact it is sufficient to assume that every point is covered no more than  $N$  times, where  $N$  is

independent of  $\epsilon$ , because then the measure of  $A$  must be greater than  $1 - \epsilon N$ . Since  $\epsilon$  can be chosen arbitrarily small, this implies that  $A$  has full measure.

The uniform distribution of each orbit under Lebesgue measure lets us determine how often a given string of digits occurs as the initial string of digits of powers of  $k \in \mathbb{N}$ , i.e., how often  $p \leq k^n 10^{-i} < p + 1$  for some  $i \in \mathbb{N}$ :

**Proposition 1.7.75 (Benford–Newcomb Law)**<sup>34</sup> *If  $k \in \mathbb{N}$  is not a power of 10,  $p \in \mathbb{N}$ , and  $\lg = \log_{10}$  is the logarithm to base 10, then*

$$F(N) := \frac{1}{N} \text{card}\{0 \leq n < N \mid p \leq k^n 10^{-i} < p + 1 \text{ for some } i \in \mathbb{N}\} \xrightarrow{N \rightarrow \infty} \lg \frac{p + 1}{p}.$$

*Remark 1.7.76* For  $k = 2$  and  $p < 10$  this gives the statistics of the first digit of  $2^n$ . Over numbers with  $\lceil \lg p \rceil$  digits, these sum to 1:  $\sum_{i=\lceil \lg p \rceil}^{\lceil \lg p \rceil - 1} \lg \frac{i+1}{i} = \lceil \lg p \rceil - \lfloor \lg p \rfloor = 1$ .

*Proof* The salient event is  $\lg p \leq n \lg k - i < \lg(p + 1)$ . Subtract  $m := \lfloor \lg p \rfloor$  to get

$$0 \leq \lg \frac{p}{10^m} = \lg p - m \leq \underbrace{n \lg k - i - m}_{=\{n \lg k\}, \text{ fractional part}} < \lg(p + 1) - m = \lg \frac{p + 1}{10^m} \leq 1.$$

The rotation  $R_{\lg k}$  is ergodic,<sup>35</sup> so  $\lim_{N \rightarrow \infty} F(N) = \int \chi_{[\lg \frac{p}{10^m}, \lg \frac{p+1}{10^m})} = \lg \frac{p + 1}{p}$ .  $\square$

## 1.7.11 Circle Homeomorphisms

We can apply our insight into circle rotations beyond homogeneous systems because irrational rotations more generally represent the dynamics of many circle homeomorphisms. The Poincaré classification (Theorem 1.7.79) establishes how so,

<sup>34</sup>From Wikipedia: The discovery of Benford's law goes back to 1881, when the American astronomer Simon Newcomb noticed that in logarithm tables the earlier pages (that started with 1) were much more worn than the other pages. Newcomb's published result is the first known instance of this observation and includes a distribution on the second digit, as well. Newcomb proposed a law that the probability of a single number  $N$  being the first digit of a number was equal to  $\log(N + 1) - \log(N)$ .

The phenomenon was again noted in 1938 by the physicist Frank Benford, who tested it on data from 20 different domains and was credited for it. His data set included the surface areas of 335 rivers, the sizes of 3259 US populations, 104 physical constants, 1800 molecular weights, 5000 entries from a mathematical handbook, 308 numbers contained in an issue of Reader's Digest, the street addresses of the first 342 persons listed in American Men of Science and 418 death rates. The total number of observations used in the paper was 20,229.

<sup>35</sup>If  $\lg k = p/q$  then  $2^p 5^q = 10^p = k^q = 2^{mq} 5^{nq}$  using prime factorization. Then  $n = m$  and  $k = 10^m$ .

and for which circle homeomorphisms. The salient parameter is the average amount by which a point is being rotated (or translated) by the homeomorphism, called the *rotation number*.

The natural projection  $\pi: \mathbb{R} \rightarrow S^1 = \mathbb{R}/\mathbb{Z}, x \mapsto x + \mathbb{Z}$  provides a lift of a homeomorphism  $f: S^1 \rightarrow S^1$  to a homeomorphism  $F: \mathbb{R} \rightarrow \mathbb{R}$  with the property

$$f \circ \pi = \pi \circ F.$$

Such a lift  $F$  is unique up to an additive integer constant.

**Proposition 1.7.77** *Let  $f: S^1 \rightarrow S^1$  be an orientation-preserving homeomorphism and  $F: \mathbb{R} \rightarrow \mathbb{R}$  a lift of  $f$ . Then*

1.  $\rho(F) := \lim_{|n| \rightarrow \infty} \frac{1}{n}(F^n(x) - x)$  exists for all  $x \in \mathbb{R}$ ,
2. is independent of  $x$ , and
3. is well-defined up to an integer, i.e., if  $F_1, F_2$  are lifts of  $f$  then  $\rho(F_1) - \rho(F_2) = F_1 - F_2 \in \mathbb{Z}$ .
4.  $\rho(F^n) = n\rho(F)$ .
5.  $\rho(F) \in \mathbb{Q}$  if and only if  $f$  has a periodic point.
6. If  $\rho(F) \in \mathbb{Q}$  then all periodic orbits have the same period.
7. If  $h: S^1 \rightarrow S^1$  is an orientation-preserving homeomorphism then  $\rho(h^{-1}fh) = \rho(f)$ .

This justifies the following terminology:

**Definition 1.7.78**  $\rho(f) := \pi(\rho(F))$  is called the *rotation number* of  $f$ .

*Proof* 1. Take  $x \in \mathbb{R}$  and let  $x_n = F^n(x)$ ,  $a_n := x_n - x$ ,  $k := \lfloor a_n \rfloor$ . Then

$$\begin{aligned} a_{m+n} &= F^{m+n}(x) - x = F^m(x_n) - x_n + x_n - x \\ &= \underbrace{[F^m(x+k) - (x+k)]}_{=a_m} + \underbrace{[x_n - x]}_{=a_n} + \underbrace{[F^m(x_n) - F^m(x+k)]}_{\leq 1} - \underbrace{[x_n - x - k]}_{=a_n - \lfloor a_n \rfloor \geq 0} \\ &\leq a_m + a_n + 1, \end{aligned}$$

so Proposition 1.6.10 shows that  $a_n/n$  converges because

$$\frac{a_n}{n} = \frac{1}{n} \sum_{i=0}^{n-1} \underbrace{(F^{i+1}(x) - F^i(x))}_{=F(x_i) - x_i} \geq \min_{0 \leq y \leq 1} F(y) - y > -\infty.$$



2. Since  $f$  is an orientation-preserving homeomorphism,  $F(x+1) = F(x) + 1$ , and  $|F(y) - F(x)| < 1$  for  $x, y \in [0, 1)$ . Consequently

$$\left| \frac{1}{n} |F^n(x) - x| - \frac{1}{n} |F^n(y) - y| \right| \leq \frac{1}{n} (|F^n(x) - F^n(y)| + |x - y|) \leq \frac{2}{n}$$

and the rotation numbers of  $x$  and  $y$  coincide.

3.  $\rho(F + k) = \rho(F) + k$  for  $k \in \mathbb{Z}$   
 4.  $\rho(F^m) = \lim_{n \rightarrow \infty} \frac{1}{n} ((F^m)^n(x) - x) = m \lim_{n \rightarrow \infty} \frac{1}{mn} (F^{mn}(x) - x) = m\rho(F)$ .  
 5. If  $f$  has a  $q$ -periodic point  $\pi(x)$ , then  $F^q(x) = x + p$  for some  $p \in \mathbb{Z}$ , hence

$$\frac{F^{mq}(x) - x}{mq} = \frac{1}{mq} \sum_{i=0}^{m-1} F^q(F^{iq}(x)) - F^{iq}(x) = \frac{mp}{mq} = \frac{p}{q},$$

for  $m \in \mathbb{N}$ , and  $\rho(F) = p/q$ .

If  $\rho(F) = p/q \in \mathbb{Q}$ , then  $\rho(f^q) = \pi(\rho(F^q)) = \pi(q\rho(F)) = \pi(p) = 0$ , so, passing to  $f^q$ , it suffices to show that if  $\rho(f) = 0$  then  $f$  has a fixed point.

If  $f$  has no fixed point and  $F$  is a lift such that  $F(0) \in [0, 1)$ , then  $F(x) - x \in \mathbb{R} \setminus \mathbb{Z}$  for all  $x \in \mathbb{R}$ ,<sup>36</sup> so  $0 < F(x) - x < 1$  by the Intermediate-Value Theorem. Since  $F - \text{Id}$  is continuous on  $[0, 1]$  it attains its minimum and maximum and therefore there exists a  $\delta > 0$  such that

$$0 < \delta \leq F(x) - x \leq 1 - \delta < 1$$

for all  $x \in \mathbb{R}$  since  $F - \text{Id}$  is periodic. Taking  $x = F^i(0)$  and summing from  $i = 0$  to  $n - 1$  gives  $n\delta \leq F^n(0) \leq (1 - \delta)n$  or  $\delta \leq \frac{F^n(0)}{n} \leq 1 - \delta$ , so  $\delta < \rho(F) \leq 1 - \delta$ .

6. If  $\rho(f) = p/q$  with  $p, q \in \mathbb{Z}$  relatively prime and  $\pi(x)$  is periodic, then we need to show that there is a lift  $F$  of  $f$  for which  $F^q(x) = x + p$ . If  $F$  is a lift such that  $\rho(F) = \frac{p}{q}$ , then  $F^r(x) = x + s$  for some  $r, s \in \mathbb{Z}$ , and

$$\frac{p}{q} = \rho(F) = \lim_{n \rightarrow \infty} \frac{F^{nr}(x) - x}{nr} = \lim_{n \rightarrow \infty} \frac{ns}{nr} = \frac{s}{r},$$

so  $s = mp$  and  $r = mq$ . If  $F^q(x) - p > x$  then by monotonicity

$$\underbrace{F^r(x) - s}_{=F^{mq}(x)-mp} = \underbrace{F^{(m-1)q} \left( \overbrace{F^q(x) - p}^{> x} \right)}_{\geq F^{(m-1)q}(x) - (m-1)p} - (m-1)p \geq \cdots \geq F^q(x) - p > x,$$

contrary to the assumption, so  $F^q(x) - p \leq x$ . Similarly,  $F^q(x) - p \geq x$  as claimed.

<sup>36</sup>  $F(x) - x \in \mathbb{Z}$  implies that  $\pi(x)$  is a fixed point for  $f$ .

7. If  $F$  and  $H$  are lifts of  $f$  and  $h$ , respectively, i.e.,  $\pi F = f\pi$  and  $\pi H = h\pi$ , then

- (a)  $H^{-1}$  is a lift of  $h^{-1}$ :  $\pi H^{-1} = h^{-1}\pi$ .  $H^{-1}H = \text{id}$ ,  $H^{-1}H^{-1} = h^{-1}\pi H^{-1} = h^{-1}\pi$ .
- (b)  $H^{-1}FH$  is a lift of  $h^{-1}fh$ :  $\pi H^{-1}FH = h^{-1}\pi FH = h^{-1}f\pi H = h^{-1}fh\pi$ .

We need to estimate  $|H^{-1}F^nH(x) - F^n(x)| = |(H^{-1}FH)^n(x) - F^n(x)|$ .

- (a) If  $H(0) \in [0, 1) \ni x$ , then  $0 - 1 < H(x) - x < H(x) < H(1) < 2$  and by periodicity  $|H(x) - x| < 2$  for  $x \in \mathbb{R}$ .
- (b) Similarly,  $|H^{-1}(x) - x| < 2$  for  $x \in \mathbb{R}$ .
- (c) If  $|y - x| < 2$  then  $|F^n(y) - F^n(x)| < 3$  since  $|\lfloor y \rfloor - \lfloor x \rfloor| \leq 2$  and thus

$$\begin{aligned} -3 &\leq \lfloor y \rfloor - \lfloor x \rfloor - 1 = F^n(\lfloor y \rfloor) - F^n(\lfloor x \rfloor + 1) < F^n(y) - F^n(x) \\ &< F^n(\lfloor y \rfloor + 1) - F^n(\lfloor x \rfloor) = \lfloor y \rfloor + 1 - \lfloor x \rfloor \leq 3, \end{aligned}$$

$$\text{so } \frac{|H^{-1}F^nH(x) - F^n(x)|}{n} \leq \frac{\overbrace{|H^{-1}F^nH(x) - F^nH(x)|}^{<2} + \overbrace{|F^nH(x) - F^n(x)|}^{<3}}{n} \rightarrow 0. \quad \square$$

**Theorem 1.7.79 (Poincaré Classification Theorem [KaHa95, Theorem 11.2.7])**

For an orientation-preserving homeomorphism  $f: S^1 \rightarrow S^1$  without periodic points

- 1. if  $f$  is transitive then  $f$  is conjugate to the rotation  $R_{\rho(f)}$ ,
- 2. otherwise,  $h \circ f = R_{\rho(f)} \circ h$  for a noninvertible continuous monotone map  $h: S^1 \rightarrow S^1$ .

This implies

**Theorem 1.7.80** A circle homeomorphism without periodic points is uniquely ergodic and measure-theoretically isomorphic to an irrational rotation.

**Definition 1.7.81** We say that a map  $f$  of a measure space  $(X, \mu)$  is *nonsingular* and that  $\mu$  is quasi-invariant under  $f$  if  $\mu(A) = 0$  if and only if  $\mu(f^{-1}(A)) = 0$ . Equivalently the pullback  $f_*\mu$  is equivalent to  $\mu$  and hence, by the Radon–Nikodym Theorem, is given by a positive density.

For a topologically transitive homeomorphism (and hence for a minimal one) there is no open set that is wandering, i.e., with pairwise disjoint images, and by the Poincaré Recurrence Theorem 1.7.11 there is also no set of positive measure of this kind—so long as the measure is invariant and finite. However, when a probability measure is only *quasi*-invariant, then there might be sets of the latter type, and we presently demonstrate something even stronger for Lipschitz-continuous circle maps and Lebesgue measure.

**Proposition 1.7.82 (Kodama–Matsumoto)** There is a minimal uniquely ergodic bi-Lipschitz homeomorphism  $f$  of  $S^1$  with a measurable fundamental domain  $A$ :  $f^n(A) \cap f^m(A) = \emptyset$  for  $n \neq m$ , and  $S^1 \setminus \bigcup_{i \in \mathbb{Z}} f^i(A)$  is a Lebesgue null set.

*Remark 1.7.83* Clearly,  $\mu(A) = 0$  for any *invariant* Borel probability measure; the interest here lies in the fact that this can happen for the natural *quasi-invariant* measure. Lebesgue measure is here nonergodic in a strong sense; for instance, *any* Borel function on  $A$  can be extended to an  $f$ -invariant measurable function.

One can arrange for  $f$  to be topologically conjugate to any given irrational rotation, but we choose a Diophantine one to avert the need for a more careful construction of the set in question. A number  $\alpha$  is said to be *Diophantine* of type  $(c, d)$  if for any nonzero  $p, q \in \mathbb{Z}$  we have  $|q\alpha - p| > cq^{-d}$ .

*Proof* For the rotation  $R := R_\varphi$  with  $\varphi$  Diophantine of type  $(c, 2)$  for some  $c > 0$  we choose as a fundamental domain the Cantor set

$$C := \left\{ \sum_{i \in \mathbb{N}} \frac{\omega_i}{2^{4^i}} \quad \text{with } \omega \in \{0, 1\}^{\mathbb{N}} \right\}.$$

To check disjointness, write a point of  $R^n(C) \cap R^m(C)$  as  $x + n\varphi = y + m\varphi$  with  $x, y \in C$  and hence

$$(n - m)\varphi = y - x \in C - C = \left\{ \sum_{i \in \mathbb{N}} \frac{\beta_i}{2^{4^i}} \quad \text{with } \beta \in \{0, 1, -1\}^{\mathbb{N}} \right\}.$$

With partial sums as approximants  $\frac{p}{q}$  we have  $q = 2^{4^{k-1}}$  in

$$|(n - m)\varphi - \frac{p}{q}| = \left| \sum_{i \geq k} \frac{\beta_i}{2^{4^i}} \right| \leq \sum_{i \geq k} \frac{1}{2^{4^i}} \leq \sum_{i \geq 4^k} \frac{1}{2^i} = \frac{2}{2^{4^k}} = \frac{2}{q^{4^k/4^{k-1}}} = \frac{2}{q^4}.$$

This implies  $n = m$  because no nontrivial integer multiple of  $\varphi$  has this property. Now let  $\mu_0$  be a nonatomic probability measure supported on  $C$ . For instance, define  $\mu_0(S^1 \setminus C) = 0$  and

$$\mu_0 \left( \left[ \sum_{i=1}^k \frac{\omega_i}{2^{4^i}}, \sum_{i=1}^k \frac{\omega_i}{2^{4^i}} + \sum_{i > k} \frac{1}{2^{4^i}} \right] \right) = 2^{-k}.$$

$\mu := \frac{1}{2} \sum_{i \in \mathbb{Z}} 3^{-|i|} R_*^i \mu_0$  is a probability measure ( $\sum_{i \in \mathbb{Z}} 3^{-|i|} = 2$ ) with full support ( $\bigcup_{i \in \mathbb{Z}} R^i C$  is dense) and quasi-invariant:  $\left[ \frac{dR_*^{-1} \mu}{d\mu} \right] = \frac{1}{3} \chi_{\bigcup_{i \geq 0} R^i(C)} + 3 \chi_{\bigcup_{i < 0} R^i(C)} \in L^\infty(\mu)$ . Take  $A := h^{-1}(C)$  with  $h(\mu([0, y])) := y$  (a homeomorphism with  $h_* \lambda = \mu$ , where  $\lambda$  is Lebesgue measure); the homeomorphism  $f := h^{-1} \circ R \circ h$  is bi-Lipschitz because

$$\left[ \frac{df_*^{-1} \lambda}{d\lambda} \right] = \left[ \frac{dR_*^{-1} \mu}{d\mu} \right] \circ h = \frac{1}{3} \chi_{\bigcup_{i \geq 0} f^i(A)} + 3 \chi_{\bigcup_{i < 0} f^i(A)} \in L^\infty(\lambda).$$

□

### 1.7.12 Extensions of Rotations

We now describe a class of examples closely connected to rotations.

The first instance of these should be classified as elliptic, and it contains minimal nonergodic examples. (The second instance appears in (1.33).)

**Proposition 1.7.84** *Consider a map  $f: (x, y) \mapsto (x + \alpha, y + \varphi(x))$  of  $\mathbb{T}^2$  with  $\alpha \in \mathbb{R} \setminus \mathbb{Q}$  and  $\varphi: S^1 \rightarrow \mathbb{R}$ . If  $\varphi(x) = \Phi(x + \alpha) - \Phi(x)$  for some Lebesgue measurable function  $\Phi: S^1 \rightarrow \mathbb{R}$  then for any ergodic invariant measure  $f$  is measure-theoretically isomorphic to the rotation  $R_\alpha$  and there are uncountably many different ergodic invariant measures.*

*Proof* Take  $h(x, y) = (x, y + \Phi(x))$ . Then  $h^{-1} \circ f \circ h(x, y) = (x + \alpha, y)$ . Since the rotation is uniquely ergodic any invariant measure for  $f$  projects to Lebesgue measure on the circle and hence  $h$  defines a measure-theoretic isomorphism for any such measure. Thus the invariant ergodic measures for  $f$  are exactly the measures induced from measures on circles. There are uncountably many of these because the graph of  $\Phi + c$  for any  $c \in \mathbb{R}$  supports such a measure.  $\square$

**Proposition 1.7.85** *Consider the torus  $\mathbb{T}^2$ , a function  $\varphi: S^1 \rightarrow \mathbb{R}$ , and a map  $f: (x, y) \mapsto (x + \alpha, y + \varphi(x))$  of  $\mathbb{T}^2$ . Then either  $\varphi(x) = \Phi(x + \alpha) - \Phi(x) + r_1\alpha + r_2$  for some continuous  $\Phi: S^1 \rightarrow \mathbb{R}$  and  $r_1, r_2 \in \mathbb{Q}$ , or  $f$  is minimal.*

*Remark 1.7.86* If  $\Phi$  is continuous and  $(r_1, r_2) = 0$ , then  $h$  is a topological conjugacy to  $R_\alpha \times \text{Id}$  on  $\mathbb{T}^2$ .

*Proof* One can show that there is an invariant minimal set  $M$  for  $f$  and the projection of this set to the first coordinate is invariant, hence is  $S^1$ . Consider the intersection of  $M$  with the fiber  $\{x\} \times S^1$ . We show that if it contains two points  $y$  and  $y + \tau$  then it is invariant under translation by  $\tau$  in the fiber. Namely, by minimality there exist points  $z_i = f^{N_i}(x, y) \xrightarrow{i \rightarrow \infty} (x, y + \tau)$ , so the points  $f^{kN_i}(x, y) \xrightarrow{i \rightarrow \infty} (x, y + k\tau)$ , which are hence in  $M$ . Thus the closed set  $M \cap (\{x\} \times S^1)$  is either a coset of a finite subgroup of  $\{x\} \times S^1$  generated by  $r \in \mathbb{Q}$  or equal to  $\{x\} \times S^1$ . Since  $M$  is closed the same case occurs for all  $x$  and by continuity we obtain the same subgroup for all  $x$ , hence giving either minimality or a collection of invariant closed curves for  $f$ .

In the first case we factor the second coordinate modulo  $1/q$ . Thus in this factor the set intersects every vertical exactly once and is hence the graph of a continuous function. On the universal cover it lifts to the graph of a function  $\Phi'$  with  $\Phi'(x + 1) = \Phi'(x) + k$  and all its integer translates. Invariance under the lift  $F$  yields  $(x + \alpha, \Phi'(x) + \varphi(x)) = F(x, \Phi'(x)) = (x + \alpha, \Phi'(x + \alpha) + n)$  for some  $n \in \mathbb{Z}$ . Thus we obtain  $\varphi(x) = \Phi'(x + \alpha) - \Phi'(x) + n$ . Recalling that we factored the second coordinate by  $1/q$  and writing  $\Phi'(x) = \Phi(x) + kx$  we obtain Proposition 1.7.85 with  $r_1 = k/q$  and  $r_2 = n/q$ .  $\square$

One interesting application of the preceding two results is that a circle extension of the above form is a minimal nonergodic diffeomorphism if one can write  $\varphi(x) =$

$\Phi(x + \alpha) - \Phi(x)$  for some measurable  $\Phi$  but not  $\varphi(x) = \Phi(x + \alpha) - \Phi(x) + r$  for any continuous  $\Phi$ .

By taking  $\varphi(x) = x$  above, we now consider a map of the two-torus  $\mathbb{T}^2$  that is somewhat similar in form to a translation, can be analyzed by similar methods, shares some common features (e.g., minimality) but also exhibits different features. The map, which depends on a parameter  $\alpha$ , has the form

$$A_\alpha(x, y) = (x + \alpha, y + x) \pmod{1}. \quad (1.33)$$

It is a prototype of smooth dynamical systems on compact manifolds with *parabolic behavior*. Similarly to translations the map  $A_\alpha$  is “integrable” in the sense that there is a closed formula for its iterates:

$$A_\alpha^n(x, y) = (x + n\alpha, y + nx + n(n-1)\alpha/2) \pmod{1}.$$

This map is an example of a *skew product*. The evolution of the first coordinate depends only on itself. The partition of the torus into circles  $x = \text{const.}$  is invariant under  $A_\alpha$ , and if one identifies each circle with the corresponding value of  $x$ , the elements of this partition are mapped according to the rotation  $R_\alpha$ . The difference with the Cartesian product is that the way in which each element maps to its image changes from one element to another.

Unlike toral translations, these maps are not isometries. They have a suitably weakened property related to the *relative behavior* of orbits.

**Definition 1.7.87 (Distality)** A homeomorphism  $f$  of a compact metric space is *distal* if for  $x \neq y$  there exists a  $\delta > 0$  such that  $d(f^n(x), f^n(y)) > \delta$  for all  $n \in \mathbb{Z}$ . Obviously a map is distal if it is an isometry, or more generally, if the collection of its iterates is equicontinuous. Skew-products provide more interesting (and typical) examples of a topologically transitive distal maps.

**Proposition 1.7.88** *The affine map  $A_\alpha$  from (1.33) of the torus is distal.*

*Proof* If  $p = (x, y), p' = (x', y') \in \mathbb{T}^2$  then there are 2 cases.

- $x = x'$ :  $d(A_\alpha^n(p), A_\alpha^n(p')) = |y - y'| =: \delta$ .
- $x \neq x'$ :  $d(A_\alpha^n(p), A_\alpha^n(p')) \geq |(x + n\alpha) - (x' + n\alpha)| = |x - x'| =: \delta$ .

□

**Proposition 1.7.89** *If  $\alpha \notin \mathbb{Q}$  then  $A_\alpha$  is minimal.*

*Proof* We show that this follows from topological transitivity, which by Proposition 1.7.66 is a consequence of Proposition 1.7.90 below. We argue by contraposition. Assume  $A_\alpha$  is not minimal, i.e., for some point  $(x, y) \in T^2$  the closure of its orbit under the iterates of  $A_\alpha$  is not the whole torus. The map  $A_\alpha$  commutes with every “vertical” translation  $T_{(0,s)}$  and hence the orbit closure of any point  $(x, y + s)$  is the translation of the orbit closure of  $(x, y)$  and is also not dense. Now consider the union  $X$  of such orbit closures for all  $s$ . This set is  $A_\alpha$ -invariant

and closed and consists of whole vertical circles. No point in  $X$  has dense orbit (each is in a nondense invariant orbit closure). If  $X = \mathbb{T}^2$  this implies that  $A_\alpha$  is not topologically transitive. If  $X \neq \mathbb{T}^2$ , then  $X$  projects onto a closed invariant proper subset of the rotation  $R_\alpha$ , so  $R_\alpha$  is not minimal and hence  $\alpha \in \mathbb{Q}$ .  $\square$

**Proposition 1.7.90** *The affine map  $A_\alpha: \mathbb{T}^2 \rightarrow \mathbb{T}^2$ ,  $A_\alpha(x, y) = (x + \alpha, y + x) \pmod{1}$  is ergodic with respect to Lebesgue measure if  $\alpha \notin \mathbb{Q}$ : bounded measurable  $A_\alpha$ -invariant functions  $\varphi: \mathbb{T}^n \rightarrow \mathbb{C}$  are constant almost everywhere.*

*Proof* Write the Fourier decomposition of  $\varphi$  as

$$\varphi(x, y) = \sum_{(m,n) \in \mathbb{Z}^2} \varphi_{m,n} \exp(2\pi i(mx + ny)).$$

The invariance condition  $\varphi(x, y) = \varphi(A_\alpha(x, y))$  implies

$$\varphi_{m+n,n} = \exp(2\pi i m \alpha) \varphi_{m,n}. \quad (1.34)$$

If  $n = 0$  then  $\varphi_{m,0} = \exp(2\pi i m \alpha) \varphi_{m,0}$ , so  $\varphi_{m,0} = 0$  for  $m \neq 0$  since  $\alpha \notin \mathbb{Q}$ .

If  $n \neq 0$  then (1.34) gives infinitely many different Fourier coefficients with the same absolute value; since  $\varphi$  is bounded and hence in  $L^2$ , this absolute value is 0. Thus  $\varphi_{m,n} = 0$  unless  $m = n = 0$ , so  $\varphi$  is constant a.e.  $\square$

**Proposition 1.7.91** *The affine map  $A_\alpha: \mathbb{T}^2 \rightarrow \mathbb{T}^2$ ,  $A_\alpha(x, y) = (x + \alpha, y + x) \pmod{1}$  is uniquely ergodic if  $\alpha \notin \mathbb{Q}$ .*

*Proof* We use the fact that  $A_\alpha$  commutes with all vertical translations  $T_{(0,t)}$  and apply the method of Proposition 1.7.74 although with some modifications to show another version of the argument.

We will use the same letter  $\lambda$  for Lebesgue measures on the torus, on the circle and on each fiber  $\{x\} \times S^1$ . Every ergodic invariant measure  $\mu$  for  $A_\alpha$  projects to the Lebesgue measure  $\lambda$  in the  $x$ -coordinate. The pullback  $\mu_t := T_{(0,t)}^* \mu$  is also ergodic and  $A_\alpha$ -invariant as is the average  $\mu_E = \frac{1}{\lambda(E)} \int_E \mu_t dt$  for any Lebesgue measurable set  $E \subset S^1$ . Note that for every set of positive measure  $\mu_E$  is absolutely continuous because it projects to  $\lambda$  in the first coordinate and its conditionals are convolutions of the conditional measure for  $\mu$  with an absolutely continuous measure  $\chi_E \lambda$ . Hence  $\mu_E$  is absolutely continuous invariant measure and its density with respect to Lebesgue measure must be invariant, hence constant. Thus  $\mu_E = \lambda$ . Assume that  $\mu \neq \lambda$ . Since for any continuous function  $\varphi$  the integral  $\int_{\mathbb{T}^2} \varphi d\mu_t$  is continuous in  $t$  one can find a  $\varphi$  and an interval  $I$  around zero such that

$$\frac{1}{\lambda(I)} \int_I \varphi d\mu_t dt \neq \int_{\mathbb{T}^2} \varphi dx dy$$

and hence  $\mu_I \neq \lambda$ , a contradiction.  $\square$

Similarly to Proposition 1.7.75 (as well as Corollaries 1.7.97, 1.7.110 and 1.7.111 below), this has number-theoretic implications.

**Proposition 1.7.92** *If  $\alpha \notin \mathbb{Q}$  then the fractional part of any quadratic polynomial  $\alpha n^2 + \beta n + \gamma$  on  $\mathbb{Z}$  is uniformly distributed on  $[0, 1]$ , i.e.,  $[a, b) \subset [0, 1] \Rightarrow$*

$$\frac{1}{n} \text{card}\{0 \leq i < n \mid \alpha i^2 + \beta i + \gamma - \lfloor \alpha i^2 + \beta i + \gamma \rfloor \in [a, b)\} \xrightarrow{n \rightarrow \infty} b - a.$$

*Proof*  $A_{2\alpha}^n(\beta + \alpha, \gamma) = (\beta + (4n + 1)\alpha, \alpha n^2 + \beta n + \gamma)$  is equidistributed on  $\mathbb{T}^2$  by Proposition 1.7.91 and Corollary 1.7.59, and so is the second coordinate on  $S^1$ .  $\square$

### 1.7.13 Ergodicity of Expanding Maps and Toral Automorphisms

Now we study the linear expanding map  $E_m: S^1 \rightarrow S^1$ ,  $x \mapsto mx \pmod{1}$  for  $m \neq 0$ . Algebraically this is endomorphism of the group  $S^1 = \mathbb{R}/\mathbb{Z}$  onto itself. Geometrically it is an  $m$ -fold cover of  $S^1$ . Writing  $P_n(f) := \text{card Fix}(f^n)$ , we have

**Proposition 1.7.93** *If  $m \geq 2$  then  $P_n(E_m) = m^n - 1$  and periodic points of  $E_m$  are dense in  $S^1$ .*

*Proof*  $E_m^n(z) = z \Leftrightarrow z^{m^n} = z \Leftrightarrow z^{m^n-1} = 1$ , so  $\text{Fix}(E_m^n)$  consists of the  $m^n - 1$  roots of unity of order  $m^n - 1$ , hence is  $\frac{1}{m^n-1}$ -dense.  $\square$

$E_m$  preserves Lebesgue measure  $\lambda$  because the preimage of any interval of length  $l$  consists of  $|m|$  disjoint intervals of length  $l/m$ .

**Proposition 1.7.94** *Lebesgue measure is ergodic for  $E_m$  with  $|m| \geq 2$ .*

We give two proofs of this fact which are related to our second and third proofs for the toral translations.

*First Proof* Let  $\varphi$  be a measurable bounded  $E_m$ -invariant function. Using the Fourier expansion

$$\varphi(x) = \sum_{k \in \mathbb{Z}} \varphi_k \exp(2\pi i k x)$$

we obtain  $\varphi(E_m(x)) = \sum_{k \in \mathbb{Z}} \varphi_k \exp(2\pi i k m x)$ . Since  $\varphi(x) \stackrel{\text{a.c.}}{=} \varphi(E_m(x))$ , we have

$$\varphi_k = \varphi_{k \cdot m}, \quad m \in \mathbb{N}.$$

Since  $\varphi \in L^1$  and hence  $|\varphi_k| \xrightarrow{k \rightarrow \infty} 0$ , this implies  $\varphi_k = 0$  for  $k \neq 0$  and  $\varphi \stackrel{\text{a.c.}}{=} \varphi_0$ .  $\square$

*Second Proof* Let  $A \subset S^1$  be a measurable  $E_m$ -invariant set of positive Lebesgue measure.  $E_m^{-1}(A) = A$  implies forward-invariance of  $S^1 \setminus A = E_m(S^1 \setminus A)$ . As in the

third proof for the toral translations, fix  $\epsilon > 0$  and find an open interval  $\Delta$  of length  $|m|^{-n}$  for some  $n$  such that

$$\lambda(\Delta \setminus A) > (1 - \epsilon)\lambda(\Delta) = (1 - \epsilon)|m|^{-n}.$$

Since  $E_m$  has constant derivative it expands the Lebesgue measure of any set on which it is injective exactly  $|m|$  times. Thus  $\lambda(E_m^n(\Delta) \setminus A) = |m|^n \lambda(\Delta \setminus A) > 1 - \epsilon$ .  $\square$

The first proof of Proposition 1.7.94 adapts to give

**Proposition 1.7.95** *Consider  $A \in \text{GL}(m, \mathbb{Z})$ , i.e., an  $m \times m$ -matrix with integer entries and determinant  $\pm 1$ , and assume that no eigenvalue of  $A$  is a root of unity. Then the toral automorphism  $F_A: \mathbb{T}^m \rightarrow \mathbb{T}^m$  induced by  $A$  is ergodic.*

*Remark 1.7.96* Concrete such examples are given by Examples 1.1.1 and 1.1.3. The converse holds as well.

*Proof* Let  $\varphi$  be a measurable bounded  $F_A$ -invariant function. Expressing invariance through the Fourier expansion gives

$$\sum_{k \in \mathbb{Z}^m} \varphi_k \exp(2\pi i \langle k, x \rangle) = \varphi(x) \stackrel{\text{a.e.}}{=} \varphi(F_A(x)) = \sum_{k \in \mathbb{Z}^m} \varphi_k \exp(2\pi i \underbrace{\langle k, Ax \rangle}_{= \langle A^t k, x \rangle}).$$

Uniqueness of the Fourier expansion then gives

$$\varphi_k = \varphi_{(A^t)^n k} \quad \text{for } n \in \mathbb{N}.$$

If  $k \neq 0$ , then the  $(A^t)^n k$  (for  $n \in \mathbb{Z}$ ) are pairwise distinct since  $A$ , and hence  $A^t$ , has no roots of unity as eigenvalues. Thus, there are infinitely many  $l \in \mathbb{Z}^m$  with  $\varphi_l = \varphi_k$ . But  $\varphi \in L^1 \Rightarrow |\varphi_l| \xrightarrow{|l| \rightarrow \infty} 0$ , so  $\varphi_k = 0$  when  $k \neq 0$ , and  $\varphi \stackrel{\text{a.e.}}{=} \varphi_0$ .  $\square$

**Corollary 1.7.97 (of Proposition 1.7.94)** *Lebesgue-a.e.  $x \in \mathbb{R}$  is normal with respect to any base  $b > 1$ , i.e., each digit is equally frequent in the base- $b$  expansion.*

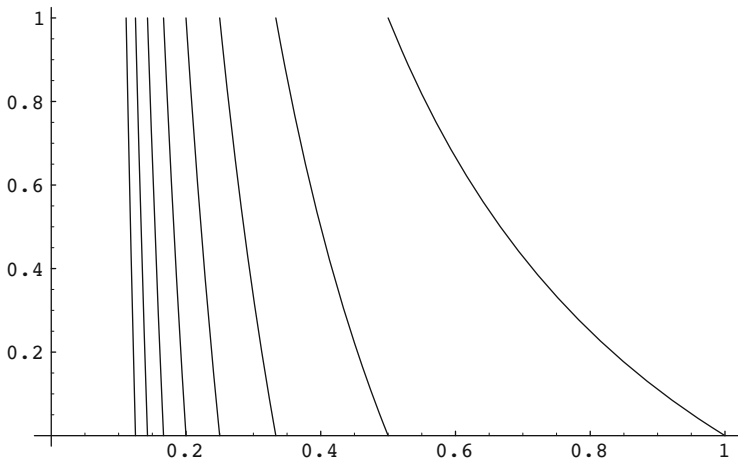
*Proof* The base- $b$  digit  $a < b$  appears in the  $k$ th place iff  $E_b^k(x) \in [\frac{a}{b}, \frac{a+1}{b})$ . By the Birkhoff Ergodic Theorem, the fraction of those  $k$  tends to  $\frac{a+1}{b} - \frac{a}{b} = 1/b$ .  $\square$

## 1.7.14 The Gauss Map

**Definition 1.7.98** The map  $G: [0, 1] \rightarrow [0, 1)$  defined by  $x \mapsto \{1/x\}$  (fractional part) is called the *Gauss map*.

The special importance of this map lies in its close connection with the classical continued-fraction algorithm, analogously to the way the doubling map is connected to binary expansion. The Gauss map has jump discontinuities at  $1/i$  for  $i \in \mathbb{N}$  which





**Fig. 1.12** The Gauss map (©Cambridge University Press, reprinted from [HaKa03] with permission)

can be removed if one identifies the end-points of the interval and makes it into a circle, and a discontinuity without right-sided limit at 0. Projecting to the circle  $S^1 = \mathbb{R}/\mathbb{Z}$  would fix the discontinuities away from zero, but as a dynamical system this map is most naturally defined on the interval. The restrictions  $x \mapsto \frac{1}{x} - i$  of  $G$  to the intervals  $(\frac{1}{i+1}, \frac{1}{i}]$  of continuity of  $G$  are called the *branches* of  $G$ . It is apparent that  $\min(G^2)' > 1$ , so the Gauss map (or, more precisely its square) can be viewed as expanding although it is not everywhere differentiable. In fact, we have a stronger statement (Fig. 1.12).

**Lemma 1.7.99**  $(G^2)' > 4$  wherever defined.

*Proof*  $G'(x) \leq -1$  wherever defined, and if  $x < 1/2$  then  $G'(x) = -1/x^2 < -4$ , so if  $x < 1/2$  or  $G(x) < 1/2$  then  $(G^2)'(x) = G'(G(x))G'(x) > 4$ . Otherwise,  $x > 1/2$  and  $G(x) > 1/2$ , so  $G^2(x) = \frac{1}{(1/x)-1} - 1 = -\frac{1}{1-x}$  and  $(G^2)'(x) = (1-x)^{-2} > 4$ .  $\square$

*Remark 1.7.100* Observe that the condition  $x > 1/2$  and  $G(x) > 1/2$  defines the longest interval of continuity of  $G^2$ , which is  $(1/2, 2/3)$  and hence has length  $1/6 < 1/4$ . Since the longest interval of continuity of  $G$  has length  $1/2$ , Lemma 1.7.99 shows that the intervals of continuity of  $G^n$  have length at most  $2^{-n}$ .

Analogously to Proposition 1.7.93 we obtain

**Proposition 1.7.101**  $P_n(G) = \infty$  and periodic points of  $G$  are dense in  $(0, 1]$ .

*Proof* Each branch of  $G$  and hence of  $G^n$  is onto, so by the Intermediate-Value Theorem  $G^n$  has a fixed point in each interval of continuity. Thus, by Remark 1.7.100, the union of these is dense, and  $P_n(G) = \infty$ .  $\square$

That  $G(x) = \frac{1}{x} - a$  implies  $a = \lfloor 1/x \rfloor$  and  $x = \frac{1}{a + G(x)}$  leads to an explicit description of periodic points. Fixed points satisfy  $x = \frac{1}{a + x}$ . Likewise,  $x = \frac{1}{a_1 + \frac{1}{a_2 + G^2(x)}}$  implies that a period-2 point satisfies  $x = \frac{1}{a_1 + \frac{1}{a_2 + x}}$ .

Generally:

**Proposition 1.7.102** *A period- $n$  point satisfies the equation*

$$x = \frac{1}{a_1 + \frac{1}{\ddots + \frac{1}{a_n + x}}} \quad (1.35)$$

and hence is of the form

$$x = \frac{1}{a_1 + \frac{1}{a_2 + \dots}} := \lim_{m \rightarrow \infty} \frac{1}{a_1 + \frac{1}{\ddots + \frac{1}{a_m}}}$$

with  $a_{i+n} = a_i$  for all  $i \in \mathbb{N}$ .

*Proof* That the limit exists follows from Remark 1.7.100 because the local inverse

$$F_{a_1 \dots a_n}: [0, 1] \rightarrow [0, 1], \quad t \mapsto \frac{1}{a_1 + \frac{1}{\ddots + \frac{1}{a_n + t}}} \quad (= t \text{ if } n = 0) \quad (1.36)$$

of  $G^n$  parametrizes the closure  $I_{a_1 \dots a_n}$  of an interval of continuity of  $G^n$ , and the lengths of these go to 0 as  $n \rightarrow \infty$ .  $\square$

**Remark 1.7.103** By (1.35), every periodic point is a *quadratic irrationality*, i.e., it satisfies a quadratic equation with integer coefficients because  $a + \frac{bx+c}{dx+e}$  is a ratio of linear expressions in  $x$  and hence so is its reciprocal:

$$\frac{1}{a + \frac{bx+c}{dx+e}} = \frac{1}{\frac{(ad+b)x+ae+c}{dx+e}} = \frac{dx+e}{(ad+b)x + (ae+c)}$$

recursively leads from (1.35) to a quadratic equation of the form  $x = \frac{ax+b}{cx+d}$ .

The branch of  $G$  on  $\Delta_n := \left(\frac{1}{n+1}, \frac{1}{n}\right)$  is  $x \mapsto \frac{1}{x} - n$  with  $n = \left\lfloor \frac{1}{x} \right\rfloor \in \mathbb{N}$ , so:

**Theorem 1.7.104 (Continued-Fraction Representation)** For  $\alpha \in \mathbb{R}$  define  $(a_i)_{i=0}^\infty$  and  $(\alpha_i)_{i=1}^\infty$  recursively by

$$a_0 := \lfloor \alpha \rfloor, \quad \alpha_1 := \{\alpha\}, \quad a_i := \left\lfloor \frac{1}{\alpha_i} \right\rfloor, \quad \alpha_{i+1} := \left\{ \frac{1}{\alpha_i} \right\}$$

and set

$$\frac{p_n}{q_n} := a_0 + \frac{1}{a_1 + \frac{1}{\ddots + \frac{1}{a_n}}} \quad (1.37)$$

in lowest terms and with  $q_n > 0$ . ( $\frac{p_n}{q_n}$  are called the convergents.) If  $\alpha \in \mathbb{Q}$  then the recursion terminates with  $\alpha_{i+1} = 0$  for some  $i$  (because  $\{\alpha_i\}$  is a proper fraction and hence  $\alpha_{i+1} = \left\{ \frac{1}{\alpha_i} \right\}$  has smaller numerator), and

$$\alpha = a_0 + \frac{1}{a_1 + \frac{1}{\ddots + \frac{1}{a_i}}}.$$

Otherwise,  $\alpha = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \dots}} := \lim_{n \rightarrow \infty} a_0 + \frac{1}{a_1 + \frac{1}{\ddots + \frac{1}{a_n}}}$ .

**Proposition 1.7.105**  $(p_n)_{n \geq -1}$  and  $(q_n)_{n \geq -1}$  both satisfy the two-step recursion

$$x_{n+1} = x_{n-1} + a_{n+1}x_n$$

with initial values  $p_{-1} = 1, q_{-1} = 0, p_0 = a_0, q_0 = 1$ .

*Proof* We establish this by defining  $p_{-1} = 1, q_{-1} = 0, p_0 = a_0, q_0 = 1$ ,

$$p_{n+1} = p_{n-1} + a_{n+1}p_n \quad \text{and} \quad q_{n+1} = q_{n-1} + a_{n+1}q_n,$$

and by inductively showing that  $p_n$  and  $q_n$

1. are relatively prime and
2. satisfy (1.37).

$p_n$  and  $q_n$  are coprime because  $p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1}$  (so any common divisor of  $p_n$  and  $q_n$  divides  $-1$ ). For  $n = 0$  this follows from  $p_{-1} = 1$ ,  $q_{-1} = 0$ ,  $q_0 = 1$ , and the induction step is

$$p_{n+1} q_n - p_n q_{n+1} = (p_{n-1} + a_{n+1} p_n) q_n - p_n (q_{n-1} + a_{n+1} q_n) = -(p_n q_{n-1} - p_{n-1} q_n).$$

$p_n$  and  $q_n$  satisfy (1.37) because  $F_{a_1 \dots a_n}$  from (1.36) satisfies

$$a_0 + F_{a_1 \dots a_n}(1/a_{n+1}) = \frac{p_{n-1} + a_{n+1} p_n}{q_{n-1} + a_{n+1} q_n} = \frac{p_{n+1}}{q_{n+1}},$$

which is the special case  $t = 1/a_{n+1}$  of

$$a_0 + F_{a_1 \dots a_n}(t) = \frac{p_n + t p_{n-1}}{q_n + t q_{n-1}} \quad \text{for } t \in [0, 1). \quad (1.38)$$

This follows inductively from  $F_{a_1 \dots a_{n+1}}(t) = F_{a_1 \dots a_n}(\frac{1}{a_{n+1} + t})$ : It is clear for  $n = 0$  because  $\frac{p_n + t p_{n-1}}{q_n + t q_{n-1}} = \frac{a_0 + t}{1 + 0}$ , and if  $n$  is such that (1.38) holds, then

$$\begin{aligned} a_0 + F_{a_1 \dots a_{n+1}}(t) &= a_0 + F_{a_1 \dots a_n}\left(\frac{1}{a_{n+1} + t}\right) = \frac{p_n + \frac{1}{a_{n+1} + t} p_{n-1}}{q_n + \frac{1}{a_{n+1} + t} q_{n-1}} = \frac{p_{n+1} + t p_n}{q_{n+1} + t q_n}. \\ &= \frac{p_{n-1} + a_{n+1} p_n + t p_n}{q_{n-1} + a_{n+1} q_n + t q_n} \end{aligned}$$

□

**Corollary 1.7.106**  $q_{n+1} \geq q_n \geq 2^{(n-1)/2}$ .

*Proof* Proposition 1.7.105 gives the first inequality, and together, they imply  $q_n \geq q_{n-1} + q_{n-2} \geq 2q_{n-2}$ . Recursively,  $q_{2n} \geq 2^n q_0 = 2^n$  and  $q_{2n+1} \geq 2^n q_1 \geq 2^n$ . □

If  $(a_n)_{n=0}^\infty$  is a sequence of natural numbers, then (1.37) defines a convergent sequence of rational numbers whose limit  $\alpha$  is irrational, and those numbers are convergents for  $\alpha$ . Thus, the Gauss map is related to continued-fraction expansion in the same way the map  $E_m$  is related to expansion in base  $m$ .

**Proposition 1.7.107** *Lebesgue measure is ergodic for the Gauss map.*

We establish this using *partial mixing*:

**Lemma 1.7.108**  $\mu$  is ergodic if for all measurable  $A, B$  there are  $n \in \mathbb{N}$  and  $c > 0$  with  $\mu(f^{-n}(A) \cap B) \geq c\mu(A)\mu(B)$ .

*Proof*  $A = f^{-1}(A) \Rightarrow c\mu(A)\mu(X \setminus A) \leq \mu(f^{-n}(A) \cap X \setminus A) = \mu(A \cap (X \setminus A)) = 0$ .

□

*Proof of Proposition 1.7.107* It suffices to establish the hypotheses of Lemma 1.7.108 for  $A := [u, v] \subset [0, 1]$  and  $B := I_{a_1 \dots a_n}$  as in the proof of Proposition 1.7.102 (the closure of an interval of continuity of  $G^n$ ) because the proof of Proposition 1.7.102 shows that intervals of this type generate the Borel  $\sigma$ -algebra.

The end-points of the interval  $I_{a_1 \dots a_n} \cap G^{-1}([u, v])$  are  $F_{a_1 \dots a_n}(u)$  and  $F_{a_1 \dots a_n}(v)$ , where  $F_{a_1 \dots a_n}$  is the local inverse in (1.36). Denoting Lebesgue measure by  $m$  and using (1.38) this gives

$$\begin{aligned}
 \frac{m(I_{a_1 \dots a_n} \cap G^{-1}([u, v]))}{m(I_{a_1 \dots a_n})} &= \frac{F_{a_1 \dots a_n}(v) - F_{a_1 \dots a_n}(u)}{F_{a_1 \dots a_n}(1) - F_{a_1 \dots a_n}(0)} \\
 &= \frac{\frac{p_n + vp_{n-1}}{q_n + vq_{n-1}} - \frac{p_n + up_{n-1}}{q_n + uq_{n-1}}}{\frac{p_n + p_{n-1}}{q_n + q_{n-1}} - \frac{p_n}{q_n}} \\
 &= \frac{(v - u)}{m([u, v])} \underbrace{\frac{q_n}{q_n + vq_{n-1}}}_{>1/2} \underbrace{\frac{q_n + q_{n-1}}{q_n + uq_{n-1}}}_{\geq 1} \\
 &> m([u, v])/2 \quad \text{since } 0 < q_{n-1} < q_n.
 \end{aligned}$$

□

While Lebesgue measure is not invariant under the Gauss map, there is an absolutely continuous invariant measure, with an explicitly given smooth positive density:

**Proposition 1.7.109** *The measure  $m_G$  with density  $\frac{1}{\log 2} \frac{1}{1+x}$  (i.e.,  $m_G([a, b]) = \log_2(b+1) - \log_2(a+1)$ ) is invariant under the Gauss map and ergodic.*

*Proof* Ergodicity follows from Propositions 1.7.107 and 1.7.41, and  $m_G(G^{-1}([a, b])) = \sum_{n \in \mathbb{N}} \underbrace{\log_2(1 + \frac{1}{a+n}) - \log_2(1 + \frac{1}{b+n})}_{=\log_2(a+n+1) - \log_2(a+n)} = \log_2 \frac{b+1}{a+1} = m_G([a, b])$ . □

**Corollary 1.7.110** *Almost no  $\alpha \in [0, 1]$  has bounded continued-fraction coefficients.*

*Proof* Since  $a_i = \lfloor 1/G^{i-1}(\alpha) \rfloor$  these are the points whose  $G$ -orbits are bounded away from zero, hence not uniformly distributed. By ergodicity, this is a null set. □

**Corollary 1.7.111** *For almost every  $x$  the number  $n \in \mathbb{N}$  occurs as a continued-fraction coefficient with asymptotic frequency  $\log_2(1 + \frac{1}{n}) - \log_2(1 + \frac{1}{n+1})$ .*

*Proof* By ergodicity of  $m_G$  the orbits of  $G$  are uniformly distributed, so the probability of  $a_i = \lfloor 1/G^{i-1}(\alpha) \rfloor = n$ , i.e.,  $\frac{1}{n+1} < G^{i-1}(\alpha) \leq \frac{1}{n}$ , is  $m_G((\frac{1}{n+1}, \frac{1}{n}])$ . □

### 1.7.15 Bernoulli Shifts

**Definition 1.7.112** A Bernoulli shift is the shift transformation  $\sigma_N$  from Definition 1.4.10 on the probability space  $\Omega_N = N^{\mathbb{Z}}$  or  $\Omega_N^R = N^{\mathbb{N}}$  with a product measure  $\mu_p$  generated from a probability vector  $p = (\mu(\{i\}))_{i=0}^{N-1}$  on  $\{0, \dots, N-1\}$  by setting

$$\mu_p(C_{\alpha_1, \dots, \alpha_k}^{n_1, \dots, n_k}) = \prod_{i=0}^k p_{\alpha_i}$$

for cylinders as defined by (1.6) and then extending to the  $\sigma$ -algebra of all Borel sets.

*Remark 1.7.113* It is readily apparent that the product measure is shift-invariant (for a shifted cylinders one computes the product of the same  $\alpha_i$ ) and that the intersection of finitely many cylinders is another cylinder; when their index sets  $\{n_1, \dots, n_k\}$  are pairwise disjoint, its measure is the product of the measures of the intersecting cylinders. This immediately implies Proposition 1.7.120 below.

This is a special case of the following.

**Definition 1.7.114** If  $(Y, \mathcal{T}, \nu)$  is a probability space then the product measure  $\mu$  of  $(X, \mathcal{B}, \mu) = \prod_{i \in \mathbb{Z}} (Y, \mathcal{T}, \nu)$  or  $(X, \mathcal{B}, \mu) = \prod_{i \in \mathbb{N}_0} (Y, \mathcal{T}, \nu)$  is called the *Bernoulli measure* and the one- or two-sided shift  $\sigma: X \rightarrow X$  defined by  $(\sigma(x))_n = x_{n+1}$  considered as  $\mu$ -preserving transformation is called a *Bernoulli shift*.

Here we used

**Definition 1.7.115 (Product Measures)** If  $(X_i, \mathcal{T}_i, \mu_i)_{i \in I}$  are probability spaces, then we define their *product*  $(X, \mathcal{B}, \mu) := \prod_{i \in I} (X_i, \mathcal{T}_i, \mu_i)$  as the cartesian product  $X := \prod_{i \in I} X_i$  with the probability measure  $\mu$  defined on cylinders

$$C_{A_1, \dots, A_k}^{i_1, \dots, i_k} := \{X \in X \mid x_{i_j} \in A_j \subset X_{i_j} \text{ for } j = 1, \dots, k\}$$

by

$$\mu(C_{A_1, \dots, A_k}^{i_1, \dots, i_k}) := \prod_{j=1}^k \mu_{i_j}(A_j)$$

and extended to a measure on a  $\sigma$ -algebra  $\mathcal{B}$ .

The importance of these systems motivates the following notion:

**Definition 1.7.116** A measure-preserving dynamical system is said to be a *Bernoulli system* if it is measure-theoretically isomorphic to a Bernoulli shift.

**Proposition 1.7.117** When  $m > 1$ , the expanding map  $E_m$  from Sect. 1.7.13 (with Lebesgue measure) is a Bernoulli system: it is measure-theoretically isomorphic to

the shift on  $m^{\mathbb{N}}$  with the uniform measure  $\mu_{(1/m, \dots, 1/m)}$ ; for  $m = 2$  so is the tent map  $T: x \mapsto 1 - 2|x - \frac{1}{2}|$  on  $[0, 1]$ .

*Proof* Let  $x = 0.x_1x_2\dots$  be the base- $m$  representation of  $x \in [0, 1]$ . Then  $mx = x_1.x_2x_3\dots = 0.x_2x_3\dots \pmod{1}$ . Thus

$$E_m(x) = 0.x_2x_3\dots \pmod{1},$$

and  $.x_1x_2\dots \mapsto (x_1, x_2, \dots)$  is the desired isomorphism.

For the tent map check that  $x \mapsto (\text{card}(\{T^i(x) \cap (\frac{1}{2}, 1)\}))_{i \in \mathbb{N}}$  works.  $\square$

### 1.7.16 Mixing

Analogously to the difference between topological transitivity and mixing, the difference in the probabilistic behavior of orbits between the simplest representatives of our two groups of examples, an irrational rotation  $R_\alpha$ , and a linear expanding map  $E_m$  lies in the following properties.

**Definition 1.7.118** A measure-preserving transformation  $f: (X, \mu) \rightarrow (X, \mu)$  is said to be *weakly mixing* if for any two measurable sets  $A, B$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} |\mu(f^{-k}(A) \cap B) - \mu(A)\mu(B)| = 0. \quad (1.39)$$

It is said to be *mixing* if for any two measurable sets  $A, B$

$$\mu(f^{-n}(A) \cap B) \rightarrow \mu(A) \cdot \mu(B) \quad \text{as } n \rightarrow \infty. \quad (1.40)$$

It is said to be *mixing of order  $N$*  if for any  $N+1$  measurable sets  $A_i$  and with  $n_0 := 0$

$$\mu\left(\bigcap_{i=0}^N f^{-n_i}(A_i)\right) \xrightarrow{n_i - n_{i-1} \rightarrow \infty} \prod_{i=0}^N \mu(A_i). \quad (1.41)$$

*Remark 1.7.119*

- Mixing is mixing of order 1.
- One can restate (1.40) as  $\mu_B(f^{-n}(A)) \xrightarrow{n \rightarrow \infty} \mu(A)$ , i.e., asymptotically  $f^{-n}(A)$  and  $B$  are independent sets.
- Clearly mixing implies weak mixing, so weak mixing is a weakened (average) version of the statement about asymptotic independence.
- By taking  $A$  invariant and  $B := X \setminus A$  (or by comparing (1.39) and (1.28), or from Proposition 1.7.138) we find that weak mixing implies ergodicity. Thus, ergodicity is the weakest statement of this sort.

- To clarify the intent of (1.41), we rewrite it for  $N = 2$  as

$$\mu(f^{-n}(A) \cap f^{-m}(B) \cap C) \xrightarrow{m \rightarrow \infty \text{ and } m-n \rightarrow \infty} \mu(A)\mu(B)\mu(C).$$

**Proposition 1.7.120** *Bernoulli shifts are mixing of all orders.*

*Proof* Since cylinders form a sufficient collection of  $\mu$ -measurable sets it is enough to check (1.40) for these, and we check only mixing of order 1. If  $C_{A_1, \dots, A_k}^{i_1, \dots, i_k}$  and  $C_{B_1, \dots, B_l}^{j_1, \dots, j_l}$  are cylinders and  $n > \max\{j_1, \dots, j_l\} - \min\{i_1, \dots, i_k\}$ , then

$$\begin{aligned} \mu(\sigma^{-n}(C_{A_1, \dots, A_k}^{i_1, \dots, i_k}) \cap C_{B_1, \dots, B_l}^{j_1, \dots, j_l}) &= \mu(C_{A_1, \dots, A_k}^{i_1+n, \dots, i_k+n} \cap C_{B_1, \dots, B_l}^{j_1, \dots, j_l}) \\ &= \mu(C_{A_1, \dots, A_k, B_1, \dots, B_l}^{i_1+n, \dots, i_k+n, j_1, \dots, j_l}) = \mu(C_{A_1, \dots, A_k}^{i_1, \dots, i_k} \cap C_{B_1, \dots, B_l}^{j_1, \dots, j_l}). \end{aligned}$$

□

If a measure-preserving transformation is mixing, then the von Neumann Mean Ergodic Theorem can be strengthened: the conclusion still holds if the iterates are sampled “haphazardly” rather than successively:

**Proposition 1.7.121** *Let  $f: (X, \mu) \rightarrow (X, \mu)$  be a mixing probability-preserving transformation,  $\varphi \in L^2(X, \mu)$ ,  $(k_i)_{i \in \mathbb{N}}$  a strictly increasing sequence in  $\mathbb{N}$ . Then*

$$\frac{1}{n} \sum_{i=0}^{n-1} \varphi \circ f^{k_i} \xrightarrow[n \rightarrow \infty]{L^2} P_f(\varphi) = \int \varphi d\mu.$$

*Proof* As usual, suffices to prove this for a sufficient family of functions, e.g., when  $\varphi = \chi_A$  is the characteristic function of a measurable set. To that end note that

$$\begin{aligned} &\int_X \left| \frac{1}{n} \sum_{i=0}^{n-1} (\chi_A \circ f^{k_i} - \mu(A)) \right|^2 d\mu \\ &= \frac{1}{n^2} \sum_{i,j=0}^{n-1} \int_X \underbrace{(\chi_A \circ f^{k_i} - \mu(A))(\chi_A \circ f^{k_j} - \mu(A))}_{= \chi_A \circ f^{k_i} \cdot \chi_A \circ f^{k_j} - \mu(A)(\chi_A \circ f^{k_j} + \chi_A \circ f^{k_i}) + \mu(A)^2} d\mu \\ &= \frac{1}{n^2} \sum_{\substack{i,j=0 \\ i \neq j}}^{n-1} \underbrace{\mu(f^{-k_i}(A) \cap f^{-k_j}(A))}_{= \mu(f^{k_j-k_i}(A) \cap A) \xrightarrow{|i-j| \rightarrow \infty} \mu(A)^2} - \mu(A)^2 \end{aligned}$$

by the mixing assumption, so the claim follows from Proposition 1.7.122 below. □

**Proposition 1.7.122** *If  $(a_{ij})_{i,j \in \mathbb{N}_0}$  is a bounded sequence with  $\lim_{|i-j| \rightarrow \infty} a_{ij} = 0$ , then  $\lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{i,j=0}^{n-1} a_{ij} = 0$ .*



*Proof* For  $\epsilon > 0$  take  $N \in \mathbb{N}$  such that  $|a_{ij}| < \epsilon/2$  when  $|i - j| > N$ . For  $n > N$  there are at most  $n(2N + 1)$  pairs  $(i, j)$  with  $0 \leq i, j < n$  and  $|i - j| \leq N$ , so if  $n > 2(2N + 1) \sup_{ij} |a_{ij}|/\epsilon$  as well, then

$$\left| \frac{1}{n^2} \sum_{i,j=0}^{n-1} a_{ij} \right| \leq \frac{2N+1}{n} \sup_{ij} |a_{ij}| + \frac{1}{n^2} \sum_{\substack{0 \leq i,j < n \\ |i-j| > N}} |a_{ij}| < \frac{\epsilon}{2} + \frac{\epsilon}{2}.$$

□

Weak mixing can be interpreted as a mixing condition in which one ignores a “negligible” set of times:

**Proposition 1.7.123** *A measure-preserving transformation is weakly mixing if and only if for any two measurable sets  $A, B$*

$$\text{there is an } E \subset \mathbb{N} \text{ of density 0 with } \lim_{E \nrightarrow \infty} \mu(f^{-n}(A) \cap B) = \mu(A) \cdot \mu(B). \quad (1.42)$$

Here we used the following notion and fact:

**Definition 1.7.124** If  $\text{card}(E \cap \{1, \dots, n\}) = dn + o(n)$ , then we say that  $E$  has density  $d$ . In particular, a set  $E \subset \mathbb{N}$  has density 0 if  $\text{card}(E \cap \{1, \dots, n\}) = o(n)$ .

**Lemma 1.7.125 (Koopman–von Neumann)** *If  $(a_n)_{n \in \mathbb{N}}$  is a bounded sequence, then  $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} |a_i| = 0$  if and only if there is an  $E \subset \mathbb{N}$  of density 0 such that*

$$\lim_{E \nrightarrow \infty} a_n = 0, \quad \text{i.e.,} \quad 0 = \lim_{n \rightarrow \infty} \begin{cases} a_n & \text{if } n \notin E \\ 0 & \text{if } n \in E. \end{cases}$$

**Corollary 1.7.126** *If  $(a_n)_{n \in \mathbb{N}}$  is bounded, then*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} |a_i| = 0 \quad \text{if and only if} \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} a_i^2 = 0.$$

*Proof*  $\lim_{E \nrightarrow \infty} a_n = 0$  if and only if  $\lim_{E \nrightarrow \infty} a_n^2 = 0$ . □

*Proof of Theorem 1.7.125 “if”:* Take  $M$  to be an upper bound of  $(|a_n|)_{n \in \mathbb{N}}$ . For  $\epsilon > 0$  there is an  $N \in \mathbb{N}$  such that for  $n \geq N$  we have

- $n \notin E \Rightarrow |a_n| < \frac{\epsilon}{M+1}$  and
- $d_n(E) := \frac{1}{n} \text{card}(E \cap \{1, \dots, n\}) < \frac{\epsilon}{M+1}$

and hence  $\frac{1}{n} \sum_{i < n} |a_i| = \frac{1}{n} \left( \sum_{n > i \in E} |a_i| + \sum_{n > i \notin E} |a_i| \right) < M d_n(E) + \frac{\epsilon}{M+1} < \epsilon$ .

“only if”: Since  $E_k := \{i \in \mathbb{N} \mid |a_i| \geq 1/k\} \subset E_{k+1}$  satisfies

$$d_n(E_k) = \frac{1}{n} \text{card } E_k \leq \frac{k}{n} \sum_{i=0}^{n-1} |a_i| \xrightarrow{n \rightarrow \infty} 0,$$

recursively take  $l_k \geq l_{k-1}$  such that  $d_n(E_k) < 1/k$  for  $n \geq l_k$ . Let  $E := \bigcup_{k \in \mathbb{N}} E_k \cap [l_{k-1}, l_k)$  and  $\epsilon > 0$ . If  $k > 1/\epsilon$  and  $l_{k-1} < n \notin E$ , then  $n \notin E_k$ , and  $|a_n| < 1/k < \epsilon$ . Now take  $K > 2/\epsilon$ ,  $n \geq l_K$  and  $k \geq K$  such that  $l_k \leq n < l_{k+1}$ . Since

$$E \cap [0, n) = (E \cap [0, l_k)) \cup (E \cap [l_k, n)) \subset (E_k \cap [0, l_k)) \cup (E_{k+1} \cap [l_k, n)),$$

we get  $d_n(E) \leq \frac{1}{n} \left( \underbrace{l_k d_{l_k}(E_k)}_{=\text{card}(E_k \cap \{1, \dots, l_k\}) \leq \text{card}(E_k \cap \{1, \dots, n\}) = n d_n(E_k) < n/k} + n \overbrace{d_n(E_{k+1})}^{< \frac{1}{k+1} < \frac{1}{k}} \right) < \frac{2}{k} < \epsilon$ . □

Clearly, mixing and weak mixing are invariants of measure-theoretic isomorphism (Definition 1.7.62). Furthermore, an argument similar to that for ergodicity in Sect. 1.7.8 shows

**Proposition 1.7.127** *If a map is (weakly) mixing, then so is any factor.*

**Proposition 1.7.128** *If a continuous map  $f$  has a mixing invariant measure  $\mu$  then  $f|_{\text{supp } \mu}$  is topologically mixing (Definition 1.3.34).*

*Proof* If  $A, B \subset \text{supp } \mu$  are open and  $n$  is sufficiently large then  $\mu(f^{-n}(A) \cap B)$  is positive and hence the intersection is nonempty. □

The converse is not true: A topologically mixing map, even a minimal one, may fail to have a mixing invariant measure with full support. This phenomenon is, however, atypical—similarly to the situation with other properties such as topological transitivity and ergodicity which we discussed at the end of Sect. 1.7.12. As we will soon see, our topologically mixing examples are mixing with respect to natural invariant measures.

Now we prove a criterion of mixing that allows us to avoid tedious approximation arguments when checking mixing for specific dynamical systems.

**Definition 1.7.129** A collection  $\mathcal{C} \subset \mathcal{S}$  in a measure space  $(X, \mathcal{S}, \mu)$  is said to be *sufficient* if finite disjoint unions of elements of  $\mathcal{C}$  form a dense collection with respect to the symmetric-difference metric

$$d(A, B) := d_\mu(A, B) := \mu(A \triangle B) \in (0, \infty]$$

**Proposition 1.7.130** *Suppose  $\mathcal{C}$  is a sufficient collection of sets. Then*

1.  *$f$  is mixing if (1.40) holds for any  $A, B \in \mathcal{C}$ ,*

2.  $f$  is weakly mixing if (1.39) or (1.42) holds for any  $A, B \in \mathcal{C}$ ,
3.  $f$  is ergodic if (1.29) holds for any  $A, B \in \mathcal{C}$ ,
4.  $f$  is mixing of order  $N$  if (1.41) holds for any  $A_i \in \mathcal{C}$ .

*Proof* We prove (1) using Proposition 1.7.46; the other parts have like proofs. Let

$$A_1, \dots, A_k, B_1, \dots, B_l \in \mathcal{C}, A_i \cap A_{i'} = \emptyset \text{ for } i \neq i', B_j \cap B_{j'} = \emptyset \text{ for } j \neq j'$$

and  $A := \bigcup_{i=1}^k A_i, B := \bigcup_{j=1}^l B_j$ . Then  $\mu(A) = \sum_{i=1}^k \mu(A_i), \mu(B) = \sum_{j=1}^l \mu(B_j)$ , and

$$\mu(f^{-n}(A) \cap B) = \sum_{i=1}^k \sum_{j=1}^l \mu(f^{-n}(A_i) \cap B_j) \rightarrow \sum_{i=1}^k \sum_{j=1}^l \mu(A_i) \cdot \mu(B_j) = \mu(A) \cdot \mu(B).$$

Thus (1.40) holds for any elements of the dense collection  $\mathfrak{U}$  formed by finite disjoint unions of elements of  $\mathcal{C}$ . Now let  $A, B$  be arbitrary measurable sets. Find  $A', B' \in \mathfrak{U}$  such that  $\mu(A \Delta A') < \epsilon/4, \mu(B \Delta B') < \epsilon/4$ . By the triangle inequality

$$\begin{aligned} |\mu(f^{-n}(A) \cap B) - \mu(A)\mu(B)| &\leq \mu(f^{-n}(A \Delta A') \cap B) + \mu(f^{-n}(A') \cap (B \Delta B')) \\ &\quad + |\mu(f^{-n}(A') \cap B') - \mu(A')\mu(B')| \\ &\quad + \mu(A) \cdot \mu(B \Delta B') + \mu(B') \cdot \mu(A \Delta A') \\ &\leq |\mu(f^{-n}(A') \cap B') - \mu(A') \cdot \mu(B')| + \epsilon. \end{aligned}$$

Since  $\epsilon > 0$  can be chosen arbitrarily small, this implies (1.40).  $\square$

It is not only with respect to the sets in question, but also in the conclusion that suitable approximation is good enough:

**Proposition 1.7.131** *Let  $f$  be a homeomorphism of a compact metric space  $X$  and  $\mu$  an  $f$ -invariant Borel probability measure with constants  $c, C > 0$  such that*

$$c\mu(P)\mu(Q) \leq \varliminf_{n \rightarrow \infty} \mu(P \cap f^{-n}(Q)) \leq \overline{\varlimsup}_{n \rightarrow \infty} \mu(P \cap f^{-n}(Q)) \leq C\mu(P)\mu(Q) \quad (1.43)$$

*for all Borel sets  $P, Q \subset X$ . Then  $\mu$  is mixing.*

*Remark 1.7.132* In fact, this is true for measure-preserving transformations of a measure space. In the proof one has to replace the use of the weak\* topology by some purely measure-theoretic considerations.

*Proof* We first show that the left inequality in (1.43) implies that the product  $f \times f$  is ergodic with respect to  $\mu \times \mu$ . Let  $A, B, C, D \subset X$  be Borel sets. Then

$$\varliminf_{n \rightarrow \infty} \underbrace{(\mu \times \mu)((f \times f)^n(A \times C) \cap (B \times D))}_{\mu(f^n(A) \cap B) \cdot \mu(f^n(C) \cap D)} \geq c^2 \underbrace{\mu(A) \cdot \mu(B) \cdot \mu(C) \cdot \mu(D)}_{(\mu \times \mu)(A \times B) \cdot (\mu \times \mu)(C \times D)}.$$

The same holds if we replace  $A \times C$  and  $B \times D$  by finite disjoint unions of product sets. Since such sets approximate every measurable  $P, Q \subset X \times X$ , we have

$$\varliminf_{n \rightarrow \infty} (\mu \times \mu)((f \times f)^n(P) \cap Q) \geq c^2(\mu \times \mu)(P) \cdot (\mu \times \mu)(Q),$$

and  $f \times f$  is ergodic with respect to  $\mu \times \mu$ .

Now let  $\nu$  be the diagonal measure in  $X \times X$  given by  $\nu(E) = \mu(\pi_1(E \cap \Delta))$ , where  $\Delta = \{(x, x) \mid x \in X\}$  and  $\pi_1: X \times X \rightarrow X$  is the projection to the first coordinate. The measure  $\nu$  and its shift  $\nu_n$  under the map  $f^n \times \text{Id}$  are  $(f \times f)$ -invariant. Explicitly,  $\nu_n(A \times B) = \mu(f^n(A) \cap B)$ . By the right inequality in (1.43) we have

$$\varlimsup_{n \rightarrow \infty} \nu_n(A \times B) = \varlimsup_{n \rightarrow \infty} \mu(f^n(A) \cap B) < C\mu(A) \cdot \mu(B) = C(\mu \times \mu)(A \times B). \quad (1.44)$$

Let  $\eta$  be any weak limit point of the sequence  $\nu_n$ . If  $A, B \subset X$  are closed sets then  $\eta(A \times B) \leq C(\mu \times \mu)(A \times B)$  by (1.44). Taking disjoint unions of products of closed sets and using approximation we deduce that  $\eta(P) < C(\mu \times \mu)(P)$  for any Borel set  $P \subset X \times X$  and hence  $\eta$  is absolutely continuous with respect to  $\mu \times \mu$ . Since  $\eta$  is  $(f \times f)$ -invariant and  $\mu \times \mu$  is ergodic we have  $\eta = \mu \times \mu$  by Proposition 1.7.41, so for any closed sets  $A, B$  with  $\mu(\partial A) = \mu(\partial B) = 0$  we have

$$\lim_{n \rightarrow \infty} \mu(f^n(A) \cap B) = \lim_{n \rightarrow \infty} \nu_n(A \times B) = (\mu \times \mu)(A \times B) = \mu(A) \cdot \mu(B).$$

Since the collection of all such sets is sufficient,  $f$  is mixing with respect to  $\mu$  by Proposition 1.7.130.  $\square$

The notions of mixing and weak mixing transfer to products:

**Proposition 1.7.133** *A measure-preserving transformation  $f: (X, \mu) \rightarrow (X, \mu)$  is mixing (weakly mixing) if and only if  $f \times f$  is.*

*Proof* If  $f \times f$  is weakly mixing and  $A, B \subset X$  then by Proposition 1.7.123 there is a set  $E \subset \mathbb{N}$  of density 0 such that

$$\begin{aligned} \mu(f^{-n}(A) \cap B) &= (\mu \times \mu)\left((f \times f)^{-n}(A \times X) \cap (B \times X)\right) \\ &\xrightarrow{E \not\rightarrow \infty} (\mu \times \mu)(A \times X) \cdot (\mu \times \mu)(B \times X) = \mu(A)\mu(B), \end{aligned}$$

so  $f$  is weakly mixing. Taking  $E = \emptyset$  proves that  $f \times f$  mixing  $\Rightarrow f$  mixing.

Suppose now that  $f$  is weakly mixing. Then for measurable  $A_1, A_2, B_1, B_2 \subset X$  there exist sets  $E_1, E_2 \subset \mathbb{N}$  of density 0 such that

$$\lim_{E_i \not\rightarrow \infty} \mu(f^{-n}(A_i) \cap B_i) = \mu(A_i) \cdot \mu(B_i)$$

for  $i = 1, 2$ . Taking  $E := E_1 \cup E_2$  we find that

$$\underbrace{(\mu \times \mu) \left( \overbrace{(f \times f)^{-n}(A_1 \times A_2) \cap (B_1 \times B_2)}^{=(f^{-n}(A_1) \cap B_1) \times (f^{-n}(A_2) \cap B_2)} \right)}_{=\mu(f^{-n}(A_1) \cap B_1)\mu(f^{-n}(A_2) \cap B_2)} \xrightarrow{E \not\rightarrow \infty} \underbrace{\mu(A_1)\mu(B_1)\mu(A_2)\mu(B_2)}_{=(\mu \times \mu)(A_1 \times A_2)(\mu \times \mu)(B_1 \times B_2)}.$$

Since the sets  $A \times B$  form a sufficient collection,  $f \times f$  is weakly mixing by Proposition 1.7.130(2). Taking  $E_1 = E_2 = \emptyset$  shows  $f$  mixing  $\Rightarrow f \times f$  mixing.  $\square$  One of the implications in Proposition 1.7.133 is easy to strengthen:

**Proposition 1.7.134** *If  $f: X \rightarrow X$  is a measure-preserving transformation and  $f \times f$  is ergodic, then  $f$  is weakly mixing.*

*Proof* Take  $A, B$  measurable and suppose  $f \times f$  is ergodic. We will show that

$$\begin{aligned} & \frac{1}{n} \sum_{k=0}^{n-1} \left( \mu(f^{-k}(A) \cap B) - \mu(A)\mu(B) \right)^2 \\ &= \frac{1}{n} \sum_{k=0}^{n-1} \left( \mu(f^{-k}(A) \cap B)^2 - 2\mu(f^{-k}(A) \cap B)\mu(A)\mu(B) + \mu(A)^2\mu(B)^2 \right) \xrightarrow{n \rightarrow \infty} 0. \end{aligned} \quad (1.45)$$

This implies the claim by Proposition 1.7.46 and Corollary 1.7.126. (1.45) follows from ergodicity of  $f \times f$  which by Proposition 1.7.46 implies

$$\begin{aligned} \frac{1}{n} \sum_{k=0}^{n-1} \mu(f^{-k}(A) \cap B) &= \frac{1}{n} \sum_{k=0}^{n-1} (\mu \times \mu)((f \times f)^{-k}(A \times X) \cap (B \times X)) \\ &\xrightarrow{n \rightarrow \infty} (\mu \times \mu)(A \times X)(\mu \times \mu)(B \times X) = \mu(A)\mu(B) \end{aligned}$$

and

$$\begin{aligned} \frac{1}{n} \sum_{k=0}^{n-1} \mu(f^{-k}(A) \cap B)^2 &= \frac{1}{n} \sum_{k=0}^{n-1} (\mu \times \mu)((f \times f)^{-k}(A \times A) \cap (B \times B)) \\ &\xrightarrow{n \rightarrow \infty} (\mu \times \mu)(A \times A)(\mu \times \mu)(B \times B) = \mu(A)^2\mu(B)^2. \end{aligned}$$

$\square$

Just a little effort beyond Propositions 1.7.133 and 1.7.134 gives

**Theorem 1.7.135 (Product-Characterization of Weak Mixing)** *For a probability-preserving  $f$  the following are equivalent:*

1.  $f$  is weakly mixing,
2.  $f \times f$  is weakly mixing,

3.  $f \times g$  is ergodic whenever  $g$  is ergodic,  
 4.  $f \times f$  is ergodic.

*Proof* Propositions 1.7.133 and 1.7.134 with Remark 1.7.119 give  $1. \Leftrightarrow 2. \Rightarrow 4. \Rightarrow 1.$  It now suffices to show  $1. \Rightarrow 3. \Rightarrow 4.$  Here,  $3. \Rightarrow 4.$  is easy: take  $g = 0$  on  $\{0\}$  in 3. to deduce that  $f$  is ergodic, then take  $g = f$  in 3. to get 4.

To prove  $1. \Rightarrow 3.$  we use Proposition 1.7.130.

$$\begin{aligned}
 & \left| \frac{1}{n} \sum_{k=0}^{n-1} (\mu \times \nu)((f \times g)^{-k}(A_1 \times A_2) \cap B_1 \times B_2) - (\mu \times \nu)(A_1 \times A_2)(\mu \times \nu)(B_1 \times B_2) \right| \\
 &= \frac{1}{n} \left| \sum_{k=0}^{n-1} \underbrace{\mu(f^{-k}(A_1) \cap B_1)}_{=:x_k} \underbrace{\nu(g^{-k}(A_2) \cap B_2)}_{=:y_k} - \underbrace{\mu(A_1)\mu(B_1)}_{=:x} \underbrace{\nu(A_2)\nu(B_2)}_{=:y} \right| \\
 &= \frac{1}{n} \left| \sum_{k=0}^{n-1} x_k y_k - x y \right| \leq \frac{1}{n} \sum_{k=0}^{n-1} |x_k - x| \cdot y_k + x \cdot \left| \frac{1}{n} \sum_{k=0}^{n-1} y_k - y \right| \xrightarrow{n \rightarrow \infty} 0. \\
 &\quad \leq \max_k y_k \frac{1}{n} \sum_{k=0}^{n-1} |x_k - x| \rightarrow 0 \text{ (} f \text{ weakly mixing)} \quad \rightarrow 0 \text{ (ergodicity of } g)
 \end{aligned}$$

□

Just as ergodicity can be expressed in terms of functions rather than sets, so can the various notions of mixing. In probabilistic terms, sets are events and functions are random variables. The preceding notions of ergodicity and mixing involve various forms of eventual independence of events, and they can be recast in terms of asymptotic independence of random variables using the *covariance* of  $L^2$ -functions.

**Definition 1.7.136** The *covariance* of  $\varphi, \psi \in L^2$  is defined as

$$\begin{aligned}
 \text{cov}(\varphi, \psi) &:= \langle \varphi - \langle \varphi, 1 \rangle, \psi - \langle \psi, 1 \rangle \rangle = \langle \varphi, \psi \rangle - \langle \varphi, 1 \rangle \langle 1, \psi \rangle \\
 &= \int (\varphi - \int \varphi)(\psi - \int \psi) = \int \varphi \bar{\psi} - \int \varphi \int \bar{\psi}.
 \end{aligned}$$

That is, we project both functions to the orthocomplement  $1^\perp \subset L^2$  of the constant functions by subtracting their average (to focus on their variation) and then take the inner product.

*Remark 1.7.137* Like the inner product itself, the covariance is sesquilinear (linear in the first entry and antilinear in the second) and invariant under isometric operators (i.e.,  $\langle U\cdot, U\cdot \rangle = \langle \cdot, \cdot \rangle \Rightarrow \text{cov}(U\cdot, U\cdot) = \text{cov}$ ). If either of the functions is constant, then the covariance is zero, so it is unaffected by the addition of constants to either function. For many statements about covariance, this allows us to assume without loss of generality that the functions in question have zero average, i.e., are in  $1^\perp$ .

Indeed, “polarization”<sup>37</sup> allows us to consider the same function in both entries:

$$\text{cov}(\varphi, \psi) = \frac{1}{4}[\text{cov}(\varphi + \psi, \varphi + \psi) - \text{cov}(\varphi - \psi, \varphi - \psi)].$$

Finally, the covariance satisfies the Cauchy–Schwarz inequality:

$$|\text{cov}(\varphi, \psi)| \leq \|\varphi\| \|\psi\|.$$

**Proposition 1.7.138** *If  $\Phi \subset L^2$  is a complete system, i.e.,  $\overline{\text{span}(\Phi)} = L^2$ , then*

- *$f$  is ergodic if and only if  $\frac{1}{n} \sum_{k=0}^{n-1} \text{cov}(U_f^n(\varphi), \psi) \xrightarrow{n \rightarrow \infty} 0$  for all  $\varphi, \psi \in \Phi$ ,*
- *$f$  is weakly mixing if and only if*

$$\frac{1}{n} \sum_{k=0}^{n-1} |\text{cov}(U_f^n(\varphi), \psi)| \xrightarrow{n \rightarrow \infty} 0 \quad (1.46)$$

*for all  $\varphi, \psi \in \Phi$ ,*

- *$f$  is weakly mixing if and only if for all  $\varphi, \psi \in \Phi$ , there exists an  $E \subset \mathbb{N}$  of density 0 (Definition 1.7.124) such that  $\text{cov}(U_f^n(\varphi), \psi) \xrightarrow{E \not\ni n \rightarrow \infty} 0$ ,*
- *$f$  is mixing if and only if  $\text{cov}(U_f^n(\varphi), \psi) \xrightarrow{n \rightarrow \infty} 0$  for all  $\varphi, \psi \in \Phi$ .*
- *$f$  is mixing of order  $N$  if  $\int \prod_{i=0}^N \varphi_i \circ f^{n_i} d\mu \xrightarrow{n_i - n_{i-1} \rightarrow \infty} \prod_{i=0}^N \int \varphi_i d\mu$  for  $\{\varphi_0, \dots, \varphi_N\} \subset \Phi$ .*

*Proof* To see how to pass from a complete system to  $L^2$  note first that sesquilinearity of covariance means that checking any of these statements for all  $\varphi, \psi \in \Phi$  implies the same for all  $\varphi, \psi \in \text{span}(\Phi)$ . Now take arbitrary  $\varphi, \psi \in L^2$  and  $\varphi', \psi' \in \text{span}(\Phi)$  such that  $\|\psi - \psi'\| < \epsilon/2\|\varphi\|$  and  $\|\varphi - \varphi'\| < \epsilon/2\|\psi'\|$ . Then

$$\begin{aligned} |\text{cov}(U_f^n(\varphi), \psi)| &= |\text{cov}(U_f^n(\varphi), \psi - \psi') + \text{cov}(U_f^n(\varphi) - U_f^n(\varphi'), \psi') \\ &\quad + \text{cov}(U_f^n(\varphi'), \psi')| \leq |\text{cov}(U_f^n(\varphi'), \psi')| + \epsilon. \end{aligned}$$

Now, for each of these statements, knowing it for all  $\varphi, \psi \in L^2$  implies the corresponding mixing property by taking  $\varphi = \chi_A, \psi = \chi_B$  for measurable sets  $A, B$ .

To see the converse note that characteristic functions of measurable sets (or of only a sufficient collection) form a complete system in  $L^2$  for which the statement about covariance boils down to the respective mixing property.  $\square$

<sup>37</sup>  $\|u + v\|^2 - \|u - v\|^2 = 4\langle u, v \rangle$ .

*Remark 1.7.139* Note that we have in particular reproved Proposition 1.7.130.

By Definition 1.7.136 this characterization of mixing can be restated:

**Proposition 1.7.140** *If  $\Phi \subset L^2$  is a complete system, i.e.,  $\overline{\text{span}(\Phi)} = L^2(\mu)$ , then  $f$  is mixing if and only if  $U_f^n(\varphi) \xrightarrow{\text{weakly}} \int \varphi$  for all  $\varphi \in \Phi$ .*

*Remark 1.7.141* This last restatement motivates the notion of mixing as follows. It implies that when  $\varphi \geq 0$  with  $\|\varphi\|_1 = 1$  represents the probability density of “material” that is being redistributed as its points evolve under iteration of  $f$ , the density evens out:  $f_*^n(\varphi\mu) \xrightarrow{n \rightarrow \infty} \mu$ . Thus, the material spreads out perfectly evenly and becomes equidistributed with respect to  $\mu$ .

Remark 1.7.137 suggests

**Proposition 1.7.142** *In each of the statements in Proposition 1.7.138 one can replace  $\text{cov}(U_f^n(\varphi), \psi)$  by  $\text{cov}(U_f^n(\varphi), \varphi)$  or by  $\langle U_f^n(\varphi), \varphi \rangle$  if  $\langle \varphi, 1 \rangle = 0$ . For instance,  $f$  is mixing if and only if*

$$\text{cov}(U_f^n(\varphi), \varphi) \xrightarrow{n \rightarrow \infty} 0$$

for all  $\varphi$  in a complete set (in  $L^2$  or in  $1^\perp$ ), which happens if and only if

$$\langle U_f^n(\varphi), \varphi \rangle \xrightarrow{n \rightarrow \infty} 0$$

for all  $\varphi$  in a complete set for  $1^\perp$ .

*Proof* While Remark 1.7.137 applies if the hypothesis is known for all  $\varphi \in L^2$ , the step from a complete system to  $L^2$  requires attention because  $\varphi \mapsto \text{cov}(U_f^n(\varphi), \varphi)$  is not linear. The following lemma covers the mixing case, and the others are analogous. The last statement follows directly from Remark 1.7.137.  $\square$

**Lemma 1.7.143** *If  $\text{cov}(U_f^n(\varphi), \varphi) \rightarrow 0$ , then  $\text{cov}(U_f^n(\varphi), \psi) \rightarrow 0$  for all  $\psi \in L^2$ .*

*Proof*  $M_\varphi := \{\psi \in L^2 \mid \text{cov}(U_f^n(\varphi), \psi) \xrightarrow{n \rightarrow \infty} 0\}$  is closed in  $L^2$ , contains  $\{1, \varphi\}$ , and  $U_f M_\varphi \subset M_\varphi$ : If  $\psi \in M_\varphi$ , then  $\langle U_f^n(\varphi), U_f(\psi) \rangle = \langle U_f(U_f^{n-1}(\varphi)), U_f(\psi) \rangle = \langle U_f^{n-1}(\varphi), \psi \rangle$  since  $U_f$  is an isometry, so  $\text{cov}(U_f^n(\varphi), U_f(\psi)) \rightarrow 0$ . Thus,

$$M_\varphi \supset m_\varphi := \bigcup \{E \subset L^2 \text{ closed} \mid 1, \varphi \in E, U_f(E) \subset E\} \supset U_f(m_\varphi).$$

If  $\psi \in m_\varphi^\perp$ , then  $\langle 1, \psi \rangle = 0$  and  $\langle U_f^n(\varphi), \psi \rangle = 0$  for all  $n$  since  $U_f^n(\varphi) \in U_f^n(m_\varphi) \subset m_\varphi$ , so  $\psi \in M_\varphi$ . Thus,  $L^2 = m_\varphi \oplus m_\varphi^\perp \subset M_\varphi$ .  $\square$

**Proposition 1.7.144** *Eigenfunctions of a weakly mixing transformation are constant, i.e., if  $\varphi \in L^2$  and  $\varphi \circ f = \lambda \varphi$  for some  $\lambda \in \mathbb{C}$  then  $\varphi = \text{const.}$ —and hence  $\lambda = 1$ , so  $f$  has only one eigenvalue. Thus “weakly mixing implies no eigenfunctions” (in  $1^\perp$ ).*



*Remark 1.7.145* We will eventually be able to prove the converse; see Definition 1.7.174 and Proposition 1.7.176.

*Proof* If  $\varphi \in L^2$  and  $\varphi \circ f = \lambda\varphi$ , then  $|\lambda| = 1$  and either  $\lambda = 1$ , so  $\varphi = \text{const.}$  by ergodicity (which follows from weak mixing), or  $\lambda \neq 1$ , in which case

$$\langle \varphi, 1 \rangle = \int \varphi = \int \varphi \circ f = \int \lambda\varphi = \lambda \langle \varphi, 1 \rangle,$$

so  $\langle \varphi, 1 \rangle = 0$  and

$$\int |\varphi|^2 = \frac{1}{n} \sum_{k=0}^{n-1} |\lambda|^k \int \varphi \bar{\varphi} = \frac{1}{n} \sum_{k=0}^{n-1} \left| \int \lambda^k \varphi \bar{\varphi} \right| = \frac{1}{n} \sum_{k=0}^{n-1} \left| \int (\varphi \circ f^k) \bar{\varphi} \right| \xrightarrow{n \rightarrow \infty} 0$$

by (1.46) since  $f$  is weakly mixing.  $\square$

### 1.7.17 Toral Translations and Expanding Maps

#### Proposition 1.7.146

1. No translation  $T_\gamma$  of the torus is weakly mixing with respect to Lebesgue measure.
2. Every expanding endomorphism  $E_m$ ,  $|m| \geq 2$ , is mixing of order  $N$  for any  $N \in \mathbb{N}$  with respect to Lebesgue measure.

*Remark 1.7.147* Item (2) follows from the Bernoulli property (Proposition 1.7.117), but it is instructive to study direct proofs.

*Proof*

1. It is convenient to use multiplicative notation:  $T_{(\gamma_1, \dots, \gamma_n)}(z_1, \dots, z_n) = (\gamma_1 z_1, \dots, \gamma_n z_n)$ . Then  $\varphi: \mathbb{T}^n \rightarrow \mathbb{C}$ ,  $(z_1, \dots, z_n) \mapsto z_1$  is a nonconstant eigenfunction for the eigenvalue  $\gamma_1$ . Now apply Proposition 1.7.144.
2. By Proposition 1.7.130 it is enough to establish (1.40) for intervals of the form  $\Delta_{i,k} = (i/|m|^k, (i+1)/|m|^k)$ . Consider a collection  $\{\Delta_{i,j,k_j} \mid 0 \leq j < N\}$  of these, and let  $K := \max_j k_j$ . The central observation is that for  $n \geq K$ ,

$$E_m^{-n}(\Delta_{i,j,k_j}) \cap \Delta_{i_l,k_l}$$

consists of  $|m|^{n-k_l}$  translates of  $\Delta_{0,n+k_j}$ . On one hand, this directly implies mixing of order 1 because this set has measure

$$|m|^{n-k_l} / |m|^{n+k_j} = |m|^{-k_j-k_l} = \lambda(\Delta_{i,j,k_j}) \cdot \lambda(\Delta_{i_l,k_l}).$$

On the other hand, this implies mixing of order  $N$  by working recursively from the back (keeping in mind the structure of these sets rather than just their measure), i.e., starting with  $l = N$ ,  $j = N - 1$  above, and then intersecting the preimage of this intersection with  $\Delta_{i_{N-2}, k_{N-2}}$ , and so on.  $\square$

*Second Proof of Mixing of Order 1 of  $E_m$*  Fourier analysis gives

$$\int U_{E_m}^n \varphi \cdot \psi = \sum_{k,l \in \mathbb{Z}} \varphi_k \psi_l \int \exp(2\pi i(m^n k + l)x) = \sum_{k \neq 0} \varphi_k \psi_{-m^n k} \xrightarrow{n \rightarrow \infty} 0 \quad (1.47)$$

for any  $L^2$  functions  $\varphi(x) = \sum_{k \in \mathbb{Z}} \varphi_k \exp(2\pi i k x)$  and  $\psi(x) = \sum_{l \in \mathbb{Z}} \psi_l \exp(2\pi i l x)$  with (without loss of generality)  $\int \psi = 0$ .  $\square$

### 1.7.18 Rates of Mixing and Decay of Correlations

In terms of functions (or observables or random variables), the various mixing properties are expressed in terms of covariance. Specifically, mixing means that  $\text{cov}(U_f^n(\varphi), \psi) \xrightarrow{n \rightarrow \infty} 0$  for any  $\varphi, \psi \in L^2$ , and we now digress to the question of how rapid this convergence might be. Since covariance is closely related to correlation, this is known as the rate of *decay of correlations*. We do so in a particular context.

**Proposition 1.7.148** *Consider the expanding endomorphism  $E_m$  for  $|m| \geq 2$  with Lebesgue measure and suppose  $\varphi, \psi: [0, 1] \rightarrow \mathbb{R}$  are  $\alpha$ -Hölder-continuous functions with coefficient  $L$ , i.e.,  $|\varphi(x) - \varphi(y)| \leq L|x - y|^\alpha$ . Then correlations decay with the exponential rate  $m^{-\alpha n}$ :*

$$|\text{cov}(U_{E_m}^n(\varphi), \psi)| \leq L m^{-\alpha n} \|\varphi\|.$$

*Remark 1.7.149* We note that the parameters that affect the decay rate of correlations are the expansion rate of the transformation as well as the Hölder exponent of the functions under consideration. In particular, for Lipschitz-continuous functions the decay rate is the reciprocal of the expansion rate of the transformation.

*Proof* Assume without loss of generality that  $\int_0^1 \varphi = 0$ . Then

$$\begin{aligned} |\text{cov}(U_{E_m}^n(\varphi), \psi)| &= \left| \sum_{k=0}^{m^n-1} \int_{\frac{k}{m^n}}^{\frac{k+1}{m^n}} \varphi \circ E_m^n \cdot \tilde{\psi} \right| \\ &\leq \left| \sum_{k=0}^{m^n-1} \int_{\frac{k}{m^n}}^{\frac{k+1}{m^n}} \varphi \circ E_m^n \cdot \left( \psi - \int_{\frac{k}{m^n}}^{\frac{k+1}{m^n}} \psi \right) \right| + \left| \sum_{k=0}^{m^n-1} \int_{\frac{k}{m^n}}^{\frac{k+1}{m^n}} \varphi \circ E_m^n \cdot \int_{\frac{k}{m^n}}^{\frac{k+1}{m^n}} \psi \right| \\ &\leq \frac{L}{m^{\alpha n}} \sum_{k=0}^{m^n-1} \int_{\frac{k}{m^n}}^{\frac{k+1}{m^n}} |\varphi \circ E_m^n| + \left| \int_{\frac{k}{m^n}}^{\frac{k+1}{m^n}} \psi \sum_{k=0}^{m^n-1} \underbrace{\int_0^1 \frac{\varphi}{m^n}}_{=0} \right| = L m^{-\alpha n} \|\varphi\|. \end{aligned}$$

$\square$

Closer study of (1.47) reveals that for smoother functions we get even more rapid decay of correlations.

**Proposition 1.7.150** *Consider the expanding endomorphism  $E_m$  for  $|m| \geq 2$  with Lebesgue measure and suppose  $\varphi, \psi: [0, 1] \rightarrow \mathbb{R}$  are  $C^r$  functions. Then correlations decay with the exponential rate  $m^{-r}$ :  $\text{cov}(U_{E_m}^n(\varphi), \psi) = O(m^{-rn})$ . (See Remark 1.6.21.) In particular, analytic functions have superexponential decay of correlations.*

*Proof*  $|\psi_l| \leq |l|^{-r}$  in (1.47), hence  $|\psi_{-m^n k}| \leq C(m^n |k|)^{-r} = m^{-rn} |k|^{-r}$ , and

$$\left\| \sum \varphi_k \psi_{-m^n k} \right\|_{\ell^2} \leq \sqrt{\sum |\varphi_k|^2} \sqrt{\sum |\psi_{-m^n k}|^2} \leq \|\varphi\|_2 \cdot C' m^{-rn}.$$

□

We motivate this in terms of another convergence rate. In Corollary 1.7.44 we noted that for ergodic measures  $\sum_{k=0}^{n-1} \varphi(f^k(x)) - n \int_X \varphi d\mu = o(n)$   $\mu$ -a.e. The right information about correlation decay improves this to  $O(\sqrt{n})$  on average:

**Proposition 1.7.151**  $\sum_{k \in \mathbb{Z}} \text{cov}(U_f^k \varphi, \varphi) < \infty \Rightarrow \left\| \sum_{k=0}^{n-1} \varphi(f^k(x)) - n \int \varphi \right\|_2 = O(\sqrt{n}).$

*Proof* We will use that  $\text{cov}(U_f \cdot, U_f \cdot) = \text{cov}$ . Let  $\sigma^2 := \sum_{k \in \mathbb{Z}} \text{cov}(U_f^k \varphi, \varphi)$ . Then

$$\begin{aligned} \left\| \sum_{k=0}^{n-1} \varphi(f^k(x)) - n \int \varphi \right\|_2^2 &= \int \left( \sum_{k=0}^{n-1} \varphi(f^k(x)) - n \int \varphi \right) \overline{\left( \sum_{k=0}^{n-1} \varphi(f^k(x)) - n \int \varphi \right)} \\ &= \int \underbrace{\sum_{k=0}^{n-1} (\varphi(f^k(x)) - \int \varphi) \sum_{k=0}^{n-1} (\varphi(f^k(x)) - \int \varphi)}_{\sum_{k=0}^{n-1} a_k \sum_{k=0}^{n-1} \overline{a_k} = \sum_{i=0}^{n-1} a_i \bar{a}_i + \sum_{k=1}^{n-1} \sum_{i=0}^{n-k-1} a_{k+i} \bar{a}_i + a_i \bar{a}_{k+i}} \\ &= \sum_{i=0}^{n-1} \text{cov}(U_f^i \varphi, U_f^i \varphi) \\ &\quad + \sum_{k=1}^{n-1} \sum_{i=0}^{n-k-1} \text{cov}(U_f^{k+i} \varphi, U_f^i \varphi) + \text{cov}(U_f^i \varphi, U_f^{k+i} \varphi) \\ &= \sum_{|k| < n} (n - |k|) \text{cov}(U_f^k \varphi, \varphi) = n\sigma^2 + o(n). \end{aligned}$$

□

The decay of correlations is of interest in smooth dynamics but is a digression here because, as Propositions 1.7.148 and 1.7.150 show, the rate in question depends on the regularity of the functions and is hence not a meaningful notion on  $L^2$ . In particular, unlike the various mixing notions, it is not an invariant under measure-theoretic isomorphism, being, rather, a quantity associated with a smooth dynamical system.

### 1.7.19 Spectral Isomorphism and Invariants

One can study measure-preserving transformations by spectral analysis, i.e., via the Koopman operator  $U_f$  (Definition 1.7.31). We first note some of its basic properties.

**Proposition 1.7.152** *If  $f$  is a probability-preserving transformation, then*

1. *The eigenvalues of  $U_f$  lie on the unit circle.*
2. *The spectrum of  $U_f$  lies on the unit circle if  $f$  is invertible.*
3. *The eigenvalues of  $U_f$  form a subgroup of the unit circle.*
4. *Eigenfunctions of  $U_f$  for different eigenvalues are orthogonal.*

*Proof*

1. If  $A$  is an isometry and  $Av = \lambda v$ , then  $\|v\| = \|Av\| = \|\lambda v\| = |\lambda| \|v\|$ .
2. If  $A$  is unitary then  $r(A^{\pm 1}) \leq \|A^{\pm 1}\| = 1$ , so  $\sigma(A^{\pm 1}) \subset \{\lambda \mid |\lambda| \leq 1\}$ .  $A \in \text{Aut}(V)$  implies  $0 \notin \sigma(A)$  and hence  $\sigma(A^{-1}) = \{\lambda^{-1} \mid \lambda \in \sigma(A)\}$  because  $(1/\lambda)I - A^{-1}$  is invertible if and only if  $-\lambda A[(1/\lambda)I - A^{-1}] = \lambda I - A$  is.
3. If  $U_f(\varphi) = \lambda\varphi$  and  $U_f(\psi) = \mu\psi$ , then  $\mu\lambda^{-1}$  is also an eigenvalue:

$$U_f(\varphi \cdot \bar{\psi}) = U_f(\varphi) \overline{U_f(\psi)} = \mu \bar{\lambda} \cdot \varphi \cdot \bar{\psi} = \mu \lambda^{-1} \cdot \varphi \cdot \bar{\psi},$$

This shows closure under inverses (take  $\mu = 1$ ) and then under multiplication.

4. If  $U_f(\varphi) = \lambda\varphi$  and  $U_f(\psi) = \mu\psi$ , then

$$\langle \varphi, \psi \rangle = \langle U_f(\varphi), U_f(\psi) \rangle = \langle \lambda\varphi, \mu\psi \rangle = \lambda \bar{\mu} \langle \varphi, \psi \rangle = \lambda \mu^{-1} \langle \varphi, \psi \rangle,$$

so  $\lambda \mu^{-1} = 1$  or  $\langle \varphi, \psi \rangle = 0$ . □

If  $f, g$  are measure-theoretically isomorphic via  $h$  (i.e.,  $g = h \circ f \circ h^{-1}$ ) then  $U_f$ , and  $U_g$  are unitarily equivalent:

$$U_g = U_h^{-1} \circ U_f \circ U_h.$$

Thus spectral invariants of  $U_f$ , e.g., eigenvalues with their multiplicities or the spectrum, are invariants of measure-theoretic isomorphism of  $f$ .

**Definition 1.7.153** Two measure-preserving transformations are said to be *spectrally isomorphic* if their Koopman operators are unitarily equivalent. An invariant of spectral isomorphism is called a *spectral invariant*.

Since ergodicity is equivalent to 1 being a simple eigenvalue of the Koopman operator, we conclude

**Proposition 1.7.154** *Ergodicity is a spectral invariant.*

Indeed, ergodicity provides further information about eigenspaces.

**Proposition 1.7.155** *A probability-preserving transformation  $f$  is ergodic iff*

1. *All eigenfunctions have constant absolute value.*
2. *All eigenspaces are 1-dimensional.*

*Proof*

1.  $U_f(|\varphi|) = |U_f(\varphi)| = |\lambda||\varphi| = |\varphi|$ , so  $|\varphi|$  is invariant, hence constant a.e.
2. If  $\varphi, \psi$  are nonzero eigenfunctions for  $\lambda$ , then they are nonzero a.e. by 1., so  $\varphi/\psi$  is a well-defined invariant function, hence constant a.e.  $\square$

It is also easy to see the following.

**Proposition 1.7.156** *Mixing is a spectral invariant (Definition 1.7.153).*

*Proof* Suppose  $f: (X, \mu) \rightarrow (X, \mu)$  is mixing,  $g: (Y, \nu) \rightarrow (Y, \nu)$ ,  $W \circ U_f = U_g \circ W$ ,  $W$  unitary, and  $\varphi_i = W(\psi_i) \in L^2(Y, \nu)$  ( $i = 1, 2$ ). Then

$$\begin{aligned} \langle U_g^n(\varphi_1), \varphi_2 \rangle &= \langle U_g^n(W(\psi_1)), W(\psi_2) \rangle = \langle W(U_f^n(\psi_1)), W(\psi_2) \rangle = \langle U_f^n(\psi_1), \psi_2 \rangle \\ &\xrightarrow{n \rightarrow \infty} \langle \psi_1, \psi_2 \rangle = \langle W\psi_1, W\psi_2 \rangle = \langle \varphi_1, \varphi_2 \rangle. \end{aligned}$$

$\square$

The most obvious examples of ergodic measure-preserving transformations that are not weakly mixing are rotations and transformations built from them by an extension process. It is easy to see that this is no accident.

**Proposition 1.7.157** *Suppose a measure-preserving transformation  $f: (X, \mu) \rightarrow (X, \mu)$  is ergodic but not weakly mixing. Then  $f$  has a measure-theoretic factor (Definition 1.7.62) that is a rotation of either a circle or a finite set.*

*Proof* Since  $f$  is not weakly mixing, there is a nonconstant eigenfunction  $\varphi$  (Proposition 1.7.144); denote the corresponding eigenvalue by  $e^{i\alpha} \in S^1$ . Then  $|\varphi| = \text{const.}$  by Proposition 1.7.155. After scaling, we may assume  $|\varphi| = 1$ . Then  $\varphi: X \rightarrow S^1$  is the desired measure-theoretic factor map. To see this let  $\nu := \varphi_*\mu$  be as in (1.24). Since  $\varphi \circ f = \lambda\varphi$ , we find that  $\nu$  is an invariant ergodic measure for the rotation  $R_\alpha$  and  $\varphi(X) \subset S^1$  is  $R_\alpha$ -invariant. If  $\alpha \notin \mathbb{Q}$ , then unique ergodicity of  $R_\alpha$  implies that  $\nu$  is Lebesgue measure on  $S^1$  (and  $\varphi(X) \stackrel{\text{a.e.}}{=} S^1$ ), so  $R_\alpha$  is the desired measure-theoretic factor via  $\varphi$ . If  $\alpha \in \mathbb{Q}$  then the only ergodic  $R_\alpha$ -invariant measures are concentrated on a periodic point, and  $\nu$  must be one of them. Thus  $\varphi$  defines a measure-theoretic factor map to a periodic orbit of  $R_\alpha$  with the unique ergodic invariant measure.  $\square$

The following notion is natural for describing a situation in which a measure-preserving transformation is “spectrally rigid”:

**Definition 1.7.158** We say that  $f$  has *pure point spectrum* or *discrete spectrum* if  $f$  is ergodic and there is a basis of eigenfunctions of  $U_f$ .

*Remark 1.7.159* This clearly implies that  $U_f$  and hence  $f$  is invertible. The terminology goes back to that in Definition 1.6.7 in that the spectrum consists entirely of eigenvalues. Note also that by Proposition 1.7.155 these  $\lambda$  are pairwise distinct; this produces enough information for spectral isomorphism.

**Proposition 1.7.160** *Ergodic measure-preserving transformations with discrete spectrum and with the same eigenvalues are spectrally isomorphic.*

*Proof* For each eigenvalue map the corresponding eigenfunction for one transformation to that for the other (see Proposition 1.7.155); extend by linearity and continuity.  $\square$

*Remark 1.7.161* In this case the dynamics of  $U_f$  consists of a product of rotations of the eigenspaces; the essential information is contained in what happens to normalized eigenfunctions. This can be exploited to show that, in fact, here the eigenvalues determine  $f$  up to a measure-theoretic isomorphism.

To go substantially beyond the topological groups  $S^1$  and  $\mathbb{T}^n$ , we introduce characters in suitable generality.

**Definition 1.7.162 (Characters)** A *topological group* is a group endowed with a topology with respect to which all *left translations*  $L_{g_0}: g \mapsto g_0g$  and *right translations*  $R_{g_0}: g \mapsto gg_0$  as well as  $g \mapsto g^{-1}$  are homeomorphisms. If  $G$  is a locally compact abelian group then the group  $\hat{G}$  of *characters* is defined as the group of continuous homomorphisms  $G \rightarrow S^1 = \{z \in \mathbb{C} \mid |z| = 1\}$  with the topology of uniform convergence on compact sets (i.e., the compact-open topology).

For locally compact abelian groups we have the following:

**Theorem 1.7.163**

1.  $\hat{S}^1 = \mathbb{Z}$ ; every character is of the form  $z \mapsto z^k$ .
2.  $\hat{\mathbb{T}^n} = \mathbb{Z}^n$ ; every character is of the form  $(z_1, \dots, z_n) \mapsto z_1^{k_1} \dots z_n^{k_n}$ .
3.  $G$  is compact if and only if  $\hat{G}$  is discrete.
4.  $G$  is second countable if and only if  $\hat{G}$  is.
5.  $\hat{\hat{G}} \cong G$  via  $G \ni a \mapsto \alpha: g \mapsto g(a)$  (Pontryagin duality).
6.  $\widehat{G_1 \times G_2} = \hat{G}_1 \times \hat{G}_2$ .
7. If  $G$  is compact then  $\hat{G}$  is a complete orthonormal<sup>38</sup> system for  $L^2(G, \text{Haar})$ .

---

<sup>38</sup>  $\int \chi(g) d\lambda_G = \int \chi(hg) d\lambda_G = \int \chi(h)\chi(g) d\lambda_G = \chi(h) \int \chi(g) d\lambda_G$  for all  $h \in G$ , so  $\chi(\cdot) \equiv 1$  or  $\int \chi = 0$ ; since a ratio of characters is a character, this gives  $\langle \chi_1, \chi_2 \rangle^2 = \int \chi_1 / \chi_2 = \begin{cases} 1 & \text{if } \chi_1 = \chi_2, \\ 0 & \text{otherwise.} \end{cases}$

A translation  $T_{g_0}: g \mapsto g_0 g$  of a compact abelian group  $G$  preserves Haar measure  $\lambda_G$ , and characters are eigenfunctions. Thus we have

**Proposition 1.7.164** *Translations of compact abelian groups have discrete spectrum, i.e., there is a basis of eigenfunctions.*

The converse is one of the classical facts of ergodic theory.

**Theorem 1.7.165 (von Neumann Discrete Spectrum Theorem)** *Any two ergodic measure-preserving transformations with discrete spectrum that are spectrally isomorphic (i.e., have the same groups of eigenvalues) are measure-theoretically isomorphic. A complete system of invariants is given by the countable subgroup  $\Gamma < S^1$  of eigenvalues: A transformation whose group of eigenvalues is  $\Gamma$  is measure-theoretically isomorphic to the translation on the compact group  $\Gamma^*$  of characters of  $\Gamma$ , considered as a discrete group, by the character  $s_0$  that defines the inclusion  $\Gamma \hookrightarrow S^1$ . The invariant measure is Haar measure.*

*Proof* Let  $f: (X, \mu) \rightarrow (X, \mu)$  be an ergodic measure-preserving transformation with discrete spectrum and  $\Gamma$  the group of eigenvalues of  $U_f$ . Let  $x_0$  be a common Lebesgue point<sup>39</sup> for all eigenfunctions of  $U_f$ . For each eigenvalue  $\gamma \in \Gamma$  denote by  $\varphi_\gamma$  the unique eigenfunction whose Lebesgue value at  $x_0$  is 1. Then

$$\varphi_{\gamma_1 \gamma_2} = \varphi_{\gamma_1} \varphi_{\gamma_2}. \quad (1.48)$$

Identify  $\Gamma$  with the group of characters of the compact dual group  $\Gamma^*$  and denote the character on  $\Gamma^*$  corresponding to the evaluation at  $\gamma$  by  $\chi_\gamma$ . This gives orthonormal bases  $\{\varphi_\gamma\}_{\gamma \in \Gamma}$  and  $\{\chi_\gamma\}_{\gamma \in \Gamma}$  in the Hilbert spaces  $L^2(X, \mu)$  and  $L^2(\Gamma^*, \lambda)$  correspondingly, where  $\lambda$  is the normalized Haar measure.

Now,  $\varphi_\gamma \mapsto \chi_\gamma$  extends linearly to a unitary operator  $V: L^2(X, \mu) \rightarrow L^2(\Gamma^*, \lambda)$ , which is multiplicative on the eigenfunctions by (1.48). Their finite linear combinations are dense in  $L^2(X, \mu)$ , so  $V$  is generated by a measure-preserving invertible transformation  $h: (X, \mu) \rightarrow (\Gamma^*, \lambda)$ , and  $VU_fV^{-1}\chi_\gamma(s) = \gamma\chi_\gamma(s) = \chi_\gamma(s_0s)$  for any  $s \in \Gamma^*$ , hence  $hfh^{-1} = L_{s_0}$ .  $\square$

**Remark 1.7.166** Even invertible spectrally isomorphic measure-preserving ergodic transformations may fail to be measure-theoretically isomorphic.

Several notions that follow relate to a collection of spectral invariants called spectral measures. To introduce these, it is useful to define  $U_f^n$  for negative  $n$  even if  $f$  is not invertible: When  $n \in \mathbb{N}$  define  $U_f^{-n} := (U_f^*)^n$ , where “ $*$ ” denotes the adjoint defined by  $\langle U^* \varphi, \psi \rangle = \langle \varphi, U \psi \rangle$  for all  $\varphi, \psi \in L^2$ . (If  $f$ , hence  $U_f$ , is invertible, this coincides with the corresponding iterate of the inverse.)

---

<sup>39</sup> $\varphi_i(x_0) = \lim_{h \rightarrow 0} \frac{1}{2h} \int_{x_0-h}^{x_0+h} \varphi_i(x) dx.$

**Definition 1.7.167** If  $\mu$  is an invariant Borel probability measure for  $f: X \rightarrow X$ , then the *spectral measure* of  $\varphi \in L^2(\mu) \setminus \{0\}$  is the unique measure  $\nu_\varphi$  on  $S^1$  with Fourier coefficients  $\varphi_{(n)} := \langle U_f^n(\varphi), \varphi \rangle = \int_{S^1} z^n d\nu_\varphi$  for  $n \in \mathbb{Z}$ .

That the  $\varphi_{(n)}$  are the Fourier coefficients of a measure on  $S^1$  follows from

**Theorem 1.7.168 (Carathéodory–Herglotz)**  $(b_n)_{n \in \mathbb{Z}}$  are the Fourier coefficients of a measure on  $S^1$  iff  $b_n = \overline{b_{-n}}$  and  $\sum_{|n|, |m| \leq N} b_{n-m} a_m \overline{a_n} \geq 0$  for  $N \in \mathbb{N}$  and  $(a_n)_{n \in \mathbb{Z}}$ .

It applies because  $\langle U_f^n(\varphi), \varphi \rangle = \overline{\langle \varphi, U_f^n(\varphi) \rangle} = \overline{\langle U_f^{-n}(\varphi), \varphi \rangle}$  and

$$0 \leq \left\| \sum_{|n| \leq N} \overline{a_n} U_f^n(\varphi) \right\|^2 = \left\langle \sum_{|n| \leq N} \overline{a_n} U_f^n(\varphi), \sum_{|m| \leq N} \overline{a_m} U_f^m(\varphi) \right\rangle = \sum_{|n|, |m| \leq N} \langle U_f^{n-m}(\varphi), \varphi \rangle \overline{a_n} a_m.$$

**Remark 1.7.169** If  $U_f \varphi = \lambda \varphi$ , then  $\nu_\varphi = \delta_\lambda$  because  $\lambda^n = \langle U_f^n \varphi, \varphi \rangle = \int_{S^1} z^n d\nu_\varphi$ . Thus, when  $f$  has discrete spectrum,  $U_f$  is completely determined by spectral measures that are discrete (or point) measures, another reason for the terminology. The description of the dynamics of  $U_f$  for  $f$  with discrete spectrum suggests that having discrete spectrum is a strengthening of the notion of ergodicity analogous to minimality as a strengthening of transitivity. We now present a notion that corresponds to strengthening transitivity in the direction of mixing.

**Definition 1.7.170** An invertible probability-preserving transformation  $f$  of a Lebesgue space is said to have *countable Lebesgue spectrum* (see Definition 1.6.7) if there is an orthonormal set  $1 = \varphi_0, \varphi_1, \dots$  in  $L^2$  such that  $\{1\} \cup \{U_f^n(\varphi_i) \mid i \in \mathbb{N}, n \in \mathbb{Z}\}$  is a complete orthonormal set in  $L^2$ —so the  $\varphi_{ik} := U_f^n(\varphi_i) \in 1^\perp$  are pairwise distinct and orthogonal.

**Remark 1.7.171** The terminology indicates that each  $\nu_{\varphi_i}$  is Lebesgue measure. Interestingly (and obviously) these are all spectrally equivalent.

**Proposition 1.7.172** Any pair of invertible probability-preserving transformations of Lebesgue spaces that have countable Lebesgue spectrum are spectrally isomorphic.

*Proof* Map the  $\varphi_i$  for one of the measure-preserving transformations to the  $\varphi_i$  for the other; extend linearly.  $\square$

**Proposition 1.7.173** Measure-preserving transformations with countable Lebesgue spectrum are mixing.

*Proof* By Proposition 1.7.142 it suffices to consider  $\varphi_{i,k} := U_f^k(\varphi_i)$  as in Definition 1.7.170 and to note that  $\langle U_f^n(\varphi_{i,k}), \varphi_{i,k} \rangle \xrightarrow[n \rightarrow \infty]{} 0$ .  $\square$

$\underbrace{\hspace{10em}}_{=0 \text{ when } n \neq 0}$

Weak mixing (Definition 1.7.118), an intermediate property between ergodicity and mixing, turns out to be a spectral invariant as well. The first step towards seeing this was Proposition 1.7.144. We conclude by proving the converse.



**Definition 1.7.174** We say that  $f$  has *continuous spectrum* if all eigenfunctions of the Koopman operator  $U_f$  of  $f$  are constant.

*Remark 1.7.175* A motivation for the term “continuous spectrum” is that one can in this case show that all spectral measures are nonatomic (for functions in  $1^\perp$ ): If  $\lambda \in S^1$  then a weak accumulation point  $\psi$  of  $1/n \sum_{k=0}^{n-1} \lambda^{-k} U_f^k(\varphi)$  is an eigenfunction for  $\lambda$ , and it is nonzero if  $0 \neq \langle \psi, \varphi \rangle = \nu_f(\lambda)$ .

**Proposition 1.7.176** *A probability-preserving transformation is weakly mixing if and only if it has continuous spectrum.*

*Proof* We will use that  $\Delta := \{(z, z) \mid z \in S^1\}$  is a  $(\nu_\varphi \times \nu_\varphi)$ -null set by absolute continuity of  $\nu_\varphi$  and the Fubini Theorem. If  $f$  has continuous spectrum, then

$$\begin{aligned} \varphi \in 1^\perp &\Rightarrow \frac{1}{n} \sum_{k=0}^{n-1} \left| \int_X \varphi \circ f^k \bar{\varphi} \right|^2 = \frac{1}{n} \sum_{k=0}^{n-1} \left| \int_{S^1} z^k d\nu_\varphi(z) \right|^2 \\ &= \frac{1}{n} \sum_{k=0}^{n-1} \underbrace{\int_{S^1} z^k d\nu_\varphi(z) \int_{S^1} \bar{w}^k d\nu_\varphi(w)}_{= \int_{S^1 \times S^1} (z\bar{w})^k d(\nu_\varphi \times \nu_\varphi)(z, w)} \\ &= \int_{S^1 \times S^1} \underbrace{\frac{1}{n} \sum_{k=0}^{n-1} (z\bar{w})^k}_{= \frac{(z\bar{w})^{n+1} - 1}{n(z\bar{w} - 1)} \rightarrow 0 \text{ off } \Delta \text{ \& bounded}} d(\nu_\varphi \times \nu_\varphi)(z, w) \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

□

**Acknowledgement** Partially supported by the Committee on Faculty Research Awards of Tufts University.

## References

- [Al68] V.M. Alekseev, Quasirandom dynamical systems. I. Quasirandom diffeomorphisms. *Matematicheskii Sbornik*. (N. S.) **76**(118), 72–134 (1968); Invariant Markov subsets of diffeomorphisms. *Akademiya Nauk SSSR i Moskovskoe Matematicheskoe Obshchestvo. Uspekhi Matematicheskikh Nauk* **23**(2) (140), 209–210 (1968)
- [AnPo37] A. Andronov, L. Pontrjagin, Systèmes grossiers. *Comptes Rendus (Doklady) de l’Académie des Sciences de l’URSS* **14**(5), 247–250 (1937)
- [An69] D.V. Anosov, *Geodesic Flows on Closed Riemann Manifolds with Negative Curvature*. Trudy Mat. Institute. V.A. Steklova, vol. 90 (American Mathematical Society, Providence, RI, 1969/1967)
- [AnSi67] D.V. Anosov, Y. Sinai, Some smooth ergodic systems. *Russ. Math. Surv.* **22**(5), 103–167 (1967)
- [ArAv68] V.I. Arnol’d, A. Avez, *Problèmes ergodiques de la mécanique classique*. Monographies Internationales de Mathématiques Modernes, vol. 9 (Gauthier–Villars, Paris, 1967); Ergodic problems of classical mechanics. Translated from the French by André Avez (W. A. Benjamin, New York, Amsterdam, 1968)

- [Ba95] W. Ballmann, *Lectures on Spaces of Nonpositive Curvature, With an Appendix by Misha Brin*. DMV Seminar, vol. 25 (Birkhäuser, Basel, 1995)
- [Ba97] J. Barrow-Green, *Poincaré and the Three Body Problem*. American Mathematical Society/London Mathematical Society History of Mathematics, vol. 11 (American Mathematical Society, Providence, RI, 1997)
- [BaPe01] L. Barreira, Y.B. Pesin, Lectures on Lyapunov exponents and smooth ergodic theory, in *American Mathematical Society Proceedings of Symposia in Pure Mathematics*, Summer Research Institute, Seattle, WA, 1999 (2001)
- [BaPe07] L. Barreira, Y.B. Pesin, *Nonuniform Hyperbolicity: Dynamics of Systems with Nonzero Lyapunov Exponents*. Encyclopedia of Mathematics and its Applications, vol. 115 (Cambridge University Press, New York, 2007)
- [BaPe13] L. Barreira, Y.B. Pesin, *Introduction to Smooth Ergodic Theory*. Graduate Studies in Mathematics, vol. 148 (American Mathematical Society, Providence, RI, 2013)
- [Bi27] G.D. Birkhoff, On the periodic motions of dynamical systems. *Acta Math.* **50**, 359–379 (1927)
- [Bi35] G.D. Birkhoff, Nouvelles recherches sur les systèmes dynamique. *Memoriae Pont. Acad. Sci. Novi Lyncaei* 1(3), 85–216 (1935)
- [Bo75] R. Bowen,  $\omega$ -limit sets for axiom A diffeomorphisms. *J. Differ. Equ.* **18**, 333–339 (1975)
- [Bo78] R. Bowen, *On Axiom A Diffeomorphisms*. Regional Conference Series in Mathematics, vol. 35 (American Mathematical Society, Providence, RI, 1978)
- [Br02] M. Brin, G. Stuck, *Introduction to Dynamical Systems* (Cambridge University Press, Cambridge, 2002)
- [BuMaWi] K. Burns, H. Masur, A. Wilkinson, The Weil–Petersson geodesic flow is ergodic. *Ann. Math.* (2) **175**(2), 835–908 (2012)
- [BuSi73] L.A. Bunimovich, Y.G. Sinai, The fundamental theorem of the theory of scattering billiards. *Mat. Sb. (N.S.)* **90**(132), 415–431, 479 (1973)
- [Ca50] M.L. Cartwright, Forced oscillations in nonlinear systems, in *Contributions to the Theory of Nonlinear Oscillations*. Annals of Mathematics Studies, vol. 20 (Princeton University Press, Princeton, NJ, 1950), pp. 149–241
- [CaLi45] M.L. Cartwright, J.E. Littlewood, On non-linear differential equations of the second order. I. The equation  $\ddot{y} - k(1 - y^2)y + y = b\lambda k \cos(\lambda t + a)$ ,  $k$  large. *J. Lond. Math. Soc.* **20**, 180–189 (1945)
- [ChMa06] N. Chernov, R. Makarian, *Chaotic Billiards*. Mathematical Surveys and Monographs, vol. 127 (American Mathematical Society, Providence, RI, 2006)
- [ChTr98] N. Chernov, S. Troubetzkoy, Ergodicity of billiards in polygons with pockets. *Nonlinearity* **11**, 1095–1102 (1998)
- [Co07] Y. Coudène, On invariant distributions and mixing. *Ergod. Theory Dyn. Syst.* **27**(1), 109–112 (2007)
- [Co16] Y. Coudène, *Théorie ergodique et systèmes dynamiques* (EDP Sciences, Les Ulis, 2013); English translation: *Ergodic Theory and Dynamical Systems*. Universitext (Springer, Berlin, 2016)
- [CoHaTr] Y. Coudène, B. Hasselblatt, S. Troubetzkoy, Multiple mixing from weak hyperbolicity by the Hopf argument, in *Proceedings of the Conference “Probability in Dynamics” Rio 2014*. Stochastics and Dynamics, vol. 16(2) (2016)
- [CoNi08] E.M. Coven, Z.H. Nitecki, On the genesis of symbolic dynamics as we know it. *Colloq. Math.* **110**(2), 227–242 (2008)
- [Du91] P.M.M. Duhem, *La Théorie Physique: Son Objet, Sa Structure* (Marcel Rivière & Cie., Paris, 1906, 1914); transl. *The Aim and Structure of Physical Theory* (Princeton Science Library, Princeton University Press, Princeton, NJ, 1991)
- [EhEh] P. Ehrenfest, T. Ehrenfest, Begriffliche Grundlagen der statistischen Auffassung in der Mechanik, in *Encyklopaedie der Mathematischen Wissenschaften*, 4: Art. 32 (Teubner, Leipzig, 1912), pp. 1–90
- [Gr39] A.M.C. Grant, Surfaces of negative curvature and permanent regional transitivity. *Duke Math. J.* **5**(2), 207–229 (1939)

- [GrLa09] J. Green, J. LaDuke, *Pioneering Women in American Mathematics: The Pre-1940 PhD's*. History of Mathematics, vol. 34 (American Mathematical Society, Providence, RI, 2009)
- [Ha98] J.S. Hadamard, Les surfaces à courbures opposées et leurs lignes géodésiques. *Journal de Mathématiques pures et appliquées* (5) **4**, 27–73 (1898)
- [Ha01] J.S. Hadamard, Sur l'itération et les solutions asymptotiques des équations différentielles. *Bulletin de la Société Mathématique de France* **29**, 224–228 (1901). Translated into English in this volume
- [HaKa03] B. Hasselblatt, A. Katok, *Dynamics: A First Course — With a Panorama of Recent Developments* (Cambridge University Press, Cambridge, 2003)
- [He36] G.A. Hedlund, Fuchsian groups and transitive horocycles. *Duke Math. J.* **2**, 530–542 (1936)
- [He39a] G.A. Hedlund, Fuchsian groups and mixtures. *Ann. Math. (2)* **40**(2), 370–383 (1939)
- [He39b] G.A. Hedlund, The dynamics of geodesic flows. *Bull. Am. Math. Soc.* **45**, 241–260 (1939)
- [Ho36] E. Hopf, Fuchsian groups and ergodic theory. *Trans. Am. Math. Soc.* **39**, 299–314 (1936)
- [Ho39] E. Hopf, *Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung*. Berichte über die Verhandlungen der Sächsischen Akademie der Wissenschaften zu Leipzig, Mathematisch-Physikalische Klasse, vol. 91 (Hirzel, Leipzig, 1939), pp. 261–304
- [Ka79] A. Katok, Bernoulli diffeomorphisms on surfaces. *Ann. Math. (2)* **110**(3), 529–547 (1979)
- [Ka81] A. Katok, Dynamical systems with hyperbolic structure. *Am. Math. Soc. Transl. (2)* **116**, 43–95 (1981)
- [Ka94] A. Katok, with the collaboration of Keith Burns, Infinitesimal Lyapunov functions, invariant cone families and stochastic properties of smooth dynamical systems. *Ergod. Theory Dyn. Syst.* **14**, 757–785 (1994)
- [KaHa95] A. Katok, B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*. Encyclopedia of Mathematics and its Applications, vol. 54 (Cambridge University Press, Cambridge, 1995)
- [Ko29] P. Koebe, Riemannsche Mannigfaltigkeiten und nicht euklidische Raumformen. Sitzungsberichte der Preußischen Akademie der Wissenschaften **I**, 164–196 (1927); **II**, **III**, 345–442 (1928); **IV**, 414–557 (1929); **V**, **VI**, 304–364, 504–541 (1930); **VII**, 506–534 (1931)
- [La95] P.S. de Laplace, *Essai philosophique sur les probabilités (Philosophical Essay on Probabilities)* (1812); translated from the fifth French edition of 1925 by Andrew I. Dale (Springer, New York, 1995)
- [Le49] N. Levinson, A second order differential equation with singular solutions. *Ann. Math. (2)* **50**, 127–153 (1949)
- [Li57] J.E. Littlewood, On non-linear differential equations of the second order. IV. Th. general equation  $\ddot{y} + kf(y)\dot{y} + g(y) = bkp(\varphi)$ ,  $\varphi = t + \alpha$ . *Acta Math.* **98**, 1–110 (1957)
- [Lo29] F. Löbell, Über die geodätischen Linien der Clifford–Kleinschen Flächen. *Math. Z.* **30**, 572–607 (1929)
- [Ma70] G. Margulis, Certain measures that are connected with Y-flows on compact manifolds. *Funkcional. Anal. i Priložen.* **4**(1), 62–76 (1970)
- [OrWe98] D. Ornstein, B. Weiss, On the Bernoulli nature of systems with some hyperbolic structure. *Ergod. Theory Dyn. Syst.* **18**(2), 441–56 (1998)
- [Os68] V.I. Oseledets, A Multiplicative Ergodic Theorem. Liapunov characteristic numbers for dynamical systems. *Trudy Moskovskogo Matematicheskogo Obščestva* **19**, 179–210 (1968); *Trans. Moscow Math. Soc.* **19**, 197–221 (1968)
- [Pe28] O. Perron, *Über Stabilität und asymptotisches Verhalten der Integrale von Differentialgleichungssystemen*. *Math. Z.* **29**(1), 129–160 (1928)

- [Pe62] M.M. Peixoto, On structural stability. *Ann. Math. (2)* **69**, 199–222 (1959); Structural stability on two-dimensional manifolds. *Topology* **1**, 101–120 (1962)
- [Pe76] Y.B. Pesin, Families of invariant manifolds corresponding to nonzero characteristic exponents. *Mathematics of the USSR, Izvestia* **10**(6), 1261–1305 (1976)
- [PeSeZh] Y.B. Pesin, S. Senti, K. Zhang, Thermodynamics of the Katok map. Preprint. arXiv:1603.08556. <https://www.cambridge.org/core/journals/ergodic-theory-and-dynamical-systems/article/thermodynamics-of-the-katok-map/8F39D44317F6CA98A7BC043ACF06D33A>
- [Pl13] M. Plancherel, Beweis der Unmöglichkeit ergodischer mechanischer Systeme. *Ann. Phys. (4)* **42**, 1061–1063 (1913)
- [Po90] H. Poincaré, Sur le problème des trois corps et les equations de la dynamique. *Acta Math.* **13**, 1–270 (1890)
- [Po94] H. Poincaré, Sur la théorie cinétique des gas. *Revue Générale des Sciences pures et appliquées* **5**, 513–521 (1894)
- [PuSh72] C.C. Pugh, M. Shub, Ergodicity of Anosov actions. *Invent. Math.* **15**, 1–23 (1972)
- [Ro13] A. Rosenthal, Beweis der Unmöglichkeit ergodischer Gassysteme. *Ann. Phys. (4)* **42**, 796–806 (1913)
- [Se35] W.P. Seidel, On a metric property of Fuchsian groups. *Proc. Natl. Acad. Sci.* **21**, 475–478 (1935)
- [Si70] Y.G. Sinai, Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards. *Uspehi Mat. Nauk* **25**(2) (152), 141–192 (1970)
- [Sm60] S. Smale, On dynamical systems. *Boletín de la Sociedad Matemática Mexicana* (2) **5**, 195–198 (1960)
- [Sm63] S. Smale, A structurally stable differentiable homeomorphism with an infinite number of periodic points, in *Qualitative Methods in the Theory of Non-linear Vibrations*. Proceedings of International Symposium on Non-linear Vibrations, vol. II, 1961 (Izdat. Akad. Nauk Ukrain. SSR, Kiev, 1963), pp. 365–366
- [Sm65] S. Smale, Diffeomorphisms with many periodic points, in *Differential and Combinatorial Topology*. A Symposium in Honor of Marston Morse, ed. by S.S. Cairns (Princeton University Press, Princeton, NJ, 1965), pp. 63–80
- [Sm67] S. Smale, Differentiable dynamical systems. *Bull. Am. Math. Soc.* **73**, 747–817 (1967)
- [Sm98] S. Smale, Finding a horseshoe on the beaches of Rio. *Math. Intell.* **20**(1), 39–44 (1998)
- [Wa78] P. Walters, On the pseudo-orbit tracing property and its relationship to stability, in *The Structure of Attractors in Dynamical Systems*. Proceedings of Conference on North Dakota State University, Fargo, ND, 1977. *Lecture Notes in Mathematics*, vol. 668 (Springer, Berlin, 1978), pp. 231–244
- [Yo95] J.-C. Yoccoz, Introduction to hyperbolic dynamics, in *Real and Complex Dynamical Systems*. Proceedings of the NATO Advanced Study Institute held in Hillerød, June 20–July 2, 1993, ed. by B. Branner, P. Hjorth. NATO Advanced Science Institutes Series C: Mathematical and Physical Sciences, vol. 464 (Kluwer Academic, Dordrecht, 1995), pp. 265–291
- [Yo98] L.-S. Young, Developments in chaotic dynamics. *Not. Am. Math. Soc.* **45**(10), 1318–1328 (1998)
- [Zu02] J. Zund, George David Birkhoff and John Von Neumann: a question of priority and the ergodic theorems, 1931–1932. *Hist. Math.* **29**(2), 138–156 (2002)

<http://www.springer.com/978-3-319-43058-4>

Ergodic Theory and Negative Curvature

CIRM Jean-Morlet Chair, Fall 2013

Hasselblatt, B. (Ed.)

2017, VII, 328 p. 68 illus., 17 illus. in color., Softcover

ISBN: 978-3-319-43058-4