HANDBOOK OF DYNAMICAL SYSTEMS

VOLUME 1B

Edited by
B. Hasselblatt
A. Katok

ELSEVIER
NORTH-HOLLAND

HANDBOOK OF DYNAMICAL SYSTEMS

Volume 1B

This page intentionally left blank

HANDBOOK OF DYNAMICAL SYSTEMS

Volume 1B

Edited by

B. HASSELBLATT
Tufts University, Medford, MA 02155, USA

A. KATOK
The Pennsylvania State University, University Park, PA 16802, USA

2006

ELSEVIER
Amsterdam • Boston • Heidelberg • London • New York • Oxford • Paris
San Diego • San Francisco • Singapore • Sydney • Tokyo

ELSEVIER B.V.
Radarweg 29
P.O. Box 211, 1000 AE
Amsterdam, The Netherlands

ELSEVIER Inc.
525 B Street, Suite 1900
San Diego, CA 92101-4495
USA

ELSEVIER Ltd
The Boulevard, Langford Lane
Kidlington, Oxford OX5 1GB
UK

ELSEVIER Ltd
84 Theobalds Road
London WC1X 8RR
UK

First edition 2006

Library of Congress Cataloging-in-Publication Data
A catalog record is available from the Library of Congress.

British Library Cataloguing in Publication Data
A catalogue record is available from the British Library.

ISBN: 0-444-52055-4

♾ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).
Printed in The Netherlands.

Preface

This second half of Volume 1 of this Handbook follows Volume 1A, which was published in 2002. The contents of these two tightly integrated parts taken together come close to a realization of the program formulated in the introductory survey "Principal Structures" of Volume 1A.

The present volume contains surveys on subjects in four areas of dynamical systems: Hyperbolic dynamics, parabolic dynamics, ergodic theory and infinite-dimensional dynamical systems (partial differential equations). These areas, with the exception of the last one, are also represented in Volume 1A.

In Volume 1A the chapters in hyperbolic dynamical systems cover uniformly hyperbolic dynamical systems (general properties, Markov partitions and Gibbs measures, periodic orbits and ζ-functions) and hyperbolic dynamical systems arising in Riemannian geometry. The present volume (1B) contains chapters on nonuniformly hyperbolic dynamical systems (to which the survey on Hyperbolic dynamics and Riemannian geometry in Volume 1A is closely related), on partially hyperbolic dynamical systems and on homoclinic bifurcations, dominated splitting and robust transitivity (both of which have developed rapidly in the last few years), as well as an account of random dynamics, which covers aspects of an area related to hyperbolic dynamics and complements the survey on random dynamics in Volume 1A. Taken together, this volume and Volume 1A thereby provide a comprehensive overview of both the foundations and the current state of art in hyperbolic dynamics and immediately adjacent areas.

In addition to an overview in the chapter "Principal Structures", parabolic dynamics is represented in Volume 1A by a detailed discussion of unipotent homogeneous systems in Section 3 of the chapter on dynamics of subgroup actions on homogeneous spaces and by the entire chapter on rational billiards and flat structures. The latter area has experienced explosive growth in recent years and the existing expository literature is far from sufficient. Anton Zorich showed great vision and exercised spirited leadership resulting in a cluster of chapters in the present volume on the subject of parabolic dynamics written by leading researchers in the area.

Volume 1A covers several aspects of ergodic theory, including the core subjects of entropy, isomorphisms and Kakutani equivalence as well as the ergodic theory of smooth or algebraic dynamical systems, and the chapter on actions of "large" groups. The present volume expands the treatment of ergodic theory with four additional chapters covering spectral theory, joinings and combinatorial constructions, ergodic theorems, multiple recurrence and related topics, and relations with topological dynamics. The coverage of ergodic theory in these two parts of Volume 1, while somewhat less comprehensive than that

of hyperbolic dynamics, is considerably more broad and thorough than that provided in other existing sources.

The final cluster of chapters in the present volume, for which Sergei Kuksin provided inspiration and leadership, discusses partial differential equations from the point of view of dynamical systems. The first of these is about attractors, the other two are about Hamiltonian PDE in finite and infinite volume, respectively.

Some of the subjects introduced and outlined in the survey "Principal Structures" in Volume 1A will be covered in the forthcoming Volume 3 of this Handbook. Among those are certain aspects of elliptic dynamics, such as KAM theory and its applications, as well as complex dynamics.

We would like to thank the authors of the chapters in this pair of volumes for investing their time so generously in this project, and for writing surveys of such high quality. We also owe much gratitude to Sergei Kuksin and Anton Zorich for the efforts they invested in the sections of the present volume on infinite-dimensional and parabolic dynamics, respectively. Numerous other mathematicians took interest in the project and read drafts of various surveys or major portions thereof. This resulted in numerous valuable suggestions. This interest also provided great encouragement for the authors and editors and helped to bring this extensive project to successful completion. We are also grateful for the expertise and craftsmanship that Elsevier and VTeX employed to produce volumes of the highest quality.

We are indebted to Kathleen Hasselblatt and Svetlana Katok for their support and patience while we worked on this volume.

Boris Hasselblatt and Anatole Katok

List of Contributors

Babin, A.V., *University of California, Irvine, CA* (Ch. 14)
Bambusi, D., *Politecnico di Milano, 20133 Milano, Italy* (Ch. 15/Appendix)
Barreira, L., *Instituto Superior Técnico, Lisboa, Portugal* (Ch. 2)
Bergelson, V., *The Ohio State University, Columbus OH* (Ch. 12)
Eskin, A., *University of Chicago, Chicago, IL* (Ch. 9)
Forni, G., *Northwestern University, Evanston, IL*; *Université de Paris-Sud, Orsay, France* (Ch. 8)
Glasner, E., *Tel Aviv University, Tel Aviv, Israel* (Ch. 10)
Hasselblatt, B., *Tufts University, Medford, MA* (Ch. 1)
Hubert, P., *Institut de Mathématiques de Luminy, Marseille, France* (Ch. 6)
Katok, A., *The Pennsylvania State University, University Park, PA* (Ch. 11)
Kifer, Y., *The Hebrew University, Jerusalem, Israel* (Ch. 5)
Kuksin, S.B., *Heriot-Watt University, Edinburgh, UK*; *Steklov Institute of Mathematics, Moscow, Russia* (Ch. 15)
Leibman, A., *The Ohio State University, OH* (Ch. 12/Appendix A)
Liu, P.-D., *Peking University, Beijing, PR China* (Ch. 5)
Luzzatto, O., *Imperial College, London, UK* (Ch. 3)
Masur, H., *UIC, Chicago, IL* (Ch. 7)
Nevo, A., *Technion, Haifa, Israel* (Ch. 13)
Pesin, Ya., *The Pennsylvania State University, University Park, PA* (Ch. 1, 2)
Pujals, E.R., *IMPA, Rio de Janeiro, RJ, Brazil* (Ch. 4)
Quas, A., *University of Memphis, Memphis, TN* (Ch. 12/Appendix B)
Sambarino, M., *CMAT-Facultad de Ciencias, Montevideo, Uruguay* (Ch. 4)
Sarig, O., *The Pennsylvania State University, University Park, PA* (Ch. 2/Appendix)
Schmidt, T., *Oregon State University, Corvallis, OR* (Ch. 6)
Thouvenot, J.-P., *Université de Paris VI, Paris, France* (Ch. 11)
Weinstein, M.I., *Columbia University, New York, NY* (Ch. 16)
Weiss, B., *Hebrew University of Jerusalem, Jerusalem, Israel* (Ch. 10)
Wierdl, H., *University of Memphis, Memphis, TN* (Ch. 12/Appendix B)

This page intentionally left blank

Contents

Contents of Volume 1A

CHAPTER 1

Partially Hyperbolic Dynamical Systems

Boris Hasselblatt

Department of Mathematics, Tufts University, Medford, MA 02144, USA
E-mail: *boris.hasselblatt@tufts.edu*
url: *http://www.tufts.edu/~bhasselb*

Yakov Pesin

Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, USA
E-mail: *pesin@math.psu.edu*
url: *http://www.math.psu.edu/pesin/*

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

1. Introduction

1.1. *Motivation*

1.1.1. *Smooth ergodic systems* The flows and maps that arise from equations of motion in classical mechanics preserve volume on the phase space, and their study led to the development of ergodic theory.

In statistical physics, the Boltzmann–Maxwell ergodic hypothesis, designed to help describe equilibrium and nonequilibrium systems of many particles, prompted a search for ergodic mechanical systems. In geometry, the quest for ergodicity led to the study of geodesic flows on negatively curved manifolds, where Eberhard Hopf provided the first and still only argument to establish ergodicity in the case of nonconstantly negatively curved surfaces [57]. Anosov and Sinai, in their aptly entitled work "Some smooth Ergodic Systems" [10] proved ergodicity of geodesic flows on negatively curved manifolds of any dimension.

With the development of the modern theory of dynamical systems and the availability of the Birkhoff ergodic theorem the impetus to find ergodic dynamical systems and to establish their prevalence grew stronger. Birkhoff conjectured that volume-preserving homeomorphisms of a compact manifold are generically ergodic.

1.1.2. *Hyperbolicity* The latter 1960s saw a confluence of the investigation of ergodic properties with the Smale program of studying structural stability, or, more broadly, the understanding of the orbit structure of generic diffeomorphisms. The aim of classifying (possibly generic) dynamical systems has not been realized, and there are differing views of whether it will be. Current efforts in this direction are related to the Palis conjecture (see [4]). A promising step towards understanding generic smooth systems would clearly be an understanding of structurally stable ones, and one of the high points in the theory of smooth dynamical systems is that this has been achieved: Structural stability has been found to characterize hyperbolic dynamical systems [2].

Structural stability implies that all topological properties of the orbit structure are robust. Of these, topological transitivity has a particularly natural measurable analog, namely, ergodicity. On one hand, then, robust topological transitivity of hyperbolic dynamical systems motivated the search for broader classes of dynamical systems that are robustly transitive [29]. On the other hand, this, and the fact that volume-preserving hyperbolic dynamical systems are ergodic (with respect to volume) may have led Pugh and Shub to pose a question at the end of [56] that amounts to asking whether ergodic toral automorphisms are *stably ergodic*, i.e., whether all their volume-preserving C^1 perturbations are ergodic. They later conjectured that stable ergodicity is open and dense among volume-preserving partially hyperbolic C^2 diffeomorphisms of a compact manifold.

1.1.3. *Partial hyperbolicity* In this chapter we aim to give an account of significant results about partially hyperbolic systems. The pervasive guiding principle in this theory is that hyperbolicity in the system provides the mechanism that produces complicated dynamics in both the topological and statistical sense, and that, with respect to ergodic properties,

it does so in essence by overcoming the effects of whatever nonhyperbolic dynamics may be present in the system.

We should point out from the start that the desired dynamical qualities (such as transitivity and ergodicity) are of an indecomposability type and evidently fail, for example, for the Cartesian product of an Anosov diffeomorphism with the identity. Accordingly, suitable extra hypotheses of some sort will always be present to exclude this obvious reducibility and other less obvious ones (Section 7).

The ideas and methods in the study of partially hyperbolic dynamical systems extend those in the theory of uniformly hyperbolic dynamical systems, parts of which are briefly presented in [2], and go well beyond that theory in several aspects. Outside of these handbook volumes, accounts of uniformly hyperbolic dynamical systems from many points of view abound; [62] provides a textbook exposition that provides sufficient background. A condensed version of the material from [62] on hyperbolicity is contained in [30], which adds a useful account of absolute continuity and ergodicity in that context. Partial hyperbolicity is also surveyed in [34,79], with different emphasis. The one source that provides the most proofs of results only stated here is [71], which we recommend for further study of the subject. The study of partial hyperbolicity developed with two objects in mind: stable ergodicity and robust transitivity. We pay more attention to the first of these, and the recent book by Bonatti, Díaz and Viana [16] covers the second one in more detail.

1.1.4. *Extensions of classical complete hyperbolicity* While classical hyperbolicity appeared in the 1960s, partial hyperbolicity was introduced in the early 1970s by Brin and Pesin [29] motivated by the study of frame flows, and it also arose naturally from the work of Hirsch, Pugh and Shub on normal hyperbolicity [56].

Partial hyperbolicity is but one possible extension of the notion of classical (complete) hyperbolicity, or, in fact, a pair of extensions. Classical hyperbolicity can be described as requiring that the possible uniform rates of exponential relative behavior of orbits come in two collections on either side of 1 or by requiring that the Mather spectrum of the system (Section 2.3) consists of parts inside and outside of the unit circle. Partial hyperbolicity (in the broad sense, Definition 2.1) merely requires that the Mather spectrum consists of two parts that are separated by some circle centered at the origin (not necessarily the unit circle), and partial hyperbolicity (in the now prevalent narrower sense, Definition 2.7) requires that the Mather spectrum has 3 annular parts of which the inner one lies inside the unit circle and the outer one lies outside of the unit circle (Theorem 2.16).

There has been some shift in terminology over time, and what exactly is meant by "partial hyperbolicity" without a further attribute often has to be inferred from the context. Furthermore, there are still minor variations in naming the various "flavors" of this notion. Partial hyperbolicity in the broad sense is the concept for which one can extend the theory of invariant distributions and foliations from the context of classical complete hyperbolicity in the most direct way (the corresponding results are presented in Sections 2 and 3). These are also the results that describe the stability of trajectories and usually precede the study of topological and ergodic properties of the system. Accordingly, in those early days, "partially hyperbolic" by itself referred to the broader notion in Definition 2.1. (This notion is also known in the study of ordinary differential equations as the presence of a *dichotomy*. We will not use this synonym.) The results on central and intermediate distributions and

foliations as well as accessibility are of interest principally when one considers partially hyperbolic systems in the narrower sense of Definition 2.7, where a central direction is present (as well as two more directions that have stronger contraction and expansion, respectively). This therefore used to be called "partial hyperbolicity in the narrow sense", but recent work has so much focused on this situation that "in the narrow sense" is usually dropped. In this respect we conform to the current majority choice, but retain the option of emphasizing greater generality by using the notion of partial hyperbolicity in the broad sense of Definition 2.1.

One can weaken the notion of partial hyperbolicity in the broad sense to a semiuniform one, namely to that of having a *dominated splitting* [4], where the rates are separated by a uniform factor but the location of the gap is allowed to vary with the point (see p. 44). Put differently, partial hyperbolicity directly constrains the Mather spectrum (see Section 2.3) and the presence of a dominated splitting does not. Since partial hyperbolicity implies the presence of a dominated splitting, results proved for dynamical systems with a dominated splitting apply to those as in Definition 2.1. Dominated splittings appear in work directed at the (strong) Palis conjecture (that the C^1-generic diffeomorphism is either Ω-stable or the limit of diffeomorphisms with homoclinic tangencies or hetero-dimensional cycles) [4], but the objectives of this chapter versus those in that by Pujals and Sambarino [4] are fairly different. As one intriguing connection one might point out that it seems reasonable to suppose that the presence of a dominated splitting may be necessary for stable ergodicity [4,79]; indeed, this is true for an open dense set of such systems ([14], see also [79, Theorem 19.1]). On the other hand, stably ergodic systems need not be partially hyperbolic [89]. Close on the heels of the introduction of partial hyperbolicity came another fundamental extension of the theory of uniformly hyperbolic dynamical systems in a different direction. Relaxing the assumption on uniform rates by hypotheses on the Lyapunov exponents leads to the study of the much broader class of nonuniformly hyperbolic dynamical systems, which has flourished since the 1970s and is presented in [1,13]. While this is an extension in a different direction, there are significant points of intersection with the theory of partial hyperbolicity, some of which we mention in due course (see Definition 6.3), and whose importance is likely to grow.

Indeed, it is natural to proceed further to the study of systems in which both uniformity and completeness of hyperbolicity are dropped, and this theory of nonuniformly partially hyperbolic dynamical systems is described in [1].

1.2. *Outline*

This chapter consists of three major portions, each of which is summarized below. The first of these introduces the basic definitions and those parts of the theory that are most directly analogous to corresponding ones in the theory of uniformly hyperbolic systems. The second part examines the central and intermediate distributions and foliations, which has a quite different character. The third part explores accessibility, ergodicity and stable ergodicity.

1.2.1. *Basic notions and results (Sections 2 and 3)* We first (in Section 2) present various definitions of partial hyperbolicity as well as basic examples. Conceptually the most

"compact" way of thinking about uniform partial versus complete hyperbolicity is in terms of the Mather spectrum (Section 2.3), where partial hyperbolicity amounts to having other possibilities for the radii and number of rings.

We then proceed to a discussion of the invariant structures associated with the various spectral rings. These come in two fundamentally distinct classes. Section 3 is an unsurprising generalization of the stable manifold theory for uniformly hyperbolic dynamical systems to partially hyperbolic ones. It produces, in the presence of different rates of contraction or expansion, a hierarchy of fast stable or fast unstable manifolds that corresponds to collections of "inner" or "outer" rings of the Mather spectrum, respectively. We briefly discuss their regularity, including absolute continuity, which is important for the ergodic theory of partially hyperbolic systems. Neither the phenomena nor the methods here are particularly unexpected given any familiarity with the classical stable manifold theory.

1.2.2. *Central and intermediate foliations (Sections 4–6)* The study of the central distribution turns out to be quite a different matter. The Hirsch–Pugh–Shub theory of normal hyperbolicity helps control both the (moderate) regularity of its leaves and provide some robustness under perturbation—once the central foliation is known to exist. Existence is a rather delicate matter and is known only under several rather stringent assumptions, while nonexistence is an open property. We present some weak forms of integrability that are more easily obtained. Here the integral manifolds for different points may intersect without coinciding, i.e., one does not obtain a foliation in the proper sense.

Considering this as the study of the invariant structures associated with the central ring of the Mather spectrum, it is natural to do the same with other rings in the Mather spectrum as well, and the associated intermediate distributions and foliations turn out to be even more delicate.

Sections 4 and 5 study primarily topological aspects of these distributions and foliations, and in Section 6 we turn to measurable aspects. We discuss, using examples and general constructions, the possible failure of the central foliation to be absolutely continuous. On one hand we present results to the effect that even when the central distribution is integrable to a foliation with smooth leaves, absolute continuity may indeed fail in the worst possible way: There is a set of full measure that intersects almost every leaf in a bounded number of points only. On the other hand, there is evidence to support the widely held surmise that singularity of the central foliation is not only possible, but indeed typical. It would not be an overstatement to say that for a partially hyperbolic system to be stably ergodic its central foliation has to fail to be absolutely continuous in most cases. See Section 6.2 for details.

1.2.3. *Accessibility and ergodicity* As we will explain more carefully, the previously mentioned Hopf argument to establish ergodicity relies on the local product structure; in a uniformly hyperbolic dynamical system any two nearby points have a heteroclinic point, i.e., the local stable leaf of one point intersects the local unstable of the other. In particular, one can take a short curve in the local stable leaf of the first point to the intersection point and concatenate it with a short arc in the unstable manifold of the second point to join the points by what one then calls a us-path (Definition 7.1). In a partially hyperbolic system this certainly fails when the two foliations are jointly integrable, such as in the

case of (Anosov $\times$ identity). A priori there could be a whole spectrum of intermediate possibilities:

- the foliations are jointly integrable in some places but not in others,
- the foliations are both subordinate to a common foliation with leaves of dimension larger than the sum of stable and unstable dimensions,
- nearby points might only be connectible by us-paths with long arcs,
- it might take multiple concatenations of us-paths.

Whether any of these possibilities are indeed realizable remains to be seen. But it is conjectured that generically only one possibility occurs: Any two points are accessible, i.e., can be connected by a us-path with finitely many legs.

Accordingly, 2 sections of this chapter are devoted to the notion of *accessibility*, which has become a central idea in the study of partially hyperbolic systems. Section 7 presents this concept, and this enables us to present next the Pugh–Shub ergodicity theory of partially hyperbolic dynamical systems in Section 8.

Finally, we discuss Sinai–Ruelle–Bowen measures (or "physical measures") in the last section.

1.3. *Other sources*

The definitive account of the theory of partially hyperbolic dynamical systems at this point is the book [71] by Pesin, and much of this chapter follow parts of that book closely. Other important recent surveys (with a slightly different emphasis) include [34,79]. Much of the technical foundation for the subtler results about the invariant foliations in partially hyperbolic dynamical systems is provided by [56] (see also [55]). This chapter concentrates more on stable ergodicity than robust transitivity, and [16] covers the latter in more detail.

As we mentioned above, the more basic theory of uniformly hyperbolic dynamical systems is surveyed in part in [2], and [62] provides a textbook exposition that provides sufficient background.

The theory of nonuniformly hyperbolic dynamical systems is a different extension of the theory of uniformly hyperbolic dynamical systems and is presented in [1].

2. Definitions and examples

2.1. *Definition of partial hyperbolicity*

Our basic definitions require 2-sided estimates of the norms of images of linear maps, and it will be convenient to have a compact notation at our disposal. Suppose V, W are normed linear spaces, $A : V \to W$ a linear map and $U \subset V$. Then we define the *norm* and *conorm* of A *restricted* to U by

$$\|A \restriction U\| := \sup\bigl\{\|Av\|/\|v\| \mid v \in U \smallsetminus \{0\}\bigr\},$$
$$\|\lfloor A \restriction U \rfloor\| := \inf\bigl\{\|Av\|/\|v\| \mid v \in U \smallsetminus \{0\}\bigr\}.$$

2.1.1. *Partial hyperbolicity in the broad sense* The first and broader definition of partial hyperbolicity is modeled on that of hyperbolicity, where the rates of exponential behavior are separated by the unit circle, by considering separation by a different circle:

DEFINITION 2.1. Consider a manifold M, an open subset U and an embedding $f : U \to M$ with an invariant set Λ. Then f is said to be *partially hyperbolic* (in the broad sense) on Λ, or Λ is said to be a partially hyperbolic invariant set of f in the broad sense [29] if Λ is closed and there exist numbers $0 < \lambda < \mu$, $c > 0$, and subspaces $E_1(x)$ and $E_2(x)$ for all $x \in \Lambda$, such that

(1) $E_1(x)$ and $E_2(x)$ form an invariant splitting of the tangent space, i.e.,

$$\begin{aligned} &T_x M = E_1(x) \oplus E_2(x), \\ &d_x f E_1(x) = E_1\big(f(x)\big), \qquad d_x f E_2(x) = E_2\big(f(x)\big); \end{aligned} \tag{2.1}$$

(2) if $n \in \mathbb{N}$ then $\|d_x f^n \restriction E_1(x)\| \leqslant c\lambda^n$ and $c^{-1}\mu^n \leqslant \|\lfloor d_x f^n \restriction E_2(x)\rfloor\|$.

If $\lambda < 1$ the subspace $E_1(x)$ is stable (in the usual sense [2]) and will be denoted by $E^s(x)$. If $\mu > 1$ the subspace $E_2(x)$ is unstable, and we use the notation $E^u(x)$.

Clearly, either $\lambda < 1$ or $\mu > 1$ (or both) and without loss of generality we assume the former.

REMARK 2.2. In [55, p. 53] this is called (absolute) "pseudo-hyperbolicity", and in [94], "(λ, μ)-splitting".

A diffeomorphism f of a smooth compact Riemannian manifold is said to be *partially hyperbolic in the broad sense* if the whole manifold is a partially hyperbolic set for f in the broad sense.

2.1.2. *Lyapunov metrics* If $0 < \lambda < \lambda' < \mu' < \mu$ define the *Lyapunov inner product* or *Lyapunov metric* $\langle \cdot, \cdot \rangle'$ by

$$\begin{aligned} \langle v, w\rangle'_x &:= \sum_{k=0}^{\infty} \big\langle df^k v, df^k w\big\rangle_{f^k(x)} \lambda'^{-2k} \quad \text{for } v, w \in E_1(x), \\ \langle v, w\rangle'_x &:= \sum_{k=0}^{\infty} \big\langle df^{-k} v, df^{-k} w\big\rangle_{f^{-k}(x)} \mu'^{2k} \quad \text{for } v, w \in E_2(x), \\ \langle v, w\rangle'_x &:= \langle v_1, w_1\rangle'_x + \langle v_2, w_2\rangle'_x \end{aligned}$$

for $v = v_1 + v_2 \in T_x M$ and $w = w_1 + w_2 \in T_x M$ with $v_1, w_1 \in E_1(x)$ and $v_2, w_2 \in E_2(x)$. The induced *Lyapunov norm* in $T_x M$ is denoted by $\|\cdot\|'_x$. Then $\measuredangle(E_1(x), E_2(x))' = \pi/2$, $\|v\|_x/\sqrt{2} \leqslant \|v\|'_x \leqslant c\|v\|_x$, and $\|df \restriction E_1(x)\|' \leqslant \lambda'$, $\|df^{-1} \restriction E_2(x)\|' \leqslant (\mu')^{-1}$.

PROPOSITION 2.3. *An embedding f is partially hyperbolic in the broad sense if and only if there are a* (*not necessarily smooth*) *Riemannian metric $\|\cdot\|$, numbers*

$$0 < \lambda_1 \leqslant \mu_1 < \lambda_2 \leqslant \mu_2 \quad \text{with } \mu_1 < 1, \tag{2.2}$$

and an invariant splitting

$$T_x M = E_1(x) \oplus E_2(x), \qquad df\, E_i(x) = E_i\big(f(x)\big) \quad \text{for } i = 1, 2 \tag{2.3}$$

of the tangent bundle such that $E_1(x) \perp E_2(x)$ for every $x \in \Lambda$ and

$$\lambda_1 \leqslant \|\lfloor df \restriction E_1(x) \rfloor\| \leqslant \|df \restriction E_1(x)\| \leqslant \mu_1,$$
$$\lambda_2 \leqslant \|\lfloor df \restriction E_2(x) \rfloor\| \leqslant \|df \restriction E_2(x)\| \leqslant \mu_2.$$

2.1.3. *Invariant distributions* A few basic observations are quite easy to make:

PROPOSITION 2.4. *Consider a manifold M, an open set $U \subset M$ and an embedding $f : M \to U$ with a compact partially hyperbolic invariant set Λ. Then, using the notations of Definition* 2.1,

(1) $E_1(x) = \{v \in T_x M \mid \exists a > 0,\ \gamma \in [\lambda, \mu)\ \forall n \in \mathbb{N},\ \|d_x f^n v\| \leqslant a\gamma^n \|v\|\}$.
(2) $E_2(x) = \{v \in T_x M \mid \exists b > 0,\ \kappa \in (\lambda, \mu]\ \forall n \in \mathbb{N},\ \|d_x f^n v\| \geqslant b\kappa^n \|v\|\}$.
(3) *$E_1(x)$ and $E_2(x)$ are continuous, so*
(4) *there exists $k > 0$ such that $\measuredangle(E_1(x), E_2(x)) \geqslant k$ for all $x \in \Lambda$.*
(5) *There exists $\varepsilon > 0$ such that if $\widetilde{E} \subset TM$ is an invariant distribution for which $\dim \widetilde{E}_1(x) = \dim E_1(x)$ and $\measuredangle(\widetilde{E}_1(x), E_1(x)) \leqslant \varepsilon$ for every $x \in \Lambda$ then $\widetilde{E}_1(x) = E_1(x)$, and likewise for $E_2(x)$.*

On the other hand, going beyond continuity is a rather more substantial achievement. This goes back to Anosov in the hyperbolic case and to Brin and Pesin [29] in the present context. The most general version is in [24]:

THEOREM 2.5. *$E_1(x)$ and $E_2(x)$ are Hölder continuous, i.e., there exist $C, \alpha > 0$ such that $\measuredangle(E_i(x), E_i(y)) \leqslant C\rho(x, y)^\alpha$ for all $x, y \in \Lambda$ and $i = 1, 2$. Indeed, the Hölder exponent can be controlled through the hyperbolicity estimates*: *In the context of Proposition* 2.3 *any*

$$\alpha < \frac{\log \lambda_2 - \log \mu_1}{\log \mu_2} \tag{2.4}$$

admits a $C > 0$ for which $\measuredangle(E_1(x), E_1(y)) \leqslant C\rho(x, y)^\alpha$, and there is an analogous estimate for the Hölder exponent of E_2.

One should not expect the distribution E_1 to be smooth even in the case of Anosov diffeomorphisms. The first example of a nonsmooth stable distribution was constructed by Anosov in [9]. Hasselblatt [52] has shown that for a "typical" Anosov diffeomorphism

the stable and unstable distributions are only Hölder continuous with Hölder exponent no larger than that in (2.4) (see also [53]). Moreover, high regularity has in several classes of hyperbolic systems been shown to occur only for algebraic systems [2]. Nevertheless, there are situations where these distributions are C^1 (see, e.g., [2,52,56]):

(1) under the *pinching condition* $\frac{\mu_1}{\lambda_2}\mu_2 < 1$,[1]
(2) if the distribution E_1 is of *codimension one*.

REMARK 2.6. Our definition of partial hyperbolicity (in the broad sense) corresponds to what is also known as *absolute partial hyperbolicity* (absolute pseudo-hyperbolicity in [55, p. 53]) as opposed to a weaker *relative* (*or pointwise*) *partial hyperbolicity* (relative pseudo-hyperbolicity in [55, p. 62f]). While for the former we have

$$\sup_{x\in M} \|df \restriction E_1(x)\| \Big(\inf_{x\in M} \big\| \big(df \restriction E_2(x)\big)^{-1} \big\| \Big)^{-1} < 1,$$

the latter is defined such that

$$\sup_{x\in M} \|df \restriction E_1(x)\| \big(\big\| \big(df \restriction E_2(x)\big)^{-1} \big\| \big)^{-1} < 1.$$

See [56] where other refined versions of absolute and relative hyperbolicity are introduced. It should be stressed that one can develop essentially the whole stability theory of partially hyperbolic systems assuming only relative partial hyperbolicity (or, rather, the presence of a *dominated splitting*, see p. 44 and [4]). However, the study of ergodic and topological properties of partially hyperbolic systems needs the stronger assumption of absolute partial hyperbolicity.[2]

2.1.4. *Partial hyperbolicity* The study of partially hyperbolic systems with a view to ergodicity has concentrated on those with a triple splitting that includes a central direction of weakest contraction and expansion:

DEFINITION 2.7. An embedding f is said to be *partially hyperbolic* on Λ if there exist numbers $C > 0$,

$$0 < \lambda_1 \leqslant \mu_1 < \lambda_2 \leqslant \mu_2 < \lambda_3 \leqslant \mu_3 \quad \text{with } \mu_1 < 1 < \lambda_3 \tag{2.5}$$

and an invariant splitting into stable, central and unstable directions

$$\begin{aligned} &T_x M = E^s(x) \oplus E^c(x) \oplus E^u(x), \\ &d_x f E^\tau(x) = E^\tau\big(f(x)\big), \quad \tau = s, c, u, \end{aligned} \tag{2.6}$$

[1] By (2.4), E_1 is Lipschitz in this case.
[2] We would like to thank M. Viana for pointing this out to us.

such that if $n \in \mathbb{N}$ then

$$C^{-1}\lambda_1^n \leqslant \|\lfloor d_x f^n \upharpoonright E^s(x) \rfloor\| \leqslant \|d_x f^n \upharpoonright E^s(x)\| \leqslant C\mu_1^n,$$
$$C^{-1}\lambda_2^n \leqslant \|\lfloor d_x f^n \upharpoonright E^c(x) \rfloor\| \leqslant \|d_x f^n \upharpoonright E^c(x)\| \leqslant C\mu_2^n,$$
$$C^{-1}\lambda_3^n \leqslant \|\lfloor d_x f^n \upharpoonright E^u(x) \rfloor\| \leqslant \|d_x f^n \upharpoonright E^u(x)\| \leqslant C\mu_3^n.$$

In this case we set $E^{cs} := E^c \oplus E^s$ and $E^{cu} := E^c \oplus E^u$.

REMARK 2.8. By Theorem 2.5 each subbundle E^τ for $\tau = u, s, c, cu, cs$ is Hölder continuous.

There is a Lyapunov metric that is fully adapted to this situation:

PROPOSITION 2.9. *An embedding is partially hyperbolic if and only if there exists a Riemannian metric for which there are numbers λ_i, μ_i, $i = 1, 2, 3$, as in* (2.5) *and an invariant splitting* (2.6) *into pairwise orthogonal subspaces $E^s(x)$, $E^c(x)$ and $E^u(x)$ such that*

$$\lambda_1 \leqslant \|\lfloor d_x f \upharpoonright E^s(x) \rfloor\| \leqslant \|d_x f \upharpoonright E^s(x)\| \leqslant \mu_1,$$
$$\lambda_2 \leqslant \|\lfloor d_x f \upharpoonright E^c(x) \rfloor\| \leqslant \|d_x f \upharpoonright E^c(x)\| \leqslant \mu_2,$$
$$\lambda_3 \leqslant \|\lfloor d_x f \upharpoonright E^u(x) \rfloor\| \leqslant \|d_x f \upharpoonright E^u(x)\| \leqslant \mu_3. \tag{2.7}$$

2.1.5. *The cone criterion* Verifying partial hyperbolicity appears to require finding the invariant distributions first, so it is useful to have a more obviously robust criterion that is easier to verify. For hyperbolic dynamical systems this goes back principally to Alekseev [7,2].

Given a point $x \in M$, a subspace $E \subset T_x M$ and a number $\alpha > 0$, define the *cone* at x centered around E of angle α by

$$C(x, E, \alpha) = \big\{ v \in T_x M \mid \measuredangle(v, E) < \alpha \big\}.$$

PROPOSITION 2.10. *An embedding f is partially hyperbolic in the broad sense if and only if there are $\alpha > 0$ and two continuous cone families $C_1(x, \alpha) = C(x, E_1(x), \alpha)$ and $C_2(x, \alpha) = C(x, E_2(x), \alpha)$ for which*

$$d_x f^{-1}\big(C_1(x, \alpha)\big) \subset C_1\big(f^{-1}(x), \alpha\big), \qquad d_x f\big(C_2(x, \alpha)\big) \subset C_2\big(f(x), \alpha\big) \tag{2.8}$$

as well as

$$\|d_x f \upharpoonright C_1(x, \alpha)\| \leqslant \mu_1 < \lambda_2 \leqslant \|\lfloor d_x f \upharpoonright C_2(x, \alpha) \rfloor\|. \tag{2.9}$$

The evident advantage of this definition is that one can verify it having only approximations of E and F in Definition 2.1, and these approximations need not be invariant in order for suitable cones around them to satisfy (2.8) and (2.9).

PROPOSITION 2.11. *An embedding f is partially hyperbolic if and only if there are families of stable and unstable cones*

$$C^s(x,\alpha) = C\big(x, E^s(x), \alpha\big), \qquad C^u(x,\alpha) = C\big(x, E^u(x), \alpha\big)$$

and of center-stable cones or center-unstable cones

$$C^{cs}(x,\alpha) = C\big(x, E^{cs}(x), \alpha\big), \qquad C^{cu}(x,\alpha) = C\big(x, E^{cu}(x), \alpha\big),$$

where

$$E^{cs}(x) = E^c(x) \oplus E^s(x), \qquad E^{cu}(x) = E^c(x) \oplus E^u(x),$$

such that

$$\begin{aligned} &d_x f^{-1}\big(C^s(x,\alpha)\big) \subset C^s\big(f^{-1}(x),\alpha\big), \qquad d_x f\big(C^u(x,\alpha)\big) \subset C^u\big(f(x),\alpha\big),\\ &d_x f^{-1}\big(C^{cs}(x,\alpha)\big) \subset C^{cs}\big(f^{-1}(x),\alpha\big), \qquad d_x f\big(C^{cu}(x,\alpha)\big) \subset C^{cu}\big(f(x),\alpha\big) \end{aligned} \tag{2.10}$$

and there are $0 < \mu_1 < \lambda_2 \leqslant \mu_2 < \lambda_3$ with $\mu_1 < 1 < \lambda_3$ such that

$$\begin{aligned} &\|d_x f \restriction C^s(x,\alpha)\| \leqslant \mu_1, \qquad \lambda_3 \leqslant \|\lfloor d_x f \restriction C^u(x,\alpha)\rfloor\|,\\ &\|d_x f \restriction C^{cs}(x,\alpha)\| \leqslant \mu_2, \qquad \lambda_2 \leqslant \|\lfloor d_x f \restriction C^{cu}(x,\alpha)\rfloor\|. \end{aligned} \tag{2.11}$$

2.2. *Examples of partially hyperbolic systems*

2.2.1. *The time-t map of a hyperbolic flow* Let φ_t be a flow on a compact smooth Riemannian manifold M with a hyperbolic invariant set Λ. Given $t \in \mathbb{R}$, the map φ_t is partially hyperbolic on Λ with 1-dimensional central direction generated by the vector field.

2.2.2. *Frame flows* Let V be a closed oriented n-dimensional manifold of negative sectional curvature and $M = SV$ the unit tangent bundle of V. Let also N be the space of positively oriented orthonormal n-frames in TV. This produces a fiber bundle $\pi : N \to M$ where the natural projection π takes a frame into its first vector. The associated structure group $SO(n-1)$ acts on fibers by rotating the frames, keeping the first vector fixed. Therefore, we can identify each fiber N_x with $SO(n-1)$ where g_t is the geodesic flow. The *frame flow* Φ_t acts on frames by moving their first vectors according to the geodesic flow and moving the other vectors by parallel translation along the geodesic defined by the first vector. For each t, we have that $\pi \circ \Phi_t = g_t \circ \pi$. The frame flow Φ_t preserves the measure that is locally the product of the Liouville measure with normalized Haar measure on $SO(n-1)$. The time-t map of the frame flow is a partially hyperbolic diffeomorphism (for $t \neq 0$). The center bundle has dimension $1 + \dim SO(n-1)$ and is spanned by the flow direction and the fiber direction.

2.2.3. *Direct products* Let $f : U \to M$ be an embedding with a compact hyperbolic set $\Lambda \subset M$ and $E^s_f(x)$, $E^u_f(x)$ the stable and unstable subspaces at $x \in \Lambda$. Also, let $g : U' \to N$ be an embedding with a compact invariant set K such that

$$\max_{x\in\Lambda} \|df \restriction E^s_f(x)\| < \min_{y\in K} \|\lfloor dg(y) \rfloor\| \leqslant \max_{y\in K} \|dg(y)\| < \min_{x\in\Lambda} \|\lfloor df \restriction E^u_f(x) \rfloor\|.$$

Then $F : M \times N \to M \times N$, $F(x, y) = (f(x), g(y))$ is partially hyperbolic on $\Lambda \times K$.

Particular cases are g being the identity map of N or a rotation of $N = S^1$.

2.2.4. *Skew products* Let $f : U \to M$ be an embedding with a compact hyperbolic set $\Lambda \subset M$ and $E^s_f(x)$, $E^u_f(x)$ the stable and unstable subspaces at $x \in \Lambda$. Also, let $g_x : U_x \to N$ be a family of embeddings of $U_x \subset N$ that depend smoothly on $x \in \Lambda$ and have a common compact invariant set K such that

$$\begin{aligned}\max_{x\in\Lambda} \|df \restriction E^s_f(x)\| &< \min_{x\in\Lambda} \min_{y\in K} \|\lfloor dg_x(y) \rfloor\| \leqslant \max_{x\in\Lambda} \max_{y\in K} \|dg_x(y)\| \\ &< \min_{x\in\Lambda} \|\lfloor df \restriction E^u_f(x) \rfloor\|. \end{aligned} \tag{2.12}$$

The map $F : \Lambda \times \bigcap_x U_x \to M \times N$ given by $F(x, y) = (f(x), g_x(y))$ is partially hyperbolic on $\Lambda \times K$.

A particular case is obtained by taking $\Lambda = M$, $K = N = S^1$, $\alpha : M \to M$ smooth and $g_x = R_{\alpha(x)}$ (rotation by $\alpha(x)$). The map

$$\begin{aligned} &F = F_\alpha : M \times S^1 \to M \times S^1, \\ &F(x, y) = \big(f(x), R_{\alpha(x)}(y)\big), \quad x \in M,\ y \in S^1, \end{aligned}$$

is partially hyperbolic with 1-dimensional central direction.

2.2.5. *Group extensions* An "algebraic" version of the previous example is a group extension over an Anosov diffeomorphism. Let G be a compact Lie group and $\varphi : M \to G$ a smooth function on M with values in G. Define the map $F = F_\varphi : M \times G \to M \times G$ by

$$F(x, y) = \big(f(x), \varphi(x)g\big), \quad x \in M,\ g \in G.$$

The map F is partially hyperbolic since left translations are isometries of G in the bi-invariant metric. If f preserves a smooth probability measure ν then F preserves the smooth probability measure $\nu \times \nu_G$ where ν_G is the (normalized) Haar measure on G.

2.2.6. *Partially hyperbolic systems on* 3*-dimensional manifolds* It is an open problem to describe compact smooth Riemannian manifolds that admit partially hyperbolic diffeomorphisms. To admit the splitting into stable, unstable and center distribution the dimension of the manifold must be at least three, so it is natural to inquire first, which 3-manifolds support partially hyperbolic diffeomorphisms. The torus $\mathbb{T}^3$ does, because an automorphism

given by an integer matrix with eigenvalues λ, 1, λ^{-1}, where $|\lambda| \neq 1$, is partially hyperbolic, as is any sufficiently small perturbation (Corollary 2.18). Recently, Brin, Burago and Ivanov have begun a study of partially hyperbolic dynamical systems on 3-manifolds, and interesting results have already been obtained.

THEOREM 2.12 (Brin, Burago and Ivanov [26]). *A compact* 3*-dimensional manifold whose fundamental group is finite does not carry a partially hyperbolic diffeomorphism.*

This implies that there are no partially hyperbolic diffeomorphisms on the 3-dimensional sphere $\mathbb{S}^3$.

We should mention a result that can be viewed as a precursor to Theorem 2.12. L. Díaz, E. Pujals and R. Ures showed that a robustly transitive diffeomorphism of a 3-manifold M (i.e., a diffeomorphism all of whose C^1 perturbations are topologically transitive) is generically partially hyperbolic in the broad sense, and if the center-unstable bundle is integrable then the fundamental group of M is infinite [40]. This was extended to arbitrary dimension by Bonatti, Díaz and Pujals [15]: Generically the homoclinic class of any periodic saddle is either contained in the closure of an infinite set of sinks or sources (Newhouse phenomenon), or admits a dominated splitting; in particular, robust transitivity implies dominated splitting (see also p. 44 and [4, Section 5]).

One may ask a question complementary to the previous one: Of what type can partially hyperbolic diffeomorphisms of 3-manifolds be? The known robustly transitive or stably ergodic ones are

- perturbations of skew-products over an Anosov diffeomorphism on $\mathbb{T}^2$,
- perturbations of the time-1 map of a transitive Anosov flow,
- some derived-from-Anosov diffeomorphisms on $\mathbb{T}^3$.

Pujals has speculated (see [20]) that this is indeed a complete list, and recent work by Bonatti and Wilkinson [20] makes it plausible that such a classification of transitive partially hyperbolic diffeomorphisms of 3-manifolds might hold: They show that the homoclinic geometry of a single periodic orbit can determine much of the global orbit structure. (Note that volume-preserving such diffeomorphisms are generically transitive by Theorems 7.9 and 7.12.)

Specifically, in the case of a skew product a periodic orbit arises from a periodic point for the base diffeomorphism and hence comes with nearby homoclinic periodic orbits that are also embedded circles. Their first result turns this observation around:

THEOREM 2.13 (Bonatti–Wilkinson [20]). *Let f be a transitive partially hyperbolic diffeomorphism of a compact* 3*-manifold M with an embedded invariant circle γ such that there is some* (*sufficiently large*) δ *for which $W^s_\delta(\gamma) \cap W^u_\delta(\gamma) \smallsetminus \gamma$ has a connected component that is a circle. Then, possibly after passing to on orientable cover, M is a circle bundle over $\mathbb{T}^2$ and f is conjugate to a topological skew-product over a linear Anosov map A of $\mathbb{T}^2$, i.e., to a map of M that preserves the fibration and projects to A.*

In the case of the time-1 map of a transitive Anosov flow the homoclinic curves to an invariant circle are noncompact. The corresponding result is a little less complete than the previous one.

THEOREM 2.14 (Bonatti–Wilkinson [20]). *Let f be a dynamically coherent* (*Definition* 4.4) *partially hyperbolic diffeomorphism of a compact* 3*-manifold M with a closed periodic center leaf γ such that each center leaf in $W^s_{\mathrm{loc}}(\gamma)$ is periodic for f. Then the center foliation supports a continuous flow conjugate to a transitive expansive flow.*

It is conjectured and seems likely to be true that the expansive flows that arise here are in turn topologically conjugate to Anosov flows.

2.3. *The Mather spectrum*

An embedding f with a compact invariant set Λ generates a continuous linear operator f_* on the Banach space $\Gamma^0(T_\Lambda M)$ of continuous vector fields $\mathbf{v}$ on Λ by the formula

$$(f_*\mathbf{v})(x) = df\mathbf{v}\big(f^{-1}(x)\big).$$

The spectrum $Q = Q_f$ of the complexification of f_* is called the *Mather spectrum* of the dynamical system f on Λ, and it provides alternative ways of expressing our various hyperbolicity conditions as well as more detailed information about separation of expansion and contraction rates:

THEOREM 2.15 (Mather [64,71]). *If nonperiodic orbits of f are dense in Λ then*

(1) *any connected component of the spectrum Q is a ring* (*or annulus*) $Q_i = \{z \in \mathbb{C} \mid \lambda_i \leqslant |z| \leqslant \mu_i\}$ *around* 0 *with radii λ_i and μ_i, where $0 < \lambda_1 \leqslant \mu_1 < \cdots < \lambda_t \leqslant \mu_t$ and $t \leqslant \dim M$;*
(2) *the invariant subspace $H_i \in \Gamma^0(TM)$ of f_* corresponding to the component Q_i of the spectrum is a module over the ring of continuous functions;*
(3) *the collection of the subspaces $E_i(x) = \{\mathbf{v}(x) \in T_x \mid \mathbf{v} \in H_i\}$ constitutes a df-invariant continuous distribution on M and*

$$T_xM = \bigoplus_{i=1}^{t} E_i(x) \quad \text{for all } x \in M.$$

Since density of nonperiodic orbits is an easy consequence of hyperbolicity assumptions, one can characterize various classes of dynamical systems using their Mather spectra.

THEOREM 2.16 (Mather [64,71]).

(1) *A diffeomorphism f is Anosov if and only if* 1 *is not contained in its Mather spectrum Q.*
(2) *A diffeomorphism f is partially hyperbolic on Λ in the broad sense if and only if its Mather spectrum* (*over Λ*) *is contained in a disjoint union of two nonempty rings, $Q \subset Q_1 \cup Q_2$ with Q_1 lying inside of the unit disk or Q_2 lying outside of the unit disk.*
(3) *A diffeomorphism f is partially hyperbolic on Λ if and only if its Mather spectrum* (*over Λ*) *is contained in a disjoint union of three nonempty rings, $Q \subset Q_1 \cup Q_2 \cup Q_3$ with Q_1 lying inside of the unit disk and Q_3 lying outside of the unit disk.*

While [64,71] state these results only for $\Lambda = M$, the proofs readily extend to invariant subsets.

It is natural to expect that the Mather spectrum is stable under small perturbations of dynamical systems, and the most straightforward approach to establishing this would be to show that the action induced on vector fields by a perturbation is close to the original such action. Unfortunately this is always false. The expected result about the Mather spectrum nevertheless turns out to be true:

THEOREM 2.17 (Pesin [70,71]). *Let M be a compact manifold, $f : M \to M$ a diffeomorphism whose nonperiodic orbits are dense. Let*

$$Q_f = \bigcup_{i=1}^{t} Q_{f,i}$$

be the decomposition of its Mather spectrum into nonempty disjoint rings $Q_{f,i}$ with radii

$$0 < \lambda_{f,1} \leqslant \mu_{f,1} < \cdots < \lambda_{f,t} \leqslant \mu_{f,t}.$$

Let also

$$TM = \bigoplus_{i=1}^{t} E_{f,i}$$

be the corresponding decomposition of the tangent bundle into df-invariant subbundles $E_{f,i}$, $i = 1, \ldots, t$. Then for any sufficiently small $\varepsilon > 0$ there exists a neighborhood η of f in $\mathrm{Diff}^1(M)$ *such that for any $g \in \eta$:*

(1) *the Mather spectrum Q_g is a union of disjoint components $Q_{g,i}$, each being contained in a ring with radii $\lambda_{g,i} \leqslant \mu_{g,i}$ satisfying*

$$|\lambda_{f,i} - \lambda_{g,i}| \leqslant \varepsilon, \qquad |\mu_{f,i} - \mu_{g,i}| \leqslant \varepsilon;$$

(2) *the distribution $E_{g,i}$ corresponding to the component $Q_{g,i}$ satisfies*

$$\max_{x \in M} \measuredangle\big(E_{f,i}(x), E_{g,i}(x)\big) \leqslant L\delta^{\alpha} \leqslant \varepsilon,$$

where $\delta = d_{C^1}(f, g)$ and $L > 0$, $\alpha > 0$ are constants.

As usual, $\mathrm{Diff}^q(M)$ is the space of C^q diffeomorphisms with the C^q topology.

COROLLARY 2.18. *Anosov systems, partially hyperbolic systems, and partially hyperbolic diffeomorphisms form open subsets in* $\mathrm{Diff}^q(M)$, $q \geqslant 1$.

REMARK 2.19. While a component $Q_{f,i}$ of the spectrum of f may be a ring, the corresponding component $Q_{g,i}$ of the spectrum of g may consist of several rings. To illustrate

this situation consider an Anosov flow φ_t on a smooth manifold M and observe that the Mather spectrum of $\varphi_0 = \mathrm{Id}$ is the unit circle while the Mather spectrum of φ_t for $t \neq 0$ (which is partially hyperbolic) contains at least two more additional rings.

REMARK 2.20. There are 3 general situations in which partial hyperbolicity is known to be stable. First, we just saw that this is the case when the entire manifold is a partially hyperbolic set (Corollary 2.18). Second, Theorem 2.17 extends to partially hyperbolic attractors because attractors are stable under perturbation, so partially hyperbolic attractors are also stably partially hyperbolic. Finally, when the partially hyperbolic set is a normally hyperbolic manifold then Theorem 4.3 below together with Theorem 2.17 gives persistence of partial hyperbolicity.

There are also some particular cases when partially hyperbolic sets survive under small perturbations, such as when a partially hyperbolic set Λ is the direct product of a locally maximal hyperbolic set and a compact manifold. Indeed, Λ is foliated by leaves of its center foliation and can be viewed as a normally hyperbolic lamination in the sense of [56]. Its stability follows from Theorem 4.11. Partially hyperbolic sets of this type appear in bifurcation theory (see [48,59]).

3. Filtrations of stable and unstable foliations

3.1. *Existence and subfoliation*

For hyperbolic dynamical systems the classical Stable-Manifold Theorem [2] establishes that the stable and unstable distributions are each tangent to a unique foliation. A moderate adaptation of the Stable-Manifold Theorem yields analogous but more finely stratified information when the Mather spectrum consists of a larger number of rings (see [71] and the references therein). We should mention that the word foliation is used here in a looser sense than in differential geometry. Even in the case of Anosov diffeomorphisms these foliations are partitions into smooth manifolds that may only admit (Hölder) continuous foliation charts; for hyperbolic sets the foliation locally only fills a Cantor set times a disk (see [2,80]).

DEFINITION 3.1. A partition W of M is called a *foliation of M with smooth leaves* or simply *foliation* if there exist $\delta > 0$ and $\ell > 0$ such that for each $x \in M$,

1. the element $W(x)$ of the partition W containing x is a smooth ℓ-dimensional injectively immersed submanifold; it is called the (*global*) *leaf* of the foliation at x; the connected component of the intersection $W(x) \cap B(x, \delta)$ that contains x is called the *local leaf* at x and is denoted by $V(x)$;
2. there exists a continuous map $\varphi_x : B(x, \delta) \to C^1(D, M)$ (where $D \subset \mathbb{R}^\ell$ is the unit ball) such that for every $y \in M \cap B(x, \delta)$ the manifold $V(y)$ is the image of the map $\varphi_x(y) : D \to M$.

The function $\varphi_x(y, z) = \varphi_x(y)(z)$ is called the *foliation coordinate chart*. This function is continuous and has continuous derivative $\frac{\partial}{\partial z}\varphi_x$.

A continuous k-dimensional distribution E on M is said to be

(1) *weakly integrable* if for each point $x \in M$ there is an immersed complete C^1 manifold $W(x)$ which contains x and is everywhere tangent to E, i.e., $T_y W(y) = E(y)$ for each $y \in W(x)$ [26]. We call $W(x)$ an integral manifold of E through x (note that a priori the integral manifolds $W(x)$ may be self-intersecting and may not form a partition of M);
(2) *integrable* if there is a foliation whose tangent bundle is E;
(3) *uniquely integrable* if there is a foliation W with k-dimensional leaves such that any C^1 curve $\sigma : \mathbb{R} \to M$ satisfying $\dot\sigma(t) \in E(\sigma(t))$ for all t, is contained in $W(\sigma(0))$ (in particular, $T_x W(x) = E(x)$ for all $x \in M$);
(4) *locally uniquely integrable* if for each $x \in M$ there is a k-dimensional smooth submanifold $W_{\mathrm{loc}}(x)$ and $\alpha(x) > 0$ such that a piecewise C^1 curve $\sigma : [0,1] \to M$ is contained in $W_{\mathrm{loc}}(x)$ so long as $\sigma(0) = x$, $\dot\sigma(t) \in E(\sigma(t))$ for $t \in [0,1]$ and $\operatorname{length}\sigma < \alpha(x)$. (In this case E is integrable and the integral foliation is unique.)

THEOREM 3.2 (Hirsch, Pugh and Shub [55], Brin and Pesin [29], [71]). *Suppose f is an embedding with a compact invariant set Λ on which the tangent space admits a df-invariant splitting*

$$T_\Lambda M = \bigoplus_{i=1}^{t} E_i \tag{3.1}$$

with

$$\lambda_i < \|\lfloor df \restriction E_i(x) \rfloor\| \leqslant \|df \restriction E_i(x)\| < \mu_i \tag{3.2}$$

for all $x \in \Lambda$, where

$$0 < \lambda_1 \leqslant \mu_1 < \cdots < \lambda_t \leqslant \mu_t. \tag{3.3}$$

(1) *If $\mu_k < 1$ then the distribution*

$$F_k^s = \bigoplus_{i=1}^{k} E_i$$

is uniquely integrable and the maximal integral manifolds of this distribution generate a foliation W_k^s of M. The global leaf $W_k^s(x)$ through $x \in M$ is a C^1-immersed submanifold of M.
(2) *If $\lambda_k > 1$ then an analogous statement holds for the distribution*

$$F_k^u = \bigoplus_{i=k}^{t} E_i;$$

the corresponding foliation is W_k^u and its leaves are $W_k^u(x)$, $x \in M$.

(3) *The foliation W_k^s is f-invariant and contracting, i.e., for any $x \in M$, $y \in W_k^s(x)$ and $n \geqslant 0$,*

$$\rho_k^s\big(f^n(x), f^n(y)\big) \leqslant C(\lambda_k + \varepsilon)^n \rho_k^s(x, y),$$

where ε is such that $0 < \varepsilon < \min\{\lambda_{k+1} - \mu_k, 1 - \mu_k\}$, $C = C(\varepsilon) > 0$ is a constant independent of x, y and n, and ρ_k^s is the distance in $W_k^s(x)$ induced by the Riemannian metric.

(4) *The foliation W_k^u is f-invariant and contracting under f^{-1}.*

(5) *If f is C^q then $W_k^s(x)$ and $W_k^u(x)$ are C^q.*

Thus, for an embedding f with a compact invariant set Λ satisfying (3.1) the two filtrations of distributions

$$F_1^s \subset F_2^s \subset \cdots \subset F_\ell^s, \qquad F_m^u \supset F_{m+1}^u \supset \cdots \supset F_t^u$$

integrate to filtrations of foliations

$$W_1^s \subset W_2^s \subset \cdots \subset W_\ell^s, \qquad W_m^u \supset W_{m+1}^u \supset \cdots \supset W_t^u,$$

where ℓ is maximal and m is minimal such that $\mu_\ell < 1$ and $\lambda_m > 1$ (note that $m = \ell + 1$ or $\ell + 2$). W_k^s is called the *k-stable foliation* and W_k^u the *k-unstable foliation* for f. If f is partially hyperbolic the foliations $W^s = W_\ell^s$ and $W^u = W_m^u$ are called the *stable* and *unstable* foliations.

THEOREM 3.3 (Hirsch, Pugh and Shub [56, Theorem 6.1]). *Under the assumptions of Theorem* 3.2 *and with $1 \leqslant k < \ell$ (respectively, $m \leqslant k < t$), the foliation W_k^s* subfoliates *the foliation W_{k+1}^s (respectively, W_{k+1}^u subfoliates W_k^u). For every $x \in M$ the leaves $W_k^s(y)$ depend C^{n_k} smoothly on $y \in W_{k+1}^s(x)$, where n_k is the largest integer such that $\mu_k < \lambda_{k+1}^{n_k}$. An analogous statement holds for W_k^u.*

3.2. *Absolute continuity*

The fact that the stable and unstable foliations may not admit smooth local foliation charts prevents us from applying the classical Fubini theorem to conclude that a set that intersects each local leaf in a set of full (leaf-) measure must itself be of full measure. Anosov [9] identified this as the central technical point in the ergodic theory of hyperbolic dynamical systems (in the Hopf argument, see [57,10,30] and Section 7, p. 34). In this subsection and the next we explain that while, due to the absence of smooth foliation charts, it seems possible that the foliations might be singular in the measure-theoretic sense (see Section 6), the needed property of absolute continuity still holds for the stable and unstable foliations. We will later see that the central direction is much less well behaved.

The first step is absolute continuity of the holonomy maps.

DEFINITION 3.4. Let W be a foliation of a manifold M with smooth leaves and for $x \in M$, $r > 0$ consider the family

$$\mathcal{L}(x) = \{V(w)\colon\ w \in B(x, r)\} \tag{3.4}$$

of local manifolds, where $V(w)$ is the connected component containing w of $W(w) \cap B(x, r)$ and $B(x, r)$ is the ball centered at x of radius r.

Choose two local disks D^1 and D^2 that are transverse to the family $\mathcal{L}(x)$, and define the *holonomy map* $\pi = \pi(x, W)\colon D^1 \to D^2$ (generated by the family of local manifolds) by setting

$$\pi(y) = D^2 \cap V(w) \quad \text{if } y = D^1 \cap V(w) \text{ and } w \in B(x, r).$$

The holonomy map π is a homeomorphism onto its image.

Let m denote the Riemannian volume. Given a submanifold D in M, let m_D be the Riemannian volume on D induced by the restriction of the Riemannian metric to D.

THEOREM 3.5 (Brin and Pesin [29], Pugh and Shub [75,77], [13,71]). *Let f be a partially hyperbolic C^2 diffeomorphism of a compact smooth manifold M. Given $x \in M$ and two transverse disks D^1 and D^2 to the family $\mathcal{L}(x)$ of local stable manifolds $V(y)$, $y \in B(x, r)$, the holonomy map π is absolutely continuous (with respect to the measures m_{D^1} and m_{D^2}) and the Jacobian* $\mathrm{Jac}(\pi) := dm_{D^2}/d(\pi_* m_{D^1})$ *(Radon–Nikodym derivative) is bounded from above and bounded away from zero.*

The Jacobian of the holonomy map at a point $y \in D^1$ can be computed by the following formula:

$$\mathrm{Jac}(\pi)(y) = \prod_{k=0}^{\infty} \frac{\mathrm{Jac}(d_{f^k(\pi(y))} f^{-1} | T_{f^k(\pi(y))} f^k(D^2))}{\mathrm{Jac}(d_{f^k(y)} f^{-1} | T_{f^k(y)} f^k(D^1))}.$$

In particular, the infinite product on the right-hand side converges.

The issue of absolute continuity as it affects the ergodic theory of hyperbolic and partially hyperbolic dynamical systems can be put in this form: If $E \subset B(x, q)$ is a Borel set of positive volume, can the intersection $E \cap V(y)$ have zero Lebesgue measure (with respect to the Riemannian volume on $V(y)$) for almost every $y \in E$?

Theorem 3.5 is the main step towards ruling out this pathology for the stable and unstable foliations.

THEOREM 3.6. *Let f be a partially hyperbolic C^2 diffeomorphism of a compact smooth manifold M, ν a smooth f-invariant probability measure on M. Then the conditional measures on each $V(w)$ are absolutely continuous with respect to the induced Riemannian volume, and likewise for transversals.*

To state this more precisely, let $\nu_V(w)$ be the conditional measures on $V(w)$ for $w \in B(x,r)$ and consider the measurable partition ξ of

$$Q(x) := \bigcup_{w \in B(x,r)} V(w)$$

into local manifolds, identifying the factor space $Q(x)/\xi$ with an open transverse disk D. Denote by $\hat{\nu}_D$ the factor measure generated by the partition ξ (supported on D), by $m_V(w)$ the Riemannian volume on $V(w)$, and by m_D the Riemannian volume on D. Then Theorem 3.6 is meant to say that

(1) the measures $\nu_V(w)$ and $m_V(w)$ are equivalent for ν-almost every $w \in B(x,r)$;
(2) the factor measure $\hat{\nu}_D$ is equivalent to the measure m_D.

As a consequence

$$d\nu_V(w)(y) = \kappa(w,y)\, dm_V(w)(y)$$

for every $w \in B(x,r)$ and $y \in V(w)$, where $\kappa(w,y)$ is continuous and satisfies the homological equation

$$\kappa\big(f(w), f(y)\big) = \frac{\mathrm{Jac}(df \restriction E(y))}{\mathrm{Jac}(df \restriction E(w))}\, \kappa(w,y).$$

It follows that

$$\kappa(w,y) = \prod_{i=0}^{\infty} \frac{\mathrm{Jac}(df \restriction E(f^i(y)))}{\mathrm{Jac}(df \restriction E(f^i(w)))}.$$

By the Hölder continuity of the distribution E (see Theorem 2.5), the product converges.

The absolute continuity property of local stable manifold described in Theorem 3.6 follows from absolute continuity of the holonomy map (see Theorem 3.5). However, the converse does not hold.

4. Central foliations

4.1. *Normal hyperbolicity*

The notion of normal hyperbolicity in dynamical systems was introduced by Hirsch, Pugh and Shub in [55] (see also [56]; a particular case of normal hyperbolicity was considered by R. Sacker [84] in his work on partial differential equations). The theories of normal hyperbolicity and partial hyperbolicity are closely related in their results and methods. Moreover, the former provides techniques to study integrability of the central distribution and robustness of the central foliation for partially hyperbolic systems (see Sections 4.3 and 4.4).

DEFINITION 4.1. Let $q \geqslant 1$, M a C^q compact connected Riemannian manifold (without boundary), $U \subset M$ open, $f : U \to M$ a C^q embedding, and $N = N_f$ a compact C^1 f-invariant submanifold of M, i.e., $f(N) = N$. The map f is said to be *normally hyperbolic to* N if f is partially hyperbolic on (or "along") N, i.e., there is an invariant splitting

$$\begin{aligned} &T_x M = E^s(x) \oplus T_x N \oplus E^u(x) \\ &\quad \text{with } df\, E^s(x) = E^s\big(f(x)\big) \text{ and } df\, E^u(x) = E^u\big(f(x)\big) \end{aligned} \tag{4.1}$$

for every $x \in N$, such that

$$\begin{aligned} &\lambda_1 \leqslant \|\lfloor df \restriction E^s(x) \rfloor\| \leqslant \|df \restriction E^s(x)\| \leqslant \mu_1, \\ &\lambda_2 \leqslant \|\lfloor df \restriction T_x N \rfloor\| \leqslant \|df \restriction T_x N\| \leqslant \mu_2, \\ &\lambda_3 \leqslant \|\lfloor df \restriction E^u(x) \rfloor\| \leqslant \|df \restriction E^u(x)\| \leqslant \mu_3, \end{aligned} \tag{4.2}$$

where $0 < \lambda_1 \leqslant \mu_1 < \lambda_2 \leqslant \mu_2 < \lambda_3 \leqslant \mu_3$ and $\mu_1 < 1 < \lambda_3$.

Similarly to Theorem 2.5, the splitting (4.1) is Hölder continuous.

By the Local-Stable-Manifold Theorem one can construct, for every $x \in N$, local stable and unstable manifolds, $V^s(x)$ and $V^u(x)$, respectively, at x, such that

(1) $x \in V^s(x)$, $x \in V^u(x)$;
(2) $T_x V^s(x) = E^s(x)$, $T_x V^u(x) = E^u(x)$;
(3) if $n \in \mathbb{N}$ then

$$\begin{aligned} &\rho\big(f^n(x), f^n(y)\big) \leqslant C(\mu_1 + \varepsilon)^n \rho(x, y) \quad \text{for } y \in V^s(x), \\ &\rho\big(f^{-n}(x), f^{-n}(y)\big) \leqslant C(\lambda_3 - \varepsilon)^n \rho(x, y) \quad \text{for } y \in V^u(x), \end{aligned}$$

where $C > 0$ is a constant and $\varepsilon > 0$ is sufficiently small.

Set

$$V^{so}(N) = \bigcup_{x \in N} V^s(x) \quad \text{and} \quad V^{uo}(N) = \bigcup_{x \in N} V^u(x). \tag{4.3}$$

These are topological manifolds called local *stable* and *unstable* manifolds of N. They are f-invariant and

$$N = V^{so}(N) \cap V^{uo}(N).$$

THEOREM 4.2 (Hirsch, Pugh and Shub [56, Proposition 5.7, Theorem 3.5]). *$V^{uo}(N)$ and $V^{so}(N)$ are Lipschitz continuous and indeed smooth submanifolds of M.*

In [56], Hirsch, Pugh and Shub used the Hadamard method for constructing local stable and unstable manifolds through N. Their approach does not rely on the existence of local stable manifolds through individual points $x \in N$ but instead, builds local stable and unstable manifolds through N as a whole. Of course, a posteriori one can derive (4.3). Hirsch,

Pugh and Shub obtained more complete information on local stable and unstable manifolds through N. In particular, they showed that a normally hyperbolic manifold N survives under small perturbation of the system, thus establishing stability of normal hyperbolicity.

THEOREM 4.3 (Hirsch, Pugh and Shub [56, Sections 4–6]). *Let f be a C^q embedding with $q \geqslant 1$ that is normally hyperbolic to a compact smooth manifold N,*

$$\begin{aligned} \ell_u &:= \max\{j \in \{0,\dots,q\} \mid \mu_1 < \lambda_2^j\}, \\ \ell_s &:= \max\{j \in \{0,\dots,q\} \mid \mu_2^j < \lambda_3\}, \end{aligned} \qquad \ell := \min\{\ell_s, \ell_u\}, \tag{4.4}$$

where λ_i and μ_i, $i = 1, 2, 3$, are as in Definition 4.1. *Then*

(1) *Existence*: *there exist locally f-invariant submanifolds $V^{so}(N)$ and $V^{uo}(N)$ tangent to $E^s \oplus TN$ and $E^u \oplus TN$, respectively.*
(2) *Uniqueness*: *if N' is an f-invariant set which lies in an ε-neighborhood $U_\varepsilon(N)$ of N, for sufficiently small ε, then $N' \subset V^{so}(N) \cup V^{uo}(N)$.*
(3) *Characterization*: *$V^{so}(N)$ (respectively, $V^{uo}(N)$) consists of all points y for which $\rho(f^n(y), N) \leqslant r$ for all $n \geqslant 0$ (respectively, $n \leqslant 0$) and some small $r > 0$; indeed, $\rho(f^n(y), N) \to 0$ exponentially as $n \to +\infty$ (respectively, as $n \to -\infty$).*
(4) *Smoothness*: *$V^{so}(N)$ and $V^{uo}(N)$ are submanifolds in M of class C^{ℓ_s} and C^{ℓ_u}, respectively; in particular, N is a C^ℓ submanifold.*
(5) *Lamination*: *$V^{so}(N)$ and $V^{uo}(N)$ are fibered by $V^s(x)$ and $V^u(x)$, $x \in N$; see* (4.3).

For every $\delta > 0$ there exist $r > 0$ and $\varepsilon > 0$ such that

(6) *for every embedding g of class C^q with $d_{C^1}(f, g) \leqslant \varepsilon$, there exists a smooth submanifold N_g, invariant under g, to which g is normally hyperbolic; N_g lies in an r-neighborhood $U_r(N_f)$ of N_f;*
(7) *$V_g^{so}(N) \in C^{\ell_s}$, $V_g^{uo}(N) \in C^{\ell_u}$ and $N_g \in C^\ell$ (where the numbers ℓ_s, ℓ_u, and ℓ are given by* (4.4)); *they depend continuously on g in the C^1 topology;*
(8) *there exists a homeomorphism $H : U_r \to M$ which is δ-close to the identity map in the C^0 topology and such that $H(N_f) = N_g$.*

4.2. *Integrability of the central foliation and dynamical coherence*

Let M be a compact smooth Riemannian manifold and $f : M \to M$ a diffeomorphism that is partially hyperbolic with a df-invariant splitting of the tangent bundle (2.6) satisfying (2.7). The central distribution E^c may not, in general, be integrable as Section 5.1 below illustrates. Nonintegrability is an open property (Theorem 4.9).

We describe some conditions that guarantee integrability of the central distribution. In fact, these conditions guarantee the stronger property of unique integrability, which we introduce now.

DEFINITION 4.4. A partially hyperbolic embedding is said to be *dynamically coherent* if E^{cs} and E^{cu} are integrable to foliations W^{cs} and W^{cu}, respectively.

In this case E^c is integrable to the *central foliation* W^c for which $W^c(x) = W^{cu}(x) \cap W^{cs}(x)$, each leaf of W^{cs} is foliated by leaves of W^c and W^s, and each leaf of W^{cu} is foliated by leaves of W^c and W^u.

Note that these assumptions do not imply that the integral foliations are unique, so it is not clear whether the central subbundle is uniquely integrable in this case. [17] demonstrates that Hölder continuous distributions may have many different integral foliations. On the other hand, there is no known example in which a central subbundle is integrable and not uniquely integrable.

Brin communicated the following result, whose proof is essentially contained in [26]:

THEOREM 4.5. *If the central subbundle is uniquely integrable then the system is dynamically coherent, and the center-stable and center-unstable subbundles are also uniquely integrable.*

DEFINITION 4.6. A foliation W of M is said to be *quasi-isometric* if there are $a > 0$ and $b > 0$ such that $\rho_W(x, y) \leqslant a \cdot \rho(x, y) + b$ for every $x \in M$ and every $y \in W(x)$, where ρ_W is the distance along the leaves of W.

A partially hyperbolic embedding f is said to be *center-isometric* if it acts isometrically in the central direction, i.e., $\|df(x)v\| = \|v\|$ for every $x \in \Lambda$ and $v \in E^c(x)$.

Denote by $\widetilde{M}$ the universal cover of M and by $\widetilde{W}^s$ and $\widetilde{W}^u$ the lifts of the stable and unstable foliations to $\widetilde{M}$.

THEOREM 4.7 (Brin [25,71]). *Let M be a compact smooth Riemannian manifold and $f : M \to M$ a diffeomorphism which is partially hyperbolic with a df-invariant splitting of the tangent bundle* (2.6) *satisfying* (2.7).

If $\widetilde{W}^s$ and $\widetilde{W}^u$ are quasi-isometric in the universal cover $\widetilde{M}$, then the distributions E^{cs}, E^{cu} and E^c are locally uniquely integrable.

If f is center-isometric then the central distribution E^c is locally uniquely integrable.

As we mentioned above the central distribution may not be integrable. However, it is often weakly integrable (Definition 3.1), and this weak integrability persists under small perturbations:

THEOREM 4.8 (Brin, Burago and Ivanov [26]). *Let f be a partially hyperbolic diffeomorphism of M. Assume the distributions E^{cs}_f, E^{cu}_f and E^c_f are weakly integrable. Then there is a C^1 neighborhood $\mathcal{U}$ of f such that every $g \in \mathcal{U}$ is a partially hyperbolic diffeomorphism whose distributions E^{cs}_g, E^{cu}_g and E^c_g are weakly integrable.*

So long as one stays safely within the partially hyperbolic context this weak integrability is also a closed property:

THEOREM 4.9 (Brin, Burago and Ivanov [26]). *Let $\{f_n\}_{n \geqslant 0}$ be a sequence of partially hyperbolic diffeomorphisms of M. Assume that*

(1) *$f_n \to g$ in the C^1 topology;*

(2) *the distributions* $E^{cs}_{f_n}$, $E^{cu}_{f_n}$ *and* $E^{c}_{f_n}$ *are weakly integrable for all* n;
(3) *all* f_n *have the same hyperbolicity constants* (2.5).
Then g *is partially hyperbolic and the distributions* E^{cs}_g, E^{cu}_g *and* E^{c}_g *are weakly integrable.*

4.3. *Smoothness of central leaves via normal hyperbolicity*

Theorem 4.3 gives a fair amount of information about normally hyperbolic submanifolds. Since the center foliation, if defined, is "essentially" normally hyperbolic in that the strong contraction and expansion act transversely to it, one would like to apply Theorem 4.3 to this situation. Hirsch, Pugh and Shub developed a construction which allows one to do this.

Let f be a partially hyperbolic diffeomorphism with a df-invariant splitting of the tangent bundle (2.6) satisfying (2.7) and with E^c integrable. For $r > 0$ let $U_r(W^c(x)) \subset M$ be the tubular neighborhood of radius r of the leaf $W^c(x)$. Consider the manifold that is the *disjoint* union

$$\mathcal{M}_r = \bigcup_{x\in M} U_r\big(W^c(x)\big). \tag{4.5}$$

For sufficiently small ε, $0 < \varepsilon < r$, the map f induces a diffeomorphism $F : \mathcal{M}_\varepsilon \to \mathcal{M}_r$ which is normally hyperbolic to the submanifold

$$\mathcal{N} = \bigcup_{x\in M} W^c(x) \subset \mathcal{M}_\varepsilon.$$

The manifolds $\mathcal{M}_\varepsilon$ and $\mathcal{N}$ are not compact but complete. Theorem 4.3 extends to this situation because the proof relies only on the existence of a tubular neighborhood of the normally hyperbolic manifold, a uniform lower bound for the radius of injectivity of the exponential map, and uniform estimates (4.2). We have all this at our disposal since the manifold M is compact and the central foliation W^c is integrable. This yields a corollary of Theorem 4.3:

THEOREM 4.10. *Let* $f : M \to M$ *be a partially hyperbolic embedding with integrable central distribution. Then* $W^c(x) \in C^\ell$ *for every* $x \in M$, *where* ℓ *is as in* (4.4).

4.4. *Robustness of the central foliation*

A far less straightforward application of the Hirsch–Pugh–Shub construction can be used to produce robustness of the central foliation W^c under small perturbation of the system; the subtlety of the matter is evidenced by the requirement that this foliation be smooth.

THEOREM 4.11 (Hirsch, Pugh and Shub [56, Theorem 7.5]). *Assume that the central distribution* E^c *for* f *is integrable, that the corresponding foliation* W^c *is smooth and that* g *is a* C^q *diffeomorphism sufficiently close to* f *in the* C^1 *topology. Then* g *is partially hyperbolic with integrable central distribution* E^c_g.

The direct approach suggested above is carried out in [71], but this result is usually obtained as an immediate corollary of Theorems 4.13 and 4.15 below, whose proofs exploit plaque expansivity and pseudo-orbits.

In general, we do not have a smooth central foliation. Moreover, even if the central foliation for f were smooth, the central foliation for a "typical" perturbation of f would not be. Thus the assumptions of integrability and smoothness are not jointly robust; only integrability persists if we assume both at the outset. In [56], Hirsch, Pugh and Shub introduced a property of the central foliation for f called *plaque expansivity* that is weaker than smoothness (see Theorem 4.13 below) but still guarantees integrability of the central distribution for sufficiently small perturbations of f and furthermore persists itself under small perturbations (see Theorem 4.15 below).

DEFINITION 4.12. Let W be a foliation of a compact smooth manifold M whose leaves are C^r smooth immersed submanifolds of dimension k. Given a point $x \in M$, we call the set $P(x) \subset W(x)$ a C^r *plaque* of W at x if $P(x)$ is the image of a C^r embedding of the unit ball $D \subset \mathbb{R}^k$ into $W(x)$. A *plaquation* $\mathcal{P}$ for W is a collection of plaques such that every point $x \in M$ is contained in a plaque $P \in \mathcal{P}$.

Let $\{x_n\}_{n\in\mathbb{Z}}$ be a pseudo-orbit for f (see [3,2]). We say that the pseudo-orbit *respects* a plaquation $\mathcal{P}$ for W if for every $n \in \mathbb{Z}$ the points $f(x_n)$ and x_{n+1} lie in a common plaque $P \in \mathcal{P}$.

Assume that the foliation W is invariant under a diffeomorphism f of M. We say that f is *plaque expansive* with respect to W if there exists $\varepsilon > 0$ with the following property: if $\{x_n\}_{n\in\mathbb{Z}}$ and $\{y_n\}_{n\in\mathbb{Z}}$ are ε-pseudo-orbits which respect W and if $\rho(x_n, y_n) \leqslant \varepsilon$ for all $n \in \mathbb{Z}$ then x_n and y_n lie in a common plaque for all $n \in \mathbb{Z}$.

Note that plaque expansivity does not depend on the choice of either the Riemannian structure in M or the plaquation $\mathcal{P}$ for W. It is indeed weaker than smoothness:

THEOREM 4.13 (Hirsch, Pugh and Shub [56, Theorem 7.2]). *Let f be a partially hyperbolic diffeomorphism. Assume that the central distribution E^c is integrable and the central foliation W^c is smooth. Then W^c is plaque expansive.*

REMARK 4.14. If $df \restriction E^c(x)$ acts as an isometry for every $x \in M$ then the central distribution E^c is integrable by Theorem 4.7, and the central foliation W^c is plaque expansive (see [56, Section 7]).

THEOREM 4.15 (Hirsch, Pugh and Shub [56, Theorem 7.1]). *Let $f : M \to M$. If f is partially hyperbolic with the central distribution E^c_f for f integrable and f plaque expansive with respect to the central foliation W^c_f then the same holds for any sufficiently C^1-close diffeomorphism g (with respect to E^c_g and W^c_g).*

5. Intermediate foliations

The central distribution we studied in the previous section corresponds to the central ring in the Mather spectrum, and it is now natural to study the structures associated with other intermediate rings (as opposed to the inner- and outermost ones, which figured in Section 3).

Consider a diffeomorphism f of class C^q of a compact Riemannian manifold M admitting a df-invariant splitting (3.1) satisfying (3.2) and (3.3). Given $1 < k < t$ with $\mu_k < 1$ we now discuss the integrability problem for the invariant distribution E_k, called the *intermediate* distribution.

5.1. *Nonintegrability of intermediate distributions*

In general, E_k is not integrable as we now illustrate with an example that goes back to Smale [88] and appears in [62, Section 17.3] as well as [92, p. 1549], where it provides an example of a diffeomorphism that is normally hyperbolic with respect to a smooth, 1-dimensional foliation and not conjugate to the time-1 map of any Anosov flow (and can be shown to be stably ergodic using the methods of [49]).

Consider the *Heisenberg group* of matrices

$$H = \left\{ \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} : (x, y, z) \in \mathbb{R} \right\}$$

with the usual matrix multiplication: in (x, y, z) coordinates it is given by

$$(x_1, y_1, z_1) \times (x_2, y_2, z_2) = (x_1 + x_2, y_1 + y_2, z_1 + z_2 + x_1 y_2).$$

The center of H is the 1-parameter subgroup

$$\begin{pmatrix} 1 & 0 & z \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus, H is a 3-dimensional, simply connected, non-Abelian nilpotent group. Its Lie algebra is

$$\mathcal{L}(H) = \left\{ \begin{pmatrix} 0 & x & z \\ 0 & 0 & y \\ 0 & 0 & 0 \end{pmatrix} : (x, y, z) \in \mathbb{R} \right\}$$

with generators

$$X = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Then $[X, Y] = Z$ while all other Lie brackets of generators are zero.

Let $G = H \times H$ be the Lie group with generators $X_1, Y_1, Z_1, X_2, Y_2, Z_2$ such that $[X_i, Y_i] = Z_i$ and all other brackets of generators are zero. Its Lie algebra is

$$\mathcal{L}(G) = \left\{ \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \mid A, B \in \mathcal{L}(H) \right\}.$$

The group H has an obvious integer lattice of matrices with entries in $\mathbb{Z}$ which generates an integer lattice in G. We need another lattice in G however.

Consider the number field $\mathbb{K} = \{a + b\sqrt{5} \mid a, b \in \mathbb{Q}\}$. It possesses a unique nontrivial automorphism σ such that $\sigma(a + b\sqrt{5}) = a - b\sqrt{5}$.

Let Γ be the subgroup of G given by $\exp_{\mathrm{Id}} \gamma$, where $\exp_{\mathrm{Id}} : \mathcal{L}(G) \to G$ is the exponential map and

$$\gamma := \left\{ \begin{pmatrix} A & 0 \\ 0 & \sigma(A) \end{pmatrix} \mid A \in \mathcal{L}(H) \text{ with entries in the algebraic integers in } \mathbb{K} \right\} \subset \mathcal{L}(G)$$

with $\sigma(A)_{ij} = \sigma(A_{ij})$. It can be shown that Γ is a lattice [62, Section 17.3]. Define a Lie algebra automorphism Φ on $\mathcal{L}(G)$ by

$$\Phi(X_1) = \lambda_1 X_1, \quad \Phi(Y_1) = \lambda_1^2 Y_1, \quad \Phi(Z_1) = \lambda_1^3 Z_1,$$
$$\Phi(X_2) = \lambda_1^{-1} X_2, \quad \Phi(Y_2) = \lambda_1^{-2} Y_2, \quad \Phi(Z_2) = \lambda_1^{-3} Z_2,$$

where $\lambda_1 = \frac{3+\sqrt{5}}{2}$ and $\lambda_2 = \frac{3-\sqrt{5}}{2}$. There exists a unique automorphism $F : G \to G$ with $dF_{|\mathrm{Id}} = \Phi$. Since λ_1 and λ_2 are units in $\mathbb{K}$, that is integers whose inverses are also integers, and $\sigma(\lambda_1) = \lambda_2$ we have $F(\Gamma) = \Gamma$. Thus, F projects to an Anosov diffeomorphism f of $\Gamma \setminus G$.

The invariant splitting for f is $T(\Gamma \setminus G) = E^s \oplus E^u$, where E^s is the 3-dimensional distribution generated by X_2, Y_2 and Z_2 and E^u is the 3-dimensional distribution generated by X_1, Y_1 and Z_1. Observe that $E^u = P \oplus Q$ where P is the 2-dimensional distribution generated by X_1, Y_1 and Q is the 1-dimensional distribution generated by Z_1. The distribution P is intermediate and is not integrable. To see this note that the generators X_1, Y_1 and Z_1 induce three vector fields x_1, y_1 and z_1 on $g \in \Gamma \setminus G$ such that $x_1(g), y_1(g) \in P(g)$ and $z_1(g) \in Q(g)$ for any $g \in \Gamma \setminus G$. Since the distribution P is smooth, by the Frobenius theorem, its integrability would imply that the Lie bracket $[x_1, y_1]$ of vector fields x_1 and y_1 lies in P, contrary to $[X_1, Y_1] = Z_1$.

It follows from Theorem 4.9 that nonintegrability in this example is an open property.

5.2. *Invariant families of local manifolds*

Recall that we assume there is a df-invariant splitting (3.1) satisfying (3.2) and (3.3) and we consider $1 < k < t$ for which $\mu_k < 1$. The preceding example notwithstanding, there

are positive results for the integrability problem for the invariant distribution E_k. After all, if $1 < k < t$ then the intermediate distribution E_k is the central distribution in the splitting

$$TM = \left(\bigoplus_{j=1}^{k-1} E_j\right) \oplus E_k \oplus \left(\bigoplus_{j=k+1}^{t} E_j\right),$$

so we can apply results of Sections 4.3 and 4.4.

REMARK 5.1. For $\bigoplus_{j=1}^{k-1} E_j$ and $\bigoplus_{j=k+1}^{t} E_j$ to be integrable we need this to correspond to a standard partially hyperbolic situation, which by (2.5) requires $\mu_{k-1} < 1$ (this follows from our assumption $\mu_k < 1$) as well as $\lambda_{k+1} > 1$. By (3.2) and (3.3) this means that f is an Anosov diffeomorphism.

Theorem 4.10 gives the class of smoothness of the leaves of the foliation W_k when E_k is integrable:

THEOREM 5.2. *With the notations of* (3.2) *and* (3.3), *suppose* η_k *and* m_k *are the largest integers such that* $\mu_{k-1} < \lambda_k^{\eta_k}$ *and* $\mu_k < \lambda_{k+1}^{m_k}$, *respectively, and let* $n_k = \min\{\eta_k, m_k\}$. *If* E_k *is integrable then the leaves of the corresponding intermediate invariant foliation* W_k *are* C^{n_k}.

Note that the assumptions are closely related to those of Theorem 3.3, but the conclusion is complementary. The present result asserts smoothness of leaves, whereas Theorem 3.3 is about smooth dependence of the leaves on a base point (when those leaves are known to be as smooth as the diffeomorphism by Theorem 3.2).

As to robustness of the integral foliation W_k, we wish to apply Theorem 4.15.

THEOREM 5.3. *Assume* W_k *is plaque expansive* (*e.g., smooth*) *and that* $\lambda_{k+1} > 1$. *Let* g *be a* C^q *diffeomorphism sufficiently close to* f *in the* C^1 *topology. By Theorem* 2.17, g *possesses an invariant distribution* $(E_k)_g$ *corresponding to* E_k. *This distribution is integrable and the corresponding foliation* $(W_k)_g$ *is plaque expansive.*

Since the diffeomorphism f in the last theorem is Anosov by Remark 5.1 so is g. By the structural stability theorem, f and g are topologically conjugate by a Hölder homeomorphism h which is close to the identity map. It follows that $h(W_k)$ is a g-invariant foliation whose leaves are Hölder continuous submanifolds. Theorem 5.3 shows that the leaves of this foliation are indeed smooth of class C^ℓ.[3]

Even if the distribution E_k is integrable its leaves may not be C^q. To explain this phenomenon consider the linear map $A(x, y) = (\lambda x, \mu y)$ of the plane, where $0 < \lambda < \mu < 1$. The origin is an attracting fixed point. The x-axis can be geometrically characterized as consisting of points P for which

$$\rho(0, A^n P) \leqslant \lambda^n \rho(0, P).$$

[3] $h(W_k)$ is an integral foliation because by a lemma of Hirsch–Pugh–Shub, normally hyperbolic manifolds are unique and robust in the C^0 topology.

On the other hand any curve $\gamma_C = \{(x, y)\colon\ x = Cy^{\log\lambda/\log\mu}\}$ is invariant under A and consists of points P for which

$$\rho(0, A^n P) \leqslant \mu^n \rho(0, P).$$

Note that for $\log\lambda/\log\mu \notin \mathbb{N}$ (nonresonance) these curves are only finitely differentiable except for the y-axis (corresponding to $C = 0$). Therefore, there is no "obvious" choice of a local leaf and it seems unlikely that the intermediate foliation will happen to include the leaf that is infinitely differentiable.

However, if $\mu_k < 1$ and some special *nonresonance* condition holds, smooth leaves are realizable: the distribution E_k admits an invariant family of local manifolds $\{V_k(x)\}_{x\in M}$ which are as smooth as the map f is—but they may not constitute a foliation.

THEOREM 5.4 (Pesin [70]). *Fix k such that $0 < \lambda_k \leqslant \mu_k < 1$ and assume the* nonresonance condition $N := [\log\lambda_1/\log\mu_k] + 1 \leqslant q$ *and if $j = 1, \dots, N$, $1 \leqslant i < k$ then $[(\lambda_k)^j, (\mu_k)^j] \cap [\lambda_i, \mu_i] = \emptyset$.*

Then for every $x \in M$ there exists a local submanifold $V_k(x)$ such that:

(1) $x \in V_k(x)$ *and* $T_x V_k(x) = E_k(x)$;
(2) $f(V_k(x)) \subset V_k(f(x))$;
(3) $V_k(x) \in C^q$;
(4) *for any $x \in M$ the collection of local manifolds $\{V_k(x)\}_{x\in M}$ is the only collection of C^N local manifolds that satisfies $T_x V_k(x) = E_k(x)$, $f(V_k(x)) \subset V_k(f(x))$, and*

$$\sup_{1\leqslant s\leqslant N} \sup_{x\in M} \|d^s V_k(x)\| \leqslant \text{const}.$$

REMARK 5.5. The nonintegrable intermediate distribution P for the diffeomorphism in Section 5.1 does not satisfy the nonresonance condition.

5.3. *Lack of smoothness of the intermediate foliations*

The following example illustrates the possible lack of smoothness of leaves for intermediate distributions. Consider an automorphism A of the torus $\mathbb{T}^3$ with eigenvalues λ_i, $i = 1, 2, 3$, such that $0 < \lambda_1 < \lambda_2 < 1 < \lambda_3$. We have an invariant splitting

$$T\mathbb{T}^3 = \bigoplus_{i=1}^{3} E_{i,A}.$$

Assume $\log\lambda_1/\log\lambda_2 \notin \mathbb{Z}$ (nonresonance), and let $N = [\log\lambda_1/\log\lambda_2] + 1$.

Consider the foliation $W_{2,A}$ associated to $E_{2,A}$. By Theorem 5.2 any C^∞ diffeomorphism f sufficiently C^1-close to A possesses an invariant foliation $W_{2,f}$ tangent to $E_{2,f}$ and with C^{N-1} leaves. In general, the leaves $W_{2,f}(x)$ cannot be more than C^{N-1} smooth for a "large" set of points $x \in M$. Hence, they are different from the local submanifolds given by the preceding theorem, since these submanifolds are of class C^N (indeed, of class C^∞ in this particular case).

THEOREM 5.6 (Jiang, de la Llave and Pesin [61]). *In any neighborhood η of A in the space* $\mathrm{Diff}^1(\mathbb{T}^3)$ *there exists $G \in \eta$ such that*

(1) *G is a C^∞ diffeomorphism and topologically conjugate to A;*

(2) *G admits an invariant splitting*

$$T\mathbb{T}^3 = \bigoplus_{i=1}^{3} E_{i,G}$$

with $E_{i,G}$ close to $E_{i,A}$ and integrable; the integral manifold $W_{i,G}(x)$ passing through x is of class C^{N-1} but not C^N for some $x \in \mathbb{T}^3$;

(3) *the set of points $\{x \mid W_{i,G}(x)$ is not of class $C^N\}$ is a residual subset of $\mathbb{T}^3$.*

6. Failure of absolute continuity

Let W be a foliation of M with smooth leaves and $V(x)$, $x \in M$, the local leaf passing through x. In our discussion of the stable and unstable foliations in Section 3 we discussed the question of absolute continuity:

> If $E \subset B(x, q)$ is a Borel set of positive volume, can the intersection $E \cap V(y)$ have zero Lebesgue measure (with respect to the Riemannian volume on $V(y)$) for almost every $y \in E$?

6.1. *An example of a foliation that is not absolutely continuous*

We describe a scheme due to Katok for producing partially hyperbolic maps whose central foliation fails to be absolutely continuous in the strongest possible way: there is a set of full measure that intersects each leaf of the foliation in at most one point. This phenomenon is known as "Fubini's nightmare" since the Fubini theorem fails with respect to this foliation in the strongest possible way. (An example of this construction on an annulus was widely circulated from 1992 [32], and in 1997 a version on the square was published [65].) We thank Keith Burns for providing the presentation rendered here.

Let A be the hyperbolic automorphism of the torus $\mathbb{T}^2$ defined by the matrix

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

There is a family $\{f_t \mid t \in [0, 1]\}$ of diffeomorphisms preserving the area m and satisfying the following conditions:

(1) f_t is a small perturbation of A for every $t \in [0, 1]$;

(2) f_t depends smoothly on t;

(3) $l'(t) \neq 0$, where $l(t)$ is the larger eigenvalue of the derivative of f_t at its fixed point.

The diffeomorphisms f_t are all Anosov, conjugate to A, and ergodic with respect to m. For any s and t in $[0, 1]$, the maps f_s and f_t are conjugate via a unique homeomorphism h_{st} close to the identity, i.e., $f_t = h_{st} \circ f_s \circ h_{st}^{-1}$. The homeomorphism h_{st} is Hölder continuous.

Let m_{st} be the pushforward of m by h_{st}. Then m_{st} is an ergodic invariant measure for f_t. Using the condition on $l(t)$ and the following lemma, we see that $m \neq m_{st}$ unless $s = t$.

LEMMA 6.1 (de la Llave [39]). *Suppose $f, g : \mathbb{T}^2 \to \mathbb{T}^2$ are smooth area-preserving Anosov diffeomorphisms that are conjugate via an area-preserving homeomorphism h. Let p be a periodic point for f with least period k. Then $Df^k(p)$ and $Dg^k(h(p))$ have the same eigenvalues up to sign.*

PROOF. Let λ and λ' be the eigenvalues of $Df^k(p)$ and $Dg^k(h(p))$, respectively, that lie inside the unit circle. Since f and g are area-preserving, the other eigenvalues of $Df^k(p)$ and $Dg^k(h(p))$ are $1/\lambda$ and $1/\lambda'$, respectively. Choose $x \in W^u_{\mathrm{loc}}(p; f) \smallsetminus \{p\}$ and $y \in W^s_{\mathrm{loc}}(p; f) \smallsetminus \{p\}$ that are not equal to p. Let R_n be the smallest "rectangle" bounded by (parts of) $W^s_{\mathrm{loc}}(p; f)$, $W^s_{\mathrm{loc}}(f^{-kn}(x); f)$, $W^u_{\mathrm{loc}}(p; f)$, and $W^u_{\mathrm{loc}}(f^{kn}(y); f)$. Let R'_n be the smallest "rectangle" bounded by (parts of) $W^s_{\mathrm{loc}}(h(p); g)$, $W^s_{\mathrm{loc}}(h(f^{-kn}(x)); g)$, $W^u_{\mathrm{loc}}(h(p); g)$, and $W^u_{\mathrm{loc}}(h(f^{kn}(y)); g)$. Then

$$\lim_{n\to\infty} \frac{\operatorname{area}(R_{n+1})}{\operatorname{area}(R_n)} = \lambda^{2k} \quad \text{and} \quad \lim_{n\to\infty} \frac{\operatorname{area}(R'_{n+1})}{\operatorname{area}(R_n)} = \lambda'^{2k}.$$

On the other hand, the conjugacy h takes R_n to R'_n for any n. Since h is area-preserving, it follows that $\lambda = \pm\lambda'$. □

A point is *generic* with respect to an invariant measure if the forward and backward Birkhoff averages of any continuous function are defined at the point and are equal to integral of the function with respect to the measure. If x is generic for f_s with respect to m, then $h_{st}(x)$ is generic for f_t with respect to m_{st} and hence is not generic for f_t with respect to m, unless $s = t$. (To see this, note that the Birkhoff averages of a continuous function φ along the f_t-orbit of $h_{st}(x)$ are the same as the Birkhoff averages of $\varphi \circ h_{st}$ along the f_s orbit of x.)

Now consider the diffeomorphism $F : \mathbb{T}^2 \times [0,1] \to \mathbb{T}^2 \times [0,1]$ given by $F(x,t) = (f_t(x), t)$. We have just observed that for any $x \in \mathbb{T}^2$ the set $H(x) = \{(h_{0t}(x), t) \mid t \in [0,1]\}$ contains at most one element of the set $\mathcal{G}$ of points $(y,t) \in \mathbb{T}^2 \times [0,1]$ such that y is generic for f_t with respect to m.

Now, F is a small perturbation of $A \times \mathrm{Id}_{[0,1]}$ and thus partially hyperbolic. It follows from Theorem 4.11 that F has a center foliation whose leaves are small perturbations of the intervals $\{x\} \times [0,1]$ for $x \in \mathbb{T}^2$. Since F maps the tori $\mathbb{T}^2 \times \{t\}$ into themselves, it is easily seen that the leaves of W^c_F are ℓ-normally hyperbolic for any ℓ, and hence are C^∞ by Theorem 4.10. On the other hand, for each $x \in \mathbb{T}^2$, the leaf of W^c_F that passes through $(x,0) \in \mathbb{T}^2 \times [0,1]$ is $H(x)$.

The set $\mathcal{G}$ of generic points for F has full measure with respect to m in each torus $\mathbb{T}^2 \times \{t\}$ and hence has full Lebesgue measure in $\mathbb{T}^2 \times [0,1]$, but, as observed above, it intersects each center leaf in at most one point.

To construct an analogous example on $\mathbb{T}^2 \times S^1$ use two periodic points simultaneously instead of the one fixed point. The example here is constructed in such a

way that $l(t) = l(s) \Rightarrow t = s$. For a continuous parametrization using $t \in S^1$ this will not work, but starting from the map A^2 instead, which has several fixed points, we use perturbations for which the largest eigenvalues $l_1(t)$ and $l_2(t)$ at two fixed points $x_1(t)$ and $x_2(t)$ satisfy $l_1(t) = l_1(s)$ and $l_2(t) = l_2(s) \Rightarrow t = s \pmod 1$. For example, make $l_1'(t) > 0$ on $(0, 1/2)$, $l_1'(t) < 0$ on $(1/2, 1)$ and $l_2'(t) = 0$ on $(0, 1/2)$, $l_1'(t) > 0$ on $(1/2, 3/4)$, $l_2'(t) < 0$ on $(3/4, 1)$.

6.2. *Pathological foliations*

We saw earlier that even in terms of existence, uniqueness and smoothness of leaves the central foliation is a rather more delicate entity than the members of the stable and unstable filtrations, and the preceding example shows that if there is a central foliation at all it may fail to be absolutely continuous. It turns out that this is not at all exceptional.

Let A be an area-preserving linear hyperbolic automorphism of the 2-dimensional torus $\mathbb{T}^2$. Consider the map $F = A \times \mathrm{Id}$ of the 3-dimensional torus $\mathbb{T}^3 = \mathbb{T}^2 \times S^1$. Any sufficiently small C^1 perturbation G of F is uniformly partially hyperbolic with 1-dimensional central distribution. The latter is integrable to a continuous foliation W^c of M with compact leaves (they are diffeomorphic to S^1; this foliation can be shown to be Hölder continuous [80]). There is a perturbation G of F which preserves volume and has nonzero Lyapunov exponents in the central direction [87] (see also [42]). In this case the central foliation is not absolutely continuous: for almost every $x \in M$ the conditional measure (generated by the Riemannian volume) on the leaf $W^c(x)$ of the central foliation passing through x has finite support [83].

We describe a more general version of this result. Let (X, ν) be a probability space and $f : X \to X$ an invertible transformation that preserves the measure ν and is ergodic with respect to ν. Let M be an n-dimensional smooth compact Riemannian manifold and $\varphi : X \to \mathrm{Diff}^{1+\alpha}(M)$. Assume that the skew-product transformation

$$F : X \times M \to X \times M, \quad F(x, y) = \big(f(x), \varphi_x(y)\big)$$

is Borel measurable and possesses an invariant ergodic measure μ on $X \times M$ such that $\pi_* \mu = \nu$, where $\pi : X \times M \to X$ is the projection.

For $x \in X$ and $k \in \mathbb{Z}$ define $\varphi_x^{(k)} : M \to M$ by

$$\varphi_x^{(k+1)} = \varphi_{f^k(x)} \circ \varphi_x^{(k)},$$

where $\varphi_x^{(0)} = \mathrm{Id}$. Since the tangent bundle to M is measurably trivial the derivative map of φ along the M direction gives a cocycle

$$\mathcal{A} : X \times M \times \mathbb{Z} \to GL(n, \mathbb{R}),$$

where $\mathcal{A}(x, y, k) = d_y \varphi_x^{(k)}$. If $\log^+ \|d\varphi\| \in L^1(X \times M, \mu)$ then the Multiplicative Ergodic Theorem and ergodicity of μ imply that the Lyapunov exponents $\chi_1 < \cdots < \chi_\ell$ of this cocycle are constant for μ-almost every (x, y).

THEOREM 6.2 (Ruelle and Wilkinson [83]). *If for some $\gamma > 0$ the function φ satisfies*

$$\log^+ \|d\varphi\|_\gamma \in L^1(X, \nu), \tag{6.1}$$

where $\|\cdot\|_\gamma$ is the γ-Hölder norm, and if $\chi_\ell < 0$ then there exists a set $S \subset X \times M$ of full measure and $k \in \mathbb{N}$ such that $\operatorname{card}(S \cap (\{x\} \times M)) = k$ *for almost every $x \in X$.*

This phenomenon is rather typical.

DEFINITION 6.3. A partially hyperbolic diffeomorphism that preserves a smooth measure is said to have *negative central exponents* if the Lyapunov exponents in the central direction are negative almost everywhere.

CONJECTURE 6.4. *The central foliation of a "typical" partially hyperbolic diffeomorphism with negative central exponents is not absolutely continuous.*

Mañé proved (unpublished) that if the central foliation is one-dimensional and has compact leaves then this foliation is not absolutely continuous provided the Lyapunov exponent in the central direction is nonzero on a set of positive measure.[4] Hirayama and Pesin [54] showed that the central foliation is not absolutely continuous if it has compact leaves and f is "central dissipative", i.e., the sum of the central exponents is nonzero on a set of positive measure (here, negative, positive or zero exponents can be present). Note that partially hyperbolic central dissipative diffeomorphisms whose central foliation has compact leaves form an open set in the space of C^1 diffeomorphisms and that any partially hyperbolic diffeomorphism whose central foliation has compact leaves can be perturbed to become central dissipative.

This motivates the question whether one can perturb a partially hyperbolic system with all central Lyapunov exponents zero to a system with negative central exponents. This has been shown to be true in some particular cases (see [8,43,12,14]) but remains unknown otherwise.

CONJECTURE 6.5. *Given a partially hyperbolic dynamical system f whose central Lyapunov exponents are zero, there exists a partially hyperbolic dynamical system with negative central exponents arbitrarily close to f.*

7. Accessibility and stable accessibility

We now begin our study of the ergodic theory of partially hyperbolic dynamical systems. The strategy for establishing ergodicity is based on suitable extensions of the Hopf argument [57], see also [62, p. 217], and we describe it here in order to explain the main object of the present section.

The Hopf argument establishes ergodicity of a uniformly hyperbolic diffeomorphism as follows. By the Birkhoff Ergodic Theorem, ergodicity means that for every L^1-function

[4] We thank A. Wilkinson for providing us with this information.

φ (by L^1-density φ is without loss of generality continuous, hence uniformly continuous by compactness) the time averages or Birkhoff averages $\varphi_n := \frac{1}{n}\sum_{i=0}^{n-1} \varphi \circ f^n$ converge to a constant a.e. Uniform continuity of φ and the contraction of stable leaves imply that the limit function is constant on stable leaves, and likewise for "backwards" time averages (obtained analogously from f^{-1}) on unstable leaves. Since the Birkhoff Ergodic Theorem implies that the forward and backward limits exist and agree a.e., one deduces that these are constant a.e. from the fact that this holds on stable and unstable leaves separately, using absolute continuity: "Almost everywhere on almost every leaf" is the same as "almost everywhere".

For partially hyperbolic dynamical systems the same argument can be attempted, but first of all, one cannot use all three foliations because the Hopf argument relies on contraction in either forward or backward time to conclude that an invariant function is constant on leaves. The center foliation lacks this feature (and may, moreover, fail to be absolutely continuous as we have seen, which would cause problems in the later stage of the argument). On the other hand, in this case it is not clear that any two nearby points have a heteroclinic point. Put differently, in the hyperbolic situation one can join any two nearby points by a path consisting of two short segments, one each in a stable and an unstable leaf. (We call such a path a us-path.) This may not be the case in a partially hyperbolic system, as one sees, for example, in the case of Cartesian products of a hyperbolic dynamical system with the identity, which are evidently not ergodic. More to the point, joint integrability of the stable and unstable foliations limits these connections to pairs of points that lie in the same joint stable–unstable leaf. This motivates interest in how joint integrability can fail. It is conceivable, for example, that there are situations in which the foliations are jointly integrable in some places but not in others, or cases in which they are not jointly integrable but nevertheless both subordinate to a common foliation the dimension of whose leaves is larger than the sum of stable and unstable dimensions. Whether these are possible is not very well understood, and the question of which of these situations may occur in examples is of interest in its own right.

In terms of salvaging the Hopf argument, say, it would be natural to make the assumption that any two nearby points can be joined by a us-path consisting of two short segments in a stable and unstable leaf, respectively ("accessibility by a us-path with short legs"). This should be relatively easy to use. Under the name of "local transitivity" it was imposed by Brin and Pesin [29], but it turned out to be too restrictive to be widely applicable. Therefore one wishes to explore weaker assumptions that are still strong enough to yield topological or measurable irreducibility. One can relax this assumption by allowing "long legs", i.e., by requiring only that two nearby points be connected by a us-path whose stable and unstable pieces may be rather long. On the other hand, one may allow the connection to be established by a path consisting of a multitude of pieces that lie alternatingly in stable and unstable leaves. If one simultaneously drops the requirement that the legs be short, one obtains the notion of accessibility that is now in use.

While it is intuitive to present this notion in terms of paths, and these are used in proving topological transitivity, they are not employed in proofs of ergodicity. The most obvious technical difficulty with these would be that the transition points between stable and unstable segments must have the same forward and backward Birkhoff averages for the function at hand in order for the Hopf argument to work. But this may be tricky to arrange. There-

fore one argues directly with the algebras of sets in the proofs of ergodicity, as explained in the next section.

7.1. *The accessibility property*

DEFINITION 7.1. Let f be a partially hyperbolic diffeomorphism of a compact Riemannian manifold M.

Two points $p, q \in M$ are said to be *accessible*, if there are points $z_i \in M$ with $z_0 = p$, $z_\ell = q$, such that $z_i \in V^\alpha(z_{i-1})$ for $i = 1, \dots, \ell$ and $\alpha = s$ or u. The collection of points $z_0, z_1, \dots, z_\ell$ is called the *us-path* connecting p and q and is denoted variously by $[p, q]_f = [p, q] = [z_0, z_1, \dots, z_\ell]$. (Note that there is an actual path from p to q that consists of pieces of smooth curves on local stable or unstable manifolds with the z_i as endpoints.)

Accessibility is an equivalence relation and the collection of points accessible from a given point p is called the *accessibility class* of p.

A diffeomorphism f is said to have the *accessibility property* if the accessibility class of any point is the whole manifold M, or, in other words, if any two points are accessible.

If f has the accessibility property then the distribution $E^s \oplus E^u$ is not integrable (and therefore, the stable and unstable foliations, W^s and W^u, are not jointly integrable). Otherwise, the accessibility class of any $p \in M$ would be the leaf of the corresponding foliation passing through p.

There is a weaker version of accessibility which provides a useful tool in studying topological transitivity of f.

DEFINITION 7.2. Given $\varepsilon > 0$, we say that f is *ε-accessible* if for every open ball B of radius ε the union of accessibility classes passing through B is M.

An equivalent requirement is that the accessibility class of any point should enter every open ball of radius ε, i.e., be ε-dense. Clearly, if f is accessible then it is ε-accessible for any ε. It is not hard to check that a perturbation of an accessible dynamical system is ε-accessible:

PROPOSITION 7.3. *If a partially hyperbolic diffeomorphism f has the accessibility property and $\varepsilon > 0$ then*

(1) *there exist $\ell > 0$ and $R > 0$ such that for any $p, q \in M$ one can find a us-path that starts at p, ends within distance $\varepsilon/2$ of q, and has at most ℓ legs, each of them with length at most R;*
(2) *there exists a neighborhood $\mathcal{U}$ of f in the space* $\mathrm{Diff}^2(M)$ *such that every $g \in \mathcal{U}$ is ε-accessible.*

Often, an "almost-everywhere" accessibility notion is adequate:

DEFINITION 7.4. We say that f has the *essential accessibility property* if the partition of M by the accessibility classes is trivial in the measure-theoretical sense, i.e., any measurable set that consists of accessibility classes has measure zero or one.

7.2. *Accessibility and topological transitivity*

It is not hard to see that accessibility plus volume-preservation produce a fair amount of recurrence.

DEFINITION 7.5 [3]. Given $\varepsilon > 0$ we say that an orbit is *ε-dense* if the points of the orbit form an ε-net. Clearly, a trajectory $\{f^n(x)\}_{n\in\mathbb{Z}}$ is everywhere dense in M if and only if it is ε-dense for every $\varepsilon > 0$.

We say that a point $x \in M$ is *forward* (respectively, *backward*) *recurrent* if for any $r > 0$ there exists $n > 0$ (respectively, $n < 0$) such that $f^n(x) \in B(x, r)$. If a point x is forward (respectively, backward) recurrent then for any $r > 0$ there exists a sequence $n_k \to +\infty$ (respectively, $n_k \to -\infty$) such that $f^{n_k}(x) \in B(x, r)$.

THEOREM 7.6 (Burns, Dolgopyat and Pesin [31]). *If a partially hyperbolic diffeomorphism f is ε-accessible and preserves a smooth measure then almost every orbit of f is ε-dense.*

PROOF. Fix an open ball B of radius ε. Say that a point is *good* if it has a neighborhood of which almost every point has an iterate in B. We must show that every $p \in M$ is good.

Fix $p \in M$. Since f is ε-accessible, there is a us-path $[z_0, \ldots, z_k]$ with $z_0 \in B$ and $z_k = p$. Then z_0 is good, and we show by induction on j that each z_j is good.

If z_j has a neighborhood N such that $\mathcal{O}(x) \cap B \neq \emptyset$ for almost every $x \in N$ let S be the subset of N consisting of points with this property that are also both forward and backward recurrent. By the Poincaré Recurrence Theorem [3, Theorem 3.4.1], S has full measure in N. If $x \in S$ and $y \in W^s(x) \cup W^u(x)$ then $\mathcal{O}(y) \cap B \neq \emptyset$. The absolute continuity of the foliations W^s and W^u means that $\bigcup_{x\in S}(W^s(x) \cup W^u(x))$ has full measure in the set $\bigcup_{x\in N}(W^s(x) \cup W^u(x))$, which is a neighborhood of z_{j+1}. □

COROLLARY 7.7 (Brin [22]). *Let f be a partially hyperbolic diffeomorphism of a compact Riemannian manifold M that preserves a smooth measure on M and has the accessibility property. Then for almost every point $x \in M$ the trajectory $\{f^n(x)\}_{n\in\mathbb{Z}}$ is dense in M. In particular, f is topologically transitive.*

REMARK 7.8. In fact, Brin proved this using only that every point is nonwandering. This holds in particular when the map preserves a smooth measure as well as when periodic points are dense.

One can relax accessibility to essential accessibility:

THEOREM 7.9 (Burns, Dolgopyat and Pesin [31]). *If a partially hyperbolic diffeomorphism f is essentially accessible and preserves a smooth measure then it is topologically transitive.*

The assumption that f preserves a smooth measure cannot be dropped in general.

THEOREM 7.10 (Niţică and Török [66]). *Consider $F = f \times \mathrm{Id} : M \times S^1 \to M \times S^1$, where f is a C^1 Anosov diffeomorphism of M. There exists a C^1 neighborhood of F whose elements are accessible but not topologically transitive.*

PROOF. By Theorem 7.12 below (see also Theorem 7.13) there is a C^1-open and C^1-dense set of accessible C^1-small perturbations of F, so it suffices to construct an open set of nontransitive diffeomorphisms. Choose $h \in \mathrm{Diff}^1(S^1)$ as close to the identity as desired with h having an attracting fixed point. There are open neighborhoods $U, V \subset S^1$ of this point with $h(\bar{U}) \subset V \subset \bar{V} \subset U$. If $g := f \times h$ then $g(M \times \bar{U}) \subset M \times V$ and any map that is C^0-close to g has the same property. Note that such a transformation is not topologically transitive because each positive semiorbit has at most one element in the open set $M \times (U \setminus \bar{V})$. □

7.3. *Stability of accessibility*

Accessibility allows one to salvage the Hopf argument for ergodicity. Since we are also interested in stable ergodicity, it is natural to begin by looking at stable accessibility.

DEFINITION 7.11. A diffeomorphism f is said to be *stably accessible* if there exists a neighborhood $\mathcal{U}$ of f in the space $\mathrm{Diff}^1(M)$ (or in the space $\mathrm{Diff}^1(M, \nu)$ where ν is an f-invariant Borel probability measure) such that any diffeomorphism $g \in \mathcal{U}$ has the accessibility property.

7.3.1. *General theory* The study of stable accessibility is based on the *quadrilateral argument* first introduced by Brin [23]. Roughly speaking it goes as follows (we assume for simplicity that the central distribution E^c is integrable). Given a point $p \in M$, consider a 4-legged us-path $[z_0, z_1, z_2, z_3, z_4]$ originating at $z_0 = p$. We connect z_{i-1} with z_i by a geodesic γ_i lying in the corresponding stable or unstable manifold and we obtain the curve $\Gamma_p = \bigcup_{1 \leqslant i \leqslant 4} \gamma_i$. We parameterize it by $t \in [0, 1]$ with $\Gamma_p(0) = p$.

If the distribution $E^s \oplus E^u$ were integrable (and hence, the accessibility property for f would fail) the endpoint $z_4 = \Gamma_p(1)$ would lie on the leaf of the corresponding foliation passing through p. Therefore, one can hope to achieve accessibility if one can arrange a 4-legged us-path in such a way that $\Gamma_p(1) \in W^c(p)$ and $\Gamma_p(1) \neq p$. In this case the path Γ_p can be homotoped through 4-legged us-paths originating at p to the trivial path so that the endpoints stay in $W^c(p)$ during the homotopy and form a continuous curve. Such a situation is usually persistent under small perturbations of f and hence leads to stable accessibility.

We note that all current applications of stable accessibility are to dynamically coherent systems.

The first substantial result is that the accessibility property is C^1 generic in the space of partially hyperbolic diffeomorphisms, volume-preserving or not.

THEOREM 7.12 (Dolgopyat and Wilkinson [44]). *Let $q \geqslant 1$, $f \in \mathrm{Diff}^q(M)$ (or $f \in \mathrm{Diff}^q(M, \nu)$, where ν is a smooth invariant measure on M) be partially hyperbolic. Then for every neighborhood $\mathcal{U} \subset \mathrm{Diff}^1(M)$ (respectively, $\mathcal{U} \subset \mathrm{Diff}^1(M, \nu)$) of f there exists a C^q diffeomorphism $g \in \mathcal{U}$ that is* stably *accessible.*

An outline of the proof of this theorem in the special case when the central distribution E^c is 1-dimensional and integrable can be found in [71].

In the special case when the partially hyperbolic diffeomorphism has 1-dimensional center bundle, accessibility can be shown to be an open dense property in the space of diffeomorphisms of class C^2 (see [41]).

7.3.2. *Results in special cases* Theorem 7.12 can be improved in some special cases. In the remainder of this subsection we consider skew products over Anosov diffeomorphisms satisfying (2.12), time-t maps of suspension flows and group extensions over Anosov diffeomorphisms. These systems are partially hyperbolic and hence so are small perturbations. Their central distribution is integrable and the corresponding central foliation has compact smooth leaves. The proofs of accessibility exploit various versions of Brin's quadrilateral argument, and outlines can be found in [71].

7.3.3. *Skew products over Anosov diffeomorphisms* In the context of Section 2.2.4 we get

THEOREM 7.13 (Niţică and Török [66]). *If M is a connected manifold then there is a neighborhood of F in $\mathrm{Diff}^q(M \times S^1)$ or $\mathrm{Diff}^q(M \times S^1, \nu \times m)$ in which stable accessibility is open and dense.*

7.3.4. *Special flows*

THEOREM 7.14 (Brin [21], Talitskaya [90], [71]). *Let T_t be the special flow (see* [3, Sections 1.3j, 2.2j, 5.2j, 6.5d]*) over a C^q Anosov diffeomorphism with roof function $H : M \to \mathbb{R}^+$. There exists an open and dense set $\mathcal{U}$ of C^q functions $H : M \to \mathbb{R}^+$ such that the special flow T_t is stably accessible.*

7.3.5. *Group extensions* Let G be a compact connected Lie group, $f : M \to M$ a C^q Anosov diffeomorphism, and $\varphi : M \to G$ a C^q function. Consider the G-extension

$$F = F_\varphi : M \times G \to M \times G, \quad F_\varphi(x, y) = \big(f(x), \varphi(x)y\big)$$

of f. See Section 2.2.

THEOREM 7.15 (Brin [23], Burns and Wilkinson [36]). *For every neighborhood $\mathcal{U} \subset C^q(M, G)$ of φ there is a $\psi \in \mathcal{U}$ such that F_ψ is stably accessible. In other words, stably accessible group extensions are dense in the space of C^q group extensions over the Anosov diffeomorphism f.*

7.3.6. *Time-t maps of an Anosov flow* Let φ_t be an Anosov flow on a compact smooth Riemannian manifold M. It turns out that stable accessibility of the time-1 diffeomorphism depends on whether the distribution $E^s \oplus E^u$ is integrable, i.e., whether the stable and unstable foliations, W^s and W^u, of the time-1 map are jointly integrable. First, let us comment on joint integrability.

Fix $\varepsilon > 0$. Given a point $x \in M$, consider a local smooth submanifold

$$\Pi(x) = \bigcup_{y \in B^u(x,\varepsilon)} \bigcup_{-\varepsilon \leqslant \tau \leqslant \varepsilon} \varphi_\tau(y)$$

through x. For $x, x' \in M$ let $\pi_{x,x'} : \Pi(x) \to \Pi(x')$ be the holonomy map generated by the family of local stable manifolds. The foliations W^s and W^u are *jointly integrable* if for every $y \in \Pi(x)$ the image of the local unstable leaf $V^u(y)$ under $\pi_{x,x'}$ is the local unstable leaf $V^u(\pi_{x,x'}(y))$.

THEOREM 7.16 (Burns, Pugh and Wilkinson [35]). *Assume the stable and unstable foliations of the flow are not jointly integrable. Then the time-1 map φ_1 is stably accessible.*

By verifying the hypotheses of Theorem 7.16 one can establish stable accessibility of the time-1 map for

(1) geodesic flows on negatively curved manifolds (more generally, contact flows; Katok and Kononenko [63]);
(2) C^2 volume-preserving flows on compact 3-manifolds that are not special flows with a constant height function (Burns, Pugh and Wilkinson [35]).

We close this section with two conjectures about accessibility.

CONJECTURE 7.17. *A partially hyperbolic dynamical system with the accessibility property is stably accessible.*

This conjecture fails if one replaces accessibility by essential accessibility due to an example by Brin [34].

CONJECTURE 7.18 [76]. *The space of stably accessible partially hyperbolic dynamical systems is open and dense in the C^r topology for any $r \geqslant 1$. (This is known for $r = 1$ by* [44].)

8. The Pugh–Shub ergodicity theory

8.1. *Conditions for ergodicity*

Let f be a C^2 diffeomorphism of a smooth compact Riemannian manifold M that is partially hyperbolic and that preserves a smooth measure ν. To study ergodicity of f one uses a version of the Hopf argument [2,62,30,71] adapted to the case of partially hyperbolic systems.

Let $\mathcal{B}$ be the Borel σ-algebra of M. Say that $x, y \in M$ are stably equivalent if

$$\rho\big(f^n(x), f^n(y)\big) \to 0 \quad \text{as } n \to +\infty,$$

and unstably equivalent if

$$\rho\big(f^n(x), f^n(y)\big) \to 0 \quad \text{as } n \to -\infty.$$

Stable and unstable equivalence classes induce two partitions of M, and we denote by $\mathcal{S}$ and $\mathcal{U}$ the Borel σ-algebras they generate. Recall that for an algebra $\mathcal{A} \subset \mathcal{B}$ its saturated algebra is the set

$$\operatorname{Sat}(\mathcal{A}) = \big\{B \in \mathcal{B}\colon \text{ there exists } A \in \mathcal{A} \text{ with } \nu(A \triangle B) = 0\big\}.$$

It follows from the Hopf argument that f is ergodic if

$$\operatorname{Sat}(\mathcal{S}) \cap \operatorname{Sat}(\mathcal{U}) = \mathcal{T}, \tag{8.1}$$

where $\mathcal{T}$ is the trivial algebra.

For an Anosov diffeomorphism f the stable equivalence class containing a point x is the leaf $W^s(x)$ of the stable foliation. Similarly, the unstable equivalence class containing x is the leaf $W^u(x)$ of the unstable foliation. The σ-algebra $\mathcal{S}$ consists of those Borel sets S for which $W^s(x) \subset S$ whenever $x \in S$, and the σ-algebra $\mathcal{U}$ consists of those Borel sets U for which $W^u(x) \subset U$ whenever $x \in U$. The relation (8.1) holds by absolute continuity of stable and unstable foliations, which proves ergodicity for Anosov diffeomorphisms.

If a diffeomorphism f is partially hyperbolic the stable and unstable foliations W^s and W^u of M also generate Borel σ-algebras $\mathcal{M}^s$ and $\mathcal{M}^u$, respectively, so $\mathcal{S} \subset \mathcal{M}^s$ and $\mathcal{U} \subset \mathcal{M}^u$ (note that stable and unstable sets containing a point x may be larger than $W^s(x)$ and $W^u(x)$ due to possible contractions and expansions along the central directions). It follows that

$$\operatorname{Sat}(\mathcal{S}) \cap \operatorname{Sat}(\mathcal{U}) \subset \operatorname{Sat}(\mathcal{M}^s) \cap \operatorname{Sat}(\mathcal{M}^u).$$

If f is accessible then $\operatorname{Sat}(\mathcal{M}^s \cap \mathcal{M}^u) = \mathcal{T}$. In fact, essential accessibility (Definition 7.4) is enough. If f is essentially accessible then ergodicity would follow from

$$\operatorname{Sat}(\mathcal{M}^s) \cap \operatorname{Sat}(\mathcal{M}^u) \subset \operatorname{Sat}(\mathcal{M}^s \cap \mathcal{M}^u) \tag{8.2}$$

(the opposite inclusion is obvious). We describe conditions that guarantee this.

THEOREM 8.1. *A volume-preserving essentially accessible dynamically coherent partially hyperbolic diffeomorphism is ergodic if it has Lipschitz continuous center foliation and Lipschitz continuous stable and unstable holonomy maps between center transversals.*

PROOF. (8.2) follows from the conditions of the theorem. □

The assumptions on Lipschitz continuity are very strong and "typically" fail (see Section 6.2). On one hand, the modern work in the field has found ways to circumvent the requirement that the center foliation be Lipschitz continuous, and on the other hand, in the presence of dynamical coherence, Lipschitz continuity of the holonomies between center transversals is obtained from the following condition.

DEFINITION 8.2 [37]. We say that f is *center-bunched* if $\max\{\mu_1, \lambda_3^{-1}\} < \lambda_2/\mu_2$ in (2.7).

This definition due to Burns and Wilkinson imposes a much weaker constraint than earlier versions; in fact, their results assume an even weaker condition one might call "pointwise center bunching": $\max\{\mu_1(p), \lambda_3^{-1}(p)\} < \lambda_2(p)/\mu_2(p)$ for every point p, where $\mu_i(p)$ and $\lambda_i(p)$ are pointwise bounds on rates of expansion and contraction. This pointwise condition always holds when $\dim E^c = 1$, and they show in [37] that this assumption suffices to get the following.

THEOREM 8.3. *A C^2 volume-preserving partially hyperbolic essentially accessible* (*dynamically coherent*) *center-bunched diffeomorphism is ergodic.*

Grayson, Pugh and Shub [49] proved this theorem for small perturbations of the time one map of the geodesic flow on a surface of constant negative curvature. Wilkinson in her thesis extended their result to small perturbations of the time-1 map of the geodesic flow on an arbitrary surface of negative curvature. Then Pugh and Shub in [77,78] proved the theorem assuming a stronger center bunching condition. The proof of the theorem in the form stated here (with a weaker center bunching condition) was obtained by Burns and Wilkinson in [37].

REMARK 8.4. Burns and Wilkinson recently announced that the assumption of dynamical coherence is not needed in Theorem 8.3 [38].

Together with the comments on Definition 8.2 this in particular gives the following.

COROLLARY 8.5. *A C^2 volume-preserving essentially accessible partially hyperbolic diffeomorphism with* $\dim E^c = 1$ *is ergodic.*

This corollary was also announced recently by F. Rodriguez Hertz, J. Rodriguez Hertz and R. Ures [82].

The way to establish (8.2) without absolute continuity of the center-stable and center-unstable foliation is through the use of a collection of special sets at every point $x \in M$ called *Juliennes*, $J_n(x)$.[5] We shall describe a construction of these sets which assures that

(J1) $J_n(x)$ form a basis of the topology.

(J2) $J_n(x)$ form a basis of the Borel σ-algebra. More precisely, let Z be a Borel set; a point $x \in Z$ is said to be *Julienne dense* if

$$\lim_{n\to+\infty} \frac{\nu(J_n(x) \cap Z)}{\nu(J_n(x))} = 1.$$

Let $D(Z)$ be the set of all Julienne dense points of Z. Then

$$D(Z) = Z \pmod 0.$$

(J3) If $Z \in \mathrm{Sat}(\mathcal{M}^s) \cap \mathrm{Sat}(\mathcal{M}^u)$, then $\mathcal{D}(Z) \in \mathrm{Sat}(\mathcal{M}^s \cap \mathcal{M}^u)$.

Properties (J1)–(J3) imply (8.2).

Note that the collection of balls $B(x, 1/n)$ satisfies requirements (J1) and (J2) but not (J3). Juliennes can be viewed as balls "distorted" by the dynamics in the following sense. Fix an integer $n \geqslant 0$, a point $x \in M$ and numbers τ, σ such that $0 < \tau < \sigma < 1$. Denote by

$$B_n^s(x,\tau) = \big\{y \in W^s(x) \mid \rho\big(f^{-k}(x), f^{-k}(y)\big) \leqslant \tau^k\big\},$$
$$B_n^u(x,\tau) = \big\{y \in W^u(x) \mid \rho\big(f^{k}(x), f^{k}(y)\big) \leqslant \tau^k\big\},$$

and define the Julienne

$$J_n(x) := \big[J_n^{cs}(x) \times B_n^u(x,\tau)\big] \cap \big[B_n^s(x,\tau) \times J_n^{cu}(x)\big],$$

where the local foliation products

$$J_n^{cs}(x) = B_n^s(x,\tau) \times B^c(x,\sigma^n), \qquad J_n^{cu}(x) = B_n^u(x,\tau) \times B^c(x,\sigma^n)$$

are the *center-stable* and *center-unstable Juliennes*, and $B^c(x,\sigma^n)$ is the ball in $W^c(x)$ centered at x of radius σ^n. One may think of $J_n(x)$ as a substitute for $B_n^s(x,\tau) \times B^c(x,\sigma^n) \times B_n^u(x,\tau)$, which is only well defined if the stable and unstable foliations are jointly integrable.

The proof of (J1)–(J3) is based on the following properties of Juliennes:

(1) *scaling*: if $k \geqslant 0$ then $\nu(J_n(x))/\nu(J_{n+k}(x))$ is bounded, uniformly in $n \in \mathbb{N}$;

(2) *engulfing*: there is $\ell \geqslant 0$ such that, for any $x, y \in M$, if $J_{n+\ell}(x) \cap J_{n+\ell}(y) \neq \emptyset$ then $J_{n+\ell}(x) \cup J_{n+\ell}(y) \subset J_n(x)$;

(3) *quasi-conformality*: there is $k \geqslant 0$ such that if $x, y \in M$ are connected by an arc on an unstable manifold that has length $\leqslant 1$ then the holonomy map $\pi : V^{cs}(x) \to V^{cs}(y)$ generated by the family of local unstable manifolds (see Section 3.2) satisfies $J_{n+k}^{cs}(y) \subset \pi(J_n^{cs}(x)) \subset J_{n-k}^{cs}(y)$.

[5] They resemble slivered vegetables as used in consommé Julienne, said to be attributed to the chef Jean Julien in 1722 by François Massialot (Le nouveau cuisinier royal et bourgeois ou cuisine moderne, reprint 2003 by Eibron Classics).

The properties (1) and (2) are possessed by the family of balls in Euclidean space and they underlie the proof of the Lebesgue Density Theorem. One can use these properties to show that Juliennes are density bases. The center-unstable Juliennes are a density basis on $W^{cu}(x)$ with respect to the smooth conditional measure $\nu_{W^{cu}}$ on $W^{cu}(x)$, the center-stable Juliennes are a density basis on $W^{cs}(x)$ with respect to the smooth conditional measure $\nu_{W^{cs}}$ on $W^{cs}(x)$, and the Juliennes are a density basis on M with respect to the smooth measure ν.

Juliennes, $J_n(x)$, are small but highly eccentric sets in the sense that the ratio of their diameter to their inner diameter increases with n (the inner diameter of a set is the diameter of the largest ball it contains). In general, sets of such shape may not form density bases, but Juliennes do because their elongation and eccentricity are controlled by the dynamics; in particular, they nest in a way similar to balls.

Quasi-conformality is what is needed to prove Property (J3). Roughly speaking it means that the holonomy map (almost) preserves the shape of Juliennes.

CONJECTURE 8.6. *A partially hyperbolic dynamical system preserving a smooth measure and with the accessibility property is ergodic.*

8.2. *The Pugh–Shub stable ergodicity theorem*

DEFINITION 8.7. Let $q \geqslant 1$. A C^q diffeomorphism f of a compact C^q Riemannian manifold M preserving a smooth measure ν is said to be *stably ergodic* if any C^1-small perturbation of f preserving ν is ergodic.

Stable ergodicity imposes some conditions on the map. Bochi, Fayad and Pujals [14] showed that there is an open and dense set of $C^{1+\alpha}$ stably ergodic (with respect to a smooth measure) diffeomorphisms with nonzero Lyapunov exponents (this answers a problem posed in [31]).

A stably ergodic diffeomorphism f need not be partially hyperbolic [89]. However, it possesses a *dominated splitting*, i.e., the tangent space splits into two invariant subspaces E and F such that for $n \in \mathbb{N}$,

$$\left\|df^n \restriction E(x)\right\| \left\|df^{-n} \restriction F\left(f^n(x)\right)\right\| \leqslant C\lambda^n$$

with uniform $C > 0$ and $0 < \lambda < 1$. This was proved by Arbieto and Matheus assuming that f is $C^{1+\epsilon}$ and volume preserving [11]. On the other hand if f is a symplectic stably ergodic diffeomorphism then f must be partially hyperbolic [58].

To establish stable ergodicity of a partially hyperbolic diffeomorphism f one can check whether the hypotheses of Theorem 8.3 are stable under small perturbations.

(1) If f is dynamically coherent and the central foliation W^c is of class C^1 then every diffeomorphism g which is sufficiently close to f in the C^1 topology is dynamically coherent (see Theorems 4.11 and 4.15).
(2) If f is center-bunched then every diffeomorphism g which is sufficiently close to f in the C^1 topology is center-bunched.

Thus, for dynamically coherent center-bunched partially hyperbolic diffeomorphisms, stable ergodicity follows from stable accessibility:

THEOREM 8.8 [78,79]. *A* (*dynamically coherent*) *center-bunched stably* (*essentially*) *accessible partially hyperbolic diffeomorphism that preserves a smooth measure ν and has a smooth or plaque-expansive center foliation is stably ergodic* (*and stably K* [62, Sections 3.6k, 3.7j], [36, Corollary 1.2]).

REMARK 8.9. As noted in Remark 8.4, Burns and Wilkinson recently announced that the assumption of dynamical coherence is not needed in this result [38].

CONJECTURE 8.10 [76]. *A partially hyperbolic dynamical system preserving a smooth measure and with the accessibility property is stably ergodic.* (*This would follow from Conjectures* 7.17 *and* 8.6.)

A proof of this conjecture in the case when the central distribution is one-dimensional was recently announced by F. Rodriguez Hertz, J. Rodriguez Hertz and R. Ures [82].

CONJECTURE 8.11 [76]. *Stably ergodic diffeomorphisms are open and dense in the space of C^r partially hyperbolic dynamical systems for $r \geqslant 1$.* (*This would follow from Conjectures* 7.18 *and* 8.6.)

Combining Theorem 8.8 with the results in Section 7.3 we obtain several classes of stably ergodic systems:

8.2.1. *Skew product maps over Anosov diffeomorphisms* If $F = f \times \mathrm{Id} : M \times S^1 \to M \times S^1$ then there is a neighborhood $\mathcal{U}$ of F in $\mathrm{Diff}^2(M \times S^1)$ or $\mathrm{Diff}^2(M \times S^1, \nu \times m)$ such that stable ergodicity is open and dense in $\mathcal{U}$ (here m is the length).

8.2.2. *Special flows over Anosov diffeomorphisms* There exists an open and dense set of C^q functions $H : M \to \mathbb{R}^+$ such that the special flow T_t with the roof function H is stably ergodic. Field, Melbourne and Török [46] strengthened this result.

THEOREM 8.12. *For $r > 0$, there exists a C^r-open and dense subset $\mathcal{A}$ in the space of strictly positive C^r* (*roof*) *functions such that for every $H \in \mathcal{A}$ the special flow T_t with the roof function H is stably mixing. If $r \geqslant 2$ then $\mathcal{A}$ is open in the C^2 topology and C^∞ roof functions are $C^{[r]}$-dense in $\mathcal{A}$.*

8.2.3. *Group extensions over Anosov diffeomorphisms* If $F_\varphi : M \times G \to M \times G$ is a group extension then for every neighborhood $\mathcal{U} \subset C^q(M, G)$ of the function φ there exists a function $\psi \in \mathcal{U}$ such that the diffeomorphism F_ψ is stably ergodic.

Burns and Wilkinson obtained a complete characterization of stable ergodicity for group extensions over volume-preserving Anosov diffeomorphisms. Namely, we say that the map $h : M \times Y \to M \times Y$ of class $C^{q+\alpha}$ is an *algebraic factor* of the $C^{q+\alpha}$ group extension F_φ

if $Y = H \setminus G$, where H is a closed subgroup of G, and there exists a $C^{q+\alpha}$ function $\Phi : M \to G/H$ for which the following diagram is commutative:

$$\begin{array}{ccc} M \times G & \xrightarrow{F_\varphi} & M \times G \\ {\scriptstyle \pi_\Phi}\downarrow & & \downarrow{\scriptstyle \pi_\Phi} \\ M \times H \setminus G & \xrightarrow{h} & M \times H \setminus G \end{array}$$

where $\pi_\Phi(x, g) = (x, \Phi(x)^{-1}g)$ (and $\Phi(x)^{-1} = \{g^{-1} \mid g \in \Phi(x)\}$ is an element of $H \setminus G$).

THEOREM 8.13. *Let $f : M \to M$ be a $C^{q+\alpha}$ volume-preserving Anosov diffeomorphism of an infranilmanifold, G a compact, connected Lie group and $\varphi : M \to G$ a $C^{q+\alpha}$ map. If the group extension F_φ is not stably ergodic then it has an algebraic factor $h : M \times H \setminus G \to M \times H \setminus G$, where one of the following holds*:

(1) *$H \neq G$, and h is the product of f with* $\mathrm{Id}_{H\setminus G}$;
(2) *h is normal, $H \setminus G$ is a circle, and h is the product of f with a rotation*;
(3) *h is normal, $H \setminus G$ is a d-torus, and $h = f_\psi$ where ψ is homotopic to a constant and maps M into a coset of a lower-dimensional Lie subgroup of the d-torus.*

If F_φ has an algebraic factor of type (1), *it is not ergodic*; *if F_φ has an algebraic factor of type* (2) *but none of type* (1) *then it is ergodic, but not weakly mixing*; *otherwise F_φ is Bernoulli. In addition, F_φ is stably ergodic if and only if it is stably ergodic within skew products.*

Applying this result to the case when the group G is semisimple, one can show that F_φ is stably ergodic if and only if it is ergodic.

Field, Melbourne and Török [46] studied stable ergodicity of group extensions over hyperbolic sets and generalized earlier results in [6,47,68,91]. Let f be a C^2 diffeomorphism of a compact smooth manifold M possessing a locally maximal hyperbolic set Λ which is not a periodic orbit. Let μ be the unique equilibrium measure on Λ corresponding to a Hölder continuous potential (so $f|\Lambda$ is ergodic with respect to μ). Consider a compact connected Lie group G with the Haar measure m.

THEOREM 8.14. *For $r > 0$ there exists a C^r-open and dense subset $\mathcal{A} \subset C^r(M, G)$ such that for every $\varphi \in \mathcal{A}$ the group extension F_φ is ergodic with respect to the measure $\nu = \mu \times m$. If $f|\Lambda$ is topologically mixing then F_φ is mixing with respect to ν.*

In other words, if $f|\Lambda$ is topologically transitive (respectively, topologically mixing) then the stably ergodic (respectively, stably mixing) group extensions form an open and dense set in the space of C^r group extensions for any $r > 0$.

8.2.4. *Time-t maps of Anosov flows* If the stable and unstable foliations of an Anosov flow are not jointly integrable then the time-t map for $t \neq 0$ is stably ergodic.

Theorem 7.16 provides a strong dichotomy between joint integrability of the strong foliations and (stable) accessibility. The paper in which this theorem is proved also produces a clean dichotomy between joint integrability and (stable) ergodicity:

THEOREM 8.15 (Burns, Pugh and Wilkinson [35]). *The time-*1 *map of a volume-preserving Anosov flow is stably ergodic unless the strong stable and strong unstable foliations for the flow are jointly integrable.*

The proof follows the line of argument in [49] and uses the crucial fact that the holonomy maps are not just continuous but indeed Hölder continuous with Hölder exponent close to 1 (see [80]) so that these maps do not distort Juliennes too much.

In the special case of flows on 3-manifolds one can strengthen this result and show that *the time-*1 *map is stably ergodic if and only if the flow is not a suspension flow over an Anosov diffeomorphism with a constant roof function.* In particular, *the time-*1 *map of a volume-preserving topologically mixing C^2 Anosov flow is stably ergodic*. As a corollary one has that the time-1 map of geodesic flows on a closed negatively curved Riemannian surface is stably ergodic (this result was earlier obtained by Wilkinson [92]).

8.2.5. *Frame flows* There are also several cases in which the frame flow and its time-t maps are known to be ergodic [33]:

THEOREM 8.16. *Let Φ_t be the frame flow on an n-dimensional compact smooth Riemannian manifold with sectional curvatures between $-\Lambda^2$ and $-\lambda^2$. Then in each of the following cases the flow is ergodic, K* ([62, Sections 3.6k, 3.7j], [5, Section 4.3]), *and even Bernoulli* [5, Sections 6–7], *and the time-*1 *map of the frame flow is stably ergodic and stably K*:

(1) *if the curvature is constant* (*Brin* [22]);
(2) *for a set of metrics of negative curvature which is open and dense in the C^3 topology* (*Brin* [22]);
(3) *if n is odd and $n \neq 7$* (*Brin and Gromov* [27]);
(4) *if n is even, $n \neq 8$, and $\lambda/\Lambda > 0.93$* (*Brin and Karcher* [28]);
(5) *if $n = 7$ and $\lambda/\Lambda > 0.99023\ldots$* (*Burns and Pollicott* [33]);
(6) *if $n = 8$ and $\lambda/\Lambda > 0.99023\ldots$* (*Burns and Pollicott* [33]).

Ergodicity of the frame flow was proved by the authors cited in each case; [33] pointed out the K and Bernoulli property and used [36, Corollary 1.2] (which relies on [29]) to deduce those of the time-1 maps across all cases.

8.3. *Ergodicity and stable ergodicity for toral automorphisms*

Theorem 8.16 has as a particular consequence that the time-1 map of the frame flow of a manifold with negative curvature is stably ergodic in all cases where it is known to be ergodic. At the end of [56] Hirsch, Pugh and Shub posed a question that might be interpreted as asking whether every ergodic automorphism of the n-torus is stably ergodic. In

this context ergodicity is easily characterized by the property that the automorphism has no eigenvalue that is a root of unity [51].

The dissertation of Rodriguez Hertz [81] answers the question in the affirmative for dimension up to 5:

THEOREM 8.17 (Rodriguez Hertz [81]). *Every ergodic linear automorphism of a torus of dimension up to* 5 *is stably ergodic*. (*But for dimension* 4 *only with respect to* C^{22} *perturbations*.)

This result arises, in fact, as a consequence of rather more general ones.

DEFINITION 8.18. An automorphism of $\mathbb{T}^n$ none of whose eigenvalues is a root of the unity and whose characteristic polynomial is irreducible over the integers and not a polynomial in t^i for any $i \geqslant 2$ is said to be a pseudo-Anosov automorphism.

THEOREM 8.19 (Rodriguez Hertz [81]). *If* $n \geqslant 6$ *then any pseudo-Anosov automorphism of* $\mathbb{T}^n$ *with* $\dim E^c = 2$ *is stably ergodic with respect to the* C^5 *topology, and if* $n = 4$ *then any pseudo-Anosov automorphism of* $\mathbb{T}^n$ *is stably ergodic with respect to the* C^{22} *topology.*

Rodriguez Hertz derives Theorem 8.17 from Theorem 8.19 by showing that ergodic automorphisms of $\mathbb{T}^4$ are either Anosov or pseudo-Anosov and ergodic automorphisms of $\mathbb{T}^5$ are Anosov (and hence clearly stably ergodic). In fact, the odd-dimensional case is much simplified by his observation that if n is odd and $A \in SL(n, \mathbb{Z})$ has irreducible characteristic polynomial then A is Anosov. The remaining substance of the work therefore lies in the cases $n = 4$ and $n \geqslant 6$, in each of which Rodriguez Hertz studies a dichotomy concerning accessibility. He considers the accessibility classes (lifted to the universal cover) for such a perturbation and shows that these are either all trivial (they intersect each stable leaf in a point) or else must all be equal to $\mathbb{R}^n$ [81, Theorem 5.1]. The latter implies accessibility, and in the former case an application of KAM-theory (or, for $n = 4$, a separate theorem of Moser) then establishes smooth conjugacy of the foliations to those of the linear system, which yields essential accessibility. Rodriguez Hertz can then apply Theorem 8.3 in either case.

His result prompted Pugh and Shub to make their earlier question more explicit in the following form:

PROBLEM [79]. Is every ergodic toral automorphism stably ergodic in the C^r topology for some r?

9. Partially hyperbolic attractors

A partially hyperbolic set Λ for a diffeomorphism f of a compact manifold M is called a *partially hyperbolic attractor* if there is a neighborhood U of Λ such that $\overline{f(U)} \subset U$ and

$$\Lambda = \bigcap_{n \in \mathbb{Z}} f^n(U).$$

An important property of a partially hyperbolic attractor is as follows.

THEOREM 9.1. *$W^u(x) \subset \Lambda$ for every $x \in \Lambda$.*

Any diffeomorphism g sufficiently close to f possesses a partially hyperbolic attractor which lies in a small neighborhood of Λ.

An invariant Borel probability measure μ on Λ is said to be a *u-measure* if the conditional measures $\mu^u(x)$ generated by μ on local unstable leaves $V^u(x)$ are absolutely continuous with respect to the Riemannian volume on $V^u(x)$.

Consider a smooth measure ν on U with the density function ψ with respect to the Riemannian volume m, i.e.,

$$\operatorname{supp} \psi \subset U, \qquad \int_U \psi \, dm = 1.$$

The sequence of measures

$$\nu_n = \frac{1}{n} \sum_{i=0}^{n-1} f_*^i \nu$$

is the evolution of the measure ν under the system f. Even if the sequence ν_n does not converge, any limit measure μ is supported on Λ.

THEOREM 9.2 (Pesin and Sinai [72]). *Any limit measure of the sequence of measures ν_n is an f-invariant u-measure on Λ.*

We describe another approach for constructing u-measures on Λ. For $x \in \Lambda$ and $y \in V^u(x)$ consider the function

$$\kappa(x, y) = \prod_{i=0}^{n-1} \frac{J(df \restriction E^u(f^i(y)))}{J(df \restriction E^u(f^i(x)))}.$$

Define the probability measure $\tilde{m}_n$ on $V_n(x) = f^n(V^u(x))$ by

$$d\tilde{m}_n(y) = c_n \kappa\big(f^n(x), y\big) \, dm_{V_n(x)} \quad \text{for } y \in V_n(x),$$

$\tilde{m}_n$ on $V_n^u(x) = f^n(V^u(x))$ by where c_n is normalizing factor and $m_{V_n(x)}$ is the Riemannian volume on $V_n(x)$ induced by the Riemannian metric. We define the Borel measure m_n on Λ by

$$m_n(A) = \tilde{m}_n\big(A \cap V_n(x)\big),$$

when $A \subset \Lambda$ is a Borel set. One can show that $m_n(A) = m_0(f^{-n}(A))$.

THEOREM 9.3 (Pesin and Sinai [72]). *Any limit measure of the sequence of measures m_n is an f-invariant u-measure on Λ.*

While Theorem 9.2 describes u-measures as a result of the evolution of an absolutely continuous measure in a neighborhood of the attractor, Theorem 9.3 determines u-measures as limit measures for the evolution of an absolutely continuous measure supported on a local unstable manifold. One can deduce Theorem 9.2 from Theorem 9.3 and the proof of the latter, presented in [72], exploits a method which allows one to avoid the use of Markov partitions—the classical tool to prove existence of SRB-measures for hyperbolic attractor (in general, a partially hyperbolic attractor does not have any Markov partition).

Assume now that the unstable distribution E^u splits into the sum of two invariant subdistributions $E^u = E_1 \oplus E_2$ with E_1 expanding more rapidly than E_2. One can view f as a partially hyperbolic diffeomorphism with E_1 as the new unstable distribution (and $E_2 \oplus E^c$ as the new center distribution) and construct u-measures, associated with this distribution, according to Theorem 9.3.

THEOREM 9.4. *Any u-measure associated with the distribution E^u is a u-measure associated with the distribution E_1.*

The proof of this theorem can be easily obtained from the fact that the leaves of the $W_1(y)$ depend smoothly on $y \in W^u(x)$, see Theorem 3.3.

If Λ is a hyperbolic attractor and $f \restriction \Lambda$ is topologically transitive then there exists a unique u-measure. This may not be true for a general partially hyperbolic attractor and some additional strong conditions are necessary to guarantee uniqueness.

Let us denote by $\chi(x, v)$ the Lyapunov exponent at the point $x \in \Lambda$ and the vector $v \in T_x M$.

Let ν be an invariant Borel probability measure on Λ. We say that Λ has *negative central exponents* with respect to ν if there exists a set $A \subset \Lambda$ of positive measure such that $\chi(x, v) < 0$ for every $x \in A$ and $v \in E^c(x)$.

THEOREM 9.5 (Burns, Dolgopyat and Pesin [31]). *If ν is a u-measure with negative central exponents then every ergodic component of $f \restriction A$ of positive measure is open* (mod 0).

Using this result one can provide conditions which guarantee uniqueness or that there is at most a finite number of u-measures.

THEOREM 9.6 (Burns, Dolgopyat and Pesin [31]). *Assume that there exists a u-measure ν with negative central exponents. Assume, in addition, that for almost every $x \in \Lambda$ the trajectory $\{f^n(x)\}$ is everywhere dense in Λ. Then $f \restriction \Lambda$ is ergodic with respect to ν.*

One can show that if for every $x \in \Lambda$ the global strongly unstable manifold $W^u(x)$ is dense then almost every orbit is dense. Moreover, under this assumption there is a unique u-measure which is also an SRB-measure for f.

THEOREM 9.7 (Bonatti and Viana [19]). *Let f be a C^2 diffeomorphism possessing a partially hyperbolic attractor Λ. Assume that for every $x \in \Lambda$ and every disk $D^u(x) \subset W^u(x)$ centered at x, we have that $\chi(y, v) < 0$ for a positive Lebesgue measure subset of points $y \in D^u$ and every vector $v \in E^c(y)$. Then f has at most finitely many u-measures.*

THEOREM 9.8 (Alves, Bonatti and Viana [8]). *Assume that f is nonuniformly expanding along the center-unstable direction, i.e.,*

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \log \|df^{-1} \restriction E^{cu}_{f^j(x)}\| < 0 \tag{9.1}$$

for all x in a positive Lebesgue measure set $A \subset M$. Then f has an ergodic SRB-measure supported in $\bigcap_{j=0}^{\infty} f^j(M)$. Moreover, if the limit in (9.1) *is bounded away from zero then A is contained* (mod 0) *in the union of the basins of finitely many SRB-measures.*

Let $\Lambda = \Lambda_f$ be a partially hyperbolic attractor for f. It is well known that any C^1 diffeomorphism g which is sufficiently close to f in the C^1 topology possesses a partially hyperbolic attractor Λ_g which lies in a small neighborhood of Λ_f. The following statement shows that u-measures depend continuously on the perturbation.

THEOREM 9.9 (Dolgopyat [42]). *Let f_n be a sequence of C^2 diffeomorphisms converging to a diffeomorphism f in the C^2 topology. Let also ν_n be a u-measure for f_n. Assume that the sequence of measures ν_n converges in the weak topology to a measure ν. Then ν is a u-measure for f.*

The following statement describes a version of stable ergodicity for partially hyperbolic attractors.

THEOREM 9.10. *Assume that there exist a u-measure $\nu = \nu_f$ for f with negative central exponents. Assume also that for every $x \in \Lambda_f$ the global strongly unstable manifold $W^u(x)$ is dense in Λ_f. Then any C^2 diffeomorphism g which is sufficiently close to f also has negative central exponents on a set that has positive measure with respect to a u-measure ν_g; this measure is the only SRB-measure for g and $g \restriction \Lambda_g$ is ergodic with respect to ν_g.*

The following statement provides conditions which guarantee uniqueness of u-measures.

THEOREM 9.11 (Bonatti and Viana [19]). *Let f be a C^2 diffeomorphism possessing a partially hyperbolic attractor Λ. Assume that*

(1) *there exist $x \in \Lambda$ and a disk $D^u(x) \subset W^u(x)$ centered at x for which $\chi(y, v) < 0$ for a positive Lebesgue measure subset of points $y \in D^u$ and every vector $v \in E^c(y)$;*
(2) *every leaf of the foliation W^u is dense in Λ.*

Then f has a unique u-measure and it is ergodic. The support of this measure coincides with Λ.

The measure ν in this theorem is a Sinai–Ruelle–Bowen (SRB) measure (for the definition and some relevant results on SRB-measures see [1, Section 14]).

Acknowledgements

It is a pleasure to thank Michael Brin, Keith Burns, Dmitry Dolgopyat, Marcelo Viana and Amie Wilkinson for significant help with the writing of this chapter. We should mention in particular that the presentation of the example in Section 6.1 was provided to us by Keith Burns and that Keith Burns and Amie Wilkinson generously shared recent work of theirs that clarifies and simplifies the assumptions needed for the Pugh–Shub ergodicity theorem. The work of Ya. Pesin is partially supported by the National Science Foundation, division of Mathematical Sciences.

References

Surveys in volume 1A and this volume

[1] L. Barreira and Ya. Pesin, *Smooth ergodic theory and nonuniformly hyperbolic dynamics*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 57–263.
[2] B. Hasselblatt, *Hyperbolic dynamical systems*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 239–319.
[3] B. Hasselblatt and A. Katok, *Principal structures*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 1–203.
[4] E. Pujals and M. Sambarino, *Homoclinic bifurcations, dominated splitting and robust transitivity*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 327–378.
[5] J.-P. Thouvenot, *Entropy, isomorphism and equivalence in ergodic theory*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 205–238.

Other sources

[6] R. Adler, B. Kitchens and M. Shub, *Stably ergodic skew products*, Discrete Contin. Dynam. Systems **2** (3) (1996), 349–350.
[7] V. Alekseev, *Quasirandom dynamical systems. I. Quasirandom diffeomorphisms*, Mat. Sb. (N.S.) **76** (1968), 72–134; *Invariant Markov subsets of diffeomorphisms*, Uspekhi Mat. Nauk **23** (2) (1968), 209–210.
[8] J.F. Alves, C. Bonatti and M. Viana, *SRB measures for partially hyperbolic systems whose central direction is mostly expanding*, Invent. Math. **140** (2) (2000), 351–398.
[9] D. Anosov, *Geodesic flows on closed Riemannian manifolds with negative curvature*, Proc. Steklov Inst. Math. **90** (1969), 1–235.
[10] D. Anosov and Y. Sinai, *Certain smooth ergodic systems*, Russian Math. Surveys **22** (1967), 103–167.
[11] A. Arbieto and C. Matheus, *A pasting lemma I: the case of vector fields*, Preprint IMPA (2003).
[12] A. Baraviera and C. Bonatti, *Removing zero Lyapunov exponents*, Ergodic Theory Dynam. Systems **23** (6) (2003), 1655–1670.
[13] L. Barreira and Y. Pesin, *Lyapunov Exponents and Smooth Ergodic Theory*, University Lecture Series, Vol. 23, Amer. Math. Soc., Providence, RI (2001).
[14] J. Bochi, B. Fayad and E. Pujals, *A remark on conservative diffeomorphisms*, Preprint (2004).
[15] C. Bonatti, L.J. Díaz and E.R. Pujals, *A C^1-generic dichotomy for diffeomorphisms: Weak forms of hyperbolicity or infinitely many sinks or sources*, Ann. of Math. (2) **158** (2) (2003), 355–418.

[16] C. Bonatti, L.J. Díaz and M. Viana, *Dynamics Beyond Uniform Hyperbolicity: A Global Geometric and Probabilistic Perspective*, Encyclopedia Math. Sciences, Vol. 102, Springer-Verlag (2004).
[17] C. Bonatti and J. Franks, *A Hölder continuous vector field tangent to many foliations*, Modern Dynamical Systems and Applications, B. Hasselblatt, M. Brin and Y. Pesin, eds, Cambridge Univ. Press, New York (2004), 299–306.
[18] C. Bonatti, C. Matheus, M. Viana and A. Wilkinson, *Abundance of stable ergodicity*, Comment. Math. Helv. **79** (4) (2004), 753–757.
[19] C. Bonatti and M. Viana, *SRB measures for partially hyperbolic systems whose central direction is mostly contracting*, Israel J. Math. **115** (2000), 157–193.
[20] C. Bonatti and A. Wilkinson, *Transitive partially hyperbolic diffeomorphisms on* 3*-manifolds*, Topology **44** (3) (2005), 475–508.
[21] M. Brin, *Partial hyperbolicity*, Ph.D. thesis, Moscow (1975).
[22] M. Brin, *Topological transitivity of one class of dynamical systems and flows of frames on manifolds of negative curvature*, Funct. Anal. Appl. **9** (1975), 9–19.
[23] M. Brin, *The topology of group extensions of C systems*, Mat. Zametki **18** (1975), 453–465.
[24] M. Brin, *Hölder continuity of invariant distributions*, Smooth Ergodic Theory and Its Applications, A. Katok, R. de la Llave, Y. Pesin and H. Weiss, eds, Proc. Sympos. Pure Math., Amer. Math. Soc. (2001).
[25] M. Brin, *On dynamical coherence*, Ergodic Theory Dynam. Systems **23** (2) (2003), 395–401.
[26] M. Brin, D. Burago and S. Ivanov, *On partially hyperbolic diffeomorphisms of* 3*-manifolds with commutative fundamental group*, Modern Dynamical Systems and Applications, B. Hasselblatt, M. Brin and Y. Pesin, eds, Cambridge Univ. Press, New York (2004), 307–312.
[27] M. Brin and M. Gromov, *On the ergodicity of frame flows*, Invent. Math. **60** (1980), 1–7.
[28] M. Brin and H. Karcher, *Frame flows on manifolds of negative curvature*, Compos. Math. **52** (1984), 275–297.
[29] M. Brin and Y. Pesin, *Partially hyperbolic dynamical systems*, Izv. Akad. Nauk SSSR Ser. Mat. **38** (1974), 170–212.
[30] M. Brin and G. Stuck, *Dynamical Systems*, Cambridge Univ. Press (2002).
[31] K. Burns, D. Dolgopyat and Y. Pesin, *Partial hyperbolicity, Lyapunov exponents and stable ergodicity*, J. Stat. Phys. **109** (2002), 927–942.
[32] K. Burns and L. Flaminio, *Fubini's nightmare*, Unpublished note (1992).
[33] K. Burns and M. Pollicott, *Stable ergodicity and frame flows*, Geom. Dedicata **98** (2003), 189–210.
[34] K. Burns, C. Pugh, M. Shub and A. Wilkinson, *Recent results about stable ergodicity*, Smooth Ergodic Theory and Its Applications, A. Katok, R. de la Llave, Y. Pesin and H. Weiss, eds, Proc. Sympos. Pure Math., Amer. Math. Soc. (2001).
[35] K. Burns, C. Pugh and A. Wilkinson, *Stable ergodicity and Anosov flows*, Topology **39** (2000), 149–159.
[36] K. Burns and A. Wilkinson, *Stable ergodicity of skew products*, Ann. Sci. École Norm. Sup. **32** (1999), 859–889.
[37] K. Burns and A. Wilkinson, *Better center bunching*, http://www.math.northwestern.edu/∼wilkinso/papers/bunch0221.pdf.
[38] K. Burns and A. Wilkinson, *On the ergodicity of partially hyperbolic systems*, http://www.math.northwestern.edu/∼wilkinso/papers/hopf0627.pdf, Preprint.
[39] R. de la Llave, *Smooth conjugacy and S-R-B measures for uniformly and nonuniformly hyperbolic systems*, Comm. Math. Phys. **150** (1992), 289–320.
[40] L.J. Díaz, E.R. Pujals and R. Ures, *Partial hyperbolicity and robust transitivity*, Acta Math. **183** (1) (1999), 1–43.
[41] P. Didier, *Stability of accessibility*, Ergodic Theory Dynam. Systems **23** (6) (2003), 1717–1731.
[42] D. Dolgopyat, *On dynamics of partially hyperbolic systems on three manifolds*, Preprint (1999).
[43] D. Dolgopyat and Y. Pesin, *Every compact manifold carries a completely hyperbolic diffeomorphism*, Ergodic Theory Dynam. Systems **22** (2002), 1–27.
[44] D. Dolgopyat and A. Wilkinson, *Stable accessibility is* C^1 *dense*, Geometric Methods in Dynamics. II, Astérisque, Vol. 287 (2003), xvii, 33–60.
[45] N. Fenichel, *Persistence and smoothness of invariant manifolds of slows*, Indiana Univ. Math. J. **21** (1971–1972), 193–226.

[46] M. Field, I. Melbourne and A. Török, *Stable ergodicity for smooth compact Lie group extensions of hyperbolic basic sets*, Ergodic Theory Dynam. Systems **25** (2) (2005), 517–551.
[47] M. Field and W. Parry, *Stable ergodicity of skew extensions by compact Lie groups*, Topology **38** (1) (1999), 167–187.
[48] A.S. Gorodetskiĭ and Yu.S. Ilyashenko, *Some new robust properties of invariant sets and attractors of dynamical systems*, Funktsional. Anal. i Prilozhen. **33** (2) (1999), 16–30, 95; transl. in Funct. Anal. Appl. **33** (2) (1999), 95–105.
[49] M. Grayson, C. Pugh and M. Shub, *Stably ergodic diffeomorphisms*, Ann. of Math. **140** (1994), 295–329.
[50] J. Hadamard, *Sur l'itération et les solutions asymptotiques des équations différentielles*, Bull. Soc. Math. France **29** (1901), 224–228.
[51] P. Halmos, *On automorphisms of compact groups*, Bull. Amer. Math. Soc. **49** (1943), 619–624.
[52] B. Hasselblatt, *Regularity of the Anosov splitting and of Horospheric foliations*, Ergodic Theory Dynam. Systems **14** (46) (1994), 45–666.
[53] B. Hasselblatt and A. Wilkinson, *Prevalence of non-Lipschitz Anosov foliations*, Ergodic Theory Dynam. Systems **19** (3) (1999), 643–656.
[54] M. Hirayama and Ya. Pesin, *Non-absolutely continuous invariant foliations*, Moscow Math. J. (2005), to appear.
[55] M. Hirsch, C. Pugh and M. Shub, *Invariant manifolds*, Bull. Amer. Math. Soc. **76** (1970), 1015–1019.
[56] M. Hirsch, C. Pugh and M. Shub, *Invariant Manifolds*, Lecture Notes in Math., Vol. 583, Springer-Verlag (1977).
[57] E. Hopf, *Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung*, Ber. Verh. Sächs. Akad. Wiss. Leipzig **91** (1939), 261–304.
[58] V. Horita and A. Tahzibi, *Partial hyperbolicity for symplectic diffeomorphisms*, Ann. Inst. H. Poincaré Anal. Non Linéaire (2005), to appear.
[59] Yu. Ilyashenko and W. Li, *Nonlocal Bifurcations*, Amer. Math. Soc., Providence, RI (1999).
[60] M. Irwin, *On the stable manifold theorem*, Bull. London Math. Soc. **2** (1970), 196–198.
[61] M. Jiang, R. de la Llave and Y. Pesin, *On the integrability of intermediate distributions for Anosov diffeomorphisms*, Ergodic Theory Dynam. Systems **15** (2) (1995), 317–331.
[62] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge Univ. Press, Cambridge (1995).
[63] A. Katok and A. Kononenko, *Cocycle stability for partially hyperbolic systems*, Math. Res. Lett. **3** (1996), 191–210.
[64] J. Mather, *Characterization of Anosov diffeomorphisms*, Nederl. Akad. Wetensch. Proc. Ser. A71, Indag. Math. **30** (1968), 479–483.
[65] J. Milnor, *Fubini foiled: Katok's paradoxical example in measure theory*, Math. Intelligencer **19** (1997), 30–32.
[66] V. Niţică and A. Török, *An open dense set of stably ergodic diffeomorphisms in a neighborhood of a non-ergodic one*, Topology **40** (2001), 259–278.
[67] D. Orendovici and Y. Pesin, *Chaos in traveling waves of lattice systems of unbounded media*, Numerical Methods for Bifurcation Problems and Large-Scale Dynamical Systems (Minneapolis, MN, 1997), Springer, New York (1999), 327–358.
[68] W. Parry and M. Pollicott, *Stability of mixing for toral extensions of hyperbolic sets*, Tr. Mat. Inst. Steklova **216** (1997), Din. Sist. i Smezhnye Vopr., 354–363; transl. in Proc. Steklov Inst. Math. **216** (1) (1997) 350–359.
[69] O. Perron, *Über Stabilität und asymptotisches Verhalten der Lösungen eines Systemes endlicher Differenzengleichungen*, J. Reine Angew. Math. **161** (1929), 41–64.
[70] Y. Pesin, *On the existence of invariant fiberings for a diffeomorphism of a smooth manifold*, Mat. Sb. **91** (2) (1973), 202–210.
[71] Y. Pesin, *Lectures on Partial Hyperbolicity and Stable Ergodicity*, Zürich Lectures in Advanced Mathematics, EMS (2004).
[72] Y. Pesin and Y. Sinai, *Gibbs measures for partially hyperbolic attractors*, Ergodic Theory Dynam. Systems **2** (3–4) (1982), 417–438.
[73] J. Plante, *Anosov flows*, Amer. J. Math. **94** (1972), 729–754.
[74] C. Pugh and M. Shub, *Ergodicity of Anosov actions*, Invent. Math., **15** (1972), 1–23.

[75] C. Pugh and M. Shub, *Ergodic attractors*, Trans. Amer. Math. Soc. **312** (1) (1989), 1–54.
[76] C. Pugh and M. Shub, *Stable ergodicity and partial hyperbolicity*, International Conference on Dynamical Systems (Montevideo, 1995), Pitman Res. Notes Math. Ser., Vol. 362, Longman, Harlow (1996), 182–187.
[77] C. Pugh and M. Shub, *Stably ergodic dynamical systems and partial hyperbolicity*, J. Complexity **13** (1) (1997), 125–179.
[78] C. Pugh and M. Shub, *Stable ergodicity and Julienne quasi-conformality*, J. European Math. Soc. (JEMS) **2** (1) (2000), 1–52; **6** (1) (2004), 149–151.
[79] C. Pugh and M. Shub, *Stable ergodicity*, Bull. Amer. Math. Soc. (N.S.) **41** (1) (2004), 1–41.
[80] C. Pugh, M. Shub and A. Wilkinson, *Hölder foliations*, Duke Math. J. **86** (3) (1997), 517–546; *Correction to: "Hölder foliations"*, Duke Math. J. **105** (2000), 105–106.
[81] F. Rodriguez Hertz, *Stable ergodicity of certain linear automorphisms of the torus*, Ann. of Math. **162** (1) (2005), to appear.
[82] F. Rodriguez Hertz, J. Rodriguez Hertz and R. Ures, *Accessibility and stable ergodicity for partially hyperbolic diffeomorphisms with* 1*D-center bundle*, Preprint.
[83] D. Ruelle and A. Wilkinson, *Absolutely singular dynamical foliations*, Comm. Math. Phys., **219** (2001), 481–487.
[84] R. Sacker, *A perturbation theorem for invariant Riemannian manifolds*, Differential Equations and Dynamical Systems (Proc. Internat. Sympos., Mayaguez, P.R., 1965), Academic Press, New York (1967), 43–54.
[85] M. Shub and A. Wilkinson, *A stably Bernoullian diffeomorphism that is not Anosov*, Preprint (1998).
[86] M. Shub and A. Wilkinson, *Stably ergodic approximation: two examples*, Ergodic Theory Dynam. Systems **20** (3) (2000), 875–893.
[87] M. Shub and A. Wilkinson, *Pathological foliations and removable zero exponents*, Invent. Math. **139** (3) (2000), 495–508.
[88] S. Smale, *Differentiable dynamical systems*, Bull. Amer. Math. Soc. **73** (1967), 747–817.
[89] A. Tahzibi, *Stably ergodic systems which are not partially hyperbolic*, Israel J. Math., to appear.
[90] A. Talitskaya, *Partially hyperbolic phenomena in dynamical systems with discrete and continuous time*, Ph.D. thesis, PSU (2004), http://etda.libraries.psu.edu/theses/approved/WorldWideFiles/ETD-533/Thesis.pdf.
[91] C. Walkden, *Stable ergodic properties of cocycles over hyperbolic attractors*, Comm. Math. Phys. **205** (2) (1999), 263–281.
[92] A. Wilkinson, *Stable ergodicity of the time-one map of a geodesic flow*, Ergodic Theory Dynam. Systems **18** (6) (1998), 1545–1587.
[93] H. Yamabe, *On an arcwise connected subgroup of a Lie group*, Osaka Math. J. **2** (1950), 13–14.
[94] J.-C. Yoccoz, *Introduction to hyperbolic dynamics*, Real and Complex Dynamical Systems, Proceedings of the NATO Advanced Study Institute held in Hillerød, June 20–July 2, 1993, B. Branner and P. Hjorth, eds, NATO Advanced Science Institutes Series C: Mathematical and Physical Sciences, Vol. 464, Kluwer Academic, Dordrecht (1995), 265–291.

This page intentionally left blank

CHAPTER 2

Smooth Ergodic Theory and Nonuniformly Hyperbolic Dynamics

Luis Barreira

Departamento de Matemática, Instituto Superior Técnico, 1049-001 Lisboa, Portugal
E-mail: *barreira@math.ist.utl.pt*
url: *http://www.math.ist.utl.pt/~barreira/*

Yakov Pesin

Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, USA
E-mail: *pesin@math.psu.edu*
url: *http://www.math.psu.edu/pesin/*

With an appendix by Omri Sarig

Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, USA
E-mail: *sarig@math.psu.edu*
url: *http://www.math.psu.edu/sarig/*

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

This page intentionally left blank

Introduction

The goal of this chapter is to describe the contemporary status of nonuniform hyperbolicity theory. We present the core notions and results of the theory as well as discuss recent developments and some open problems. We also describe essentially all known examples of nonuniformly hyperbolic systems. Following the principles of the handbook we include informal discussions of many results and sometimes outline their proofs.

Originated in the works of Lyapunov [171] and Perron [194,195] the nonuniform hyperbolicity theory has emerged as an independent discipline in the works of Oseledets [192] and Pesin [198]. Since then it has become one of the major parts of the general dynamical systems theory and one of the main tools in studying highly sophisticated behavior associated with "deterministic chaos". We refer the reader to the article [5] by Hasselblatt and Katok in volume 1A of the handbook for a discussion on the role of nonuniform hyperbolicity theory, its relations to and interactions with other areas of dynamics. See also the article [4] by Hasselblatt in the same volume for a brief account of nonuniform hyperbolicity theory in view of the general hyperbolicity theory, and the book by Barreira and Pesin [35] for a detailed presentation of the core of the nonuniform hyperbolicity theory.

Nonuniform hyperbolicity conditions can be expressed in terms of the Lyapunov exponents. Namely, a dynamical system is nonuniformly hyperbolic if it admits an invariant measure with nonzero Lyapunov exponents almost everywhere. This provides an efficient tool in verifying the nonuniform hyperbolicity conditions and determines the importance of the nonuniform hyperbolicity theory in applications.

We emphasize that the nonuniform hyperbolicity conditions are *weak enough* not to interfere with the topology of the phase space so that any compact smooth manifold of dimension $\geqslant 2$ admits a volume-preserving C^∞ diffeomorphism which is nonuniformly hyperbolic. On the other hand, these conditions are *strong enough* to ensure that any $C^{1+\alpha}$ nonuniformly hyperbolic diffeomorphism has positive entropy with respect to any invariant *physical* measure (by physical measure we mean either a smooth measure or a Sinai–Ruelle–Bowen (SRB) measure). In addition, any ergodic component has positive measure and up to a cyclic permutation the restriction of the map to this component is Bernoulli. Similar results hold for systems with continuous time.

It is conjectured that dynamical systems of class $C^{1+\alpha}$ with nonzero Lyapunov exponents preserving a given smooth measure are typical in some sense. This remains one of the major open problems in the field and its affirmative solution would greatly benefit and boost the applications of the nonuniform hyperbolicity theory. We stress that the systems under consideration should be of class $C^{1+\alpha}$ for some $\alpha > 0$: not only the nonuniform hyperbolicity theory for C^1 systems is substantially less interesting but one should also expect a "typical" C^1 map to have some zero Lyapunov exponents (unless the map is Anosov).

In this chapter we give a detailed account of the topics mentioned above as well as many others. Among them are:

(1) stable manifold theory (including the construction of local and global stable and unstable manifolds and their absolute continuity);
(2) local ergodicity problem (i.e., finding conditions which guarantee that every ergodic component of positive measure is open $(\mathrm{mod}\,0)$);

(3) description of the topological properties of systems with nonzero Lyapunov exponents (including the density of periodic orbits, the closing and shadowing properties, and the approximation by horseshoes); and
(4) computation of the dimension and the entropy of arbitrary hyperbolic measures.

We also describe some methods which allow one to establish that a given system has nonzero Lyapunov exponents (for example, SRB-measures) or to construct a hyperbolic measure with "good" ergodic properties (for example, the Markov extension approach). Finally, we outline a version of nonuniform hyperbolicity theory for systems with singularities (including billiards).

The nonuniform hyperbolicity theory covers an enormous area of dynamics and despite the scope of this survey there are several topics not covered or barely mentioned. Among them are nonuniformly hyperbolic one-dimensional transformations, random dynamical systems with nonzero Lyapunov exponents, billiards and related systems (for example, systems of hard balls), and numerical computation of Lyapunov exponents. For more information on these topics we refer the reader to the articles in the handbook [2,3,7–9]. Here the reader finds the ergodic theory of random transformations [8,3] (including a "random" version of Pesin's entropy formula in [8]), nonuniform one-dimensional dynamics [10,7], ergodic properties and decay of correlations for nonuniformly expanding maps [10], the dynamics of geodesic flows on compact manifolds of nonpositive curvature [9], homoclinic bifurcations and dominated splitting [11] and dynamics of partially hyperbolic systems with nonzero Lyapunov exponents [6]. Last but no least, we would like to mention the article [2] on the Teichmüller geodesic flows showing in particular, that the Kontsevich–Zorich cocycle over the Teichmüller flow is nonuniformly hyperbolic [99].

Although we included comments of historical nature concerning some main notions and basic results, the chapter is not meant to present a complete historical account of the field.

1. Lyapunov exponents of dynamical systems

Let $f^t : M \to M$ be a dynamical system with discrete time, $t \in \mathbb{Z}$, or continuous time, $t \in \mathbb{R}$, of a smooth Riemannian manifold M. Given a point $x \in M$, consider the family of linear maps $\{d_x f^t\}$ which is called *the system in variations* along the trajectory $f^t(x)$. It turns out that for a "typical" trajectory one can obtain a sufficiently complete information on stability of the trajectory based on the information on the asymptotic stability of the "zero solution" of the system in variations.

In order to characterize the asymptotic stability of the "zero solution", given a vector $v \in T_x M$, define the *Lyapunov exponent* of v at x by

$$\chi^+(x, v) = \overline{\lim_{t \to +\infty}} \frac{1}{t} \log \|d_x f^t v\|.$$

For every $\varepsilon > 0$ there exists $C = C(v, \varepsilon) > 0$ such that if $t \geqslant 0$ then

$$\|d_x f^t v\| \leqslant C e^{(\chi^+(x,v)+\varepsilon)t} \|v\|.$$

The Lyapunov exponent possesses the following basic properties:

1. $\chi^+(x, \alpha v) = \chi^+(x, v)$ for each $v \in V$ and $\alpha \in \mathbb{R} \setminus \{0\}$;
2. $\chi^+(x, v + w) \leqslant \max\{\chi^+(x, v), \chi^+(x, w)\}$ for each $v, w \in V$;
3. $\chi^+(x, 0) = -\infty$.

The study of the Lyapunov exponents can be carried out to a certain extent using only these three basic properties. This is the subject of the abstract theory of Lyapunov exponents (see [35]). As a simple consequence of the basic properties we obtain that the function $\chi^+(x, \cdot)$ attains only finitely many values on $T_x M \setminus \{0\}$. Let $p^+(x)$ be the number of distinct values and

$$\chi_1^+(x) < \cdots < \chi_{p^+(x)}^+(x),$$

the values themselves. The Lyapunov exponent $\chi^+(x, \cdot)$ generates the filtration $\mathcal{V}_x^+$ of the tangent space $T_x M$,

$$\{0\} = V_0^+(x) \subsetneqq V_1^+(x) \subsetneqq \cdots \subsetneqq V_{p^+(x)}^+(x) = T_x M,$$

where $V_i^+(x) = \{v \in T_x M\colon\ \chi^+(x, v) \leqslant \chi_i^+(x)\}$. The number

$$k_i^+(x) = \dim V_i^+(x) - \dim V_{i-1}^+(x)$$

is the *multiplicity* of the value $\chi_i^+(x)$. We have

$$\sum_{i=1}^{p^+(x)} k_i^+(x) = \dim M.$$

The collection of pairs

$$\operatorname{Sp} \chi^+(x) = \left\{\left(\chi_i^+(x), k_i^+(x)\right)\colon\ 1 \leqslant i \leqslant p^+(x)\right\}$$

is called the *Lyapunov spectrum* of the exponent $\chi^+(x, \cdot)$.

The functions $\chi_i^+(x)$, $p^+(x)$, and $k_i^+(x)$ are *invariant* under f and (Borel) *measurable* (but not necessarily continuous).

One can obtain another Lyapunov exponent for f by reversing the time. Namely, for every $x \in M$ and $v \in T_x M$ let

$$\chi^-(x, v) = \overline{\lim_{t \to -\infty}} \frac{1}{|t|} \log \left\| d_x f^t v \right\|.$$

The function $\chi^-(x, \cdot)$ possesses the same basic properties as $\chi^+(x, \cdot)$ and hence, takes on finitely many values on $T_x M \setminus \{0\}$:

$$\chi_1^-(x) > \cdots > \chi_{p^-(x)}^-(x),$$

where $p^-(x) \leqslant \dim M$. Denote by $\mathcal{V}_x^-$ the filtration of T_xM associated with $\chi^-(x,\cdot)$:

$$T_xM = V_1^-(x) \supsetneqq \cdots \supsetneqq V_{p^-(x)}^-(x) \supsetneqq V_{p^-(x)+1}^-(x) = \{0\},$$

where $V_i^-(x) = \{v \in T_xM\colon\ \chi^-(x,v) \leqslant \chi_i^-(x)\}$. The number

$$k_i^-(x) = \dim V_i^-(x) - \dim V_{i+1}^-(x)$$

is the *multiplicity* of the value $\chi_i^-(x)$. The collection of pairs

$$\operatorname{Sp}\chi^-(x) = \left\{\left(\chi_i^-(x), k_i^-(x)\right)\colon\ i = 1, \ldots, p^-(x)\right\}$$

is called the *Lyapunov spectrum* of the exponent $\chi^-(x,\cdot)$.

We now introduce the crucial concept of Lyapunov regularity. Roughly speaking it asserts that the forward and backward behavior of the system along a "typical" trajectory comply in a quite strong way.

A point x is called *Lyapunov forward regular point* if

$$\lim_{t\to+\infty} \frac{1}{t} \log\left|\det d_x f^t\right| = \sum_{i=1}^{p^+(x)} k_i^+(x)\chi_i^+(x).$$

Similarly, a point x is called *Lyapunov backward regular* if

$$\lim_{t\to-\infty} \frac{1}{|t|} \log\left|\det d_x f^t\right| = \sum_{i=1}^{p^-(x)} k_i^-(x)\chi_i^-(x).$$

If a point x is forward (backward) regular then so is any point along its trajectory and we can say that the whole trajectory is forward (backward) regular. Note that there may be trajectories which are neither forward nor backward regular and that forward (backward) regularity does not necessarily imply backward (forward) regularity (an example is a flow in $\mathbb{R}^3$ that progressively approaches zero and infinity when time goes to $+\infty$, oscillating between the two, but which tends to a given point when time goes to $-\infty$).

Given $x \in M$, we say that the filtrations $\mathcal{V}_x^+$ and $\mathcal{V}_x^-$ comply if:

1. $p^+(x) = p^-(x) \stackrel{\text{def}}{=} p(x)$;
2. the subspaces $E_i(x) = V_i^+(x) \cap V_i^-(x)$, $i = 1, \ldots, p(x)$, form a splitting of the tangent space

$$T_xM = \bigoplus_{i=1}^{p(x)} E_i(x).$$

We say that a point x is *Lyapunov regular* or simply *regular* if:

1. the filtrations $\mathcal{V}_x^+$ and $\mathcal{V}_x^-$ comply;

2. for $i = 1, \dots, p(x)$ and $v \in E_i(x) \setminus \{0\}$ we have

$$\lim_{t \to \pm\infty} \frac{1}{t} \log \|d_x f^t v\| = \chi_i^+(x) = -\chi_i^-(x) \stackrel{\text{def}}{=} \chi_i(x)$$

with uniform convergence on $\{v \in E_i(x)\colon \|v\| = 1\}$;

3. $$\lim_{t \to \pm\infty} \frac{1}{t} \log \left|\det d_x f^t\right| = \sum_{i=1}^{p(x)} \chi_i(x) \dim E_i(x).$$

Note that if x is regular then so is the point $f^t(x)$ for any t and thus, one can speak of the whole trajectory as being regular.

In order to simplify our notations in what follows, we will drop the superscript $+$ from the notation of the Lyapunov exponents and the associated quantities if it does not cause any confusion.

The following criterion for regularity is quite useful in applications. Denote by $V(v_1, \dots, v_k)$ the k-volume of the parallelepiped defined by the vectors $v_1, \dots, v_k$.

THEOREM 1.1 (see [69]). *If x is Lyapunov regular then the following statements hold*:

1. *for any vectors $v_1, \dots, v_k \in T_x M$ there exists the limit*

$$\lim_{t \to \pm\infty} \frac{1}{t} \log V\left(d_x f^t v_1, \dots, d_x f^t v_k\right);$$

if, in addition, $v_1, \dots, v_k \in E_i(x)$ and $V(v_1, \dots, v_k) \neq 0$ then

$$\lim_{t \to \pm\infty} \frac{1}{t} \log V\left(d_x f^t v_1, \dots, d_x f^t v_k\right) = \chi_i(x) k;$$

2. *if $v \in E_i(x) \setminus \{0\}$ and $w \in E_j(x) \setminus \{0\}$ with $i \neq j$ then*

$$\lim_{t \to \pm\infty} \frac{1}{t} \log \left|\sin \angle\left(d_x f^t v, d_x f^t w\right)\right| = 0.$$

Furthermore, if these properties hold then x is Lyapunov regular.

Forward and backward regularity of a trajectory does not automatically yields that the filtrations comply and hence, forward and backward regularity do not, in general, imply Lyapunov regularity. Roughly speaking the forward behavior of a trajectory may not depend on its backward behavior while Lyapunov regularity requires some compatibility between the forward and backward behavior expressed in terms of the filtrations $\mathcal{V}_{\chi^+}$ and $\mathcal{V}_{\chi^-}$. However, if a trajectory $\{f^n(x)\}$ returns infinitely often to an arbitrary small neighborhood of x as $n \to \pm\infty$ one may expect the forward and backward behavior to comply in a certain sense.

The following celebrated result of Oseledets [192] gives a rigorous mathematical description of this phenomenon and shows that regularity is "typical" from the measure-theoretical point of view.

THEOREM 1.2 (Multiplicative Ergodic Theorem). *If f is a diffeomorphism of a smooth Riemannian manifold M, then the set of Lyapunov regular points has full measure with respect to any f-invariant Borel probability measure on M.*

This theorem is a particular case of a more general statement (see Section 5.3).

The notion of Lyapunov exponent was introduced by Lyapunov [171], with background and motivation coming from his study of differential equations. A comprehensive but somewhat outdated reference for the theory of Lyapunov exponents as well as its applications to the theory of differential equations is the book of Bylov, Vinograd, Grobman and Nemyckii [69] which is available only in Russian. A part of this theory is presented in modern language in [35].

The notion of forward regularity originated in the work of Lyapunov [171] and Perron [194,195] in connection with the study of the stability properties of solutions of linear ordinary differential equations with nonconstant coefficients (see [35] for a detailed discussion).

2. Examples of systems with nonzero exponents

2.1. *Hyperbolic invariant measures*

Smooth ergodic theory studies topological and ergodic properties of smooth dynamical systems with nonzero Lyapunov exponents. Let f be a diffeomorphism of a complete (not necessarily compact) smooth Riemannian manifold M. The map f should be of class at least $C^{1+\alpha}$, $\alpha > 0$. We assume that there exists an f-invariant set Λ with the property that for every $x \in \Lambda$ the values of the Lyapunov exponent at x are nonzero. More precisely, there exists a number $s = s(x)$, $1 \leqslant s < p(x)$ such that

$$\chi_1(x) < \cdots < \chi_s(x) < 0 < \chi_{s+1}(x) < \cdots < \chi_{p(x)}(x). \tag{2.1}$$

We say that f has nonzero exponents on Λ. Let us stress that according to our definition there should always be at least one negative value and at least one positive value of the exponent.

Assume now that f preserves a Borel probability measure ν on M. We call ν *hyperbolic* if (2.1) holds for almost every $x \in M$. It is not known whether *a diffeomorphism f which has nonzero exponents on a set Λ possesses a hyperbolic measure ν with $\nu(\Lambda) = 1$.*

In the case ν is ergodic the values of the Lyapunov exponent are constant almost everywhere, i.e., $k_i(x) = k_i^\nu$ and $\chi_i(x) = \chi_i^\nu$ for $i = 1, \ldots, p(x) = p^\nu$. The collection of pairs

$$\operatorname{Sp} \chi^\nu = \left\{(\chi_i^\nu, k_i^\nu)\colon\ i = 1, \ldots, p^\nu\right\}$$

is called the *Lyapunov spectrum of the measure*. The measure ν is hyperbolic if none of the numbers χ_i^ν in its spectrum is zero.

We now discuss the case of dynamical systems with continuous time. Let f^t be a smooth flow on a smooth Riemannian manifold M. It is generated by a vector field X on M such

that $X(x) = \frac{df^t(x)}{dt}|_{t=0}$. Clearly, $\chi(x, v) = 0$ for every v in the direction of $X(x)$, i.e., for $v = \alpha X(x)$ with some $\alpha \in \mathbb{R}$.

We say that *the flow f^t has nonzero exponents on an invariant set Λ* if for every $x \in \Lambda$ all the Lyapunov exponents, but the one in the direction of the flow, are nonzero, at least one of them is negative and at least one of them is positive. More precisely, there exists a number $s = s(x)$, $1 \leqslant s < p(x) - 1$ such that

$$\chi_1(x) < \cdots < \chi_s(x) < \chi_{s+1}(x) = 0 < \chi_{s+2}(x) < \cdots < \chi_{p(x)}(x), \tag{2.2}$$

where $\chi_{s+1}(x)$ is the value of the exponent in the direction of $X(x)$.

Assume now that a flow f^t preserves a Borel probability measure ν on M. We call ν *hyperbolic* if (2.2) holds for almost every $x \in M$.

There are two classes of hyperbolic invariant measures on compact manifolds for which one can obtain a sufficiently complete description of its ergodic properties. They are:

1. *smooth measures*, i.e., measures which are equivalent to the Riemannian volume with the Radon–Nikodym derivative bounded from above and bounded away from zero (see Section 11);
2. *Sinai–Ruelle–Bowen measures* (see Section 14).

Dolgopyat and Pesin [88] proved that any compact manifold of dimension $\geqslant 2$ admits a volume-preserving diffeomorphism with nonzero Lyapunov exponents, and Hu, Pesin and Talitskaya [126] showed that any compact manifold of dimension $\geqslant 3$ admits a volume-preserving flow with nonzero Lyapunov exponents; see Section 13.1 for precise statements and further discussion. However, there are few particular examples of volume-preserving systems with nonzero Lyapunov exponents. In the following subsections we present some basic examples of such systems to illustrate some interesting phenomena associated with nonuniform hyperbolicity.

2.2. *Diffeomorphisms with nonzero exponents on the* 2*-torus*

The first example of a diffeomorphism with nonzero Lyapunov exponent, which is not an Anosov map, was constructed by Katok [134]. This is an area-preserving ergodic (indeed, Bernoulli) diffeomorphism $G_{\mathbb{T}^2}$ of the two-dimensional torus $\mathbb{T}^2$ which is obtained by a "surgery" of an area-preserving hyperbolic toral automorphism A with two eigenvalues $\lambda > 1$ and $\lambda^{-1} < 1$. The main idea of Katok's construction is to destroy the uniform hyperbolic structure associated with A by slowing down trajectories in a small neighborhood U of the origin (which is a fixed hyperbolic point for A). This means that the time, a trajectory of a "perturbed" map $G_{\mathbb{T}^2}$ stays in U, gets larger and larger the closer the trajectory passes by the origin, while the map is unchanged outside U. In particular, it can be arranged that the trajectories, starting on the stable and unstable separatrices of the origin, have zero exponents and thus, $G_{\mathbb{T}^2}$ is a not an Anosov map. Although a "typical" trajectory may spend arbitrarily long periods of time in U, the average time it stays in U is proportional to the measure of U and hence, is small. This alone does not automatically guarantee that a "typical" trajectory has nonzero exponents. Indeed, one should make sure that between the time the trajectory enters and exits U a vector in small cone around the unstable direction

of A does not turn into a vector in a small cone around the stable direction of A. If this occurs the vector may contract, while travelling outside U, so one may loose control over its length.

The construction depends upon a real-valued function ψ which is defined on the unit interval $[0, 1]$ and has the following properties:

1. ψ is a C^∞ function except at the origin;
2. $\psi(0) = 0$ and $\psi(u) = 1$ for $u \geqslant r_0$ where $0 < r_0 < 1$;
3. $\psi'(u) > 0$ for every $0 < u < r_0$;
4. the following integral converges:

$$\int_0^1 \frac{du}{\psi(u)} < \infty.$$

Consider the disk D_r centered at 0 of radius r and a coordinate system (s_1, s_2) in D_r formed by the eigendirections of A such that

$$D_r = \big\{(s_1, s_2)\colon\ s_1^2 + s_2^2 \leqslant r^2\big\}.$$

Observe that A is the time-one map of the flow generated by the following system of differential equations:

$$\dot{s}_1 = s_1 \log \lambda, \qquad \dot{s}_2 = -s_2 \log \lambda.$$

Fix a sufficiently small number $r_1 > r_0$ and consider the time-one map g generated by the following system of differential equations in D_{r_1}:

$$\dot{s}_1 = s_1 \psi\big(s_1^2 + s_2^2\big) \log \lambda, \qquad \dot{s}_2 = -s_2 \psi\big(s_1^2 + s_2^2\big) \log \lambda. \tag{2.3}$$

Our choice of the function ψ guarantees that $g(D_{r_2}) \subset D_{r_1}$ for some $r_2 < r_1$, and that g is of class C^∞ in $D_{r_1} \setminus \{0\}$ and coincides with A in some neighborhood of the boundary ∂D_{r_1}. Therefore, the map

$$G(x) = \begin{cases} A(x) & \text{if } x \in \mathbb{T}^2 \setminus D_{r_1}, \\ g(x) & \text{if } x \in D_{r_1}, \end{cases}$$

defines a homeomorphism of the torus $\mathbb{T}^2$ which is a C^∞ diffeomorphism everywhere except at the origin. The map $G(x)$ is a slowdown of the automorphism A at 0.

Denote by W^u and W^s the projections of the eigenlines in $\mathbb{R}^2$ to $\mathbb{T}^2$ corresponding to the eigenvalues λ and λ^{-1}. Set $W = W^u \cup W^s$ and $X = \mathbb{T}^2 \setminus W$. Note that the set W is everywhere dense in $\mathbb{T}^2$.

Let $x = (0, s_2) \in D_{r_1} \cap W^s$. For a vertical vector $v \in T_x\mathbb{T}^2$,

$$\chi(x, v) = \overline{\lim_{t \to +\infty}} \frac{\log|s_2(t)|}{t} = \overline{\lim_{t \to +\infty}} \big(\log\big|s_2(t)\big|\big)' = \overline{\lim_{t \to +\infty}} \big(-\psi\big(s_2(t)^2\big) \log \lambda\big),$$

where $s_2(t)$ is the solution of (2.3) with the initial condition $s_2(0) = s_2$. In view of the choice of the function ψ, we obtain that $\chi(x, v) = 0$. Similarly, $\chi(x, v) = 0$ whenever x, $v \in W^u$. In particular, G is not an Anosov diffeomorphism.

Choose $x \in X \setminus D_{r_1}$ and define the stable and unstable cones in $T_x\mathbb{T}^2 = \mathbb{R}^2$ by

$$C^s(x) = \{(v_1, v_2) \in \mathbb{R}^2\colon\ |v_1| \leqslant \alpha|v_2|\},$$
$$C^u(x) = \{(v_1, v_2) \in \mathbb{R}^2\colon\ |v_2| \leqslant \alpha|v_1|\},$$

where $v_1 \in W^u$, $v_2 \in W^s$ and $0 < \alpha < 1/4$. The formulae

$$E^s(x) = \bigcap_{j=0}^{\infty} dG^{-j} C^s\big(G^j(x)\big), \qquad E^u(x) = \bigcap_{j=0}^{\infty} dG^j C^u\big(G^{-j}(x)\big)$$

define one-dimensional subspaces at x such that $\chi(x, v) < 0$ for $v \in E^s(x)$ and $\chi(x, v) > 0$ for $v \in E^u(x)$. The map G is uniformly hyperbolic on $X \setminus D_{r_1}$: there is a number $\mu > 1$ such that for every $x \in X \setminus D_{r_1}$,

$$\big\|dG|E^s(x)\big\| \leqslant \frac{1}{\mu}, \qquad \big\|dG^{-1}|E^u(x)\big\| \leqslant \frac{1}{\mu}.$$

One can show that the stable and unstable subspaces can be extended to $W \setminus \{0\}$ to form two one-dimensional continuous distributions on $\mathbb{T}^2 \setminus \{0\}$.

The map G preserves the probability measure $d\nu = \kappa_0^{-1}\kappa\, dm$ where m is area and the density κ is a positive C^∞ function that is infinite at 0. It is defined by the formula

$$\kappa(s_1, s_2) = \begin{cases} (\psi(s_1^2 + s_2^2))^{-1} & \text{if } (s_1, s_2) \in D_{r_1}, \\ 1 & \text{otherwise,} \end{cases}$$

and

$$\kappa_0 = \int_{\mathbb{T}^2} \kappa\, dm.$$

Consider the map φ of the torus given by

$$\varphi(s_1, s_2) = \frac{1}{\sqrt{\kappa_0(s_1^2 + s_2^2)}} \left(\int_0^{s_1^2 + s_2^2} \frac{du}{\psi(u)} \right)^{1/2} (s_1, s_2) \tag{2.4}$$

in D_{r_1} and $\varphi = \mathrm{Id}$ in $\mathbb{T}^2 \setminus D_{r_1}$. It is a homeomorphism and is a C^∞ diffeomorphism except at the origin. It also commutes with the involution $I(t_1, t_2) = (1 - t_1, 1 - t_2)$. The map $G_{\mathbb{T}^2} = \varphi \circ G \circ \varphi^{-1}$ is of class C^∞, area-preserving and has nonzero Lyapunov exponents almost everywhere. One can show that $G_{\mathbb{T}^2}$ is ergodic and is a Bernoulli diffeomorphism.

2.3. *Diffeomorphisms with nonzero exponents on the 2-sphere*

Using the diffeomorphism $G_{\mathbb{T}^2}$ Katok [134] constructed a diffeomorphisms with nonzero exponents on the 2-sphere S^2. Consider a toral automorphism A of the torus $\mathbb{T}^2$ given by the matrix $A = \left(\begin{smallmatrix} 5 & 8 \\ 8 & 13 \end{smallmatrix}\right)$. It has four fixed points $x_1 = (0,0)$, $x_2 = (1/2, 0)$, $x_3 = (0, 1/2)$ and $x_4 = (1/2, 1/2)$.

For $i = 1, 2, 3, 4$ consider the disk D^i_r centered at x_i of radius r. Repeating the construction from the previous section we obtain a diffeomorphism g_i which coincides with A outside $D^i_{r_1}$. Therefore, the map

$$G_1(x) = \begin{cases} A(x) & \text{if } x \in \mathbb{T}^2 \setminus D, \\ g_i(x) & \text{if } x \in D^i_{r_1}, \end{cases}$$

defines a homeomorphism of the torus $\mathbb{T}^2$ which is a C^∞ diffeomorphism everywhere expect at the points x_i. Here $D = \bigcup_{i=1}^4 D^i_{r_1}$. The Lyapunov exponents of G_1 are nonzero almost everywhere with respect to the area.

Consider the map

$$\varphi(x) = \begin{cases} \varphi_i(x) & \text{if } x \in D^i_{r_1}, \\ x & \text{otherwise,} \end{cases}$$

where φ_i are given by (2.4) in each disk $D^i_{r_2}$. It is a homeomorphism of $\mathbb{T}^2$ which is a C^∞ diffeomorphism everywhere expect at the points x_i. The map $G_2 = \varphi \circ G_1 \circ \varphi^{-1}$ is of class C^∞, area-preserving and has nonzero Lyapunov exponents almost everywhere.

Consider the map $\zeta : \mathbb{T}^2 \to S^2$ defined by

$$\zeta(s_1, s_2) = \left(\frac{s_1^2 - s_2^2}{\sqrt{s_1^2 + s_2^2}}, \frac{2 s_1 s_2}{\sqrt{s_1^2 + s_2^2}} \right).$$

This map is a double branched covering and is C^∞ everywhere except at the points x_i, $i = 1, 2, 3, 4$, where it branches. It commutes with the involution I and preserves the area. Consider the map $G_{S^2} = \zeta \circ G_2 \circ \zeta^{-1}$. One can show that it is a C^∞ diffeomorphism which preserves the area and has nonzero Lyapunov exponents almost everywhere. Furthermore, one can show that G_{S^2} is ergodic and indeed, is a Bernoulli diffeomorphism.

2.4. *Analytic diffeomorphisms with nonzero exponents*

We describe an example due to Katok and Lewis [138] of a volume-preserving analytic diffeomorphism of a compact smooth Riemannian manifold. It is a version of the well-known blow-up procedure from algebraic geometry.

Setting $X = \{x \in \mathbb{R}^n : \|x\| > 1\}$ consider the map $\varphi : \mathbb{R}^n \setminus \{0\} \to X$ given by

$$\varphi(x) = \frac{(\|x\|^n + 1)^{1/n}}{\|x\|} x.$$

It is easy to see that φ has Jacobian 1 with respect to the standard coordinates on $\mathbb{R}^n$.

Let A be a linear hyperbolic transformation of $\mathbb{R}^n$. The diffeomorphism $f = \varphi \circ A \circ \varphi^{-1}$ extends analytically to a neighborhood of the boundary which depends on A. This follows from the formula

$$\varphi \circ A \circ \varphi^{-1}(x) = \left(\frac{\|x\|^n - 1}{\|x\|} + \frac{1}{\|Ax\|^n} \right)^{1/n} Ax.$$

Let (r, θ) be the standard polar coordinates on X so that $X = \{(r, \theta)\colon\ r > 1,\ \theta \in S^{n-1}\}$. Introducing new coordinates (s, θ), where $s = r^n - 1$, observe that these coordinates extend analytically across the boundary and have the property that the standard volume form is proportional to $ds \wedge d\theta$. Let B be the quotient of $\bar{X}$ under the identification of antipodal points on the boundary. The map f induces a map F of B which preserves the volume form $ds \wedge d\theta$, has nonzero Lyapunov exponents and is analytic.

2.5. *Pseudo-Anosov maps*

Pseudo-Anosov maps were singled out by Thurston in connection with the problem of classifying diffeomorphisms of a compact C^∞ surface M up to isotopy (see [240,95]). According to Thurston's classification, a diffeomorphism f of M is isotopic to a homeomorphism g satisfying one of the following properties (see [95, Exposé 9]):

1. g is of finite order and is an isometry with respect to a Riemannian metric of constant curvature on M;
2. g is a "reducible" diffeomorphism, that is, a diffeomorphism leaving invariant a closed curve;
3. g is a *pseudo-Anosov map*.

Pseudo-Anosov maps are surface homeomorphisms that are differentiable except at most at finitely many points called *singularities*. These maps minimize both the number of periodic points (of any given period) and the topological entropy in their isotopy classes. A pseudo-Anosov map is Bernoulli with respect to an absolutely continuous invariant measure with C^∞ density which is positive except at the singularities (see [95, Exposé 10]).

We proceed with a formal description. Let $\{x_1, \dots, x_m\}$ be a finite set of points and ν a Borel measure on M. Write $\mathcal{D}_a = \{z \in \mathbb{C}\colon\ |z| < a\}$.

We say that $(\mathcal{F}, \nu)$ is a *measured foliation* of M with *singular points* $x_1, \dots, x_m$ if $\mathcal{F}$ is a partition of M for which the following properties hold:

1. there is a collection of C^∞ charts $\varphi_k : U_k \to \mathbb{C}$ for $k = 1, \dots, \ell$ and some $\ell \geqslant m$ with $\bigcup_{k=1}^{\ell} U_k = M$;
2. for each $k = 1, \dots, m$ there is a number $p = p(k) \geqslant 3$ of elements of $\mathcal{F}$ meeting at x_k such that:
   (a) $\varphi_k(x_k) = 0$ and $\varphi_k(U_k) = \mathcal{D}_{a_k}$ for some $a_k > 0$;
   (b) if C is an element of $\mathcal{F}$ then $C \cap U_k$ is mapped by φ_k to a set

$$\left\{z\colon\ \mathrm{Im}\left(z^{p/2}\right) = \text{constant}\right\} \cap \varphi_k(U_k);$$

(c) the measure $\nu|U_k$ is the pullback under φ_k of

$$\left|\operatorname{Im}\left(dz^{p/2}\right)\right| = \left|\operatorname{Im}\left(z^{(p-2)/2}\,dz\right)\right|;$$

3. for each $k > m$ we have:
   (a) $\varphi_k(U_k) = (0, b_k) \times (0, c_k) \subset \mathbb{R}^2 \equiv \mathbb{C}$ for some $b_k, c_k > 0$;
   (b) if C is an element of $\mathcal{F}$ then $C \cap U_k$ is mapped by φ_k to a segment

$$\left\{(x, y)\colon\ y = \text{constant}\right\} \cap \varphi_k(U_k);$$

   (c) the measure $\nu|U_k$ is given by the pullback of $|dy|$ under φ_k.

The elements of $\mathcal{F}$ are called *leaves* of the foliation, and ν a *transverse* measure. For $k = 1, \ldots, m$, each point x_k is called a $p(k)$*-prong singularity* of $\mathcal{F}$ and each of the leaves of $\mathcal{F}$ meeting at x_k is called a *prong* of x_k. If, in addition, we allow single leaves of $\mathcal{F}$ to terminate in a point (called a *spine*, in which case we set $p = 1$ above), then $(\mathcal{F}, \nu)$ is called a *measured foliation with spines*.

The transverse measure is consistently defined on chart overlaps, because whenever $U_j \cap U_k \neq \emptyset$, the transition functions $\varphi_k \circ \varphi_j^{-1}$ are of the form

$$\left(\varphi_k \circ \varphi_j^{-1}\right)(x, y) = \left(h_{jk}(x, y), c_{jk} \pm y\right),$$

where h_{jk} is a function, and c_{jk} is a constant.

A surface homeomorphism f is called *pseudo-Anosov* if it satisfies the following properties:

1. f is differentiable except at a finite number of points $x_1, \ldots, x_m$;
2. there are two measured foliations $(\mathcal{F}^s, \nu^s)$ and $(\mathcal{F}^u, \nu^u)$ with the same singularities $x_1, \ldots, x_m$ and the same number of prongs $p = p(k)$ at each point x_k, for $k = 1, \ldots, m$;
3. the leaves of the foliations $\mathcal{F}^s$ and $\mathcal{F}^u$ are transversal at nonsingular points;
4. there are C^∞ charts $\varphi_k : U_k \to \mathbb{C}$ for $k = 1, \ldots, \ell$ and some $\ell \geqslant m$, such that for each k we have:
   (a) $\varphi_k(x_k) = 0$ and $\varphi_k(U_k) = \mathcal{D}_{a_k}$ for some $a_k > 0$;
   (b) leaves of $\mathcal{F}^s$ are mapped by φ_i to components of the sets

$$\left\{z\colon\ \operatorname{Re} z^{p/2} = \text{constant}\right\} \cap \mathcal{D}_{a_k};$$

   (c) leaves of $\mathcal{F}^u$ are mapped by φ_i to components of the sets

$$\left\{z\colon\ \operatorname{Im}\left(z^{p/2}\right) = \text{constant}\right\} \cap \mathcal{D}_{a_k};$$

   (d) there exists a constant $\lambda > 1$ such that

$$f(\mathcal{F}^s, \nu^s) = (\mathcal{F}^s, \nu^s/\lambda) \quad \text{and} \quad f(\mathcal{F}^u, \nu^u) = (\mathcal{F}^u, \lambda\nu^u).$$

If, in addition, $(\mathcal{F}^s, \nu^s)$ and $(\mathcal{F}^u, \nu^u)$ are measured foliations with spines (with $p = p(k) = 1$ when there is only one prong at x_k), then f is called a *generalized pseudo-Anosov homeomorphism*.

We call $\mathcal{F}^s$ and $\mathcal{F}^u$ the *stable* and *unstable foliations*, respectively. At each singular point x_k, with $p = p(k)$, the *stable* and *unstable prongs* are, respectively, given by

$$P_{kj}^s = \varphi_k{}^{-1}\left\{\rho e^{i\tau}\colon\ 0 \leqslant \rho < a_k,\ \tau = \frac{2j+1}{p}\pi\right\},$$

$$P_{kj}^u = \varphi_k{}^{-1}\left\{\rho e^{i\tau}\colon\ 0 \leqslant \rho < a_k,\ \tau = \frac{2j}{p}\pi\right\},$$

for $j = 0, 1, \ldots, p-1$. We define the *stable* and *unstable sectors* at x_k by

$$S_{kj}^s = \varphi_k{}^{-1}\left\{\rho e^{i\tau}\colon\ 0 \leqslant \rho < a_k,\ \frac{2j-1}{p}\pi \leqslant \tau \leqslant \frac{2j+1}{p}\pi\right\},$$

$$S_{kj}^u = \varphi_k{}^{-1}\left\{\rho e^{i\tau}\colon\ 0 \leqslant \rho < a_k,\ \frac{2j}{p}\pi \leqslant \tau \leqslant \frac{2j+2}{p}\pi\right\},$$

respectively, for $j = 0, 1, \ldots, p-1$.

Since f is a homeomorphism, $f(x_k) = x_{\sigma_k}$ for $k = 1, \ldots, m$, where σ is a permutation of $\{1, \ldots, m\}$ such that $p(k) = p(\sigma_k)$ and f maps the stable prongs at x_k into the stable prongs at x_{σ_k} (provided the numbers a_k are chosen such that $a_k/\lambda^{2/p} \leqslant a_{\sigma_k}$). Hence, we may assume that σ is the identity permutation, and

$$f(P_{kj}^s) \subset P_{kj}^s \quad \text{and} \quad f^{-1}(P_{kj}^u) \subset P_{kj}^u$$

for $k = 1, \ldots, m$ and $j = 0, \ldots, p-1$. Consider the map

$$\Phi_{kj} : \varphi_k(S_{kj}^s) \to \{z\colon\ \operatorname{Re} z \geqslant 0\},$$

given by

$$\Phi_{kj}(z) = 2z^{p/2}/p,$$

where $p = p(k)$. Write $\Phi_{kj}(z) = s_1 + is_2$ and $z = t_1 + it_2$, where s_1, s_2, t_1, t_2 are real numbers. Define a measure ν on each stable sector by

$$d\nu | S_{kj}^s = \varphi_k^* \Phi_{kj}^* (ds_1\, ds_2)$$

if $k = 1, \ldots, m$, $j = 0, \ldots, p(i) - 1$, and on each "nonsingular" neighborhood by

$$d\nu | U_k = \varphi_k^* (dt_1\, dt_2)$$

if $k > m$. The measure ν can be extended to an f-invariant measure with the following properties:

1. ν is equivalent to the Lebesgue measure on M; moreover, ν has a density which is smooth everywhere except at the singular points x_k, where it vanishes if $p(k) \geqslant 3$, and goes to infinity if $p(k) = 1$;
2. f is Bernoulli with respect to ν (see [95, Section 10]).

One can show that the periodic points of any pseudo-Anosov map are dense.

If M is a torus, then any pseudo-Anosov map is an Anosov diffeomorphism (see [95, Exposé 1]). However, if M has genus greater than 1, a pseudo-Anosov map cannot be made a diffeomorphism by a coordinate change which is smooth outside the singularities or even outside a sufficiently small neighborhood of the singularities (see [105]). Thus, in order to find smooth models of pseudo-Anosov maps one may have to apply some non-trivial construction which is global in nature. In [105], Gerber and Katok constructed, for every pseudo-Anosov map g, a C^∞ diffeomorphism which is topologically conjugate to f through a homeomorphism isotopic to the identity and which is Bernoulli with respect to a smooth measure (that is, a measure whose density is C^∞ and positive everywhere).

In [104], Gerber proved the existence of real analytic Bernoulli models of pseudo-Anosov maps as an application of a conditional stability result for the smooth models constructed in [105]. The proofs rely on the use of Markov partitions. The same results were obtained by Lewowicz and Lima de Sá [164] using a different approach.

2.6. *Flows with nonzero exponents*

The first example of a volume-preserving ergodic flow with nonzero Lyapunov exponents, which is not an Anosov flow, was constructed by Pesin in [196]. The construction is a "surgery" of an Anosov flow and is based on slowing down trajectories near a given trajectory of the Anosov flow.

Let φ_t be an Anosov flow on a compact three-dimensional manifold M given by a vector field X and preserving a smooth ergodic measure μ. Fix a point $p_0 \in M$. There is a coordinate system x, y, z in a ball $B(p_0, d)$ (for some $d > 0$) such that p_0 is the origin (i.e., $p_0 = 0$) and $X = \partial/\partial z$.

For each $\varepsilon > 0$, let $T_\varepsilon = S^1 \times D_\varepsilon \subset B(0, d)$ be the solid torus obtained by rotating the disk

$$D_\varepsilon = \left\{(x, y, z) \in B(0, d):\ x = 0 \text{ and } (y - d/2)^2 + z^2 \leqslant (\varepsilon d)^2\right\}$$

around the z-axis. Every point on the solid torus can be represented as (θ, y, z) with $\theta \in S^1$ and $(y, z) \in D_\varepsilon$.

For every $0 \leqslant \alpha \leqslant 2\pi$, consider the cross-section of the solid torus $\Pi_\alpha = \{(\theta, y, z): \theta = \alpha\}$. We construct a new vector field $\tilde{X}$ on $M \setminus T_\varepsilon$. We describe the construction of $\tilde{X}$ on the cross-section Π_0 and we obtain the desired vector field $\tilde{X}|\Pi_\alpha$ on an arbitrary cross-section by rotating it around the z-axis.

Consider the Hamiltonian flow given by the Hamiltonian $H(y, z) = y(\varepsilon^2 - y^2 - z^2)$. In the annulus $\varepsilon^2 \leqslant y^2 + z^2 \leqslant 4\varepsilon^2$ the flow is topologically conjugated to the one shown in Figure 1. However, the Hamiltonian vector field $(-2yz, 3y^2 + z^2 - \varepsilon^2)$ is not everywhere vertical on the circle $y^2 + z^2 = 4\varepsilon^2$. To correct this consider a C^∞ function $p: [\varepsilon, \infty) \to$

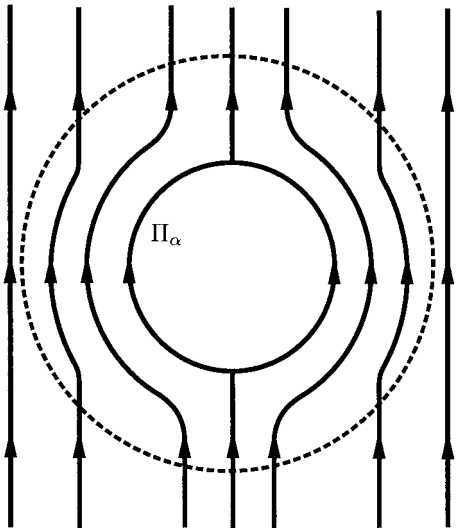


Fig. 1. A cross-section Π_α and the flow $\tilde{\varphi}_t$.

$[0, 1]$ such that $p|[\varepsilon, 3\varepsilon/2] = 1$, $p|[2\varepsilon, \infty) = 0$, and p is strictly decreasing in $(3\varepsilon/2, 2\varepsilon)$. The flow defined by the system of differential equations

$$\begin{cases} y' = -2yzp(\sqrt{y^2+z^2}), \\ z' = (3y^2+z^2-\varepsilon^2)p(\sqrt{y^2+z^2}) + 1 - p(\sqrt{y^2+z^2}), \end{cases}$$

has now the behavior shown in Figure 1. Denote by $\bar{\varphi}_t$ and $\bar{X}$ the corresponding flow and vector field in coordinates x, y, z.

By changing the time one can obtain a flow $\bar{\varphi}_t$ in the annulus $\varepsilon^2 \leqslant y^2+z^2 \leqslant 4\varepsilon^2$ so that the new flow $\tilde{\varphi}_t$ preserves the measure μ. As a result we have a smooth vector field $\tilde{X}$ on $M \setminus T_\varepsilon$ such that the flow $\tilde{\varphi}_t$ generated by $\tilde{X}$ has the following properties:

1. $\tilde{X}|(M \setminus T_{2\varepsilon}) = X|(M \setminus T_{2\varepsilon})$;
2. for any $0 \leqslant \alpha, \beta \leqslant 2\pi$, the vector field $\tilde{X}|\Pi_\beta$ is the image of the vector field $\tilde{X}|\Pi_\alpha$ under the rotation around the z-axis that moves Π_α onto Π_β;
3. for every $0 \leqslant \alpha \leqslant 2\pi$, the unique two fixed points of the flow $\tilde{\varphi}_t|\Pi_\alpha$ are those in the intersection of Π_α with the hyperplanes $z = \pm\varepsilon d$;
4. for every $0 \leqslant \alpha \leqslant 2\pi$ and $(y, z) \in D_{2\varepsilon} \setminus \operatorname{int} D_\varepsilon$, the trajectory of the flow $\tilde{\varphi}_t|\Pi_\alpha$ passing through the point (y, z) is invariant under the symmetry $(\alpha, y, z) \mapsto (\alpha, y, -z)$;
5. the flow $\tilde{\varphi}_t|\Pi_\alpha$ preserves the conditional measure induced by the measure μ on the set Π_α.

The orbits of the flows φ_t and $\tilde{\varphi}_t$ coincide on $M \setminus T_{2\varepsilon}$, the flow $\tilde{\varphi}_t$ preserves the measure μ and the only fixed points of this flow are those on the circles $\{(\theta, y, z)\colon z = -\varepsilon d\}$ and $\{(\theta, y, z)\colon z = \varepsilon d\}$.

On $T_{2\varepsilon} \setminus \operatorname{int} T_\varepsilon$ consider the new coordinates θ_1, θ_2, r with $0 \leqslant \theta_1, \theta_2 < 2\pi$ and $\varepsilon d \leqslant r \leqslant 2\varepsilon d$ such that the set of fixed points of $\tilde{\varphi}_t$ consists of those for which $r = \varepsilon d$, and $\theta_1 = 0$ or $\theta_1 = \pi$.

Define the flow on $T_{2\varepsilon} \setminus \operatorname{int} T_\varepsilon$ by

$$(\theta_1, \theta_2, r, t) \mapsto \big(\theta_1, \theta_2 + \big[2 - r/(\varepsilon d)\big]^4 t \cos\theta_1, r\big),$$

and let $\hat{X}$ be the corresponding vector field. Consider the flow ψ_t on $M \setminus \operatorname{int} T_\varepsilon$ generated by the vector field Y on $M \setminus \operatorname{int} T_\varepsilon$,

$$Y(x) = \begin{cases} X(x), & x \in M \setminus \operatorname{int} T_{2\varepsilon}, \\ \tilde{X}(x) + \hat{X}(x), & x \in \operatorname{int} T_{2\varepsilon} \setminus \operatorname{int} T_\varepsilon. \end{cases}$$

The flow ψ_t has no fixed points, preserves the measure μ and for μ-almost every $x \in M \setminus T_{2\varepsilon}$,

$$\chi(x, v) < 0 \quad \text{if } v \in E^s(x) \quad \text{and} \quad \chi(x, v) > 0 \quad \text{if } v \in E^u(x),$$

where $E^u(x)$ and $E^s(x)$ are respectively stable and unstable subspaces of the Anosov flow φ_t at x.

Set $M_1 = M \setminus T_\varepsilon$ and consider a copy $(\tilde{M}_1, \tilde{\psi}_t)$ of the flow (M_1, ψ_t). Gluing the manifolds M_1 and $\tilde{M}_1$ along their boundaries ∂T_ε one obtains a three-dimensional smooth Riemannian manifold D without boundary. We define a flow F_t on D by

$$F_t x = \begin{cases} \psi_t x, & x \in M_1, \\ \tilde{\psi}_t x, & x \in \tilde{M}_1. \end{cases}$$

It is clear that the flow F_t is smooth and preserves a smooth hyperbolic measure.

2.7. *Geodesic flows*

Our next example is the geodesic flow on a compact smooth Riemannian manifold of nonpositive curvature. Let M be a compact smooth p-dimensional Riemannian manifold with a Riemannian metric of class C^3.

The *geodesic flow* g_t acts on the tangent bundle TM by the formula

$$g_t(v) = \dot{\gamma}_v(t),$$

where $\dot{\gamma}_v(t)$ is the unit tangent vector to the geodesic $\gamma_v(t)$ defined by the vector v (i.e., $\dot{\gamma}_v(0) = v$; this geodesic is uniquely defined). The geodesic flow generates a vector field V on TM given by

$$V(v) = \frac{d(g_t(v))}{dt}\bigg|_{t=0}.$$

Since M is compact the flow g_t is well defined for all $t \in \mathbb{R}$ and is a smooth flow.

We recall some basic notions from Riemannian geometry of nonpositively curved manifolds (see [92,91] for a detailed exposition). We endow the second tangent space $T(TM)$ with a special Riemannian metric. Let $\pi : TM \to M$ be the natural projection (i.e., $\pi(x, v) = x$ for each $x \in M$ and each $v \in T_x M$) and $K : T(TM) \to TM$ the linear (connection) operator defined by $K\xi = (\nabla Z)(t)|_{t=0}$, where $Z(t)$ is any curve in TM such that $Z(0) = d\pi\xi$, $\frac{d}{dt}Z(t)|_{t=0} = \xi$ and ∇ is the covariant derivative. The *canonical metric* on $T(TM)$ is given by

$$\langle \xi, \eta \rangle_v = \langle d_v\pi\xi, d_v\pi\eta \rangle_{\pi v} + \langle K\xi, K\eta \rangle_{\pi v}.$$

The set $SM \subset TM$ of the unit vectors is invariant with respect to the geodesic flow, and is a compact manifold of dimension $2p - 1$. In what follows we consider the geodesic flow restricted to SM.

The study of hyperbolic properties of the geodesic flow is based upon the description of solutions of the variational equation for the flow. This equation along a given trajectory $g_t(v)$ of the flow is the *Jacobi equation* along the geodesic $\gamma_v(t)$:

$$Y''(t) + R_{XY}X(t) = 0. \tag{2.5}$$

Here $Y(t)$ is a vector field along $\gamma_v(t)$, $X(t) = \dot{\gamma}(t)$, and R_{XY} is the curvature operator. More precisely, the relation between the variational equations and the Jacobi equation (2.5) can be described as follows. Fix $v \in SM$ and $\xi \in T_v SM$. Let $Y_\xi(t)$ be the unique solution of (2.5) satisfying the initial conditions $Y_\xi(0) = d_v\pi\xi$ and $Y'_\xi(0) = K\xi$. One can show that the map $\xi \mapsto Y_\xi(t)$ is an isomorphism for which $d_{g_t v}\pi d_v g_t \xi = Y_\xi(t)$ and $K d_v g_t \xi = Y'_\xi(t)$. This map establishes the identification between solutions of the variational equation and solutions of the Jacobi equation (2.5).

Recall that the Fermi coordinates $\{e_i(t)\}$, for $i = 1, \ldots, p$, along the geodesic $\gamma_v(t)$ are obtained by the time t parallel translation along $\gamma_v(t)$ of an orthonormal basis $\{e_i(0)\}$ in $T_{\gamma_v(0)}M$ where $e_1(t) = \dot{\gamma}(t)$. Using these coordinates we can rewrite Equation (2.5) in the matrix form

$$\frac{d^2}{dt^2}A(t) + K(t)A(t) = 0, \tag{2.6}$$

where $A(t) = (a_{ij}(t))$ and $K(t) = (k_{ij}(t))$ are matrix functions with entries $k_{ij}(t) = K_{\gamma_v(t)}(e_i(t), e_j(t))$.

Two points $x = \gamma(t_1)$ and $y = \gamma(t_2)$ on the geodesic γ are called *conjugate* if there exists a nonidentically zero Jacobi field Y along γ such that $Y(t_1) = Y(t_2) = 0$. Two points $x = \gamma(t_1)$ and $y = \gamma(t_2)$ are called *focal* if there exists a Jacobi field Y along γ such that $Y(t_1) = 0$, $Y'(t_1) \neq 0$ and $\frac{d}{dt}\|Y(t)\|^2|_{t=t_2} = 0$.

We say that the manifold M has:

1. *no conjugate points* if on each geodesic no two points are conjugate;
2. *no focal points* if on each geodesic no two points are focal;

3. *nonpositive curvature* if for any $x \in M$ and any two vectors $v_1, v_2 \in T_x M$ the sectional curvature $K_x(v_1, v_2)$ satisfies

$$K_x(v_1, v_2) \leqslant 0. \tag{2.7}$$

If the manifold has no focal points then it has no conjugate points and if it has nonpositive curvature then it has no focal points.

From now on we consider only manifolds with no conjugate points. The boundary value problem for Equation (2.6) has a unique solution, i.e., for any numbers s_1, s_2 and any matrices A_1, A_2 there exists a unique solution $A(t)$ of (2.6) satisfying $A(s_1) = A_1$ and $A(s_2) = A_2$.

PROPOSITION 2.1 (Eberlein [90]). *Given $s \in \mathbb{R}$, let $A_s(t)$ be the unique solution of Equation* (2.6) *satisfying the boundary conditions*: $A_s(0) = \mathrm{Id}$ (*where* Id *is the identity matrix*) *and $A_s(s) = 0$. Then there exists the limit*

$$\lim_{s \to \infty} \frac{d}{dt} A_s(t)\bigg|_{t=0} = A^+.$$

We define the *positive limit solution* $A^+(t)$ of (2.6) as the solution that satisfies the initial conditions:

$$A^+(0) = \mathrm{Id} \quad \text{and} \quad \frac{d}{dt} A^+(t)\bigg|_{t=0} = A^+.$$

This solution is nondegenerate (i.e., $\det A^+(t) \neq 0$ for every $t \in \mathbb{R}$) and $A^+(t) = \lim_{s \to +\infty} A_s(t)$.

Similarly, letting $s \to -\infty$, define the *negative limit solution* $A^-(t)$ of Equation (2.6).

For every $v \in SM$ set

$$E^+(v) = \big\{\xi \in T_v SM\colon \langle \xi, V(v)\rangle = 0 \text{ and } Y_\xi(t) = A^+(t) d_v \pi \xi\big\}, \tag{2.8}$$

$$E^-(v) = \big\{\xi \in T_v SM\colon \langle \xi, V(v)\rangle = 0 \text{ and } Y_\xi(t) = A^-(t) d_v \pi \xi\big\}, \tag{2.9}$$

where V is the vector field generated by the geodesic flow.

PROPOSITION 2.2 (Eberlein [90]). *The following properties hold*:

1. *the sets $E^-(v)$ and $E^+(v)$ are linear subspaces of $T_v SM$*;
2. $\dim E^-(v) = \dim E^+(v) = p - 1$;
3. $d_v \pi E^-(v) = d_v \pi E^+(v) = \{w \in T_{\pi v} M\colon w \text{ is orthogonal to } v\}$;
4. *the subspaces $E^-(v)$ and $E^+(v)$ are invariant under the differential $d_v g_t$, i.e., $d_v g_t E^-(v) = E^-(g_t v)$ and $d_v g_t E^+(v) = E^+(g_t v)$*;
5. *if $\tau : SM \to SM$ is the involution defined by $\tau v = -v$, then*

$$E^+(-v) = d_v \tau E^-(v) \quad \text{and} \quad E^-(-v) = d_v \tau E^+(v);$$

6. *if* $K_x(v_1, v_2) \geqslant -a^2$ *for some* $a > 0$ *and all* $x \in M$, *then* $\|K\xi\| \leqslant a\|d_v\pi\xi\|$ *for every* $\xi \in E^+(v)$ *and* $\xi \in E^-(v)$;
7. *if* $\xi \in E^+(v)$ *or* $\xi \in E^-(v)$, *then* $Y_\xi(t) \neq 0$ *for every* $t \in \mathbb{R}$;
8. $\xi \in E^+(v)$ (*respectively,* $\xi \in E^-(v)$) *if and only if*

$$\langle \xi, V(v) \rangle = 0 \quad \textit{and} \quad \|d_{g_t v}\pi d_v g_t \xi\| \leqslant c$$

for every $t > 0$ (*respectively,* $t < 0$) *and some* $c > 0$;
9. *if the manifold has no focal points then for any* $\xi \in E^+(v)$ (*respectively,* $\xi \in E^-(v)$) *the function* $t \mapsto \|Y_\xi(t)\|$ *is nonincreasing* (*respectively, nondecreasing*).

In view of properties 6 and 8, we have $\xi \in E^+(v)$ (respectively, $\xi \in E^-(v)$) if and only if $\langle \xi, V(v) \rangle = 0$ and $\|d_v g_t \xi\| \leqslant c$ for $t > 0$ (respectively, $t < 0$), for some constant $c > 0$. This observation and property 4 justify to call $E^+(v)$ and $E^-(v)$ the *stable* and *unstable subspaces*.

In general, the subspaces $E^-(v)$ and $E^+(v)$ do not span the whole second tangent space $T_v SM$. Eberlein (see [90]) has shown that if they do span $T_v SM$ for every $v \in SM$, then the geodesic flow is Anosov. This is the case when the curvature is strictly negative. For a general manifold without conjugate points consider the set

$$\Delta = \left\{ v \in SM\colon \varlimsup_{t\to\infty} \frac{1}{t} \int_0^t K_{\pi(g_s v)}(g_s v, g_s w)\, ds < 0 \right. \\ \left. \text{for every } w \in SM \text{ orthogonal to } v \right\}. \tag{2.10}$$

It is easy to see that Δ is measurable and invariant under g_t. The following result shows that the Lyapunov exponents are nonzero on the set Δ.

THEOREM 2.3 (Pesin [198]). *Assume that the Riemannian manifold* M *has no conjugate points. Then for every* $v \in \Delta$ *we have* $\chi(v, \xi) < 0$ *if* $\xi \in E^+(v)$ *and* $\chi(v, \xi) > 0$ *if* $\xi \in E^-(v)$.

The geodesic flow preserves the Liouville measure μ on the tangent bundle. Denote by m the Lebesgue measure on M. It follows from Theorem 2.3 that if the set Δ has positive Liouville measure then the geodesic flow $g_t|\Delta$ has nonzero Lyapunov exponents almost everywhere. It is, therefore, crucial to find conditions which would guarantee that Δ has positive Liouville measure.

We first consider the two-dimensional case.

THEOREM 2.4 (Pesin [198]). *Let* M *be a smooth compact surface of nonpositive curvature* $K(x)$ *and genus greater than* 1. *Then* $\mu(\Delta) > 0$.

In the multi-dimensional case one can establish the following criterion for positivity of the Liouville measure of the set Δ.

THEOREM 2.5 (Pesin [198]). *Let M be a smooth compact Riemannian manifold of non-positive curvature. Assume that there exist $x \in M$ and a vector $v \in S_x M$ such that*

$$K_x(v, w) < 0$$

for any vector $w \in S_x M$ which is orthogonal to v. Then $\mu(\Delta) > 0$.

One can show that if $\mu(\Delta) > 0$ then the set Δ is open (mod 0) and is everywhere dense (see Theorem 17.7 below).

3. Lyapunov exponents associated with sequences of matrices

In studying the stability of trajectories of a dynamical system f one introduces the system of variations $\{d_x f^m,\ m \in \mathbb{Z}\}$ and uses the Lyapunov exponents for this systems (see Section 1). Consider a family of trivializations τ_x of M, i.e., linear isomorphisms $\tau_x : T_x M \to \mathbb{R}^n$ where $n = \dim M$. The sequence of matrices

$$A_m = \tau_{f^{m+1}(x)} \circ d_{f^m(x)} f \circ \tau_{f^m(x)}^{-1} : \mathbb{R}^n \to \mathbb{R}^n$$

can also be used to study the linear stability along the trajectory $f^m(x)$.

In this section we extend our study of Lyapunov exponents for sequences of matrices generated by smooth dynamical systems to arbitrary sequences of matrices. This will also serve as an important intermediate step in studying Lyapunov exponents for the even more general case of cocycles over dynamical systems.

3.1. *Definition of the Lyapunov exponent*

Let $\mathcal{A}^+ = \{A_m\}_{m \geqslant 0} \subset GL(n, \mathbb{R})$ be a one-sided sequence of matrices. Set $\mathcal{A}_m = A_{m-1} \dots A_1 A_0$ and consider the function $\chi^+ : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ given by

$$\chi^+(v) = \chi^+(v, \mathcal{A}^+) = \overline{\lim_{m \to +\infty}} \frac{1}{m} \log \|\mathcal{A}_m v\|. \tag{3.1}$$

We make the convention $\log 0 = -\infty$, so that $\chi^+(0) = -\infty$.

The function $\chi^+(v)$ is called the *forward Lyapunov exponent of v* (*with respect to the sequence $\mathcal{A}^+$*). It has the following basic properties:

1. $\chi^+(\alpha v) = \chi^+(v)$ for each $v \in \mathbb{R}^n$ and $\alpha \in \mathbb{R} \setminus \{0\}$;
2. $\chi^+(v + w) \leqslant \max\{\chi^+(v), \chi^+(w)\}$ for each $v, w \in \mathbb{R}^n$;
3. $\chi^+(0) = -\infty$.

As an immediate consequence of the basic properties we obtain that there exist a positive integer p^+, $1 \leqslant p^+ \leqslant n$, a collection of numbers $\chi_1 < \chi_2 < \dots < \chi_{p^+}$, and linear subspaces

$$\{0\} = V_0 \subsetneqq V_1 \subsetneqq V_2 \subsetneqq \dots \subsetneqq V_{p^+} = \mathbb{R}^n$$

such that $V_i = \{v \in \mathbb{R}^n\colon\ \chi^+(v) \leqslant \chi_i\}$, and if $v \in V_i \setminus V_{i-1}$, then $\chi^+(v) = \chi_i$ for each $i = 1, \dots, p^+$. The spaces V_i form the *filtration* $\mathcal{V}_{\chi^+}$ of $\mathbb{R}^n$ associated with χ^+. The number

$$k_i = \dim V_i - \dim V_{i-1}$$

is called the *multiplicity* of the value χ_i, and the collection of pairs

$$\operatorname{Sp}\chi^+ = \left\{(\chi_i, k_i)\colon\ i = 1, \dots, p^+\right\}$$

the *Lyapunov spectrum* of χ^+. We also set

$$n_i = \dim V_i = \sum_{j=1}^{i} k_j.$$

In a similar way, given a sequence of matrices $\mathcal{A}^- = \{A_m\}_{m<0}$, define the *backward Lyapunov exponent* (*with respect to the sequence* $\mathcal{A}^-$) $\chi^-\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ by

$$\chi^-(v) = \chi^-(v, \mathcal{A}^-) = \overline{\lim_{m\to-\infty}} \frac{1}{|m|} \log \|\mathcal{A}_m v\|, \tag{3.2}$$

where $\mathcal{A}_m = (A_m)^{-1} \dots (A_{-2})^{-1}(A_{-1})^{-1}$ for each $m < 0$. Let $\chi_1^- > \dots > \chi_{p^-}^-$ be the *values* of χ^-, for some integer $1 \leqslant p^- \leqslant n$. The subspaces

$$\mathbb{R}^n = V_1^- \supsetneqq \dots \supsetneqq V_{p^-}^- \supsetneqq V_{p^-+1}^- = \{0\},$$

where $V_i^- = \{v \in \mathbb{R}^n\colon\ \chi^-(v) \leqslant \chi_i^-\}$, form the *filtration* $\mathcal{V}_{\chi^-}$ of $\mathbb{R}^n$ associated with χ^-. The number

$$k_i^- = \dim V_i^- - \dim V_{i+1}^-$$

is the *multiplicity* of the value χ_i^-, and the collection of pairs

$$\operatorname{Sp}\chi^- = \left\{(\chi_i^-, k_i^-)\colon\ i = 1, \dots, p^-\right\}$$

is the *Lyapunov spectrum* of χ^-.

In the case when the sequence of matrices is obtained by iterating a given matrix A, i.e., $\mathcal{A}_m = A^m$ the Lyapunov spectrum is calculated as follows. Take all the eigenvalues with absolute value r. Then $\log r$ is a value of the Lyapunov exponent and the multiplicity is equal to the sum of the multiplicities of the exponents with this absolute value.

Equality (3.1) implies that for every $\varepsilon > 0$ there exists $C_+ = C_+(v, \varepsilon) > 0$ such that if $m \geqslant 0$ then

$$\|\mathcal{A}_m v\| \leqslant C_+ e^{(\chi^+(v)+\varepsilon)m} \|v\|. \tag{3.3}$$

Similarly, (3.2) implies that for every $\varepsilon > 0$ there exists $C_- = C_-(v, \varepsilon) > 0$ such that if $m \geqslant 0$ then

$$\|\mathcal{A}_{-m} v\| \leqslant C_- e^{(\chi^-(v)+\varepsilon)m} \|v\|. \tag{3.4}$$

Given vectors $v_1, \ldots, v_k \in \mathbb{R}^n$, we denote by $V(v_1, \ldots, v_k)$ the volume of the k-parallelepiped formed by $v_1, \ldots, v_k$. The *forward* and *backward k-dimensional Lyapunov exponents of the vectors* $v_1, \ldots, v_k$ are defined, respectively, by

$$\chi^+(v_1, \ldots, v_k) = \chi^+(v_1, \ldots, v_k, \mathcal{A}^+) = \overline{\lim_{m \to +\infty}} \frac{1}{m} \log V(\mathcal{A}_m v_1, \ldots, \mathcal{A}_m v_k),$$

$$\chi^-(v_1, \ldots, v_k) = \chi^-(v_1, \ldots, v_k, \mathcal{A}^-) = \overline{\lim_{m \to -\infty}} \frac{1}{|m|} \log V(\mathcal{A}_m v_1, \ldots, \mathcal{A}_m v_k).$$

These exponents depend only on the linear space generated by the vectors $v_1, \ldots, v_k$. Since $V(v_1, \ldots, v_k) \leqslant \prod_{i=1}^k \|v_i\|$ we obtain

$$\chi^+(v_1, \ldots, v_k) \leqslant \sum_{i=1}^k \chi^+(v_i). \tag{3.5}$$

A similar inequality holds for the backward Lyapunov exponent.

The inequality (3.5) can be strict. Indeed, consider the sequence of matrices $A_m = \left(\begin{smallmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{smallmatrix}\right)$ and the vectors $v_1 = (1, 0)$, $v_2 = (1, 1)$. We have $\chi^+(v_1) = \chi^+(v_2) = \log 2$. On the other hand, since $\det A_m = 1$, we have $\chi^+(v_1, v_2) = 0 < \chi^+(v_1) + \chi^+(v_2)$.

3.2. *Forward and backward regularity*

We say that a sequence of matrices $\mathcal{A}^+$ is *forward regular* if

$$\lim_{m \to +\infty} \frac{1}{m} \log |\det \mathcal{A}_m| = \sum_{i=1}^n \chi_i', \tag{3.6}$$

where $\chi_1', \ldots, \chi_n'$ are the finite values of the exponent χ^+ counted with their multiplicities. By (3.5), this is equivalent to

$$\underline{\lim_{m \to +\infty}} \frac{1}{m} \log |\det \mathcal{A}_m| \geqslant \sum_{i=1}^n \chi_i'.$$

The forward regularity is equivalent to the statement that there exists a positive definite symmetric matrix Λ such that

$$\lim_{m \to \infty} \|\mathcal{A}_m \Lambda^{-m}\| = 0, \qquad \lim_{m \to \infty} \left\| \Lambda^m \mathcal{A}_m^{-1} \right\| = 0. \tag{3.7}$$

Let $\mathcal{A}^+ = \{A_m\}_{m\geqslant 0}$ and $\mathcal{B}^+ = \{B_m\}_{m\geqslant 0}$ be two sequences of matrices. They are called *equivalent* if there is a nondegenerate matrix C such that $A_m = C^{-1}B_mC$ for every $m \geqslant 0$. Any sequence of linear transformations, which is equivalent to a forward regular sequence, is itself forward regular.

The Lyapunov exponent χ^+ is said to be

1. *exact with respect to the collection of vectors* $\{v_1, \dots, v_k\} \subset \mathbb{R}^n$ if

$$\chi^+(v_1, \dots, v_k) = \lim_{m\to+\infty} \frac{1}{m} \log V(\mathcal{A}_m v_1, \dots, \mathcal{A}_m v_k);$$

2. *exact* if for any $1 \leqslant k \leqslant n$, the exponent χ^+ is exact with respect to every collection of vectors $\{v_1, \dots, v_k\} \subset \mathbb{R}^n$.

If the Lyapunov exponent is exact then in particular for $v \in \mathbb{R}^n$ one has

$$\chi^+(v) = \lim_{m\to+\infty} \frac{1}{m} \log \|\mathcal{A}_m v\|;$$

equivalently (compare with (3.3) and (3.4)): for every $\varepsilon > 0$ there exists $C = C(v, \varepsilon) > 0$ such that if $m \geqslant 0$ then

$$C^{-1}e^{(\chi^+(v)-\varepsilon)m} \leqslant \|\mathcal{A}_m v\| \leqslant Ce^{(\chi^+(v)+\varepsilon)m}.$$

THEOREM 3.1 (Lyapunov [171]). *If the sequence of matrices $\mathcal{A}^+$ is forward regular, then the Lyapunov exponent χ^+ is exact.*

The following simple example demonstrates that the Lyapunov exponent χ^+ may be exact even for a sequence of matrices which is not forward regular. In other words, the existence of the limit in (3.1) does *not* guarantee that the Lyapunov exponent χ^+ is forward regular.

EXAMPLE 3.2. Let $\mathcal{A}^+ = \{A_m\}_{m\geqslant 0}$ be the sequence of matrices where $A_0 = \left(\begin{smallmatrix} 1 & 0 \\ 2 & 4 \end{smallmatrix}\right)$ and $A_m = \left(\begin{smallmatrix} 1 & 0 \\ -2^{m+1} & 4 \end{smallmatrix}\right)$ for each $m \geqslant 1$ so that $\mathcal{A}_m = \left(\begin{smallmatrix} 1 & 0 \\ 2^m & 4^m \end{smallmatrix}\right)$ for every $m \geqslant 1$. Given a vector $v = (a, b) \neq (0, 0)$ we have $\chi^+(v) = \log 2$ if $b = 0$, and $\chi^+(v) = \log 4$ if $b \neq 0$. This implies that χ^+ is exact with respect to every vector v. Let $v_1 = (1, 0)$ and $v_2 = (0, 1)$. Then $\chi^+(v_1) = \log 2$ and $\chi^+(v_2) = \log 4$. Since $\det \mathcal{A}_m = 4^m$ we obtain $\chi^+(v_1, v_2) = \log 4$. Therefore, χ^+ is exact with respect to $\{v_1, v_2\}$, and hence with respect to every collection of two vectors. On the other hand,

$$\chi^+(v_1, v_2) = \log 4 < \log 2 + \log 4 = \chi^+(v_1) + \chi^+(v_2)$$

and the sequence of matrices $\mathcal{A}^+$ is not forward regular.

In the one-dimensional case the situation is different.

PROPOSITION 3.3. *A sequence of numbers $\mathcal{A}^+ \subset GL(1, \mathbb{R}) = \mathbb{R} \setminus \{0\}$ is forward regular if and only if the Lyapunov exponent χ^+ is exact.*

We now present an important characteristic property of forward regularity which is very useful in applications. We say that a basis $\mathbf{v} = (v_1, \dots, v_n)$ of $\mathbb{R}^n$ is *normal* with respect to the filtration $\mathcal{V} = \{V_i\colon\ i = 0, \dots, p^+\}$ if for every $1 \leqslant i \leqslant p^+$ there exists a basis of V_i composed of vectors from $\{v_1, \dots, v_n\}$.

THEOREM 3.4 (see [69]). *A sequence of matrices $\mathcal{A}^+$ is forward regular if and only if for any normal basis* $\mathbf{v}$ *of $\mathbb{R}^n$ with respect to the filtration $\mathcal{V}_{\chi^+}$ and any subset $K \subset \{1, \dots, n\}$, we have*:

1. $$\chi^+\big(\{v_i\}_{i\in K}\big) = \lim_{m\to+\infty} \frac{1}{m} \log V\big(\{\mathcal{A}_m v_i\}_{i\in K}\big) = \sum_{i\in K} \chi^+(v_i);$$

2. *if σ_m is the angle between the subspaces* $\operatorname{span}\{\mathcal{A}_m v_i\colon\ i \in K\}$ *and* $\operatorname{span}\{\mathcal{A}_m v_i\colon\ i \notin K\}$, *then*

$$\lim_{m\to+\infty} \frac{1}{m} \log|\sin\sigma_m| = 0.$$

A sequence of matrices $\mathcal{A}^- = \{A_m\}_{m<0}$ is called *backward regular* if

$$\lim_{m\to-\infty} \frac{1}{|m|} \log|\det \mathcal{A}_m| = \sum_{i=1}^{n} \tilde{\chi}_i',$$

where $\tilde{\chi}_1', \dots, \tilde{\chi}_n'$ are the finite values of χ^- counted with their multiplicities.

Given a sequence $\mathcal{A}^- = \{A_m\}_{m<0}$, we construct a new sequence $\mathcal{B}^+ = \{B_m\}_{m\geqslant 0}$ by setting $B_m = (A_{-m-1})^{-1}$. The backward regularity of $\mathcal{A}^-$ is equivalent to the forward regularity of $\mathcal{B}^+$. This reduction allows one to translate any fact about forward regularity into a corresponding fact about backward regularity.

For example, the backward regularity of the sequence of matrices $\mathcal{A}^-$ implies that the Lyapunov exponent χ^- is exact. Moreover, if the Lyapunov exponent χ^- is exact (in particular, if it is backward regular) then

$$\chi^-(v) = \lim_{m\to-\infty} \frac{1}{|m|} \log\|\mathcal{A}_m v\|$$

for every $v \in \mathbb{R}^n$. This is equivalent to the following: for every $\varepsilon > 0$ there exists $C = C(v, \varepsilon) > 0$ such that if $m \geqslant 0$ then

$$C^{-1} e^{-(\chi^-(v)-\varepsilon)m} \leqslant \|\mathcal{A}_{-m} v\| \leqslant C e^{-(\chi^-(v)+\varepsilon)m}.$$

3.3. *A criterion for forward regularity of triangular matrices*

Let $\mathcal{A}^+ = \{A_m\}_{m\geqslant 0}$ be a sequence of matrices. One can write each A_m in the form $A_m = R_m T_m$, where R_m is orthogonal, and T_m is lower triangular. In general, the diagonal entries of T_m alone do *not* determine the values of the Lyapunov exponent associated

with $\mathcal{A}^+$. Indeed, let $\mathcal{A}^+ = \{A_m\}_{m\geqslant 0}$ be a sequence of matrices where $A_m = \left(\begin{smallmatrix} 0 & -1/2 \\ 2 & 0 \end{smallmatrix}\right)$ for each $m > 0$. We have $A_m = R_m T_m$, where $R_m = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ is orthogonal and $T_m = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 1/2 \end{smallmatrix}\right)$ is diagonal. Since ${A_m}^2 = -\mathrm{Id}$, we obtain $\chi^+(v, \mathcal{A}^+) = 0$ for every $v \neq 0$ whereas the values of the Lyapunov exponent for $\mathcal{T} = \{T_m\}_{m\geqslant 0}$ are equal to $\pm \log 2$.

However, in certain situations one can reduce the study of sequences of arbitrary matrices to the study of sequences of lower triangular matrices (see Section 5.3). Therefore, we shall consider sequences of lower triangular matrices and present a useful criterion of regularity of the Lyapunov exponent. This criterion is used in the proof of the Multiplicative Ergodic Theorem 5.5, which is one of the central results in smooth ergodic theory. We write $\log^+ a = \max\{\log a, 0\}$ for a positive number a.

THEOREM 3.5 (see [69]). *Let $\mathcal{A}^+ = \{(a_{ij}^m)\}_{m\geqslant 0} \subset GL(n, \mathbb{R})$ be a sequence of lower triangular matrices such that*:

1. *for each $i = 1, \ldots, n$, the following limit exists and is finite*:

$$\lim_{m\to+\infty} \frac{1}{m} \sum_{k=0}^{m} \log|a_{ii}^k| \stackrel{\text{def}}{=} \lambda_i;$$

2. *for any $i, j = 1, \ldots, n$, we have*

$$\overline{\lim_{m\to+\infty}} \frac{1}{m} \log^+|a_{ij}^m| = 0.$$

Then the sequence $\mathcal{A}^+$ is forward regular, and the numbers λ_i are the values of the Lyapunov exponent χ^+ (counted with their multiplicities but possibly not ordered).

Let us comment on the proof of this theorem. If we count each exponent according to its multiplicity we have exactly n exponents. To verify (3.6) we will produce a basis $v_1, \ldots, v_n$ which is normal with respect to the standard filtration (i.e., related with the standard basis by an upper triangular coordinate change) such that $\chi^+(v_i) = \lambda_i$.

If the exponents are ordered so that $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$ then the standard basis is in fact normal. To see this notice that while multiplying lower triangular matrices one obtains a matrix whose off-diagonal entries contain a polynomially growing number of terms each of which can be estimated by the growth of the product of diagonal terms below.

However, if the exponents are not ordered that way then an element e_i of the standard basis will grow according to the maximal of the exponents λ_j for $j \geqslant i$. In order to produce the right growth one has to compensate the growth caused by off-diagonal terms by subtracting from the vector e_i a certain linear combination of vectors e_j for which $\lambda_j > \lambda_i$. This can be done in a unique fashion. The detailed proof proceeds by induction.

A similar criterion of forward regularity holds for sequences of upper triangular matrices.

Using the correspondence between forward and backward sequences of matrices we immediately obtain the corresponding criterion for backward regularity.

3.4. *Lyapunov regularity*

Let $\mathcal{A} = \{A_m\}_{m\in\mathbb{Z}}$ be a sequence of matrices in $GL(n, \mathbb{R})$. Set $\mathcal{A}^+ = \{A_m\}_{m\geqslant 0}$ and $\mathcal{A}^- = \{A_m\}_{m<0}$. Consider the forward and backward Lyapunov exponents χ^+ and χ^- specified by the sequence $\mathcal{A}$, i.e., by the sequences $\mathcal{A}^+$ and $\mathcal{A}^-$, respectively; see (3.1) and (3.2). Denote by

$$\mathcal{V}_{\chi^+} = \{V_i^+ \colon i = 1, \dots, p^+\} \quad \text{and} \quad \mathcal{V}_{\chi^-} = \{V_i^- \colon i = 1, \dots, p^-\}$$

the filtrations of $\mathbb{R}^n$ associated with the Lyapunov exponents χ^+ and χ^-.

We say that the filtrations $\mathcal{V}_{\chi^+}$ and $\mathcal{V}_{\chi^-}$ *comply* if the following properties hold:

1. $p^+ = p^- \stackrel{\text{def}}{=} p$;
2. there exists a decomposition

$$\mathbb{R}^n = \bigoplus_{i=1}^{p} E_i$$

into subspaces E_i such that if $i = 1, \dots, p$ then

$$V_i^+ = \bigoplus_{j=1}^{i} E_j \quad \text{and} \quad V_i^- = \bigoplus_{j=i}^{p} E_j$$

(note that necessarily $E_i = V_i^+ \cap V_i^-$ for $i = 1, \dots, p$);

3. if $v \in E_i \setminus \{0\}$ then

$$\lim_{m\to\pm\infty} \frac{1}{m} \log \|\mathcal{A}_m v\| = \chi_i,$$

with uniform convergence on $\{v \in E_i \colon \|v\| = 1\}$.

We say that the sequence $\mathcal{A}$ is *Lyapunov regular* or simply *regular* if:

1. $\mathcal{A}$ is simultaneously forward and backward regular (i.e., $\mathcal{A}^+$ is forward regular and $\mathcal{A}^-$ is backward regular);
2. the filtrations $\mathcal{V}_{\chi^+}$ and $\mathcal{V}_{\chi^-}$ comply.

Notice that the constant cocycle generated by a single matrix A (see Section 3.1) is Lyapunov regular since

$$\sum_{i=1}^{p} \chi_i \dim E_i = \log|\det A|.$$

PROPOSITION 3.6. *If $\mathcal{A}$ is regular then*:

1. *the exponents χ^+ and χ^- are exact*;

2. $\chi_i^- = -\chi_i$, $\dim E_i = k_i$, *and*

$$\lim_{m\to\pm\infty} \frac{1}{m} \log\bigl|\det(\mathcal{A}_m|E_i)\bigr| = \chi_i k_i.$$

Simultaneous forward and backward regularity of a sequence of matrices $\mathcal{A}$ is not sufficient for Lyapunov regularity. Forward (respectively, backward) regularity does not depend on the backward (respectively, forward) behavior of $\mathcal{A}$, i.e., for $m \leqslant 0$ (respectively, $m \geqslant 0$). On the other hand, Lyapunov regularity requires some compatibility between the forward and backward behavior which is expressed in terms of the filtrations $\mathcal{V}_{\chi^+}$ and $\mathcal{V}_{\chi^-}$.

EXAMPLE 3.7. Let

$$A_m = \begin{cases} \left(\begin{smallmatrix} 2 & 0 \\ 0 & 1/2 \end{smallmatrix}\right) & \text{if } m \geqslant 0, \\ \left(\begin{smallmatrix} 5/4 & -3/4 \\ -3/4 & 5/4 \end{smallmatrix}\right) & \text{if } m < 0. \end{cases}$$

Note that

$$\begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} = R^{-1} \begin{pmatrix} 5/4 & -3/4 \\ -3/4 & 5/4 \end{pmatrix} R,$$

where R is the rotation by $\pi/4$ around 0. We have $\chi^+(1,0) = \chi^-(1,1) = \log 2$ and $\chi^+(0,1) = \chi^-(1,-1) = -\log 2$. Hence, $V_1^+ \neq V_1^-$, and thus, $\mathcal{A}$ is not regular. On the other hand, since $\det A_m = 1$, we have

$$\chi^+(v_1, v_2) = \chi^-(v_1, v_2) = \log 2 - \log 2 = 0,$$

and the exponents $\chi^+(v_1, v_2)$ and $\chi^-(v_1, v_2)$ are exact for any linearly independent vectors $v_1, v_2 \in \mathbb{R}^2$. Therefore, the sequence $\mathcal{A}$ is simultaneously forward and backward regular.

4. Cocycles and Lyapunov exponents

4.1. *Cocycles and linear extensions*

In what follows we assume that X is a measure space which is endowed with a σ-algebra of measurable subsets and that $f : X \to X$ is an invertible measurable transformation. For most substantive statements we will assume that f preserves a finite measure.

A function $\mathcal{A} : X \times \mathbb{Z} \to GL(n, \mathbb{R})$ is called a *linear multiplicative cocycle over* f or simply a *cocycle* if the following properties hold:

1. for every $x \in X$ we have $\mathcal{A}(x, 0) = \mathrm{Id}$ and if $m, k \in \mathbb{Z}$ then

$$\mathcal{A}(x, m+k) = \mathcal{A}\bigl(f^k(x), m\bigr)\mathcal{A}(x, k); \tag{4.1}$$

2. for every $m \in \mathbb{Z}$ the function $\mathcal{A}(\cdot, m) : X \to GL(n, \mathbb{R})$ is measurable.

If $\mathcal{A}$ is a cocycle, then $\mathcal{A}(f^{-m}(x), m)^{-1} = \mathcal{A}(x, -m)$ for every $x \in X$ and $m \in \mathbb{Z}$.

Given a measurable function $A : X \to GL(n, \mathbb{R})$ and $x \in X$, define the cocycle

$$\mathcal{A}(x, m) = \begin{cases} A(f^{m-1}(x)) \dots A(f(x))A(x) & \text{if } m > 0, \\ \text{Id} & \text{if } m = 0, \\ A(f^m(x))^{-1} \dots A(f^{-2}(x))^{-1} A(f^{-1}(x))^{-1} & \text{if } m < 0. \end{cases}$$

The map A is called the *generator* of the cocycle $\mathcal{A}$. One also says that the cocycle $\mathcal{A}$ is *generated* by the function A. Each cocycle $\mathcal{A}$ is generated by the function $A(\cdot) = \mathcal{A}(\cdot, 1)$.

The sequences of matrices that we discussed in the previous section are cocycles over the shift map $f : \mathbb{Z} \to \mathbb{Z}$, $f(n) = n + 1$.

A cocycle $\mathcal{A}$ over f induces a *linear extension* $F : X \times \mathbb{R}^n \to X \times \mathbb{R}^n$ of f to $X \times \mathbb{R}^n$, or a *linear skew product*, defined by

$$F(x, v) = \big(f(x), A(x)v\big).$$

In other words, the action of F on the fiber over x to the fiber over $f(x)$ is given by the linear map $A(x)$. If $\pi : X \times \mathbb{R}^n \to X$ is the projection, $\pi(x, v) = x$, then the diagram

$$\begin{array}{ccc} X \times \mathbb{R}^n & \xrightarrow{F} & X \times \mathbb{R}^n \\ {\scriptstyle \pi}\downarrow & & \downarrow{\scriptstyle \pi} \\ X & \xrightarrow{f} & X \end{array}$$

is commutative. Notice that for each $m \in \mathbb{Z}$,

$$F^m(x, v) = \big(f^m(x), \mathcal{A}(x, m)v\big).$$

Linear extensions are particular cases of bundle maps of measurable vector bundles which we now consider. Let E and X be measure spaces and $\pi : E \to X$ a measurable map. One says that E is a *measurable vector bundle* over X if for every $x \in X$ there exists a measurable subset $Y_x \subset X$ containing x such that there exists a measurable map with measurable inverse $\pi^{-1}(Y_x) \to Y_x \times \mathbb{R}^n$. A bundle map $F : E \to E$ over a measurable map $f : X \to X$ is a measurable map which makes the following diagram commutative:

$$\begin{array}{ccc} E & \xrightarrow{F} & E \\ {\scriptstyle \pi}\downarrow & & \downarrow{\scriptstyle \pi} \\ X & \xrightarrow{f} & X \end{array}$$

The following proposition shows that from the measure theory point of view every vector bundle over a compact metric space is trivial, and hence, without loss of generality, one may always assume that $E = X \times \mathbb{R}^n$. In other words every bundle map of E is essentially a linear extension provided that the base space X is a compact metric space.

PROPOSITION 4.1. *If E is a measurable vector bundle over a compact metric space (X, ν), then there is a subset $Y \subset X$ such that $\nu(Y) = 1$ and $\pi^{-1}(Y)$ is (isomorphic to) a trivial vector bundle.*

4.2. *Cohomology and tempered equivalence*

Let $A : X \to GL(n, \mathbb{R})$ be the generator of a cocycle $\mathcal{A}$ over the invertible measurable transformation $f : X \to X$. The cocycle $\mathcal{A}$ acts on the linear coordinate v_x on the fiber $\{x\} \times \mathbb{R}^n$ of $X \times \mathbb{R}^n$ by $v_{f(x)} = A(x)v_x$. Let $L(x) \in GL(n, \mathbb{R})$ be a linear coordinate change in each fiber, given by $u_x = L(x)v_x$ for each $x \in X$. We assume that the function $L : X \to GL(n, \mathbb{R})$ is measurable. Consider the function $B : X \to GL(n, \mathbb{R})$ for which $u_{f(x)} = B(x)u_x$. One can easily verify that

$$A(x) = L\big(f(x)\big)^{-1} B(x)L(x),$$

and that B generates a new cocycle $\mathcal{B}$ over f. One can naturally think of the cocycles $\mathcal{A}$ and $\mathcal{B}$ as equivalent. However, since the function L is in general only measurable, without any additional assumption on L the measure-theoretical properties of the cocycles $\mathcal{A}$ and $\mathcal{B}$ can be very different. We now introduce a sufficiently general class of coordinate changes which make the notion of equivalence productive.

Let $Y \subset X$ be an f-invariant nonempty measurable set. A measurable function $L : X \to GL(n, \mathbb{R})$ is said to be *tempered on Y with respect to f* or simply *tempered on Y* if for every $x \in Y$ we have

$$\lim_{m \to \pm\infty} \frac{1}{m} \log \big\| L\big(f^m(x)\big) \big\| = \lim_{m \to \pm\infty} \frac{1}{m} \log \big\| L\big(f^m(x)\big)^{-1} \big\| = 0.$$

A cocycle over f is said to be *tempered on Y* if its generator is tempered on Y. If the real functions $x \mapsto \|L(x)\|$, $\|L(x)^{-1}\|$ are bounded or, more generally, have finite essential supremum, then the function L is tempered with respect to any invertible transformation $f : X \to X$ on any f-invariant nonempty measurable subset $Y \subset X$. The following statement provides a more general criterion for a function L to be tempered.

PROPOSITION 4.2. *Let $f : X \to X$ be an invertible transformation preserving a probability measure ν, and $L : X \to GL(n, \mathbb{R})$ a measurable function. If*

$$\log \|L\|, \log \big\|L^{-1}\big\| \in L^1(X, \nu),$$

then L is tempered on some set of full ν-measure.

Let $A, B : X \to GL(n, \mathbb{R})$ be the generators, respectively, of two cocycles $\mathcal{A}$ and $\mathcal{B}$ over an invertible measurable transformation f, and $Y \subset X$ a measurable subset. The cocycles $\mathcal{A}$ and $\mathcal{B}$ are said to be *equivalent on Y* or *cohomologous on Y*, if there exists a measurable function $L : X \to GL(n, \mathbb{R})$ which is tempered on Y such that for every $x \in Y$, we have

$$A(x) = L\big(f(x)\big)^{-1} B(x)L(x). \tag{4.2}$$

This is clearly an equivalence relation and if two cocycles $\mathcal{A}$ and $\mathcal{B}$ are equivalent, we write $\mathcal{A} \sim_Y \mathcal{B}$. Equation (4.2) is called *cohomology equation*.

It follows from (4.2) that for any $x \in Y$ and $m \in \mathbb{Z}$,

$$\mathcal{A}(x, m) = L\big(f^m(x)\big)^{-1} \mathcal{B}(x, m) L(x). \tag{4.3}$$

Proposition 4.2 immediately implies the following.

COROLLARY 4.3. *If* $L : X \to \mathbb{R}$ *is a measurable function such that* $\log\|L\|$, $\log\|L^{-1}\| \in L^1(X, \nu)$ *then any two cocycles* $\mathcal{A}$ *and* $\mathcal{B}$ *satisfying* (4.3) *are equivalent cocycles*.

We now consider the notion of equivalence for cocycles over different transformations. Let $f : X \to X$ and $g : Y \to Y$ be invertible measurable transformations. Assume that f and g are *measurably conjugated*, i.e., that $h \circ f = g \circ h$ for some invertible measurable transformation $h : X \to Y$. Let $\mathcal{A}$ be a cocycle over f and $\mathcal{B}$ a cocycle over g. The cocycles $\mathcal{A}$ and $\mathcal{B}$ are said to be *equivalent* if there exists a measurable function $L : X \to GL(n, \mathbb{R})$ which is tempered on Y with respect to g, such that for every $x \in Y$, we have

$$A\big(h^{-1}(x)\big) = L\big(g(x)\big)^{-1} B(x) L(x).$$

4.3. *Examples and basic constructions with cocycles*

We describe various examples of measurable cocycles over dynamical systems. Perhaps the simplest example is provided by the rigid cocycles generated by a single matrix. Starting from a given cocycle one can build other cocycles using some basic constructions in ergodic theory and algebra. Thus one obtains power cocycles, induced cocycles, and exterior power cocycles.

Let $\mathcal{A}$ be a measurable cocycle over a measurable transformation f of a Lebesgue space X. We will call a cocycle $\mathcal{A}$ *rigid* if it is equivalent to a cocycle whose generator A is a constant map. Rigid cocycles naturally arise in the classical Floquet Theory (where the dynamical system in the base is a periodic flow), and among smooth cocycles over translations on the torus with rotation vector satisfying a Diophantine condition (see [93,153] and the references therein). In the setting of actions of groups other than $\mathbb{Z}$ and $\mathbb{R}$ rigid cocycles appear in the measurable setting for actions of higher rank semisimple Lie groups and lattices in such groups (see [261]), and in the smooth setting for hyperbolic actions of higher rank Abelian groups (see [140,141]).

Given $m \geqslant 1$, consider the transformation $f^m : X \to X$ and the measurable cocycle $\mathcal{A}^m$ over f^m with the generator

$$A^m(x) \stackrel{\text{def}}{=} A\big(f^{m-1}(x)\big) \dots A(x).$$

The cocycle $\mathcal{A}^m$ is called the *mth power cocycle* of $\mathcal{A}$.

Assume that f preserves a measure ν and let $Y \subset X$ be a measurable subset of positive ν-measure. By Poincaré's Recurrence Theorem the set $Z \subset Y$ of points $x \in Y$ such that $f^n(x) \in Y$ for infinitely many positive integers n, has measure $\nu(Z) = \nu(Y)$. We define the transformation $f_Y : Y \to Y \pmod 0$ as follows:

$$f_Y(x) = f^{k_Y(x)}(x), \quad \text{where } k_Y(x) = \min\bigl\{k \geqslant 1\colon\ f^k(x) \in Y\bigr\}.$$

The function k_Y and the map f_Y are measurable on Z. We call k_Y the (*first*) *return time* to Y and f_Y the (*first*) *return map* or *induced transformation* on Y.

PROPOSITION 4.4 (see, for example, [77]). *The measure ν is invariant under f_Y, the function $k_Y \in L^1(X, \nu)$ and $\int_Y k_Y \, d\nu = \nu(\bigcup_{n \geqslant 0} f^n Y)$.*

Since $k_Y \in L^1(X, \nu)$, it follows from Birkhoff's Ergodic Theorem that the function

$$\tau_Y(x) = \lim_{k \to +\infty} \frac{1}{k} \sum_{i=0}^{k-1} k_Y\bigl(f_Y^i(x)\bigr)$$

is well defined for ν-almost all $x \in Y$ and that $\tau_Y \in L^1(X, \nu)$.

If $\mathcal{A}$ is a measurable linear cocycle over f with generator A, we define the *induced cocycle* $\mathcal{A}_Y$ over f_Y to be the cocycle with the generator

$$A_Y(x) = A^{k_Y(x)}(x).$$

Finally, given a cocycle $\mathcal{A}$ we define the cocycle $\mathcal{A}^{\wedge k} : X \times \mathbb{Z} \to (GL(n, \mathbb{R}))^{\wedge k}$ by $\mathcal{A}^{\wedge k}(x, m) = \mathcal{A}(x, m)^{\wedge k}$ (see Section 5.1 for the definition of exterior power). We call $\mathcal{A}^{\wedge k}$ the *k-fold exterior power cocycle* of $\mathcal{A}$.

4.4. *Hyperbolicity of cocycles*

The crucial notion of nonuniformly hyperbolic diffeomorphisms was introduced by Pesin in [197,198]. In terms of cocycles this is the special case of derivative cocycles (see Section 6.1). Pesin's approach can readily be extended to general cocycles.

Consider a family of inner products $\langle \cdot, \cdot \rangle = \{\langle \cdot, \cdot \rangle_x\colon\ x \in X\}$ on $\mathbb{R}^n$. Given $x \in X$ we denote by $\| \cdot \|_x$ the norm and by $\sphericalangle(\cdot, \cdot)_x$ the angle induced by the inner product $\langle \cdot, \cdot \rangle_x$. In order to simplify the notation we often write $\| \cdot \|$ and $\sphericalangle(\cdot, \cdot)$ omitting the reference point x.

Let $Y \subset X$ be an f-invariant nonempty measurable subset. Let also $\lambda, \mu : Y \to (0, +\infty)$ and $\varepsilon : Y \to [0, \varepsilon_0]$, $\varepsilon_0 > 0$, be f-invariant measurable functions such that $\lambda(x) < \mu(x)$ for every $x \in Y$.

We say that a cocycle $\mathcal{A} : X \times \mathbb{Z} \to GL(n, \mathbb{R})$ over f is *nonuniformly partially hyperbolic in the broad sense* if there exist measurable functions C, $K : Y \to (0, +\infty)$ such that

1. for every $x \in Y$,

$$\text{either } \lambda(x) e^{\varepsilon(x)} < 1 \quad \text{or} \quad \mu(x) e^{-\varepsilon(x)} > 1; \tag{4.4}$$

2. there exists a decomposition $\mathbb{R}^n = E_1(x) \oplus E_2(x)$, depending measurably on $x \in Y$, such that $A(x)E_1(x) = E_1(f(x))$ and $A(x)E_2(x) = E_2(f(x))$;
3. (a) for $v \in E_1(x)$ and $m > 0$,

$$\big\|\mathcal{A}(x,m)v\big\| \leqslant C(x)\lambda(x)^m e^{\varepsilon(x)m}\|v\|;$$

(b) for $v \in E_2(x)$ and $m < 0$,

$$\big\|\mathcal{A}(x,m)v\big\| \leqslant C(x)\mu(x)^m e^{\varepsilon(x)|m|}\|v\|;$$

(c) $\sphericalangle(E_1(f^m(x)), E_2(f^m(x))) \geqslant K(f^m(x))$ for every $m \in \mathbb{Z}$;
(d) for $m \in \mathbb{Z}$,

$$C\big(f^m(x)\big) \leqslant C(x)e^{|m|\varepsilon(x)}, \qquad K\big(f^m(x)\big) \geqslant K(x)e^{-|m|\varepsilon(x)}.$$

We say that a cocycle $\mathcal{A}: X \times \mathbb{Z} \to GL(n,\mathbb{R})$ over f is *nonuniformly* (*completely*) *hyperbolic* if the requirement (4.4) is replaced by the following stronger one: for every $x \in Y$,

$$\lambda(x)e^{\varepsilon(x)} < 1 < \mu(x)e^{-\varepsilon(x)}.$$

PROPOSITION 4.5. *If a cocycle $\mathcal{A}$ over f is partially hyperbolic in the broad sense, then for every $x \in Y$,*

1. $\mathcal{A}(x,m)E_1(x) = E_1(f^m(x))$ *and* $\mathcal{A}(x,m)E_2(x) = E_2(f^m(x))$;
2. *for $v \in E_1(x)$ and $m < 0$,*

$$\big\|\mathcal{A}(x,m)v\big\| \geqslant C\big(f^m(x)\big)^{-1}\lambda(x)^m e^{-\varepsilon(x)|m|}\|v\|;$$

3. *for $v \in E_2(x)$ and $m > 0$,*

$$\big\|\mathcal{A}(x,m)v\big\| \geqslant C\big(f^m(x)\big)^{-1}\mu(x)^m e^{-\varepsilon(x)m}\|v\|.$$

The set Y is nested by the invariant sets $Y_{\lambda\mu\varepsilon}$ for which $\lambda(x) \leqslant \lambda$, $\mu(x) \leqslant \mu$ and $\varepsilon(x) \leqslant \varepsilon$, i.e., $Y = \bigcup Y_{\lambda\mu\varepsilon}$ and $Y_{\lambda'\mu'\varepsilon'} \subset Y_{\lambda''\mu''\varepsilon''}$ provided $\lambda' \leqslant \lambda''$, $\mu' \leqslant \mu''$ and $\varepsilon' \leqslant \varepsilon''$. On each of these sets the above estimates hold with $\lambda(x)$, $\mu(x)$ and $\varepsilon(x)$ replaced by λ, μ and ε, respectively.

Even when a cocycle is continuous or smooth one should expect the functions λ, μ, ε, C and K to be only measurable, the function C to be unbounded and K to have values arbitrarily close to zero.

If these functions turn to be continuous we arrive to the special case of uniformly hyperbolic cocycles. More precisely, we say that the cocycle $\mathcal{A}: X \times \mathbb{Z} \to GL(n,\mathbb{R})$ over f is *uniformly partially hyperbolic in the broad sense* if there exist $0 < \lambda < \mu < \infty$, $\lambda < 1$, and constants $c > 0$ and $\gamma > 0$ such that the following conditions hold:

1. there exists a decomposition $\mathbb{R}^n = E_1(x) \oplus E_2(x)$, depending continuously on $x \in Y$, such that $A(x)E_1(x) = E_1(f(x))$ and $A(x)E_2(x) = E_2(f(x))$;

2. (a) for $v \in E_1(x)$ and $m > 0$,

$$\big\|\mathcal{A}(x,m)v\big\| \leqslant c\lambda^m \|v\|;$$

(b) for $v \in E_2(x)$ and $m < 0$,

$$\big\|\mathcal{A}(x,m)v\big\| \leqslant c\mu^m \|v\|;$$

(c) $\sphericalangle(E_1(f^m(x)), E_2(f^m(x))) \geqslant \gamma$ for every $m \in \mathbb{Z}$.

The principal example of uniformly hyperbolic cocycles are cocycles generated by Anosov diffeomorphisms and more generally Axiom A diffeomorphisms. The principal examples of nonuniformly hyperbolic cocycles are cocycles with nonzero Lyapunov exponents.

We will see below that a nonuniformly hyperbolic cocycle on a set Y of full measure (with respect to an invariant measure) is in fact, uniformly hyperbolic on a set $Y_\delta \subset Y$ of measure at least $1 - \delta$ for arbitrarily small $\delta > 0$. This observation is crucial in studying topological and measure-theoretical properties of such cocycles. However, the "parameters" of uniform hyperbolicity, i.e., the numbers c and γ may vary with δ approaching ∞ and 0, respectively. We stress that this can only occur with a subexponential rate. We proceed with the formal description.

4.5. *Regular sets of hyperbolic cocycles*

Nonuniformly hyperbolic cocycles turn out to be uniformly hyperbolic on some compact but noninvariant subsets, called *regular sets*. They are nested and exhaust the whole space. Nonuniform hyperbolic structure appears then as a result of deterioration of the hyperbolic structure when a trajectory travels from one of these subsets to another. We first introduce regular sets for arbitrary cocycles and then establish their existence for nonuniformly hyperbolic cocycles.

Let $\mathcal{A}$ be a cocycle over X and $\varepsilon : X \to [0, +\infty)$ an f-invariant measurable function. Given $0 < \lambda < \mu < \infty$, $\lambda < 1$, and $\ell \geqslant 1$, denote by $\Lambda^\ell = \Lambda^\ell_{\lambda\mu}$ the set of points $x \in X$ for which there exists a decomposition $\mathbb{R}^n = E_{1x} \oplus E_{2x}$ such that for every $k \in \mathbb{Z}$ and $m > 0$,

1. if $v \in \mathcal{A}(x,k)E_{1x}$ then

$$\big\|\mathcal{A}\big(f^k(x), m\big)v\big\| \leqslant \ell\lambda^m e^{\varepsilon(x)(m+|k|)}\|v\|$$

and

$$\big\|\mathcal{A}\big(f^k(x), -m\big)v\big\| \geqslant \ell^{-1}\lambda^{-m} e^{-\varepsilon(x)(|k-m|+m)}\|v\|;$$

2. if $v \in \mathcal{A}(x,k)E_{2x}$ then

$$\big\|\mathcal{A}\big(f^k(x), -m\big)v\big\| \leqslant \ell\mu^{-m} e^{\varepsilon(x)(m+|k|)}\|v\|$$

and

$$\big\|\mathcal{A}\big(f^k(x),m\big)v\big\| \geqslant \ell^{-1}\mu^m e^{-\varepsilon(x)(|k+m|+m)}\|v\|;$$

3. $\sphericalangle(E_{1f^k(x)}, E_{2f^k(x)}) \geqslant \ell^{-1}e^{-\varepsilon(x)|k|}$.

The set Λ^ℓ is called a *regular set* (or a *Pesin set*).

It is easy to see that these sets have the following properties:

1. $\Lambda^\ell \subset \Lambda^{\ell+1}$;
2. for $m \in \mathbb{Z}$ we have $f^m(\Lambda^\ell) \subset \Lambda^{\ell'}$, where

$$\ell' = \ell \exp\big(|m| \sup\big\{\varepsilon(x)\colon\, x \in \Lambda^\ell\big\}\big);$$

3. the set $\Lambda = \Lambda_{\lambda\mu} \stackrel{\text{def}}{=} \bigcup_{\ell\geqslant 1}\Lambda^\ell$ is f-invariant;
4. if X is a topological space and $\mathcal{A}$ and ε are continuous then the sets Λ^ℓ are closed and the subspaces E_{1x} and E_{2x} vary continuously with $x \in \Lambda^\ell$ (with respect to the Grassmannian distance).

Every cocycle $\mathcal{A}$ over f, which is nonuniformly hyperbolic on a set $Y \subset X$, admits a nonempty regular set. Indeed, for each $0 < \lambda < \mu < \infty$, $\lambda < 1$, and each integer $\ell \geqslant 1$ let $Y^\ell \subset X$ be the set of points for which

$$\lambda(x) \leqslant \lambda < \mu \leqslant \mu(x), \qquad C(x) \leqslant \ell \quad \text{and} \quad K(x) \geqslant \ell^{-1}.$$

We have $Y^\ell \subset Y^{\ell+1}$, $Y^\ell \subset \Lambda^\ell$ and $E_{1x} = E_1(x)$, $E_{2x} = E_2(x)$ for every $x \in Y$.

4.6. *Lyapunov exponents for cocycles*

We extend the notion of Lyapunov exponent to cocycles over dynamical systems.

Let $\mathcal{A}$ be a cocycle over an invertible measurable transformation f of a measure space X with generator $A\colon X \to GL(n,\mathbb{R})$. Note that for each $x \in X$ the cocycle $\mathcal{A}$ generates a sequence of matrices $\{A_m\}_{m\in\mathbb{Z}} = \{A(f^m(x))\}_{m\in\mathbb{Z}}$. Therefore, every cocycle can be viewed as a collection of sequences of matrices which are indexed by the trajectories of f. One can associate to each of these sequences of matrices a Lyapunov exponent.

However, one should carefully examine the dependence of the Lyapunov exponent when one moves from a sequence of matrices to another one (see Proposition 4.6 below). This is what constitutes a substantial difference in studying cocycles over dynamical systems and sequences of matrices (see Sections 5.1 and 5.3 below). We now proceed with the formal definition of the Lyapunov exponent for cocycles.

Consider the generator $A\colon X \to GL(n,\mathbb{R})$ of the cocycle $\mathcal{A}$. Given a point $(x,v) \in X \times \mathbb{R}^n$, we define the *forward Lyapunov exponent of* (x,v) (*with respect to the cocycle* $\mathcal{A}$) by

$$\chi^+(x,v) = \chi^+(x,v,\mathcal{A}) = \overline{\lim_{m\to+\infty}}\,\frac{1}{m}\log\big\|\mathcal{A}(x,m)v\big\|.$$

Note that the number $\chi^+(x,v)$ does not depend on the norm $\|\cdot\|$ induced by an inner product on $\mathbb{R}^n$. With the convention $\log 0 = -\infty$, we obtain $\chi^+(x,0) = -\infty$ for every $x \in X$.

There exist a positive integer $p^+(x) \leqslant n$, a collection of numbers

$$\chi_1^+(x) < \chi_2^+(x) < \cdots < \chi_{p^+(x)}^+(x),$$

and a filtration $\mathcal{V}_x^+$ of linear subspaces

$$\{0\} = V_0^+(x) \subsetneqq V_1^+(x) \subsetneqq \cdots \subsetneqq V_{p^+(x)}^+(x) = \mathbb{R}^n,$$

such that:

1. $V_i^+(x) = \{v \in \mathbb{R}^n\colon\ \chi^+(x,v) \leqslant \chi_i^+(x)\}$;
2. if $v \in V_i^+(x) \setminus V_{i-1}^+(x)$, then $\chi^+(x,v) = \chi_i^+(x)$.

The numbers $\chi_i^+(x)$ are called the *values* of the Lyapunov exponent χ^+ at x. The number

$$k_i^+(x) = \dim V_i^+(x) - \dim V_{i-1}^+(x)$$

is called the *multiplicity* of the value $\chi_i^+(x)$. We also write

$$n_i^+(x) \stackrel{\text{def}}{=} \dim V_i^+(x) = \sum_{j=1}^{i} k_j^+(x).$$

The *Lyapunov spectrum* of χ^+ at x is the collection of pairs

$$\operatorname{Sp}_x^+ \mathcal{A} = \big\{\big(\chi_i^+(x), k_i^+(x)\big)\colon\ i = 1, \ldots, p^+(x)\big\}.$$

Observe that $k_i^+(f(x)) = k_i^+(x)$ and hence, $\operatorname{Sp}_{f(x)}^+ \mathcal{A} = \operatorname{Sp}_x^+ \mathcal{A}$.

PROPOSITION 4.6. *The following properties hold*:

1. *the functions χ^+ and p^+ are measurable*;
2. *$\chi^+ \circ f = \chi^+$ and $p^+ \circ f = p^+$*;
3. *$A(x)V_i^+(x) = V_i^+(f(x))$ and $\chi_i^+(f(x)) = \chi_i^+(x)$.*

For every $(x,v) \in X \times \mathbb{R}^n$, we set

$$\chi^-(x,v) = \chi^-(x,v,\mathcal{A}) = \varlimsup_{m \to -\infty} \frac{1}{|m|} \log \big\|\mathcal{A}(x,m)v\big\|.$$

We call $\chi^-(x,v)$ the *backward Lyapunov exponent of* (x,v) (*with respect to the cocycle* $\mathcal{A}$). One can show that for every $x \in X$ there exist a positive integer $p^-(x) \leqslant n$, the *values*

$$\chi_1^-(x) > \chi_2^-(x) > \cdots > \chi_{p^-(x)}^-(x),$$

and the *filtration* $\mathcal{V}_x^-$ of $\mathbb{R}^n$ associated with χ^- at x,

$$\mathbb{R}^n = V_1^-(x) \supsetneq \cdots \supsetneq V_{p^-(x)}^-(x) \supsetneq V_{p^-(x)+1}^-(x) = \{0\},$$

such that $V_i^-(x) = \{v \in \mathbb{R}^n\colon\ \chi^-(x,v) \leqslant \chi_i^-(x)\}$. The number

$$k_i^-(x) = \dim V_i^-(x) - \dim V_{i+1}^-(x)$$

is called the *multiplicity* of the value $\chi_i^-(x)$. We define the *Lyapunov spectrum* of χ^- at x by

$$\mathrm{Sp}_x^- \mathcal{A} = \big\{\big(\chi_i^-(x), k_i^-(x)\big)\colon\ i = 1, \ldots, p^-(x)\big\}.$$

Any nonuniformly (partially or completely) hyperbolic cocycle has nonzero Lyapunov exponents. More precisely,

1. If $\mathcal{A}$ is a nonuniformly partially hyperbolic cocycle (in the broad sense) on Y, then

$$Y \subset \big\{x \in X\colon\ \chi^+(x,v) \neq 0 \text{ for some } v \in \mathbb{R}^n \setminus \{0\}\big\}.$$

2. If $\mathcal{A}$ is a nonuniformly hyperbolic cocycle on Y, then

$$Y \subset \big\{x \in X\colon\ \chi^+(x,v) \neq 0 \text{ for all } v \in \mathbb{R}^n\big\}.$$

The converse statement is also true but is much more difficult. It is a manifestation of the Multiplicative Ergodic Theorem 5.5 of Oseledets. Namely, a cocycle whose Lyapunov exponents do not vanish almost everywhere is nonuniformly hyperbolic on a set of full measure (see Theorem 5.11).

Lyapunov exponents of a cocycle are invariants of a coordinate change which satisfies the tempering property as the following statement shows.

PROPOSITION 4.7. *Let $\mathcal{A}$ and $\mathcal{B}$ be equivalent cocycles on Y over a measurable transformation $f : X \to X$, and $L : X \to GL(n, \mathbb{R})$ a measurable function satisfying* (4.2) *that is tempered on Y. If $x \in Y$ then*:

1. *the forward and backward Lyapunov spectra coincide at x, i.e.,*

$$\mathrm{Sp}_x^+ \mathcal{A} = \mathrm{Sp}_x^+ \mathcal{B} \quad \textit{and} \quad \mathrm{Sp}_x^- \mathcal{A} = \mathrm{Sp}_x^- \mathcal{B};$$

2. *$L(x)$ preserves the forward and backward filtrations of $\mathcal{A}$ and $\mathcal{B}$, i.e.,*

$$L(x) V_i^+(x, \mathcal{A}) = V_i^+(x, \mathcal{B}), \quad i = 1, \ldots, p^+(x),$$

and

$$L(x) V_i^-(x, \mathcal{A}) = V_i^-(x, \mathcal{B}), \quad i = 1, \ldots, p^-(x).$$

5. Regularity and Multiplicative Ergodic Theorem

5.1. *Lyapunov regularity*

We extend the concept of regularity to cocycles over dynamical systems. Let $\mathcal{A}$ be a cocycle over an invertible measurable transformation f of a measure space X. As we saw, given $x \in X$, the cocycle $\mathcal{A}$ generates the sequence of matrices $\{A_m\}_{m\in\mathbb{Z}} = \{A(f^m(x))\}_{m\in\mathbb{Z}}$.

We say that x is *forward* (respectively, *backward*) *regular* for $\mathcal{A}$ if the sequence of matrices $\{A(f^m(x))\}_{m\in\mathbb{Z}}$ is forward (respectively, backward) regular.

Clearly, if x is a forward (respectively, backward) regular point for $\mathcal{A}$ then so is the point $f^m(x)$ for every $m \in \mathbb{Z}$. Furthermore, if $\mathcal{A}$ and $\mathcal{B}$ are equivalent cocycles on Y then the point $y \in Y$ is forward (respectively, backward) regular for $\mathcal{A}$ if and only if it is forward (respectively, backward) regular for $\mathcal{B}$.

Consider the filtrations $\mathcal{V}^+ = \{\mathcal{V}_x^+\}_{x\in X}$ and $\mathcal{V}^- = \{\mathcal{V}_x^-\}_{x\in X}$ of $\mathbb{R}^n$ associated with the Lyapunov exponents χ^+ and χ^- specified by the cocycle $\mathcal{A}$. For each $x \in X$ these filtrations determine filtrations $\mathcal{V}_x^+$ and $\mathcal{V}_x^-$ of the Lyapunov exponents $\chi^+(x,\cdot)$ and $\chi^-(x,\cdot)$ for the sequence of matrices $\{A_m\}_{m\in\mathbb{Z}} = \{A(f^m(x))\}_{m\in\mathbb{Z}}$.

We say that the filtrations $\mathcal{V}^+$ and $\mathcal{V}^-$ comply at a point $x \in X$ if the filtrations $\mathcal{V}_x^+$ and $\mathcal{V}_x^-$ comply with respect to the sequence of matrices $\{A(f^m(x))\}_{m\in\mathbb{Z}}$. In other words, the filtrations $\mathcal{V}^+$ and $\mathcal{V}^-$ comply at $x \in X$ if the following properties hold:

1. $p^+(x) = p^-(x) \stackrel{\text{def}}{=} p(x)$;
2. there exists a decomposition

$$\mathbb{R}^n = \bigoplus_{i=1}^{p(x)} E_i(x) \tag{5.1}$$

into subspaces $E_i(x)$ such that $A(x)E_i(x) = E_i(f(x))$ and for $i = 1, \dots, p(x)$,

$$V_i^+(x) = \bigoplus_{j=1}^{i} E_j(x) \quad \text{and} \quad V_i^-(x) = \bigoplus_{j=i}^{p(x)} E_j(x);$$

3. if $v \in E_i(x) \setminus \{0\}$ then

$$\lim_{m\to\pm\infty} \frac{1}{m} \log\|\mathcal{A}(x,m)v\| = \chi_i^+(x) = -\chi_i^-(x) \stackrel{\text{def}}{=} \chi_i(x),$$

with uniform convergence on $\{v \in E_i(x)\colon \|v\| = 1\}$.

We call the decomposition (5.1) the *Oseledets' decomposition* at the point x.

Property 2 requires some degree of compatibility between forward and backward regularity and is equivalent to the following: for $i = 1, \dots, p(x)$ the spaces

$$E_i(x) = V_i^+(x) \cap V_i^-(x) \tag{5.2}$$

satisfy (5.1).

A point $x \in X$ is said to be *Lyapunov regular* or simply *regular* for $\mathcal{A}$ if the following conditions hold:

1. x is simultaneously forward and backward regular for $\mathcal{A}$;
2. the filtrations $\mathcal{V}^+$ and $\mathcal{V}^-$ comply at x.

The set of regular points is f-invariant. If $\mathcal{A}$ and $\mathcal{B}$ are equivalent cocycles on Y then $y \in Y$ is regular for $\mathcal{A}$ if and only if it is regular for $\mathcal{B}$. Under fairly general assumptions the set of regular points has full measure with respect to any invariant measure (see Theorem 5.5).

For each integer k, $1 \leqslant k \leqslant n$, let $(\mathbb{R}^n)^{\wedge k}$ be the space of alternating k-linear forms on $\mathbb{R}^n$. For any linear transformation A of $\mathbb{R}^n$, the *k-fold exterior power* $A^{\wedge k}$ of A is the unique linear transformation $A^{\wedge k}$ of $(\mathbb{R}^n)^{\wedge k}$ such that

$$A^{\wedge k}(v_1 \wedge \cdots \wedge v_k) = Av_1 \wedge \cdots \wedge Av_k$$

for any $v_1, \ldots, v_k \in (\mathbb{R}^n)^{\wedge 1} \equiv \mathbb{R}^n$. One can define an inner product in $(\mathbb{R}^n)^{\wedge k}$ by requiring that for any $v_1 \wedge \cdots \wedge v_k$, $w_1 \wedge \cdots \wedge w_k \in (\mathbb{R}^n)^{\wedge k}$,

$$\langle v_1 \wedge \cdots \wedge v_k, w_1 \wedge \cdots \wedge w_k \rangle = \det B,$$

where $B = (b_{ij})$ is the $k \times k$ matrix with entries $b_{ij} = \langle v_i, w_j \rangle$ for each i and j. The induced norm satisfies the following properties:

1. $\|v_1 \wedge \cdots \wedge v_k\| \leqslant \|v_1 \wedge \cdots \wedge v_\ell\| \cdot \|v_{\ell+1} \wedge \cdots \wedge v_k\|$ for any $\ell < k$;
2. for every linear transformations A, B of $\mathbb{R}^n$ and $1 \leqslant k, \ell \leqslant n$ the induced operator norm in $(\mathbb{R}^n)^{\wedge k}$ satisfies:
   (a) $\|(AB)^{\wedge k}\| \leqslant \|A^{\wedge k}\| \cdot \|B^{\wedge k}\|$;
   (b) $\|A^{\wedge(k+\ell)}\| \leqslant \|A^{\wedge k}\| \cdot \|A^{\wedge \ell}\| \leqslant \|A\|^{k+\ell}$;
   (c) $\|A^{\wedge k}\| = \prod_{j=1}^k d_j$, where $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_n \geqslant 0$ are the eigenvalues of $(A^*A)^{1/2}$.

PROPOSITION 5.1. *Let $x \in X$ be a regular point for $\mathcal{A}$. The following statements hold*:

1. *the exponents $\chi^+(x, \cdot)$ and $\chi^-(x, \cdot)$ are exact*;
2. *for $i = 1, \ldots, p(x)$,*
   (a) $\dim E_i(x) = k_i^+(x) = k_i^-(x) \stackrel{\text{def}}{=} k_i(x)$;
   (b) *for any vectors $v_1, \ldots, v_{k_i(x)} \in E_i(x)$ with $V(v_1, \ldots, v_{k_i(x)}) \neq 0$,*

$$\lim_{m \to \pm\infty} \frac{1}{m} \log V\big(\mathcal{A}(x, m)v_1, \ldots, \mathcal{A}(x, m)v_{k_i(x)}\big) = \chi_i(x)k_i(x);$$

3. *for $k = 1, \ldots, n$,*

$$\lim_{m \to \pm\infty} \frac{1}{m} \log \big\|\mathcal{A}(x, m)^{\wedge k}\big\| = \sum_{j=1}^k \chi'_{n-j+1}(x).$$

Identifying the space $E_i(f^m(x))$ with $\mathbb{R}^{k_i(x)}$ one can rewrite property 2b in the following way: for $i = 1, \dots, p(x)$,

$$\lim_{m\to\pm\infty} \frac{1}{m} \log\left|\det\left(\mathcal{A}(x,m)|E_i(x)\right)\right| = \chi_i(x)k_i(x).$$

Furthermore, for every regular point $x \in X$, $1 \leqslant i,\ j \leqslant p(x)$ with $i \neq j$, and every distinct vectors $v, w \in H_i(x)$,

$$\lim_{m\to\pm\infty} \frac{1}{m} \log\left|\sin\sphericalangle\left(E_i\left(f^m(x)\right), E_j\left(f^m(x)\right)\right)\right| = 0,$$

i.e., the angles between any two spaces $E_i(x)$ and $E_j(x)$ can grow at most subexponentially along the orbit of x, and

$$\lim_{m\to\pm\infty} \frac{1}{m} \log\left|\sin\sphericalangle(d_x f^m v, d_x f^m w)\right| = 0.$$

5.2. *Lyapunov exponents and basic constructions with cocycles*

PROPOSITION 5.2. *For every $x \in X$ and every $v \in \mathbb{R}^n$, if the exponent $\chi^+(x, v, \mathcal{A})$ is exact, then $\chi^+(x, v, \mathcal{A}^m)$ is exact and*

$$\chi^+(x, v, \mathcal{A}^m) = m\chi^+(x, v, \mathcal{A}).$$

It follows that if $x \in X$ is Lyapunov regular with respect to the cocycle $\mathcal{A}$ then so it is with respect to the cocycle $\mathcal{A}^m$. Moreover, the Oseledets' decomposition at a regular point $x \in X$ for the cocycle $\mathcal{A}$ provides the Oseledets' decomposition at x for the cocycle $\mathcal{A}^m$.

PROPOSITION 5.3. *Let $\mathcal{A}$ be a measurable cocycle over f and $Y \subset X$ a measurable subset of positive ν-measure. For ν-almost every $x \in Y$ and every $v \in \mathbb{R}^n$, if $\chi^+(x, v, \mathcal{A})$ is exact then $\chi^+(x, v, \mathcal{A}_Y)$ is exact and*

$$\chi^+(x, v, \mathcal{A}_Y) = \tau_Y(x)\chi^+(x, v, \mathcal{A}).$$

It follows that ν-almost every $x \in Y$ is regular with respect to the cocycle $\mathcal{A}_Y$ if and only if it is regular with respect to the cocycle $\mathcal{A}$. Moreover, the Oseledets' decomposition at x for $\mathcal{A}$ provides the Oseledets' decomposition at x for $\mathcal{A}_Y$.

PROPOSITION 5.4. *For every $x \in X$, $k = 1, \dots, n$, and $v_1 \wedge \cdots \wedge v_k \in (\mathbb{R}^n)^{\wedge k}$, if the exponent $\chi^+(x, v_i, \mathcal{A})$ is exact for $i = 1, \dots, k$, then $\chi^+(x, v_1 \wedge \cdots \wedge v_k, \mathcal{A}^{\wedge k})$ is exact and*

$$\chi^+\left(x, v_1 \wedge \cdots \wedge v_k, \mathcal{A}^{\wedge k}\right) = \sum_{i=1}^{k} \chi^+(x, v_i, \mathcal{A}).$$

It follows that if $x \in X$ is Lyapunov regular with respect to the cocycle $\mathcal{A}$ then so it is with respect to the cocycle $\mathcal{A}^{\wedge k}$. Moreover, from the Oseledets' decomposition $\bigoplus_{i=1}^{s(x)} E_i(x)$ at a regular point $x \in X$ for the cocycle $\mathcal{A}$ we obtain the Oseledets' decomposition

$$\bigoplus_{i_1,\dots,i_k} E_{i_1}(x)^{\wedge 1} \wedge \cdots \wedge E_{i_k}(x)^{\wedge 1}$$

of $(\mathbb{R}^n)^{\wedge k}$ at x for the cocycle $\mathcal{A}^{\wedge k}$.

5.3. *Multiplicative Ergodic Theorem I: Oseledets' approach*

Lyapunov regularity is a strong condition which imposes certain requirements on the forward and backward behavior of trajectories. It is also not easy to verify this condition. Nevertheless, it turns out that Lyapunov regularity is "typical" in the measure-theoretical sense.

THEOREM 5.5 (Multiplicative Ergodic Theorem, Oseledets [192]; see also [35] and [175]). *Let f be an invertible measure preserving transformation of a Lebesgue space (X, ν) and $\mathcal{A}$ a measurable cocycle over f whose generator satisfies the following integrability condition*:

$$\log^+ \|A\|, \log^+ \left\|A^{-1}\right\| \in L^1(X, \nu), \tag{5.3}$$

where $\log^+ a = \max\{\log a, 0\}$. *Then the set of regular points for $\mathcal{A}$ has full ν-measure.*

Let us notice that property (5.3) holds for any cocycle $A : X \to GL(n, \mathbb{R})$ for which there is a positive constant c such that $\|A(x)^{\pm 1}\| \leqslant c$ for ν-almost all $x \in X$.

For one-dimensional cocycles, i.e., cocycles with values in $GL(1, \mathbb{R})$, the Multiplicative Ergodic Theorem amounts to Birkhoff's Ergodic Theorem since

$$\log\left|\mathcal{A}(x, m)\right| = \sum_{j=0}^{m-1} \log\left|A\left(f^j(x)\right)\right|.$$

The main idea of Oseledets in proving the Multiplicative Ergodic Theorem is to reduce the general case to the case of triangular cocycles and then use a version of Theorem 3.5 to establish regularity.

The reduction to triangular cocycles goes as follows. First one constructs an extension of the transformation f,

$$F : X \times SO(n, \mathbb{R}) \to X \times SO(n, \mathbb{R}),$$

where $SO(n,\mathbb{R})$ is the group of orthogonal $n \times n$ matrices. Given $(x,U) \in X \times SO(n,\mathbb{R})$ one can apply the Gram–Schmidt orthogonalization procedure to the columns of the matrix $A(x)U$ and write

$$A(x)U = R(x,U)T(x,U), \tag{5.4}$$

where $R(x,U)$ is orthogonal and $T(x,U)$ is lower triangular (with positive entries on the diagonal). The two matrices $R(x,U)$ and $T(x,U)$ are uniquely defined, and their entries are linear combinations of the entries of U. Set

$$F(x,U) = \big(f(x), R(x,U)\big).$$

Consider the projection $\pi : (x,U) \mapsto U$. By (5.4), we obtain

$$T(x,U) = \big((\pi \circ F)(x,U)\big)^{-1} A(x)\pi(x,U). \tag{5.5}$$

Let $\tilde{\mathcal{A}}$ and $\mathcal{T}$ be two cocycles over F defined respectively by $\tilde{A}(x,U) = A(x)$ and $\mathcal{T}(x,U) = T(x,U)$. Since $\|U\| = 1$ for every $U \in SO(n,\mathbb{R})$, it follows from (5.5) that the cocycles $\tilde{\mathcal{A}}$ and $\mathcal{T}$ are equivalent on $X \times SO(n,\mathbb{R})$. Therefore a point $(x,U) \in X \times SO(n,\mathbb{R})$ is regular for $\tilde{\mathcal{A}}$ if and only if it is regular for $\mathcal{T}$.

By the Representation Theorem for Lebesgue spaces we may assume that X is a compact metric space and that $f : X \to X$ is Borel measurable. Let $\mathcal{M}$ be the set of all Borel probability measures $\tilde{\nu}$ on $X \times SO(n,\mathbb{R})$ which satisfy

$$\tilde{\nu}\big(B \times SO(n,\mathbb{R})\big) = \nu(B) \tag{5.6}$$

for all measurable sets $B \subset X$. Then $\mathcal{M}$ is a compact convex subset of a locally convex topological vector space. The map $F_* : \mathcal{M} \to \mathcal{M}$ defined by

$$(F_*\tilde{\nu})(B) = \tilde{\nu}\big(F^{-1}B\big)$$

is a bounded linear operator. By the Tychonoff Fixed Point Theorem, there exists a fixed point $\tilde{\nu}_0 \in \mathcal{M}$ for the operator F_*, i.e., a measure $\tilde{\nu}_0$ such that $\tilde{\nu}_0(F^{-1}B) = \tilde{\nu}_0(B)$ for every measurable set $B \subset X \times SO(n,\mathbb{R})$. By (5.6), we conclude that the set of regular points for $\mathcal{A}$ has full ν-measure if and only if the set of regular points for $\tilde{\mathcal{A}}$ has full $\tilde{\nu}_0$-measure, and hence, if and only if the set of regular points for $\mathcal{T}$ has full $\tilde{\nu}_0$-measure.

We may now assume that $A(x) = (a_{ij}(x))$ is a lower triangular matrix (i.e., $a_{ij}(x) = 0$ if $i < j$). Write $A(x)^{-1} = (b_{ij}(x))$ and note that $b_{ii}(x) = 1/a_{ii}(x)$ for each i. By (5.3), $\log^+|a_{ij}|$, $\log^+|b_{ij}| \in L^1(X,\nu)$. It follows from Birkhoff's Ergodic Theorem that for ν-almost every $x \in X$,

$$\lim_{m\to+\infty} \frac{1}{m} \log^+\big|a_{ij}\big(f^m(x)\big)\big| = \lim_{m\to-\infty} \frac{1}{m} \log^+\big|b_{ij}\big(f^m(x)\big)\big| = 0. \tag{5.7}$$

Note that

$$\big|\log|a_{ii}|\big| = \log^+|a_{ii}| + \log^-|a_{ii}| = \log^+|a_{ii}| + \log^+|b_{ii}|. \tag{5.8}$$

By (5.3) and (5.8), we obtain $\log|a_{ii}| = -\log|b_{ii}| \in L^1(X,\nu)$. Birkhoff's Ergodic Theorem guarantees the existence of measurable functions $\lambda_i \in L^1(X,\nu)$, $i = 1,\dots,n$, such that for ν-almost every $x \in X$,

$$\lim_{m\to+\infty} \frac{1}{m}\sum_{k=0}^{m-1} \log\left|a_{ii}\left(f^k(x)\right)\right| = \lim_{m\to-\infty} \frac{1}{m}\sum_{k=m}^{-1} \log\left|b_{ii}\left(f^k(x)\right)\right| = \lambda_i(x). \quad (5.9)$$

Let $Y \subset X$ be the set of points for which (5.7) and (5.9) hold. It is a set of full ν-measure. The proof is concluded by showing that Y consists of regular points for $\mathcal{A}$. Indeed, by Theorem 3.5, the sequence $\{A(f^m(x))\}_{m\in\mathbb{Z}}$ is simultaneously forward and backward regular for every $x \in Y$. Moreover, the numbers $\lambda_i(x)$ are the forward Lyapunov exponents counted with their multiplicities (but possibly not ordered), and are the symmetric of the backward Lyapunov exponents counted with their multiplicities (but possibly not ordered either). We conclude that $p^+(x) = p^-(x) \stackrel{\text{def}}{=} p(x)$ and $\chi_i^-(x) = -\chi_i^+(x)$ for $i = 1,\dots,p(x)$. The hardest and more technical part of the proof is to show that the spaces $E_1(x),\dots,E_{p(x)}(x)$, defined by (5.2), satisfy (5.1).

Consider the set N of points which are *not* Lyapunov regular. This set has zero measure with respect to *any* invariant Borel probability measure but in general is not empty. For example, for the derivative cocycle (see Section 6.1 below) generated by a volume-preserving Anosov diffeomorphism the set of nonregular points has positive Hausdorff dimension provided that the Riemannian volume is *not* the measure of maximal entropy (see [37]).

On another end, Herman [114] (see also Section 7.3.1) and Walters [245] constructed examples of continuous cocycles with values in $SL(2,\mathbb{R})$ over *uniquely ergodic* homeomorphisms of compact metric spaces for which the set of nonregular points is not empty.

Furman [101] found additional conditions on the cocycle over a uniquely ergodic homeomorphism which guarantee that *every* point is Lyapunov regular. Namely, the generator of the cocycle should be either

(1) *continuously diagonalizable*, i.e., continuously equivalent to a diagonal matrix, or
(2) *one-point Lyapunov spectrum*, or
(3) continuously equivalent to an *eventually positive function*, i.e., for some $n \geqslant 0$ all the entries of $A(x,n)$ are positive.

5.4. *Multiplicative Ergodic Theorem II: Raghunathan's approach*

We describe another approach to the proof of the Multiplicative Ergodic Theorem due to Raghunathan [211]. It exploits the Subadditive Ergodic Theorem. The work of Raghunathan also contains an extension to local fields (such as the field $\mathbb{Q}_p$ of p-adic numbers).

Let $f : X \to X$ be a measurable transformation. A measurable function $\mathcal{B} : X \times \mathbb{Z} \to \mathbb{R}\setminus\{0\}$ is called a *subadditive cocycle over* f if for every $x \in X$ the following properties hold:

1. $\mathcal{B}(x,0) = 1$;

2. if $m, k \in \mathbb{Z}$ then

$$\mathcal{B}(x, m+k) \leqslant \mathcal{B}\big(f^k(x), m\big) + \mathcal{B}(x, k).$$

If $\mathcal{A}: X \times \mathbb{Z} \to GL(n, \mathbb{R})$ is a multiplicative cocycle over f (see (4.1)), then $\mathcal{B} = \log \|\mathcal{A}\|$ is a subadditive cocycle. Indeed, by (4.1),

$$\log\big\|\mathcal{A}(x, m+k)\big\| \leqslant \log\big\|\mathcal{A}\big(f^k(x), m\big)\big\| + \log\big\|\mathcal{A}(x, k)\big\|.$$

The following statement is an immediate consequence of Kingman's Subadditive Ergodic Theorem (see [215]).

THEOREM 5.6. *Let f be an invertible measure preserving transformation of a Lebesgue space (X, ν), and $\mathcal{A}$ a measurable multiplicative cocycle over f whose generator satisfies* (5.3). *Then there exist f-invariant measurable functions $\varphi_+ : X \to \mathbb{R}$ and $\varphi_- : X \to \mathbb{R}$ such that for ν-almost every $x \in X$,*

$$\varphi_+(x) = \lim_{m \to +\infty} \frac{1}{m} \log\big\|\mathcal{A}(x, m)\big\| = -\lim_{m \to -\infty} \frac{1}{m} \log\big\|\mathcal{A}(x, m)^{-1}\big\|,$$

$$\varphi_-(x) = \lim_{m \to -\infty} \frac{1}{m} \log\big\|\mathcal{A}(x, m)\big\| = -\lim_{m \to +\infty} \frac{1}{m} \log\big\|\mathcal{A}(x, m)^{-1}\big\|.$$

Moreover $\varphi_+, \varphi_- \in L^1(X, \nu)$ and

$$\begin{aligned}
\int_X \varphi_+ \, d\nu &= \lim_{m \to +\infty} \frac{1}{m} \int_X \log\big\|\mathcal{A}(x, m)\big\| \, d\nu(x) \\
&= -\lim_{m \to -\infty} \frac{1}{m} \int_X \log\big\|\mathcal{A}(x, m)^{-1}\big\| \, d\nu(x), \\
\int_X \varphi_- \, d\nu &= \lim_{m \to -\infty} \frac{1}{m} \int_X \log\big\|\mathcal{A}(x, m)\big\| \, d\nu(x) \\
&= -\lim_{m \to +\infty} \frac{1}{m} \int_X \log\big\|\mathcal{A}(x, m)^{-1}\big\| \, d\nu(x).
\end{aligned}$$

As an immediate corollary we obtain that the values of the Lyapunov exponents $\chi_i^+(x)$ and $\chi_i^-(x)$ are integrable functions provided that (5.3) holds.

Let $\mathcal{A}$ be a measurable multiplicative cocycle over a transformation f. For each $i = 1, \ldots, n$ the function $\log \|\mathcal{A}^{\wedge i}\|$ is a subadditive cocycle.

We present now Raghunathan's version of the Multiplicative Ergodic Theorem 5.5. Let us stress that Raghunathan considered the case of noninvertible transformations but his methods can be adapted to invertible transformations and we state the corresponding result here; we refer the reader to Section 5.7 where we consider the case of noninvertible transformations.

THEOREM 5.7 (Raghunathan [211]). *Let f be an invertible measure preserving transformation of a Lebesgue space (X, ν), and $\mathcal{A}$ a measurable multiplicative cocycle over f whose generator satisfies* (5.3). *Then there exists a set $Y \subset X$ of full ν-measure such that if $x \in Y$ then*:

1. *x is a regular point for $\mathcal{A}$;*
2. *there exist matrices A_x^+ and A_x^- such that*

$$\lim_{m \to \pm\infty} \big(\mathcal{A}(x,m)^* \mathcal{A}(x,m)\big)^{1/(2|m|)} = A_x^{\pm};$$

3. *the distinct eigenvalues of A_x^+ are the numbers $e^{\chi_1(x)}, \dots, e^{\chi_{s(x)}(x)}$;*
4. *the distinct eigenvalues of A_x^- are the numbers $e^{\chi_1(x)}, \dots, e^{\chi_{s(x)}(x)}$.*

5.5. *Tempering kernels and the Reduction Theorem*

The results in the previous sections allow one to obtain a "normal form" of a general measurable cocycle associated with its Lyapunov exponent. Let us begin with the simple particular case of a rigid cocycle $\mathcal{A}$, i.e., a cocycle whose generator is a constant map A (see Section 4.6). It is easy to see that the cocycle $\mathcal{A}$ is equivalent to the rigid cocycle $\mathcal{B}$ whose generator is the Jordan block form of the matrix A. We consider $\mathcal{B}$ as the "normal form" of $\mathcal{A}$, and say that $\mathcal{A}$ is reduced to $\mathcal{B}$.

A general measurable cocycle $\mathcal{A}$ satisfying the integrability condition (5.3) is so to speak "weakly" rigid, i.e., it can be reduced to a constant cocycle up to an arbitrarily small error. We consider this constant cocycle as a "normal form" of $\mathcal{A}$. More precisely, by the Oseledets–Pesin Reduction Theorem 5.10 below given $\varepsilon > 0$, there exists a cocycle $\mathcal{A}_\varepsilon$ which is equivalent to $\mathcal{A}$ and has block form, such that the generator A_ε^i of each block satisfies

$$e^{\chi_i(x)-\varepsilon}\|v\| \leqslant \big\|A_\varepsilon^i(x)v\big\| \leqslant e^{\chi_i(x)+\varepsilon}\|v\|$$

for each regular point x and each $v \in E_i(x)$, where $\{E_i(x)\colon\ i = 1, \dots, p(x)\}$ is the Oseledets' decomposition at x (see (5.1)). We say that $\mathcal{A}_\varepsilon$ is the *reduced form* of $\mathcal{A}$.

To proceed with the description of normal forms we first introduce a family of inner products $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_x$ on $\mathbb{R}^n$ for $x \in X$. We start with the following auxiliary result.

PROPOSITION 5.8. *For each $\varepsilon > 0$ and each regular point $x \in X$ for $\mathcal{A}$, the formula*

$$\langle u, v \rangle'_{x,i} = \sum_{m \in \mathbb{Z}} \big\langle \mathcal{A}(x,m)u, \mathcal{A}(x,m)v \big\rangle e^{-2\chi_i(x)m - 2\varepsilon|m|} \tag{5.10}$$

determines a scalar product on $E_i(x)$.

For a fixed $\varepsilon > 0$ we introduce a new inner product on $\mathbb{R}^n$ by

$$\langle u, v\rangle'_x = \sum_{i=1}^{p(x)} \langle u_i, v_i\rangle'_{x,i},$$

where u_i and v_i are the projections of the vectors u and v over $E_i(x)$ along $\bigoplus_{j\neq i} E_j(x)$. We call $\langle\cdot,\cdot\rangle'_x$ a *Lyapunov inner product at* x, and the corresponding norm $\|\cdot\|'_x$ a *Lyapunov norm at* x. The sequence of weights $\{e^{-2\chi_i(x)m-2\varepsilon|m|}\}_{m\in\mathbb{Z}}$ in (5.10) is called a *Pesin Tempering Kernel*. The value of $\langle u, v\rangle'_x$ depends on the number ε. The Lyapunov inner product has the following properties.

PROPOSITION 5.9. *The following properties hold*:

1. *The inner product* $\langle\cdot,\cdot\rangle'_x$ *depends measurably on the regular point* x.
2. *For every regular point* $x \in X$ *and* $i \neq j$, *the spaces* $E_i(x)$ *and* $E_j(x)$ *are orthogonal with respect to the Lyapunov inner product.*

A coordinate change $C_\varepsilon : X \to GL(n, \mathbb{R})$ is called a *Lyapunov change of coordinates* if for each regular point $x \in X$ and $u, v \in \mathbb{R}^n$ it satisfies:

$$\langle u, v\rangle_x = \big\langle C_\varepsilon(x)u, C_\varepsilon(x)v\big\rangle'_x. \tag{5.11}$$

Note that the identity (5.11) does not determine the function $C_\varepsilon(x)$ uniquely.

The following result known as Oseledets–Pesin Reduction Theorem provides a complete description of normal forms for cocycles.

THEOREM 5.10 (see [139]). *Let* $f : X \to X$ *be an invertible measure preserving transformation of the Lebesgue space* (X, ν), *and* $\mathcal{A}$ *a measurable cocycle over* f. *Given* $\varepsilon > 0$ *and a regular point* x,

1. *there exists a Lyapunov change of coordinates* C_ε *which sends the orthogonal decomposition* $\bigoplus_{i=1}^{p(x)} \mathbb{R}^{k_i(x)}$ *to the decomposition* $\bigoplus_{i=1}^{p(x)} E_i(x)$ *of* $\mathbb{R}^n$;
2. *the cocycle* $A_\varepsilon(x) = C_\varepsilon(f(x))^{-1} A(x) C_\varepsilon(x)$ *has the block form*

$$A_\varepsilon(x) = \begin{pmatrix} A^1_\varepsilon(x) & & \\ & \ddots & \\ & & A^{s(x)}_\varepsilon(x) \end{pmatrix}, \tag{5.12}$$

   where each block $A^i_\varepsilon(x)$ *is a* $k_i(x) \times k_i(x)$ *matrix, and the entries are zero above and below the matrices* $A^i_\varepsilon(x)$;
3. *each block* $A^i_\varepsilon(x)$ *satisfies*

$$e^{\chi_i(x)-\varepsilon} \leqslant \big\|A^i_\varepsilon(x)^{-1}\big\|^{-1} \leqslant \big\|A^i_\varepsilon(x)\big\| \leqslant e^{\chi_i(x)+\varepsilon};$$

4. *if the integrability condition* (5.3) *holds then the map* C_ε *is tempered* ν*-almost everywhere, and the spectra of* $\mathcal{A}$ *and* $\mathcal{A}_\varepsilon$ *coincide* ν*-almost everywhere.*

In the particular case of cocycles with values in $GL(2,\mathbb{R})$ Thieullen [239] showed that if the two Lyapunov exponents are equal the cocycle is conjugate to one of the following: a rotation cocycle, an upper triangular cocycle, or a diagonal cocycle modulo a rotation by $\pi/2$.

An important manifestation of the Oseledets–Pesin Reduction Theorem is a criterion of nonuniform hyperbolicity (partial or complete) of measurable cocycles via the values of their Lyapunov exponents.

THEOREM 5.11. *Let $f: X \to X$ be an invertible measure preserving transformation of the Lebesgue space (X, ν), and $\mathcal{A}$ a measurable cocycle over f whose generator satisfies* (5.3). *Then the following properties hold*:

1. *if the set*

$$Z_{ph} = \{x \in X:\ \chi^+(x,v) \neq 0 \text{ for some } v \in \mathbb{R}^n \setminus \{0\}\}$$

has measure $\nu(Z_{ph}) > 0$ then $\mathcal{A}$ is nonuniformly partially hyperbolic in the broad sense on some set $W_{ph} \subset Z_{ph}$ with $\nu(W_{ph}) = \nu(Z_{ph})$;

2. *if the set*

$$Z_h = \{x \in X:\ \chi^+(x,v) \neq 0 \text{ for all } v \in \mathbb{R}^n \setminus \{0\}\}$$

has measure $\nu(Z_h) > 0$ then $\mathcal{A}$ is nonuniformly hyperbolic on some set $W_h \subset Z_h$ with $\nu(W_h) = \nu(Z_h)$.

This theorem was first proved by Pesin in [198] for the special case of derivative cocycles (see the definition of the derivative cocycle in the next section) but the argument can readily be extended to the case of general cocycle.

The proof of this crucial statement is based upon the following observation. Given a regular point x and a small ε there exists a number $m(x,\varepsilon)$ such that for $m \geqslant m(x,\varepsilon)$,

$$\chi_i - \varepsilon \leqslant \frac{1}{n}\log\|\mathcal{A}_{ix}^m\| \leqslant \chi_i + \varepsilon, \qquad -\chi_i - \varepsilon \leqslant \frac{1}{n}\log\|\mathcal{A}_{ix}^{-m}\| \leqslant -\chi_i + \varepsilon,$$

and

$$-\chi_i - \varepsilon \leqslant \frac{1}{n}\log\|\mathcal{B}_{ix}^m\| \leqslant -\chi_i + \varepsilon, \qquad \chi_i - \varepsilon \leqslant \frac{1}{n}\log\|\mathcal{B}_{ix}^{-m}\| \leqslant \chi_i + \varepsilon,$$

where $\mathcal{A}_{ix}^m = \mathcal{A}(x,nm)|E_i(x)$ and $\mathcal{B}_{ix}^m = \mathcal{B}(x,m)|E_i^*(x)$ with $E_i^*(x)$ the dual space to $E_i(x)$. Here

$$\mathcal{B}(x,m) = \begin{cases} (A(x)^*)^{-1}(A(f(x))^*)^{-1}\cdots(A(f^{m-1}(x))^*)^{-1} & \text{if } m > 0, \\ \mathrm{Id} & \text{if } m = 0, \\ A(f^{-1}(x))^* A(f^{-2}(x))^* \cdots A(f^m(x))^* & \text{if } m < 0. \end{cases}$$

Set

$$D_1^{\pm}(x,\varepsilon) = \min_{1\leqslant i\leqslant s}\min_{0\leqslant j\leqslant m(x,\varepsilon)}\left\{1, \left\|\mathcal{A}_{ix}^{j}\right\|e^{(-\chi_i\pm\varepsilon)j}, \left\|\mathcal{B}_{ix}^{j}\right\|e^{(\chi_i\pm\varepsilon)j}\right\},$$

$$D_2^{\pm}(x,\varepsilon) = \max_{1\leqslant i\leqslant s}\max_{0\leqslant j\leqslant m(x,\varepsilon)}\left\{1, \left\|\mathcal{A}_{ix}^{j}\right\|e^{(-\chi_i\pm\varepsilon)j}, \left\|\mathcal{B}_{ix}^{j}\right\|e^{(\chi_i\pm\varepsilon)j}\right\}$$

and

$$D_1(x,\varepsilon) = \min\left\{D_1^{+}(x,\varepsilon), D_1^{-}(x,\varepsilon)\right\},$$
$$D_2(x,\varepsilon) = \max\left\{D_2^{+}(x,\varepsilon), D_2^{-}(x,\varepsilon)\right\},$$
$$D(x,\varepsilon) = \max\left\{D_1(x,\varepsilon)^{-1}, D_2(x,\varepsilon)\right\}.$$

The function $D(x,\varepsilon)$ is measurable, and if $m\geqslant 0$ and $1\leqslant i\leqslant p$ then

$$D(x,\varepsilon)^{-1}e^{(\pm\chi_i-\varepsilon)m} \leqslant \left\|\mathcal{A}_{ix}^{\pm m}\right\| \leqslant D(x,\varepsilon)e^{(\pm\chi_i+\varepsilon)m},$$
$$D(x,\varepsilon)^{-1}e^{(\pm\chi_i-\varepsilon)m} \leqslant \left\|\mathcal{B}_{ix}^{\pm m}\right\| \leqslant D(x,\varepsilon)e^{(\pm\chi_i+\varepsilon)m}. \tag{5.13}$$

Moreover, if $d\geqslant 1$ is a number for which the inequalities (5.13) hold for all $m\geqslant 0$ and $1\leqslant i\leqslant p$ with $D(x,\varepsilon)$ replaced by d then $d\geqslant D(x,\varepsilon)$. Therefore,

$$D(x,\varepsilon) = \inf\big\{d\geqslant 1\colon \text{the inequalities (5.13) hold for all } n\geqslant 0 \text{ and } 1\leqslant i\leqslant p \text{ with } D(x,\varepsilon) \text{ replaced by } d\big\}. \tag{5.14}$$

We wish to compare the values of the function $D(x,\varepsilon)$ at the points x and $f^j(x)$. We introduce the identification map $\tau_x\colon(\mathbb{R}^n)^*\to\mathbb{R}^n$ such that $\langle\tau_x(\varphi), v\rangle = \varphi(v)$ where $v\in\mathbb{R}^n$ and $\varphi\in(\mathbb{R}^n)^*$.

Let $\{v_k^m\colon k=1,\dots,\ell\}$ be a basis of $E_i(f^m(x))$ and $\{w_k^m\colon k=1,\dots,\ell\}$ the dual basis of $E_i^*(f^m(x))$. We have $\tau_{f^m(x)}(w_k^m) = v_k^m$. Denote by $A_{m,j}^i$ and $B_{m,j}^i$ the matrices corresponding to the linear maps $\mathcal{A}_{if^j(x)}^m$ and $\mathcal{B}_{if^j(x)}^m$ with respect to the above bases. We have that

$$A_{j,0}^i\big(B_{j,0}^i\big)^* = \mathrm{Id},$$

where $*$ stands for matrix transposition. Hence, for every $m>0$ the matrix corresponding to the map $\mathcal{A}_{if^j(x)}^m$ is

$$A_{m,j}^i = A_{m+j,0}^i\big(A_{j,0}^i\big)^{-1} = A_{m+j,0}^i\big(B_{j,0}^i\big)^*.$$

Therefore, in view of (5.13), we obtain that if $m>0$ then

$$\left\|\mathcal{A}_{if^j(x)}^m\right\| \leqslant D(x,\varepsilon)^2e^{(\chi_i+\varepsilon)(m+j)+(-\chi_i+\varepsilon)j} = D(x,\varepsilon)^2e^{2\varepsilon j}e^{(\chi_i+\varepsilon)m},$$
$$\left\|\mathcal{A}_{if^j(x)}^m\right\| \geqslant D(x,\varepsilon)^{-2}e^{(\chi_i-\varepsilon)(m+j)+(-\chi_i-\varepsilon)j} = D(x,\varepsilon)^{-2}e^{-2\varepsilon j}e^{(\chi_i-\varepsilon)m},$$

if $m > 0$ and $j - m \geqslant 0$ then

$$\left\|\mathcal{A}^{-m}_{if^j(x)}\right\| \leqslant D(x,\varepsilon)^2 e^{(\chi_i+\varepsilon)(j-m)+(-\chi_i+\varepsilon)j} = D(x,\varepsilon)^2 e^{2\varepsilon j} e^{(-\chi_i+\varepsilon)m},$$
$$\left\|\mathcal{A}^{-m}_{if^j(x)}\right\| \geqslant D(x,\varepsilon)^{-2} e^{(\chi_i-\varepsilon)(j-m)+(-\chi_i-\varepsilon)j} = D(x,\varepsilon)^{-2} e^{-2\varepsilon j} e^{(-\chi_i-\varepsilon)m},$$

and if $m > 0$ and $m - j \geqslant 0$ then

$$\left\|\mathcal{A}^{-m}_{if^j(x)}\right\| \leqslant D(x,\varepsilon)^2 e^{(\chi_i+\varepsilon)(m-j)+(-\chi_i+\varepsilon)j} = D(x,\varepsilon)^2 e^{2\varepsilon j} e^{(-\chi_i+\varepsilon)m},$$
$$\left\|\mathcal{A}^{-m}_{if^j(x)}\right\| \geqslant D(x,\varepsilon)^{-2} e^{(\chi_i-\varepsilon)(m-j)+(-\chi_i-\varepsilon)j} = D(x,\varepsilon)^{-2} e^{-2\varepsilon j} e^{(-\chi_i-\varepsilon)m}.$$

Similar inequalities hold for the maps $\mathcal{B}^m_{if^j(x)}$. Comparing this with the inequalities (5.13) applied to the point $f^j(x)$ and using (5.14) we conclude that if $j \geqslant 0$, then

$$D\big(f^j(x),\varepsilon\big) \leqslant D(x,\varepsilon)^2 e^{2\varepsilon j}. \tag{5.15}$$

Similar arguments show that if $j \leqslant 0$, then

$$D\big(f^{-j}(x),\varepsilon\big) \leqslant D(x,\varepsilon)^2 e^{-2\varepsilon j}. \tag{5.16}$$

It follows from (5.15) and (5.16) that if $j \in \mathbb{Z}$, then

$$D\big(f^j(x),\varepsilon\big) \leqslant D(x,\varepsilon)^2 e^{2\varepsilon|j|},$$

thus establishing the subexponential behavior of the constant along the trajectory necessary for nonuniform hyperbolicity.

Another important manifestation of the Oseledets–Pesin Reduction Theorem is a crucial property of the Lyapunov inner norms. It states that the function $x \mapsto \|v(x)\|'_x/\|v(x)\|_x$ is tempered on the set of regular points for every measurable vector field $X \ni x \mapsto v(x) \in \mathbb{R}^n \setminus \{0\}$. We recall that a positive function $K : X \to \mathbb{R}$ is called *tempered* on a set $Z \subset X$ if for any $x \in Z$,

$$\lim_{m\to\pm\infty} \frac{1}{m} \log K\big(f^m(x)\big) = 0. \tag{5.17}$$

THEOREM 5.12 (see [139]). *For every measurable vector field $X \ni x \mapsto v(x) \in \mathbb{R}^n \setminus \{0\}$, the function $x \mapsto \|v(x)\|'_x/\|v(x)\|_x$ is tempered on the set of regular points.*

The proof uses a technical but crucial statement known as the Tempering Kernel Lemma.

LEMMA 5.13 [139]. *Let $f : X \to X$ be a measurable transformation. If $K : X \to \mathbb{R}$ is a positive measurable function tempered on some subset $Z \subset X$, then for any $\varepsilon > 0$ there*

exists a positive measurable function $K_\varepsilon : Z \to \mathbb{R}$ *such that* $K(x) \leqslant K_\varepsilon(x)$ *and if* $x \in Z$ *then*

$$e^{-\varepsilon} \leqslant \frac{K_\varepsilon(f(x))}{K_\varepsilon(x)} \leqslant e^{\varepsilon}.$$

Note that if f preserves a Lebesgue measure ν on the space X, then any positive function $K : X \to \mathbb{R}$ with $\log K \in L^1(X, \nu)$ satisfies (5.17). The following is now an immediate consequence of Theorem 5.12.

THEOREM 5.14. *Given* $\varepsilon > 0$ *there is a positive measurable function* $K_\varepsilon : X \to \mathbb{R}$ *such that if* $x \in X$ *is a regular point then*:

1. $K_\varepsilon(x)e^{-\varepsilon|m|} \leqslant K_\varepsilon(f^m(x)) \leqslant K_\varepsilon(x)e^{\varepsilon|m|}$ *for every* $m \in \mathbb{Z}$;
2. $n^{-1/2}\|v\|_x \leqslant \|v\|'_x \leqslant K_\varepsilon(x)\|v\|_x$ *for every* $v \in \mathbb{R}^n$.

5.6. *The case of flows*

We briefly discuss counterparts to the results in the above sections for flows. Let (X, ν) be a Lebesgue space.

The measurable map $\varphi : \mathbb{R} \times X \to X$ is called a *measurable flow* on X if

$$\varphi_0 = \mathrm{Id}, \quad \text{and} \quad \varphi_t \circ \varphi_s = \varphi_{t+s} \quad \text{for every } t, s \in \mathbb{R}. \tag{5.18}$$

A measurable flow $\varphi : \mathbb{R} \times X \to X$ is called a *measure preserving flow* if $\varphi_t \overset{\text{def}}{=} \varphi(t, \cdot)$ is ν-invariant for every $t \in \mathbb{R}$.

We note that given a family $\{\varphi_t : t \in \mathbb{R}\}$ of measurable maps $\varphi_t : X \to X$ satisfying (5.18) one can define a measurable flow $\varphi : \mathbb{R} \times X \to X$ by $\varphi(t, x) = \varphi_t(x)$.

A measurable function $\mathcal{A} : X \times \mathbb{R} \to GL(n, \mathbb{R})$ is called a *linear multiplicative cocycle over* φ or simply a *cocycle* if for every $x \in X$ the following properties hold:

1. $\mathcal{A}(x, 0) = \mathrm{Id}$;
2. if $t, s \in \mathbb{R}$ then

$$\mathcal{A}(x, t + s) = \mathcal{A}\big(\varphi_t(x), s\big)\mathcal{A}(x, t).$$

The cocycle $\mathcal{A}$ induces *linear extensions* $F_t : X \times \mathbb{R}^n \to X \times \mathbb{R}^n$ by the formula

$$F_t(x, v) = \big(\varphi_t(x), \mathcal{A}(x, t)v\big).$$

Given $(x, v) \in X \times \mathbb{R}^n$, the *forward Lyapunov exponent of* (x, v) (*with respect to the cocycle* $\mathcal{A}$) given by

$$\chi^+(x, v) = \chi^+(x, v, \mathcal{A}) = \overline{\lim_{t \to +\infty}} \frac{1}{t} \log \big\| \mathcal{A}(x, t)v \big\|.$$

For every $x \in X$, there exist a positive integer $p^+(x) \leqslant n$, a collection of values

$$\chi_1^+(x) < \chi_2^+(x) < \cdots < \chi_{p^+(x)}^+(x),$$

and linear spaces

$$\{0\} = V_0^+(x) \subsetneqq V_1^+(x) \subsetneqq \cdots \subsetneqq V_{p^+(x)}^+(x) = \mathbb{R}^n,$$

such that:

1. $V_i^+(x) = \{v \in \mathbb{R}^n\colon\ \chi^+(x,v) \leqslant \chi_i^+(x)\}$;
2. if $v \in V_i^+(x) \setminus V_{i-1}^+(x)$, then $\chi^+(x,v) = \chi_i^+(x)$.

The number

$$k_i^+(x) = \dim V_i^+(x) - \dim V_{i-1}^+(x)$$

is the *multiplicity* of the value $\chi_i^+(x)$. In a similar way the quantity

$$\chi^-(x,v) = \chi^-(x,v,\mathcal{A}) = \overline{\lim_{t\to-\infty}} \frac{1}{|t|} \log\|\mathcal{A}(x,t)v\|$$

is the *backward Lyapunov exponent of* (x,v) (*with respect to the cocycle* $\mathcal{A}$). There exist a positive integer $p^-(x) \leqslant n$, a collection of *values*

$$\chi_1^-(x) > \cdots > \chi_{p^-(x)}^-(x)$$

and the *filtration* $\mathcal{V}_x^-$ of $\mathbb{R}^n$ associated with χ^- at x,

$$\mathbb{R}^n = V_1^-(x) \supsetneqq \cdots \supsetneqq V_{p^-(x)}^-(x) \supsetneqq V_{p^-(x)+1}^-(x) = \{0\},$$

where $V_i^-(x) = \{v \in \mathbb{R}^n\colon\ \chi^-(x,v) \leqslant \chi_i^-(x)\}$. The number

$$k_i^-(x) = \dim V_i^-(x) - \dim V_{i+1}^-(x)$$

is the *multiplicity* of the value $\chi_i^-(x)$.

Write $\mathcal{V}^+ = \{\mathcal{V}_x^+\}_{x\in X}$ and $\mathcal{V}_x^- = \{\mathcal{V}_x^-\}_{x\in X}$. The filtrations $\mathcal{V}^+$ and $\mathcal{V}^-$ *comply* at the point $x \in X$ if the following properties hold:

1. $p^+(x) = p^-(x) \stackrel{\text{def}}{=} p(x)$;
2. there exists a decomposition

$$\mathbb{R}^n = \bigoplus_{i=1}^{p(x)} E_i(x)$$

into subspaces $E_i(x)$ such that $\mathcal{A}(x,t)E_i(x) = E_i(\varphi_t x)$ for every $t \in \mathbb{R}$ and

$$V_i(x) = \bigoplus_{j=1}^{i} E_j(x) \quad \text{and} \quad V_i^-(x) = \bigoplus_{j=i}^{p(x)} E_j(x);$$

3. $\chi_i^+(x) = -\chi_i^-(x) \stackrel{\text{def}}{=} \chi_i(x)$;
4. if $v \in E_i(x) \setminus \{0\}$ then

$$\lim_{t\to\pm\infty} \frac{1}{t} \log\|\mathcal{A}(x,t)v\| = \chi_i(x),$$

with uniform convergence on $\{v \in E_i(x)\colon \|v\| = 1\}$.

A point x is *forward regular* for $\mathcal{A}$ if the following limit exists

$$\lim_{t\to+\infty} \frac{1}{t} \log\bigl|\det \mathcal{A}(x,t)\bigr| = \sum_{i=1}^{p^+(x)} \chi_i^+(x) k_i^+(x),$$

and is *backward regular* for $\mathcal{A}$ if the following limit exists:

$$\lim_{t\to-\infty} \frac{1}{|t|} \log\bigl|\det \mathcal{A}(x,t)\bigr| = \sum_{i=1}^{p^-(x)} \chi_i^-(x) k_i^-(x).$$

Finally, a point x is *Lyapunov regular* or simply *regular* for $\mathcal{A}$ if

1. x is simultaneously forward and backward regular for $\mathcal{A}$;
2. the filtrations $\mathcal{V}^+$ and $\mathcal{V}^-$ comply at x.

THEOREM 5.15 (Multiplicative Ergodic Theorem for flows). *Let φ be a measure preserving flow of a Lebesgue space (X,ν) such that φ_t is invertible for every $t \in \mathbb{R}$. Let also $\mathcal{A}$ be a measurable cocycle over φ such that*

$$\sup_{-1\leqslant t\leqslant 1} \log^+\|\mathcal{A}(\cdot,t)\| \in L^1(X,\nu). \tag{5.19}$$

Then the set of regular points for $\mathcal{A}$ has full ν-measure.

Given $\varepsilon > 0$ and a regular point $x \in X$, we introduce a family of inner products $\langle\cdot,\cdot\rangle_x$ on $\mathbb{R}^n$ by setting

$$\langle u, v\rangle_x' = \int_{\mathbb{R}} \bigl\langle \mathcal{A}(x,t)u, \mathcal{A}(x,t)v\bigr\rangle e^{-2\chi_i(x)t - 2\varepsilon|t|}\,dt$$

if $u, v \in E_i(x)$, and $\langle u, v\rangle_x' = 0$ if $u \in E_i(x)$ and $v \in E_j(x)$ with $i \neq j$. We call $\langle\cdot,\cdot\rangle_x'$ a *Lyapunov inner product at x*, and the corresponding norm $\|\cdot\|_x'$ a *Lyapunov norm at x*.

One can show that there exists a tempered function $K_\varepsilon : X \to \mathbb{R}$ such that if $x \in X$ is a regular point and $v \in \mathbb{R}^n$ then

$$n^{-1/2}\|v\|_x \leqslant \|v\|'_x \leqslant K_\varepsilon(x)\|v\|_x.$$

We recall that a positive function $K : X \to \mathbb{R}$ is called *tempered* on a set $Z \subset X$ if for any $x \in Z$,

$$\lim_{t\to\pm\infty} \frac{1}{m} \log K(\varphi_t x) = 0.$$

THEOREM 5.16 (Reduction Theorem for flows). *Let φ be a measure preserving flow of a Lebesgue space (X, ν) such that φ_t is invertible for every $t \in \mathbb{R}$. Let also $\mathcal{A}$ be a measurable cocycle over φ. Given $\varepsilon > 0$ and a regular point x, there exists a Lyapunov change of coordinates C_ε with the following properties*:

1. *the cocycle $\mathcal{A}_\varepsilon(x,t) = C_\varepsilon(\varphi_t x)^{-1}\mathcal{A}(x,t)C_\varepsilon(x)$ has the block form*

$$\mathcal{A}_\varepsilon(x,t) = \begin{pmatrix} \mathcal{A}^1_\varepsilon(x,t) & & \\ & \ddots & \\ & & \mathcal{A}^{p(x)}_\varepsilon(x,t) \end{pmatrix},$$

   where each block $\mathcal{A}^i_\varepsilon(x,t)$ is a $k_i(x) \times k_i(x)$ matrix, and the entries are zero above and below the matrices $\mathcal{A}^i_\varepsilon(x,t)$;
2. *each block $\mathcal{A}^i_\varepsilon(x)$ satisfies*

$$e^{\chi_i(x)t-\varepsilon|t|} \leqslant \big\|\mathcal{A}^i_\varepsilon(x,t)^{-1}\big\|^{-1} \leqslant \big\|\mathcal{A}^i_\varepsilon(x,t)\big\| \leqslant e^{\chi_i(x)t+\varepsilon|t|};$$

3. *if the integrability condition* (5.19) *holds then the map C_ε is tempered ν-almost everywhere, and the spectra of $\mathcal{A}$ and $\mathcal{A}_\varepsilon$ coincide ν-almost everywhere.*

5.7. *The case of noninvertible dynamical systems*

Consider a measure preserving transformation $f : X \to X$ of a Lebesgue space (X, ν) (the map f need not be invertible). We assume that ν is a probability measure. Given a measurable function $A : X \to GL(n, \mathbb{R})$ and $x \in X$, define the *one-sided cocycle* $\mathcal{A} : X \times \mathbb{N} \to GL(n, \mathbb{R})$ by

$$\mathcal{A}(x,m) = A\big(f^{m-1}(x)\big)\dots A\big(f(x)\big)A(x).$$

Note that the cocycle equation (4.1) holds for every $m, k \in \mathbb{N}$. Given $(x, v) \in X \times \mathbb{R}^n$, define the *forward Lyapunov exponent* of (x, v) (with respect to $\mathcal{A}$) by

$$\chi^+(x,v) = \chi^+(x,v,\mathcal{A}) = \overline{\lim_{m\to+\infty}} \frac{1}{m} \log\big\|\mathcal{A}(x,m)v\big\|.$$

However, since the map f and the matrices $A(x)$ may not be invertible, one may not in general define a backward Lyapunov exponent. Therefore, we can only discuss the forward regularity for $\mathcal{A}$. One can establish a Multiplicative Ergodic Theorem in this case.

THEOREM 5.17. *Let f be a measure preserving transformation of a Lebesgue space (X, ν), and $\mathcal{A}$ a measurable cocycle over f such that $\log^+ \|A\| \in L^1(X, \nu)$. Then the set of forward regular points for $\mathcal{A}$ has full ν-measure and for ν-almost every $x \in X$ and every subspace $F \subset E_i^+(x)$ such that $F \cap E_{i-1}^+(x) = \{0\}$ we have*

$$\lim_{m\to+\infty} \frac{1}{m} \log \inf_v \|\mathcal{A}(x, m)v\| = \lim_{m\to+\infty} \frac{1}{m} \log \sup_v \|\mathcal{A}(x, m)v\| = \chi_i^+(x),$$

with the infimum and supremum taken over $\{v \in F\colon \|v\| = 1\}$.

When the matrix $A(x)$ is invertible for every $x \in X$ and $\log^+ \|A\|$, $\log^+ \|A^{-1}\| \in L^1(X, \nu)$ for some f-invariant Lebesgue measure ν, one can show that for the cocycle induced by $\mathcal{A}$ on the inverse limit of f the set of *regular* points has full ν-measure.

5.8. *The case of nonpositively curved spaces*

Karlsson and Margulis [132] obtained an extension of the noninvertible case of the Multiplicative Ergodic Theorem 5.5 to some nonpositively curved spaces.

Let (Y, ρ) be a complete metric space. Y is called:

1. *convex* if any two points $x, y \in Y$ have a *midpoint*, i.e., a point z for which

$$\rho(z, x) = \rho(z, y) = \frac{1}{2}\rho(x, y);$$

2. *uniformly convex* if it is convex and there is a strictly decreasing continuous function g on $[0, 1]$ such that $g(0) = 1$ and for any $x, y, w \in Y$ and midpoint m_{xy} of x and y,

$$\frac{\rho(m_{xy}, w)}{R} \leqslant g\left(\frac{\rho(m_{xy}, w)}{R}\right),$$

where $R = \max\{\rho(x, w), \rho(y, w)\}$;

3. *nonpositively curved* (in the sense of Busemann) if it is convex and for any $x, y, z \in Y$ and any midpoints m_{xz} of x and z and m_{yz} of y and z,

$$\rho(m_{xz}, m_{yz}) \leqslant \frac{1}{2}\rho(x, y).$$

If Y is uniformly convex then midpoints are unique.

Examples of nonpositively curved spaces include uniformly convex Banach spaces (e.g., Hilbert spaces or L^p for $1 < p < \infty$), Cartan–Hadamard manifolds (e.g., Euclidean spaces,

hyperbolic spaces or $GL(n,\mathbb{R})/O(n,\mathbb{R})$), and more generally CAT(0) spaces (e.g., Euclidean buildings or $\mathbb{R}$-trees).

A continuous map $\gamma : I \to Y$ (I is an interval) is called a (unit speed minimizing) *geodesic* if for any $s, t \in I$,

$$\rho\big(\gamma(s), \gamma(t)\big) = |s - t|.$$

If Y is convex then any two points can be joined by a geodesic and if Y is uniformly convex then this geodesic is unique.

A geodesic $\gamma : [0, \infty) \to Y$ is called a *ray* if the limit $\lim_{t\to\infty} \gamma(t)$ does not exist. The two rays γ_1 and γ_2 are called *asymptotic* if $\rho(\gamma_1(t), \gamma_2(t)) \leqslant \text{const}$ for $t \geqslant 0$. We denote by $[\gamma]$ the set of all rays asymptotic to γ and by $Y(\infty)$ the ideal boundary of Y, i.e., the set of all classes of asymptotic rays.

Let $D \subset Y$ be a nonempty subset. A map $\varphi : D \to D$ is called a *semicontraction* (*or nonexpanding*) if $\rho(\varphi(v), \varphi(z)) \leqslant d(y, z)$ for all $y, z \in D$. Isometries are semicontractions.

Let us fix a semigroup S of semicontractions and equip it with the Borel σ-algebra associated with the compact-open topology on S. Fix $y \in Y$. Consider a cocycle $\mathcal{A}$ with values in S over an ergodic transformation f of a measure space (X, μ). Let $A : X \to X$ be the generator.

THEOREM 5.18 (Karlsson and Margulis [132]). *Assume that*

$$\int_X \rho\big(y, A(x)y\big)\, d\mu(x) < \infty.$$

Then for almost every $x \in X$ the following limit exists:

$$\lim_{m\to\infty} \frac{1}{m} d\big(y, \mathcal{A}(x, m)y\big) = a, \tag{5.20}$$

and if $a > 0$ then for almost every $x \in X$ there exists a unique geodesic ray $\gamma(\cdot, x)$ in Y starting at y such that

$$\lim_{m\to\infty} \frac{1}{m} d\big(\gamma(an, x), \mathcal{A}(x, m)y\big) = 0$$

and hence, $\mathcal{A}(\cdot, m)y$ converges to $[\gamma]$ in $Y \cup Y(\infty)$.

The existence of the limit in (5.20) is an easy corollary of Kingman's Subadditive Ergodic Theorem.

Consider the symmetric space $Y = GL(n,\mathbb{R})/O(n,\mathbb{R})$ and a cocycle $\mathcal{A}$ with values in $O(n,\mathbb{R})$ over an ergodic transformation f of a measure space (X, μ). Let $A : X \to X$ be the generator. Fix a point $y \in O(n,\mathbb{R})$. For $g \in GL(n,\mathbb{R})$ let λ_i be the eigenvalues of $(gg^*)^{1/2}$ where g^* is the transpose of g. The distance in Y between y and gy is

$$\rho(y, gy) = \left(\sum_{i=1}^{n} (\log \lambda_i)^2\right)^{1/2}.$$

A geodesic starting at y is of the form $\gamma(t) = e^{tH}y$ where H is a symmetric matrix. Then $\Lambda = e^H$ is a positive definite symmetric matrix. We have that

$$\lim_{n\to\infty} \frac{1}{m} d\big(\Lambda^{-m}y, \mathcal{A}(m,x)^{-1}y\big) = 0$$

for almost every $x \in X$. In view of (3.7) this means that x is forward regular.

Theorem 5.18 has interesting applications to random walks and Hilbert–Schmidt operators (see [132]). It is shown in [132] with an explicit example that there is no invertible version of Theorem 5.18, i.e., there is in general no two-sided geodesic approximating both the forward and backward orbits $m \mapsto \mathcal{A}(x, \pm m)y$.

5.9. *Notes*

The term "Multiplicative Ergodic Theorem" was introduced by Oseledets in [192] where he presented the first proof of the theorem.

In [183], Millionshchikov announced a somewhat independent proof of the Multiplicative Ergodic Theorem which is based on some subtle properties of the action of the differential with respect to the Lyapunov exponents.[1] Mañé used similar properties in his proof of the entropy formula (see Section 12.2).

Other proofs of the Multiplicative Ergodic Theorem were obtained by Ruelle [215], by Mañé [175] (see also [173]),[2] and by Goldsheid and Margulis [106]. A simpler version of the Multiplicative Ergodic Theorem was considered by Johnson, Palmer and Sell [130],[3] and related topics were discussed by Sacker and Sell [221,222,226] and by Johnson [129].

In [145], Kifer established a "random" version of the Multiplicative Ergodic Theorem—for compositions of independent identically distributed transformations of a measurable vector bundle. His proof is built on the work of Furstenberg and Kifer [102] (see also Chapter III in [146]). Under more restrictive conditions a similar result was obtained by Carverhill [70]. See the book by Arnold [23] for a detailed description of various versions of the Multiplicative Ergodic Theorem and related questions in the random dynamical systems setup.

There are also infinite-dimensional versions of the Multiplicative Ergodic Theorem. Namely, it was extended by Ruelle [216] to Hilbert spaces (following closely his finite-dimensional approach in [215]), and by Mañé in [173] to compact transformations in Banach spaces (see also Thieullen [238] for the case of not necessarily compact transformations). The proof due to Goldsheid and Margulis also extends to the infinite-dimensional case (see [106]).

[1] Millionshchikov's proof was never published as a solid piece; instead, it is scattered through a series of papers with cross-references and is difficult to comprehend.

[2] In both [215] and [175] a slightly weaker version of Lyapunov regularity, then the one we introduced in Section 5.1, is considered but the proofs contain arguments which are indeed, sufficient to establish a stronger version.

[3] They established some but not all properties of Lyapunov regularity referring the reader to the original work of Oseledets.

6. Cocycles over smooth dynamical systems

6.1. *The derivative cocycle*

Let $f : M \to M$ be a diffeomorphism of a smooth n-dimensional Riemannian manifold. Given $x \in M$, set $X = \{f^m(x)\}_{m\in\mathbb{Z}}$. Identifying the tangent spaces $T_{f^m(x)}M$ with $\mathbb{R}^n$ one can introduce the cocycle $\mathcal{A}_x = \{d_{f^m(x)}f\}_{m\in\mathbb{Z}}$ over the transformation $f : X \to X$. It is called the *derivative cocycle* associated with the diffeomorphism f and the point x. The Lyapunov exponent χ^+ of x specified by the derivative cocycle is the Lyapunov exponent specified by the diffeomorphism f at the point x.

The "individual" derivative cocycles $\mathcal{A}_x$ depend on the individual trajectories $\{f^m(x)\}_{m\in\mathbb{Z}}$. We now introduce the "global" cocycle associated with f. One can represent M as a finite union $\bigcup_i \Delta_i$ of differentiable copies Δ_i of the n-simplex such that:

1. in each Δ_i one can introduce local coordinates in such a way that $T\Delta_i$ can be identified with $\Delta_i \times \mathbb{R}^n$;
2. all the nonempty intersections $\Delta_i \cap \Delta_j$, for $i \neq j$, are $(n-1)$-dimensional manifolds.

In each Δ_i the derivative of f can be interpreted as a linear cocycle. This implies that $df : M \to \mathbb{R}^n$ can be interpreted as a measurable linear cocycle $\mathcal{A}$ with $d_x f$ to be the matrix representation of $d_x f$ in local coordinates. We call $\mathcal{A}$ the *derivative cocycle* specified by the diffeomorphism f. It does not depend on the choice of the decomposition $\{\Delta_i\}$. Indeed, if we choose another decomposition $\{\Delta_i'\}$, then the coordinate change in $\Delta_i \cap \Delta_j'$ sending one representation to the other one is effected by maps which are uniformly bounded together with their derivatives, their inverses, and the inverses of their derivatives. Hence, by Proposition 4.2, the coordinate change is tempered and the two cocycles corresponding to the two decompositions $\{\Delta_i\}$ and $\{\Delta_i'\}$ are equivalent.

We remark that if ν is an f-invariant Borel probability measure on M then the decomposition $\{\Delta_i\}$ can be chosen such that $\nu(\partial\Delta_i) = 0$ for every i.

6.2. *Nonuniformly hyperbolic diffeomorphisms*

We say that a diffeomorphism f is *nonuniformly partially hyperbolic in the broad sense* if so is the derivative cocycle generated by f. More precisely, this means[4] that f possesses an invariant Borel subset $\Lambda \subset M$ such that there exist:

(a) numbers λ and μ, $0 < \lambda < \mu$, $\lambda < 1$;
(b) a sufficiently small number $\varepsilon > 0$ and Borel functions C, $K : \Lambda \to (0, \infty)$;
(c) subspaces $E_1(x)$ and $E_2(x)$, $x \in \Lambda$,

which satisfy the following conditions:

1. the subspaces $E_1(x)$ and $E_2(x)$ depend measurably on x and form an invariant splitting of the tangent space, i.e.,

$$T_x M = E_1(x) \oplus E_2(x),$$
$$d_x f E_1(x) = E_1\big(f(x)\big), \qquad d_x f E_2(x) = E_2\big(f(x)\big); \tag{6.1}$$

[4]For simplicity, we consider here only one of the nested subsets in the definition of nonuniformly hyperbolic cocycles; see Section 4.4.

2. for $v \in E_1(x)$ and $n > 0$,

$$\|d_x f^n v\| \leqslant C(x)\lambda^n e^{\varepsilon n}\|v\|; \tag{6.2}$$

3. for $v \in E_2(x)$ and $n < 0$,

$$\|d_x f^n v\| \leqslant C(x)\mu^n e^{\varepsilon|n|}\|v\|; \tag{6.3}$$

4. the angle

$$\angle\big(E_1(x), E_2(x)\big) \geqslant K(x); \tag{6.4}$$

5. for $n \in \mathbb{Z}$,

$$C\big(f^n(x)\big) \leqslant C(x)e^{\varepsilon|n|}, \qquad K\big(f^n(x)\big) \geqslant K(x)e^{-\varepsilon|n|}. \tag{6.5}$$

Condition (6.5) means that estimates (6.2)–(6.4) may deteriorate along the trajectory with subexponential rate. We stress that the rates of contraction along stable subspaces and expansion along unstable subspaces are exponential and hence, prevail.

Furthermore, f is *nonuniformly partially hyperbolic* on an f-invariant Borel subset $\Lambda \subset M$ if there exist:

(a) numbers λ, λ', μ, and μ' such that $0 < \lambda < 1 < \mu$ and $\lambda < \lambda' \leqslant \mu' < \mu$;
(b) a sufficiently small number $\varepsilon > 0$ and Borel functions C, $K : \Lambda \to (0, \infty)$;
(c) subspaces $E^s(x)$, $E^c(x)$, and $E^u(x)$, $x \in \Lambda$,

which satisfy the following conditions:

1′. the subspaces $E^s(x)$, $E^c(x)$, and $E^u(x)$ depend measurably on x and form an invariant splitting of the tangent space, i.e.,

$$\begin{aligned} &T_x M = E^s(x) \oplus E^c(x) \oplus E^u(x), \\ &d_x f E^s(x) = E^s\big(f(x)\big), \qquad d_x f E^c(x) = E^c\big(f(x)\big), \\ &d_x f E^u(x) = E^u\big(f(x)\big); \end{aligned}$$

2′. the subspaces $E^s(x)$ and $E^u(x)$ satisfy (6.2) and (6.3); in addition, for $v \in E^c(x)$ and $n \in \mathbb{Z}$,

$$C(x)^{-1}(\lambda')^n e^{-\varepsilon n}\|v\| \leqslant \|d_x f^n v\| \leqslant C(x)(\mu')^n e^{\varepsilon n}\|v\|;$$

3′. the subspaces $E^s(x)$ and $E^u(x)$ satisfy (6.4); in addition, $\angle(E^s(x), E^c(x)) \geqslant K(x)$ and $\angle(E^u(x), E^c(x)) \geqslant K(x)$;
4′. the functions $C(x)$ and $K(x)$ satisfy (6.5).

In the case $E^c(x) = 0$ we say that f is *nonuniformly* (*completely*) *hyperbolic* on Λ.

Throughout this chapter we deal with three types of nonuniform hyperbolicity: the partial hyperbolicity in the broad sense, its stronger version of partial hyperbolicity (sometimes called partial hyperbolicity in the narrow sense), and yet the stronger complete hyperbolicity (sometimes simply called nonuniform hyperbolicity). We shall refer to subspaces

$E_1(x)$ and, respectively, $E^s(x)$ as *stable* subspaces, to $E^c(x)$ as *central* subspaces and to $E^u(x)$ as *unstable* subspaces. In the case of general nonuniform partial hyperbolicity in the broad sense the subspaces $E_2(x)$ may not be unstable as some vectors may contract under the action of df.

It should be stressed that principle results describing local behavior of the system (such as Stable Manifold Theorem 8.8 and Absolute Continuity Theorems 10.1 and 11.1) as well as some results of a global nature (such as construction of global invariant manifolds in Section 9 and of the pseudo-π-partition in Theorem 11.16 and the lower bound for the metric entropy in Theorem 12.11) need only nonuniform partial hyperbolicity in the broad sense. On the other hand, more advanced results describing ergodic and topological properties of the system require stronger nonuniform complete hyperbolicity, see Sections 11–16.

Consider a diffeomorphism f which is nonuniformly partially hyperbolic in the broad sense on an invariant set Λ. Given $\ell > 0$, we introduce the *regular set* (of level ℓ) by

$$\Lambda^\ell = \left\{ x \in \Lambda \colon C(x) \leqslant \ell,\, K(x) \geqslant \frac{1}{\ell} \right\}.$$

Without loss of generality we may assume that the sets Λ^ℓ are closed (otherwise they can be replaced by their closures $\overline{\Lambda^\ell}$).

We describe a special inner product in the tangent bundle $T\Lambda$ which is known as the *Lyapunov inner product*. It provides a convenient technical tool in studying nonuniform hyperbolicity. Choose numbers $0 < \lambda' < \mu' < \infty$ such that

$$\lambda e^{\varepsilon} < \lambda', \qquad \mu' < \mu e^{-\varepsilon}.$$

We define a new inner product $\langle \cdot, \cdot \rangle'_x$, as follows. Set

$$\langle v, w \rangle'_x = \sum_{k=0}^{\infty} \langle df^k v, df^k w \rangle_{f^k(x)} \lambda'^{-2k}$$

if $v, w \in E_1(x)$, and

$$\langle v, w \rangle'_x = \sum_{k=0}^{\infty} \langle df^{-k} v, df^{-k} w \rangle_{f^{-k}(x)} \mu'^{2k}$$

if $v, w \in E_2(x)$.

Using (6.2) and (6.3) one can verify that each series converges. We extend $\langle \cdot, \cdot \rangle'_x$ to all vectors in $T_x M$ by declaring the subspaces $E_1(x)$ and $E_2(x)$ to be mutually orthogonal with respect to $\langle \cdot, \cdot \rangle'_x$, i.e., we set

$$\langle v, w \rangle'_x = \langle v_1, w_1 \rangle'_x + \langle v_2, w_2 \rangle'_x,$$

where $v = v_1 + v_2$ and $w = w_1 + w_2$ with $v_1, w_1 \in E_1(x)$ and $v_2, w_2 \in E_2(x)$.

The norm induced by the Lyapunov inner product is called the *Lyapunov norm* and is denoted by $\|\cdot\|'$. We emphasize that the Lyapunov inner product, and hence, the norm $\|\cdot\|'$ depend on the choice of numbers λ' and μ'.

The Lyapunov inner product has several important properties:

1. the angle between the subspaces $E_1(x)$ and $E_2(x)$ in the inner product $\langle\cdot,\cdot\rangle'_x$ is $\pi/2$ for each $x \in \Lambda$;
2. $\|A_x\|' \leqslant \lambda'$ and $\|B_x{}^{-1}\|' \leqslant (\mu')^{-1}$;
3. the relation between the Lyapunov inner product and the Riemannian inner product is given by

$$\frac{1}{\sqrt{2}}\|w\|_x \leqslant \|w\|'_x \leqslant D(x)\|w\|_x,$$

where $w \in T_x M$ and

$$D(x) = C(x)K(x)^{-1}\big[(1-\lambda e^{\varepsilon}/\lambda')^{-1} + \big(1-\mu'/(\mu e^{-\varepsilon})\big)^{-1}\big]^{1/2}$$

is a measurable function satisfying (in view of (6.5))

$$D\big(f^m(x)\big) \leqslant D(x)e^{2\varepsilon|m|}, \quad m \in \mathbb{Z}. \tag{6.6}$$

Properties 1 and 2 show that the action of the differential df is *uniformly* partially hyperbolic in the broad sense with respect to the Lyapunov inner product.

For a partially hyperbolic in the broad sense $C^{1+\beta}$ diffeomorphism f the subspaces $E_1(x)$ and $E_2(x)$ depend continuously on the point x in a regular set. Indeed, one can prove a stronger statement.

THEOREM 6.1. *The distribution $E_1(x)$ depends Hölder continuously on $x \in \Lambda^{\ell}$, i.e.,*

$$d\big(E_1(x), E_2(y)\big) \leqslant C\rho(x,y)^{\alpha},$$

where $C > 0$ and $\alpha \in (0,1]$ are constants, and d is the distance in the Grassmannian bundle of TM generated by the Riemannian metric.

This theorem is a particular case of a more general result which we now state.

A k-dimensional distribution E on a subset Λ of a differentiable manifold M is a family of k-dimensional subspaces $E(x) \subset T_x M$, $x \in \Lambda$. A Riemannian metric on M naturally induces distances in TM and in the space of k-dimensional subspaces in TM. The Hölder continuity of a distribution E can be defined using these distances. However, by the Whitney Embedding Theorem, every manifold M can be embedded in $\mathbb{R}^N$ with a sufficiently large N. If M is compact, the Riemannian metric on M is equivalent to the distance $\|x-y\|$ induced by the embedding. The Hölder exponent does not change if the Riemannian metric is changed for an equivalent smooth metric, while the Hölder constant may change. We assume in Theorem 6.2, without loss of generality, that the manifold is embedded in $\mathbb{R}^N$.

For a subspace $A \subset \mathbb{R}^N$ and a vector $v \in \mathbb{R}^N$, set

$$\operatorname{dist}(v, A) = \min_{w \in A} \|v - w\|,$$

i.e., $\operatorname{dist}(v, A)$ is the length of the difference between v and its orthogonal projection to A. For subspaces A, B in $\mathbb{R}^N$, define

$$\operatorname{dist}(A, B) = \max\Big\{ \max_{v \in A,\, \|v\|=1} \operatorname{dist}(v, B), \max_{w \in B,\, \|w\|=1} \operatorname{dist}(w, A) \Big\}.$$

A k-dimensional *distribution* E defined on a set $\Lambda \subset \mathbb{R}^N$ is called *Hölder continuous* with *Hölder exponent* $\alpha \in (0, 1]$ and *Hölder constant* $L > 0$ if there exists $\varepsilon_0 > 0$ such that

$$\operatorname{dist}\big(E(x), E(y)\big) \leqslant L \|x - y\|^\alpha$$

for every $x, y \in \Lambda$ with $\|x - y\| \leqslant \varepsilon_0$.

The subspaces $E_1, E_2 \subset \mathbb{R}^N$ are said to be *κ-transverse* if $\|v_1 - v_2\| \geqslant \kappa$ for all unit vectors $v_1 \in E_1$ and $v_2 \in E_2$.

THEOREM 6.2 (Brin [59]). *Let M be a compact m-dimensional C^2 submanifold of $\mathbb{R}^N$ for some $m < N$, and $f : M \to M$ a $C^{1+\beta}$ map for some $\beta \in (0, 1)$. Assume that there exist a set $\Lambda \subset M$ and real numbers $0 < \lambda < \mu$, $c > 0$, and $\kappa > 0$ such that for each $x \in \Lambda$ there are κ-transverse subspaces $E_1(x), E_2(x) \subset T_x M$ with the following properties*:

1. *$T_x M = E_1(x) \oplus E_2(x)$;*
2. *$\|d_x f^n v_1\| \leqslant c\lambda^n \|v_1\|$ and $\|d_x f^n v_2\| \geqslant c^{-1}\mu^n \|v_2\|$ for every $v_1 \in E_1(x)$, $v_2 \in E_2(x)$, and every positive integer n.*

Then for every $a > \max_{z \in M} \|d_z f\|^{1+\beta}$, the distribution E_1 is Hölder continuous with exponent

$$\alpha = \frac{\log \mu - \log \lambda}{\log a - \log \lambda} \beta.$$

6.3. *Regularity of the derivative cocycle*

We say that a point $x \in M$ is *Lyapunov forward f-regular* (or simply *forward regular*), *Lyapunov backward f-regular* (or simply *backward f-regular*), or *Lyapunov f-regular* (or simply *regular*), respectively, if it is forward regular, backward regular, or regular with respect to the cocycle $\mathcal{A}_x$.

We recall that for any regular point $x \in M$ there exist an integer $s(x) \leqslant n$, numbers $\chi_1(x) < \cdots < \chi_{s(x)}(x)$ and a decomposition

$$T_x M = \bigoplus_{i=1}^{s(x)} E_i(x) \tag{6.7}$$

into subspaces $E_i(x)$ such that for $v \in E_i(x) \setminus \{0\}$ and $i = 1, \dots, s(x)$,

$$\lim_{m \to \pm\infty} \frac{1}{m} \log \|d_x f^m v\| = \chi_i(x)$$

with uniform convergence on $\{v \in E_i(x)\colon \|v\| = 1\}$. Write $k_i(x) = \dim H_i(x)$.

Assume that there exists $C > 0$ such that $\|d_x f\|, \|d_x f^{-1}\| \leqslant C$ for every $x \in M$. Note that this property holds when M is compact. Then the derivative cocycle satisfies the condition (5.3), and by the Multiplicative Ergodic Theorem 5.5 the set of regular points (as well as the sets of forward and backward regular points) is nonempty. Moreover, the following statement is an immediate consequence of Theorem 5.5.

THEOREM 6.3. *Let f be a diffeomorphism of a smooth Riemannian manifold. Then the set of regular points has full measure with respect to any f-invariant Borel probability measure with compact support.*

The set of points which are not regular is negligible from the measure-theoretical point of view, since it has zero measure with respect to any Borel invariant measure. However, this set may be large with respect to other characteristics. For example, it may have positive Lebesgue measure, positive Hausdorff dimension, or positive topological entropy.

Theorem 6.3 does not allow one to determine whether a given trajectory is regular (or forward regular or backward regular). We now present some criteria which guarantee forward and backward regularity of individual trajectories.

Let us first notice that if x is a fixed point or a periodic point for f then the cocycle $\mathcal{A}_x$ is rigid with generator $A = d_x f$ (if x is a fixed point) or $A = d_x f^p$ (if x is a periodic point of period p).

We now consider the case of an arbitrary point x.

PROPOSITION 6.4. *Let f be a diffeomorphism of a smooth Riemannian manifold M.*

1. *If $x \in M$ is such that*

$$\chi^+(x, v_1, \dots, v_k) = \lim_{m \to +\infty} \frac{1}{m} \log V(d_x f^m v_1, \dots, d_x f^m v_k)$$

   (that is, $\chi^+(x, v_1, \dots, v_k)$ is exact), for any choice of linearly independent vectors $v_1, \dots, v_k \in T_x M$ and $k = 1, \dots, n$, then x is forward regular.

2. *If $x \in M$ is such that*

$$\chi^-(x, v_1, \dots, v_k) = \lim_{m \to -\infty} \frac{1}{|m|} \log V(d_x f^m v_1, \dots, d_x f^m v_k)$$

   (that is, $\chi^-(x, v_1, \dots, v_k)$ is exact), for any choice of linearly independent vectors $v_1, \dots, v_k \in T_x M$ and $k = 1, \dots, n$, then x is backward regular.

We also formulate a criterion for regularity.

PROPOSITION 6.5. *Let f be a diffeomorphism of a smooth Riemannian manifold M and $x \in M$. Assume that*:

1. *$\chi^+(x, v_1, \dots, v_k)$ and $\chi^-(x, v_1, \dots, v_k)$ are exact for any choice of linearly independent vectors $v_1, \dots, v_k \in T_x M$ and $k = 1, \dots, n$;*
2. *$s^+(x) = s^-(x) \stackrel{\text{def}}{=} s(x)$ and $\chi_i^+(x) = -\chi_i^-(x)$ for $i = 1, \dots, s(x)$;*
3. *$\bigoplus_{i=1}^{s(x)} (V_i^+(x) \cap V_i^-(x)) = \mathbb{R}^n$ where $\{V_i^+\}$ and $\{V_i^-\}$ are filtrations associated with the Lyapunov exponents χ^+ and χ^-.*

Then x is regular.

The diffeomorphism f acts on the cotangent bundle T^*M by its codifferential

$$d_x^* f : T_{f(x)}^* M \to T_x^* M$$

defined by

$$d_x^* f \varphi(v) = \varphi(d_x f v), \quad v \in T_x M, \ \varphi \in T_{f(x)}^* M.$$

We denote the inverse map by

$$d_x' f = (d_x^* f)^{-1} : T_x^* M \to T_{f(x)}^* M.$$

Let ν be an ergodic f-invariant Borel measure. There exist numbers $s = s^\nu$, $\chi_i = \chi_i^\nu$, and $k_i = k_i^\nu$ for $i = 1, \dots, s$ such that

$$s(x) = s, \quad \chi_i(x) = \chi_i, \quad k_i(x) = k_i \tag{6.8}$$

for ν-almost every x. The collection of pairs

$$\operatorname{Sp} \chi(\nu) = \{(\chi_i, k_i) \colon 1 \leqslant i \leqslant s\}$$

is called the *Lyapunov spectrum* of the measure ν.

A diffeomorphism f is a *dynamical system with nonzero Lyapunov exponents* if there exists an ergodic f-invariant Borel probability measure ν on M—a *hyperbolic measure*—such that the set

$$\Lambda = \{x \in \mathcal{L} \colon \text{there exists } 1 \leqslant k(x) < s(x) \text{ with } \chi_{k(x)}(x) < 0 \text{ and } \chi_{k(x)+1}(x) > 0\}$$

has full measure.

Consider the set $\tilde{\Lambda} = \tilde{\Lambda}^\nu$ of those points in Λ which are Lyapunov regular and satisfy (6.8). By the Multiplicative Ergodic Theorem 5.5, we have $\nu(\tilde{\Lambda}) = 1$. For every $x \in \tilde{\Lambda}$, set

$$E^s(x) = \bigoplus_{i=1}^{k} E_i(x) \quad \text{and} \quad E^u(x) = \bigoplus_{i=k+1}^{s} E_i(x).$$

THEOREM 6.6. *The subspaces $E^s(x)$ and $E^u(x)$, $x \in \tilde{\Lambda}$, have the following properties*:

1. *they depend Borel measurably on x*;
2. *they form a splitting of the tangent space, i.e., $T_x M = E^s(x) \oplus E^u(x)$*;
3. *they are invariant,*

$$d_x f E^s(x) = E^s\big(f(x)\big) \quad \text{and} \quad d_x f E^u(x) = E^u\big(f(x)\big).$$

Furthermore, there exist $\varepsilon_0 > 0$, Borel functions $C(x,\varepsilon) > 0$ and $K(x,\varepsilon) > 0$, $x \in \tilde{\Lambda}$ and $0 < \varepsilon \leqslant \varepsilon_0$ such that

4. *the subspace $E^s(x)$ is* stable: *if $v \in E^s(x)$ and $n > 0$, then*

$$\|d_x f^n v\| \leqslant C(x,\varepsilon) e^{(\chi_k + \varepsilon)n} \|v\|;$$

5. *the subspace $E^u(x)$ is* unstable: *if $v \in E^u(x)$ and $n < 0$, then*

$$\|d_x f^n v\| \leqslant C(x,\varepsilon) e^{(\chi_{k+1} - \varepsilon)n} \|v\|;$$

6. $$\angle\big(E^s(x), E^u(x)\big) \geqslant K(x,\varepsilon);$$

7. *for every $m \in \mathbb{Z}$,*

$$C\big(f^m(x),\varepsilon\big) \leqslant C(x,\varepsilon) e^{\varepsilon|m|} \quad \text{and} \quad K\big(f^m(x),\varepsilon\big) \geqslant K(x,\varepsilon) e^{-\varepsilon|m|}.$$

We remark that condition 7 is crucial and is a manifestation of the regularity property. It follows from Theorem 6.6 that f is nonuniformly completely hyperbolic on $\tilde{\Lambda}$.

6.4. *Cocycles over smooth flows*

Let φ_t be a smooth flow on a smooth n-dimensional Riemannian manifold M. It is generated by the vector field $\mathcal{X}$ on M given by

$$\mathcal{X}(x) = \left.\frac{d\varphi_t(x)}{dt}\right|_{t=0}.$$

For every $x_0 \in M$ the trajectory $\{x(x_0,t) = \varphi_t(x_0)\colon\ t \in \mathbb{R}\}$ represents a solution of the nonlinear differential equation

$$\dot{v} = \mathcal{X}(v)$$

on the manifold M. This solution is uniquely determined by the initial condition $x(x_0,0) = x_0$.

Given a point $x \in M$ and the trajectory $\{\varphi_t(x)\colon\ t \in \mathbb{R}\}$ passing through x we introduce the *variational differential equation*

$$\dot{w}(t) = A(x,t)w(t), \tag{6.9}$$

where

$$A(x,t) = d\mathcal{X}\big(\varphi_t(x)\big).$$

This is a linear differential equation along the trajectory $\{\varphi_t(x)\colon\ t \in \mathbb{R}\}$ known also as the *linear variational equation*.

The Lyapunov exponent generated by the cocycle $\mathcal{A}$ is defined by

$$\chi^+(x,v) = \overline{\lim_{t\to+\infty}} \frac{1}{t} \log\|w(t)\|,$$

where $w(t)$ is the solution of (6.9) with initial condition $w(0) = v$, and is called the *Lyapunov exponent of the flow* φ_t. In particular, one can speak of trajectories which are *forward* or *backward regular*, and (*Lyapunov*) *regular*.

Note that every periodic trajectory is regular. However, this is not true in general for nonperiodic trajectories. For example, consider a flow on the unit sphere with the North and the South poles to be, respectively, attracting and repelling points, and without other fixed points. If the coefficients of contraction and expansion are different then every trajectory of the flow (except for the North and the South poles) is nonregular.

One can establish a criterion for regularity of individual trajectories (see [35]). However, it is not a simple task to apply this criterion and check whether a given trajectory is regular. On the other hand, let ν be a Borel measure which is invariant under the flow φ_t. It is easy to see that the derivative cocycle $\mathcal{A}(x,t)$ satisfies

$$\sup_{-1\leqslant t\leqslant 1} \log^+\|\mathcal{A}(\cdot,t)\| \in L^1(M,\nu).$$

The Multiplicative Ergodic Theorem for flows (see Theorem 5.15) implies that almost every trajectory with respect to ν is Lyapunov regular.

We say that a smooth flow φ_t is *nonuniformly hyperbolic* if it possesses an invariant Borel subset $\Lambda \subset M$ such that there exist:

(a) numbers $0 < \lambda < 1 < \mu$;

(b) a sufficiently small number $\varepsilon > 0$ and Borel functions $C, K : \Lambda \to (0,\infty)$;

(c) subspaces $E^s(x)$ and $E^u(x)$, $x \in \Lambda$,

which satisfy conditions $1'$–$4'$ in the definition of nonuniform partial hyperbolicity with $E^c(x) = \mathcal{X}(x)$. Note that for every t the diffeomorphism φ_t is nonuniformly partially hyperbolic with one-dimensional central subspace.

Assume that a smooth flow φ_t possesses an invariant Borel subset Λ and an invariant Borel measure ν with $\nu(\Lambda) = 1$ such that $\chi(x,v) \neq 0$ for almost every $x \in \Lambda$ and every $v \in T_xM$ not colinear with $\mathcal{X}$. Assume also that for these x there are vectors $v, w \in T_xM$ such that $\chi(x,v) > 0$ and $\chi(x,w) < 0$. Then the flow φ_t is nonuniformly hyperbolic on Λ.

7. Methods for estimating exponents

The absence of zero Lyapunov exponents implies nonuniform hyperbolicity. In fact, this seems to be one of the most "practical" universal ways to establish weak hyperbolic be-

havior. We discuss a powerful method which allows one to verify that Lyapunov exponents do not vanish. It was suggested by Wojtkowski in [248] and is a significant generalization of the initial approach by Alekseev (see [14–16]) to build an invariant family of unstable cones.

The *cone* of size $\gamma > 0$ centered around $\mathbb{R}^{n-k}$ in the product space $\mathbb{R}^n = \mathbb{R}^k \times \mathbb{R}^{n-k}$ is

$$C_\gamma = \{(v, w) \in \mathbb{R}^k \times \mathbb{R}^{n-k} \colon \|v\| < \gamma \|w\|\} \cup \{(0, 0)\}.$$

Note that $\{0\} \times \mathbb{R}^{n-k} \subset C_\gamma$ for every γ.

Consider a cocycle $\mathcal{A}$ over an invertible measurable transformation $f : X \to X$ preserving a Borel probability measure ν on X, and let $A : \mathbb{R}^n \to GL(n, \mathbb{R})$ be its generator. Assume that there exist $\gamma > 0$ and $a > 1$ such that for ν-almost every $x \in \mathbb{R}^n$:

1. $A(x)C_\gamma \subset C_\gamma$;
2. $\|A(x)v\| \geqslant a\|v\|$ for every $v \in C_\gamma$.

Then the largest Lyapunov exponent can be shown to be positive ν-almost everywhere. Indeed, $n - k$ values of the Lyapunov exponent, counted with their multiplicities, are positive.

Wojtkowski's great insight is that condition 1 alone is in fact sufficient to establish positivity of the values of the Lyapunov exponent. The importance of this observation is that condition 1 is of pure qualitative nature and thus, no estimates on the growth of vectors inside the cone are required.

It turns out that Wojtkowski's approach can be described in a more general and more convenient framework elaborated by Burns and Katok in [136]. This approach, in turn, is a further development of that by Lewowicz in [162,163] and Markarian in [179] and is based on the notion of infinitesimal Lyapunov function (see Section 7.1 below; see also Section 7.2 for the version of this approach in the case of cocycles with values in the symplectic group).

In the later work Wojtkowski himself strengthened his original approach and, using results of Potapov on monotone operators of a linear space generated by a quadratic form, obtained estimates of Lyapunov exponents for cocycles with values in the semigroup of matrices preserving the form (see [249]). These results apply to estimate Lyapunov exponents for Hamiltonian dynamical systems as well as to the Boltzmann–Sinai gas of hard spheres and the system of falling balls in one dimension (see [249] for more details and references therein).

7.1. *Cone and Lyapunov function techniques*

Let $Q : \mathbb{R}^n \to \mathbb{R}$ be a continuous function which is homogeneous of degree one (i.e., $Q(\alpha v) = \alpha Q(v)$ for any $v \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$) and takes on both positive and negative values. The subset

$$C^+(Q) \stackrel{\text{def}}{=} \{0\} \cup Q^{-1}(0, +\infty) \subset \mathbb{R}^n \tag{7.1}$$

is called the *positive* (*generalized*) *cone associated with* Q or simply the *positive cone of* Q. Similarly,

$$C^-(Q) \stackrel{\text{def}}{=} \{0\} \cup Q^{-1}(-\infty, 0) \subset \mathbb{R}^n \tag{7.2}$$

is the *negative* (*generalized*) *cone associated to* Q or the *negative cone of* Q. The maximal dimension of a linear subspace $L \subset \mathbb{R}^n$ such that $L \subset C^+(Q)$ (respectively, $L \subset C^-(Q)$) is called *positive* (respectively, *negative*) *rank* of Q and is denoted by $r^+(Q)$ (respectively, $r^-(Q)$). We clearly have $r^+(Q) + r^-(Q) \leqslant n$, and since Q takes on both positive and negative values, we have $r^+(Q) \geqslant 1$ and $r^-(Q) \geqslant 1$. We call the function Q *complete* if

$$r^+(Q) + r^-(Q) = n. \tag{7.3}$$

For example, consider the function

$$Q(v) = \operatorname{sign} K(v, v) \cdot \big|K(v, v)\big|^{1/2}, \tag{7.4}$$

where K is a nondegenerate indefinite quadratic form. Q is complete and its positive and negative ranks are equal to the number of positive and negative eigenvalues of the quadratic form K, respectively.

More generally, let λ be a positive real number and K_λ a real function on $\mathbb{R}^n$ which is homogeneous of degree λ (i.e., $K_\lambda(\alpha v) = \alpha^\lambda K_\lambda(v)$ for any $v \in \mathbb{R}^n$ and $\alpha > 0$) and takes on both positive and negative values. Define a homogeneous function Q of degree one by

$$Q(v) = \operatorname{sign} K_\lambda(v) \cdot \big|K_\lambda(v)\big|^{1/\lambda}.$$

We say that K_λ is *complete* if Q is complete, and we define the positive and negative cones, and positive and negative ranks of K_λ as those of Q.

Let $\mathcal{A}: X \times \mathbb{Z} \to GL(n, \mathbb{R})$ be a cocycle, and $F: X \times \mathbb{R}^n \to X \times \mathbb{R}^n$ its linear extension defined by

$$F(x, v) = \big(f(x), A(x)v\big),$$

where $A(x) = \mathcal{A}(x, 1)$ is the generator of $\mathcal{A}$.

A real-valued measurable function Q on $X \times \mathbb{R}^n$ is called a *Lyapunov function for the extension* F or *for the cocycle* $\mathcal{A}$ (with respect to a measure ν in X) if there exist positive integers r_Q^+ and r_Q^- such that for ν-almost every $x \in X$,

1. the function Q_x given by $Q_x(v) = Q(x, v)$ is continuous, homogeneous of degree one and takes on both positive and negative values;
2. Q_x is complete and $r^+(Q_x) = r_Q^+$ and $r^-(Q_x) = r_Q^-$;
3. $$Q_{f(x)}\big(A(x)v\big) \geqslant Q_x(v) \quad \text{for all } v \in \mathbb{R}^n. \tag{7.5}$$

The numbers r_Q^+ and r_Q^- are called the *positive* and *negative ranks* of Q.

When Q is a Lyapunov function, it follows from (7.5) that for ν-almost every $x \in X$,

$$\begin{aligned} &A(x)C^+(Q_x) \subset C^+(Q_{f(x)}),\\ &A\big(f^{-1}(x)\big)^{-1}C^-(Q_x) \subset C^-(Q_{f^{-1}(x)}). \end{aligned} \tag{7.6}$$

A Lyapunov function Q on $X \times \mathbb{R}^n$ is called *strict* if the inequality in (7.5) is strict for every $v \neq 0$ and *eventually strict* if for ν-almost every $x \in X$ there exists a positive integer $m = m(x)$ such that for every $v \in \mathbb{R}^n \setminus \{0\}$,

$$Q_{f^m(x)}\big(\mathcal{A}(x,m)v\big) > Q_x(v) \tag{7.7}$$

and

$$Q_{f^{-m}(x)}\big(\mathcal{A}(x,-m)v\big) < Q_x(v). \tag{7.8}$$

If a Lyapunov function Q is eventually strict then by (7.5), for ν-almost every $x \in X$ the inequalities (7.7) and (7.8) hold for all $m \geqslant m(x)$.

When Q is a strict Lyapunov function, it follows from (7.5) that

$$\begin{aligned} &A(x)\overline{C^+(Q_x)} \subsetneqq C^+(Q_{f(x)}),\\ &A\big(f^{-1}(x)\big)^{-1}\overline{C^-(Q_x)} \subsetneqq C^-(Q_{f^{-1}(x)}) \end{aligned} \tag{7.9}$$

for ν-almost every $x \in X$. Furthermore, if Q is an eventually strict Lyapunov function it follows from (7.7) and (7.8) that

$$\begin{aligned} &\mathcal{A}(x,m)\overline{C^+(Q_x)} \subsetneqq C^+(Q_{f^m(x)}),\\ &\mathcal{A}(x,-m)^{-1}\overline{C^-(Q_x)} \subsetneqq C^-(Q_{f^{-m}(x)}) \end{aligned} \tag{7.10}$$

for ν-almost every $x \in X$ and every $m \geqslant m(x)$.

The following result establishes a criterion for nonvanishing Lyapunov exponents.

THEOREM 7.1 (Burns and Katok [136]). *If $\mathcal{A}$ possesses an eventually strict Lyapunov function Q then*

1. *$\mathcal{A}$ has ν-almost everywhere r_Q^+ positive and r_Q^- negative values of the Lyapunov exponent counted with their multiplicities*;
2. *for ν-almost every $x \in X$ we have*

$$E^+(x) = \bigcap_{m=1}^{\infty} \mathcal{A}\big(f^{-m}(x), m\big)\overline{C^+(Q_{f^{-m}(x)})} \subset C^+(Q_x)$$

and

$$E^-(x) = \bigcap_{m=1}^{\infty} \mathcal{A}\big(f^{m}(x), -m\big)\overline{C^-(Q_{f^{m}(x)})} \subset C^-(Q_x).$$

Lyapunov functions are intimately related to the invariant families of cones. Here we give a detailed description of this relationship.

A (*generalized*) *cone* C in $\mathbb{R}^n$ is a homogeneous set (i.e., $\alpha v \in C$ whenever $v \in C$ and $\alpha \in \mathbb{R}$) such that $C \setminus \{0\}$ is open. In particular, C need not be convex and $\operatorname{int} C$ need not be connected. The *rank* of C is the maximal dimension of a linear subspace $L \subset \mathbb{R}^n$ which is contained in C. We denote it by $r(C)$. The *complementary cone* $\hat{C}$ to C is defined by

$$\hat{C} = (\mathbb{R}^n \setminus \bar{C}) \cup \{0\}.$$

Obviously the complementary cone to $\hat{C}$ is C. We have $r(C) + r(\hat{C}) \leqslant n$ and this inequality may be strict (this is the case for example, when $C \neq \mathbb{R}^n$ but $\bar{C} = \mathbb{R}^n$). A pair of complementary cones C and $\hat{C}$ is called *complete* if $r(C) + r(\hat{C}) = n$.

Let $\mathcal{A}: X \times \mathbb{Z} \to GL(n, \mathbb{R})$ be a cocycle over X with generator $A: X \to GL(n, \mathbb{R})$. Consider a measurable family of cones $C = \{C_x\colon\ x \in X\}$ in $\mathbb{R}^n$. Given a measure ν in X, we say that

1. C is *complete* if the pair of complementary cones $(C_x, \hat{C}_x)$ is complete for ν-almost every $x \in X$;
2. C is *$\mathcal{A}$-invariant* if for ν-almost every $x \in X$,

$$A(x)C_x \subset C_{f(x)}, \qquad A\big(f^{-1}(x)\big)^{-1}\hat{C}_x \subset \hat{C}_{f^{-1}(x)}.$$

Let C be an $\mathcal{A}$-invariant measurable family of cones. We say that

1. C is *strict* if for ν-almost every $x \in X$,

$$A(x)\overline{C_x} \subsetneqq C_{f(x)}, \qquad A\big(f^{-1}(x)\big)^{-1}\overline{\hat{C}_x} \subsetneqq \hat{C}_{f^{-1}(x)};$$

2. C is *eventually strict* if for ν-almost every $x \in X$ there exists $m = m(x) \in \mathbb{N}$ such that

$$\mathcal{A}(x, m)\overline{C_x} \subsetneqq C_{f^m(x)}, \qquad \mathcal{A}(x, -m)^{-1}\overline{\hat{C}_x} \subsetneqq \hat{C}_{f^{-m}(x)}.$$

Let C be a complete $\mathcal{A}$-invariant measurable family of cones in $\mathbb{R}^n$. Any Lyapunov function $Q: X \times \mathbb{R}^n \to \mathbb{R}$ satisfying

$$C_x = C^+(Q_x), \quad \hat{C}_x = C^-(Q_x) \quad \text{for } \nu\text{-almost every } x \in X$$

is called a *Lyapunov function associated with* C. Any complete $\mathcal{A}$-invariant measurable family of cones has an associated Lyapunov function. It is given by

$$Q(x, v) = \begin{cases} d(v/\|v\|, \partial C_x)\|v\| & \text{if } v \in C_x, \\ -d(v/\|v\|, \partial C_x)\|v\| & \text{if } v \in \hat{C}_x. \end{cases}$$

Furthermore, if a complete invariant family of cones is strict (respectively, eventually strict) then any of its associated Lyapunov functions is strict (respectively, eventually strict).

The above discussion allows us to rephrase Theorem 7.1 in the following fashion.

THEOREM 7.2. *If* (5.3) *holds for some f-invariant measure ν, and there exists a complete $\mathcal{A}$-invariant measurable family of cones $C = \{C_x\colon x \in X\}$, then*

1. *$\mathcal{A}$ has ν-almost everywhere r_Q^+ positive and r_Q^- negative values of the Lyapunov exponent counted with their multiplicities*;
2. *for ν-almost every $x \in X$ we have*

$$E^+(x) = \bigcap_{m=1}^{\infty} \mathcal{A}\big(f^{-m}(x), m\big)\overline{C_{f^{-m}(x)}} \subset C_x$$

and

$$E^-(x) = \bigcap_{m=1}^{\infty} \mathcal{A}\big(f^{m}(x), -m\big)\overline{\hat{C}_{f^{m}(x)}} \subset \hat{C}_x.$$

Let Q be a Lyapunov function on $X \times \mathbb{R}^n$ for a cocycle $\mathcal{A}$. We consider the family of cones $C = \{C_x\colon x \in X\}$ in $\mathbb{R}^n$ given by

$$C_x = C^+(Q_x).$$

Conditions 2 and 3 in the definition of Lyapunov function imply that C is complete and $\mathcal{A}$-invariant. Note that the complementary cone $\hat{C}_x$ is not always equal to the cone $C^-(Q_x)$, and thus Q may not be a Lyapunov function associated with C. However, we have $\hat{C}_x = C^-(Q_x)$ provided that for each v such that $Q_x(v) = 0$ one can find w arbitrarily close to v such that $Q_x(w) > 0$. Furthermore, if Q is strict (respectively, eventually strict) then C is strict (respectively, eventually strict).

7.2. *Cocycles with values in the symplectic group*

Let $\mathcal{A}\colon X \times \mathbb{Z} \to GL(n, \mathbb{R})$ be a cocycle and Q a homogeneous function of degree one on $X \times \mathbb{Z}$. Consider the corresponding families of cones $C^+(Q_x)$ and $C^-(Q_x)$ given by (7.1) and (7.2). If Q is complete (see (7.3)) and (7.6) holds, then Q is a Lyapunov function. Moreover, if (7.9) (respectively, (7.10)) holds then Q is strict (respectively, eventually strict). On the other hand, if condition (7.6) is satisfied only with respect to the family of cones $C^+(Q_x)$ then Q may not be a Lyapunov function. However, this does occur for some interesting classes of cocycles and cones. The most important case for applications involves cocycles with values in the symplectic group $\mathrm{Sp}(2m, \mathbb{R})$, $m \geqslant 1$, and the so-called symplectic cones which we define later.

We begin with the simple case of $SL(2, \mathbb{R})$ cocycles.

We call a cone in $\mathbb{R}^n$ *connected* if its projection to the projective space $\mathbb{R}P^{n-1}$ is a connected set. A connected cone in $\mathbb{R}^2$ is simply the union of two opposite sectors bounded by two different straight lines intersecting at the origin plus the origin itself. By a linear coordinate change such a cone can always be reduced to the following standard cone:

$$S = \big\{(v, w) \in \mathbb{R}\colon vw > 0\big\} \cup \big\{(0, 0)\big\}.$$

THEOREM 7.3 (see [136]). *If a cocycle with values in $SL(2, \mathbb{R})$ has an eventually strictly invariant family of connected cones $C = \{C_x\colon\ x \in X\}$, then it has an eventually strict Lyapunov function Q such that for ν-almost every $x \in X$ the function Q_x has the form* (7.4) *and its zero set coincides with the boundary of the cone C_x.*

Let us now proceed with the general symplectic case. We denote by ω the standard symplectic form in $\mathbb{R}^{2m}$,

$$\omega(v, w) = \sum_{i=1}^{m} (v_i w_{m+i} - w_i v_{m+i}),$$

and by K the following nondegenerate quadratic form of signature zero:

$$K(v) = \sum_{i=1}^{m} v_i v_{m+i}.$$

The cone

$$S = \left\{v \in \mathbb{R}^{2m}\colon\ K(v) > 0\right\} \cup \{0\}$$

is called the *standard symplectic cone*. The image of this cone under an invertible linear symplectic map (i.e., a map with values in $\mathrm{Sp}(2m, \mathbb{R})$) is called a *symplectic cone*.

Let L_1 and L_2 be two transverse Lagrangian subspaces in a $2m$-dimensional symplectic space (H, ω), i.e., complementary m-dimensional subspaces on which the symplectic form ω vanishes identically. Then for any $v \in H$ there is a unique decomposition

$$v = v_1 + v_2 \quad \text{with } v_i \in L_i \text{ for } i = 1, 2.$$

Let

$$K_{L_1,L_2}(v) = \omega(v_1, v_2) \quad \text{and} \quad C_{L_1,L_2} = K_{L_1,L_2}^{-1}((0, \infty)) \cup \{0\}.$$

Then C_{L_1,L_2} is a symplectic cone and K_{L_1,L_2} is the corresponding quadratic form.

It is easy to see (for example, by a direct calculation for the case of standard cones), that for a given symplectic cone C in a symplectic space there are exactly two isolated Lagrangian subspaces L_1 and L_2 that belong to the boundary of C and that $C = C_{L_1,L_2}$ or $C = C_{L_2,L_1}$. Thus, the cone C canonically determines the form K,

$$K(C) = K_{L_1,L_2} \quad \text{or} \quad K(C) = K_{L_2,L_1},$$

depending on whether the form K_{L_1,L_2} or the form K_{L_2,L_1} is positive on C.

For example, the standard cone S is C_{L_1,L_2}, where

$$L_1 = \left\{(x, 0)\colon\ x \in \mathbb{R}^m\right\} \quad \text{and} \quad L_2 = \left\{(0, x)\colon\ x \in \mathbb{R}^m\right\}.$$

PROPOSITION 7.4. *Let H and H' be two $2m$-dimensional spaces, $L_1, L_2 \subset H$ and $L_1', L_2' \subset H'$ pairs of transverse Lagrangian subspaces and $T : H \to H'$ a symplectic linear transformation such that $T\overline{C_{L_1,L_2}} \subset C_{L_1,L_2}$. Then for all $v \in H \setminus \{0\}$ we have*

$$K_{L_1',L_2'}(Tv) > K_{L_1,L_2}(v).$$

Proposition 7.4 immediately implies the following relation between invariant cone families and Lyapunov functions.

THEOREM 7.5 (see [136]). *Let $A : X \to \mathrm{Sp}(2m, \mathbb{R})$ be a cocycle over a measurable transformation $f : X \to X$ which preserves a measure ν. If A has an eventually strictly invariant family of symplectic cones $C = \{C_x : x \in X\}$, then it also has an eventually strict Lyapunov function Q such that for ν-almost every $x \in X$ the function Q_x has the form* (7.4) *with a quadratic form $K = K_x$ of signature zero. Furthermore, the zero set of the function Q_x coincides with the boundary of the cone C_x.*

Combining Theorem 7.5 with Theorem 7.1 we immediately obtain the following.

COROLLARY 7.6. *If a cocycle $A : X \to \mathrm{Sp}(2m, \mathbb{R})$ satisfies* (5.3) *and has an eventually strictly invariant family of symplectic cones, then the linear extension F of f has ν-almost everywhere m positive and m negative values of the Lyapunov exponent.*

7.3. *Lyapunov exponents estimates for some particular cocycles*

The cone techniques provide some general methodology for establishing positivity of Lyapunov exponents for cocycles and in particular, for dynamical systems. However, in some particular cases one can use more effective tools and obtain sharper estimates of Lyapunov exponents.

7.3.1. *Herman's method* We describe a method due to Herman [115] for obtaining a lower bound for the maximal Lyapunov exponent of a holomorphic cocycle with values in a Banach algebra, in particular, with values in $\mathbb{C}^p$. This method is based on some properties of pluri-subharmonic functions.

For $r > 0$, let

$$B^n(0, r) = \left\{(z_1, \dots, z_n) \in \mathbb{C}^n : |z_i| \leqslant r,\ 1 \leqslant i \leqslant n\right\}$$

be the closed ball and

$$\mathbb{T}_r^n = \left\{(z_1, \dots, z_n) \in \mathbb{C}^n : |z_i| = r,\ 1 \leqslant i \leqslant n\right\}$$

the torus in C^n. Let also $f : U \to \mathbb{C}^n$ be a holomorphic function in a neighborhood U of $B^n(0, r)$ satisfying $f(0) = 0$, $f(B^n(0, r)) \subset B^n(0, r)$, and $f(\mathbb{T}_r^n) \subset \mathbb{T}_r^n$. We also consider

a Banach algebra $\mathcal{B}$ over $\mathbb{C}$ and a cocycle $\mathcal{A}\colon \mathbb{T}^n_r \times \mathbb{N} \to \mathcal{B}$ over f with values in $\mathcal{B}$. We have

$$\mathcal{A}(z,m) = A\big(f^{m-1}(z)\big)\dots A\big(f(z)\big)A(z),$$

where $A\colon X \to \mathcal{B}$ is the generator of the cocycle. Denote by

$$\rho(B) = \lim_{m\to\infty} \|B^m\|^{1/m} = \inf_{m\geqslant 1} \|B^m\|^{1/m}$$

the spectral radius of the element $B \in \mathcal{B}$ (where $\|\cdot\|$ is the norm in $\mathbb{C}$).

THEOREM 7.7. *If f preserves the Lebesgue measure μ in $\mathbb{T}^n_r$, and A is a holomorphic map in a neighborhood of $B^n(0,r)$ with values in a Banach algebra $\mathcal{B}$, then the cocycle $\mathcal{A}$ over f with generator A satisfies*

$$\varliminf_{m\to\infty} \frac{1}{m} \int_{\mathbb{T}^n_r} \log\big\|\mathcal{A}(z,m)\big\|\, d\mu(z) \geqslant \log \rho\big(A(0)\big).$$

To see this set

$$a_m = \int_{\mathbb{T}^n_r} \log\big\|\mathcal{A}(z,m)\big\|\, d\mu(z).$$

Since the function $z \mapsto \log\|\mathcal{A}(z,m)\|$ is pluri-subharmonic for each m (see [122]),

$$a_m \geqslant \log\big\|\mathcal{A}(0,z)\big\| = \log\big\|A(0)^m\big\|.$$

Therefore,

$$\inf_{m\geqslant 1} \frac{a_m}{m} \geqslant \log \rho\big(A(0)\big).$$

Since f preserves μ, the sequence a_m is subadditive and thus,

$$\varliminf_{m\to\infty} \frac{a_m}{m} = \lim_{m\to\infty} \frac{a_m}{m} = \inf_{m\geqslant 1} \frac{a_m}{m}$$

and the desired result follows.

7.3.2. *Parameter-exclusion techniques* In [258], Young considered a C^1 family of cocycles over irrational rotations $R_\alpha(x)$ by $2\pi\alpha$ with generators $A_t\colon S^1 \to SL(2,\mathbb{R})$ such that $|A_t(x)| \approx \chi$ (uniformly in t and x) where $\chi > 0$ is a number. The cocycles are not uniformly hyperbolic. The statement is that *for sufficiently large χ and for a generic family the set of parameters (α, t), for which the Lyapunov exponents of (R_α, A_t) are $\approx \pm\chi$, has nearly full measure*. The proof exploits a parameter-exclusion procedure which goes

back to the work of Jacobson [128] and of Benedicks and Carleson [40]: inductively, one identifies certain regions of criticality, studies orbit segments that begin and end near those regions and tries to concatenate long blocks of matrices that have been shown to be hyperbolic; parameters are deleted to ensure the hyperbolicity of the concatenated blocks, and the induction moves forward.

The parameter-exclusion techniques is used to study hyperbolic and ergodic properties of Hénon-like attractors, see Section 14.4.

7.3.3. *Open set of nonuniformly hyperbolic cocycles with values in $SL(2,\mathbb{R})$* In [257], Young constructed an open set, in the C^1 topology, of cocycles with values in $SL(2,\mathbb{R})$ over a hyperbolic automorphism T of the 2-torus $\mathbb{T}^2$ such that every cocycle in this set has positive Lyapunov exponent but is not uniformly hyperbolic.

Choose $\lambda > \sqrt{\mu}+1$ where $\mu > 1$ is the eigenvalue of the matrix T (the other eigenvalue is μ^{-1}). Given $\varepsilon > 0$, we define a cocycle over T with the generator $A_\varepsilon : \mathbb{T}^2 \to SL(2,\mathbb{R})$ as follows. Let $0 < \beta < 2\pi$ be a number, $J_\varepsilon \subset S^1$ an interval, and $\varphi_\varepsilon : \mathbb{T}^2 \to \mathbb{R}/2\pi\mathbb{R}$ a C^1 function such that

1. $\varphi_\varepsilon \equiv 0$ outside of $J_\varepsilon \times S^1$;
2. on $J_\varepsilon \times S^1$, φ_ε increases monotonically from 0 to 2π along the leaves of W^u;
3. on $\varphi_\varepsilon^{-1}[\beta, 2\pi-\beta]$, the directional derivatives of φ_ε along the leaves of W^u are $\geqslant \frac{1}{\varepsilon}$.

The cocycle A_ε is defined to be

$$A_\varepsilon(x) = \begin{pmatrix} \lambda & 0 \\ 0 & \frac{1}{\lambda} \end{pmatrix} \circ R_{\varphi_\varepsilon(x)},$$

where R_θ is the rotation by the angle θ. The statement is that *one can choose β and, for all sufficiently small ε, the interval J_ε and a neighborhood $\mathcal{U}_\varepsilon$ of A_ε in $C^1(\mathbb{T}^2, SL(2,\mathbb{R}))$ such that for any $B \in \mathcal{U}_\varepsilon$ the cocycle over T with the generator B is not uniformly hyperbolic and has a positive Lyapunov exponent with respect to the Lebesgue measure.*

7.3.4. *Cocycles associated with the Jacobi–Perron (JP) algorithm* This algorithm is a higher-dimensional generalization of the continued fraction algorithm and is used to construct simultaneous rational approximations of real numbers (see [225,154]). The map f defining the JP algorithm acts on the d-dimensional cube I^d by the formula

$$f(x) = \left(\frac{x_2}{x_1} \bmod 1, \ldots, \frac{x_d}{x_1} \bmod 1, \frac{1}{x_1} \bmod 1 \right)$$

provided $x_1 \neq 0$. The map f preserves a probability measure ν which is absolutely continuous with respect to the Lebesgue measure in the cube and is ergodic with respect to ν.

The JP algorithm associates to almost every point $x \in I^d$ a matrix $A(x)$ such that x can be expressed as $x = a_1 \circ \cdots \circ a_n \circ f^n(x)$ where a_n are the projective maps defined by the matrices $A_n = A(f^{n-1}(x))$ in the space $\mathbb{R}^n \subset \mathbb{P}^n$. The $d+1$ points

$$J_n = a_1 \circ \cdots \circ a_n(0), \quad \ldots, \quad J_{n+d} = a_1 \circ \cdots \circ a_{n+d}(0)$$

form a simplex $\sigma_n(xa)$ in $\mathbb{R}^d$ which contains x. Its asymptotic form turns out to be determined by the Lyapunov exponents of the measure ν. The latter are closely related to the Lyapunov exponents χ_i, $i = 1, \dots, d+1$, of the cocycle over f generated by the matrix function $A = A(x)$.

In [60], Broise-Alamichel and Guivarc'h showed that for the JP algorithm:

1. $\sum_{i=1}^{d+1} \chi_i = 0$ and $\chi_1 > \chi_2 > \cdots > \chi_{d+1}$;
2. $\chi_1 + \chi_{d+1} > 0$.

In the case $d = 2$ we have that $\chi_2 < 0$.

7.3.5. *Partially hyperbolic cocycles over locally maximal hyperbolic sets* Let f be a diffeomorphism of a compact smooth manifold possessing a locally maximal hyperbolic set Λ. Assume that $f|\Lambda$ is topologically transitive. Let μ be an equilibrium measure on Λ corresponding to a Hölder continuous potential φ.

Consider a cocycle $\mathcal{A}$ over f with values in $SL(p, \mathbb{R})$ and let A be the generator of the cocycle. We assume that A depends smoothly on x and that it is dominated by the hyperbolicity of f, i.e., $A(x)$ expands vectors less than the minimum expansion induced by $d_x f$ on the unstable subbundle and $A(x)$ contracts vectors less than the minimum contraction induced by $d_x f$ on the stable subbundle. In other words, the cocycle is partially hyperbolic on $X \times \mathbb{R}^p$.

In [50], Bonatti, Gómez-Mont and Viana showed that the maximal Lyapunov exponent of μ, χ_μ, is zero only in the following very special situation: there exists a continuous family of probability measures m_x, $x \in \Lambda$, on the projective space $\mathbb{C}P^{p-1}$ which is simultaneously invariant under f, and under the holonomies along the strongly stable and strongly unstable foliations. One can deduce from here that the set of cocycles with a nonzero upper Lyapunov exponent with respect to all the equilibrium measures is an open and dense set in the C^1 topology. It is also shown that for generic C^1 families of cocycles with finitely many parameters, the set of parameters for which the upper Lyapunov exponent is zero for some equilibrium measure is discrete.

8. Local manifold theory

We consider the problem of local stability of trajectories for nonuniformly partially and completely hyperbolic systems. This includes constructing local stable and unstable manifolds and studying their properties. Let us emphasize that the construction of stable (unstable) manifolds can be carried out if only one nonuniformly hyperbolic trajectory is present, i.e., the nonuniformly partially (or completely) hyperbolic set Λ consists of a single trajectory. In particular, the construction does not involve any invariant measure.

There are two well-known methods of building local stable manifolds originated in works of Hadamard [111] and Perron [194,195]. Hadamard's approach is more geometrical and can be effected for Lipschitz (not necessarily differentiable) maps while Perron's approach allows more flexibility.

These methods work well in the case of uniform hyperbolicity and extending them to nonuniformly hyperbolic systems faces substantial problems. One of them is that the size

of local stable manifolds may deteriorate along the trajectory and indeed, may become arbitrarily small. The crucial requirement (6.5) in the definition of nonuniform hyperbolicity provides a control of the deterioration: it can occur with at most subexponential rate.

Both Hadamard and Perron methods allow substantial generalizations to sequences of local diffeomorphisms (instead of iterations of a single diffeomorphism) or maps of Banach spaces (instead of Euclidean spaces), etc.

8.1. *Nonuniformly hyperbolic sequences of diffeomorphisms*

Let $f_m : U_m \to \mathbb{R}^n$, $m \in \mathbb{Z}$ ($U_m \subset \mathbb{R}^n$ is an open set) be a (two-sided) sequence of C^1 local diffeomorphisms, and $\{\langle\cdot,\cdot\rangle_m\}_{m\in\mathbb{Z}}$ a (two-sided) sequence of metrics. Write $\mathcal{F} = \{f_m\}_{m\in\mathbb{Z}}$. We assume that $f_m(0) = 0$.

We say that $\mathcal{F}$ is *nonuniformly hyperbolic* if so is the sequence of matrices $\{A_m\}_{m\in\mathbb{Z}} = \{d_0 f_m\}_{m\in\mathbb{Z}}$.

Let $\mathbb{R}^n = E^1_m \oplus E^2_m$ be the invariant splitting associated with nonuniform hyperbolic structure. For every $m \in \mathbb{Z}$ and $(x, y) \in U_m$ one can write f_m in the form

$$f_m(x, y) = \big(A_m x + g^1_m(x, y), B_m y + g^2_m(x, y)\big),$$

where $A_m = d_0 f_m|E^1_m$ and $B_m = d_0 f_m|E^2_m$ are linear invertible transformations and $g_m = (g^1_m, g^2_m) : U_m \to \mathbb{R}^n$ are C^1 maps satisfying $g_m(0) = 0$, $d_0 g_m(0) = 0$.

Set

$$\sigma_m = \sup\big\{\|d_{(x,y)} g_m\| \colon (x, y) \in U_m\big\}, \qquad \sigma = \sup\{\sigma_m \colon m \in \mathbb{Z}\}.$$

Note that σ need not be finite in general.

Let $\mathcal{A} = \{A_m\}_{m\in\mathbb{Z}}$ and $\mathcal{B} = \{B_m\}_{m\in\mathbb{Z}}$. Define new sequences of matrices $\{\mathcal{A}_m\}_{m\in\mathbb{Z}}$ and $\{\mathcal{B}_m\}_{m\in\mathbb{Z}}$ using

$$\mathcal{A}_m = \begin{cases} A_{m-1} \dots A_1 A_0 & \text{if } m > 0, \\ \mathrm{Id} & \text{if } m = 0, \\ (A_m)^{-1} \dots (A_{-2})^{-1} (A_{-1})^{-1} & \text{if } m < 0. \end{cases}$$

We also set

$$\mathcal{F}_m = \begin{cases} f_{m-1} \circ \dots \circ f_1 \circ f_0 & \text{if } m > 0, \\ \mathrm{Id} & \text{if } m = 0, \\ (f_m)^{-1} \circ \dots \circ (f_{-2})^{-1} \circ (f_{-1})^{-1} & \text{if } m < 0, \end{cases}$$

whenever it is defined. The map $(\mathcal{A}_m, \mathcal{B}_m)$ is a linear approximation of $\mathcal{F}_m$ in a neighborhood of 0. We shall describe in the following sections how the stability of the linear approximation effects the stability of the sequence of C^1 local diffeomorphisms.

8.2. *Admissible manifolds and the graph transform*

Let $\gamma > 0$, $k, n \in \mathbb{N}$, $k < n$ be given. A map $\varphi : U \to \mathbb{R}^{n-k}$ with $U \subset \mathbb{R}^k$ is called *γ-Lipschitz* if for every $x, x' \in U$,

$$\big\|\varphi(x) - \varphi(x')\big\| \leqslant \gamma \|x - x'\|.$$

A set $V \subset \mathbb{R}^n$ is said to be

1. An *admissible (s, γ)-set* if there exists a γ-Lipschitz map $\varphi : U \subset \mathbb{R}^k \to \mathbb{R}^{n-k}$ such that

$$V = \mathrm{Graph}(\varphi) = \big\{\big(x, \varphi(x)\big) \colon x \in U\big\}.$$

   If, in addition, φ is differentiable then V is called an *admissible (s, γ)-manifold*.
2. An *admissible (u, γ)-set* if there exists a γ-Lipschitz map $\varphi : U \subset \mathbb{R}^{n-k} \to \mathbb{R}^k$ such that

$$V = \mathrm{Graph}(\varphi) = \big\{\big(\varphi(x), x\big) \colon x \in U\big\}.$$

   If, in addition, φ is differentiable then V is called an *admissible (u, γ)-manifold*.

Given $\gamma > 0$, let $\Gamma(u, \gamma)$ be the space of sequences $\{V_m\}_{m\in\mathbb{Z}}$ of admissible (u, γ)-sets such that $0 \in V_m$. We define a metric on $\Gamma(u, \gamma)$ by

$$d_{\Gamma(u,\gamma)}\big(\{V_{1m}\}_{m\in\mathbb{Z}}, \{V_{2m}\}_{m\in\mathbb{Z}}\big) = \sup\big\{d_m(\varphi_{1m}, \varphi_{2m}) \colon m \in \mathbb{Z}\big\},$$

where

$$d_m(\varphi_{1m}, \varphi_{2m}) = \sup\left\{\frac{\|\varphi_{1m}(x) - \varphi_{2m}(x)\|}{\|x\|} \colon x \in U \setminus \{0\}\right\}$$

and $V_{im} = \mathrm{Graph}(\varphi_{im})$ for each $m \in \mathbb{Z}$ and $i = 1, 2$. Since $\varphi_{1m}(0) = \varphi_{2m}(0) = 0$, and φ_{1m} and φ_{2m} are γ-Lipschitz we have $d_m(\varphi_{1m}, \varphi_{2m}) \leqslant 2\gamma$ and the metric $d_{\Gamma(u,\gamma)}$ is well defined. One can verify that $\Gamma(u, \gamma)$ is a complete metric space.

We define the *graph transform $G : \Gamma(u, \gamma) \to \Gamma(u, \gamma)$ induced by $\mathcal{F}$ on $\Gamma(u, \gamma)$* by $G(\{V_m\}_{m\in\mathbb{Z}}) = \{f_m(V_m)\}_{m\in\mathbb{Z}}$.

PROPOSITION 8.1. *If V_m is an admissible (u, γ)-set such that*

$$\sigma_m \leqslant \frac{(\mu' - \lambda')\gamma}{(1+\gamma)^2}, \tag{8.1}$$

then $f_m V_m$ is an admissible (u, γ)-set.

It follows that under assumption (8.1) the map G is well defined.

PROPOSITION 8.2. *Assume that*

$$\sigma < \frac{\mu' - \lambda'}{2(1+\gamma)}.$$

Then the graph transform G is a contraction on $\Gamma(u,\gamma)$.

As an immediate corollary we obtain existence of (u,γ)-sets.

THEOREM 8.3. *Assume that*

$$\sigma \leqslant \frac{(\mu' - \lambda')\gamma}{(1+\gamma)^2} \quad \text{and} \quad \sigma < \frac{\mu' - \lambda'}{2(1+\gamma)}. \tag{8.2}$$

Then there exists a unique family $\{V_m^u\}_{m\in\mathbb{Z}}$ of admissible (u,γ)-sets such that $0 \in V_m^u$ and $f_m(V_m^u) = V_{m+1}^u$.

Note that for $\gamma < 1$ the second inequality in (8.2) follows from the first one.

We now briefly describe how to obtain similar results for (s,γ)-manifolds. For every $m \in \mathbb{Z}$ and $(x, y) \in f_m(U_m)$ one can write f_m^{-1} in the form

$$f_m^{-1}(x,y) = \big(A_m^{-1}x + h_m^1(x,y),\, B_m^{-1}y + h_m^2(x,y)\big),$$

where $h_m = (h_m^1, h_m^2) \colon f_m(U_m) \to \mathbb{R}^n$ is a C^1 map satisfying $h_m(0) = 0$ and $d_0 h_m = 0$. Let

$$\tau_m = \sup\big\{\|d_{(x,y)}h_m\| \colon (x,y) \in f_m U_m\big\}, \qquad \tau = \sup\{\tau_m \colon m \in \mathbb{Z}\}.$$

THEOREM 8.4. *Assume that*

$$\tau \leqslant \frac{(\mu' - \lambda')\gamma}{\lambda'\mu'(1+\gamma)^2} \quad \text{and} \quad \tau < \frac{\mu' - \lambda'}{2\lambda'\mu'(1+\gamma)}.$$

Then there exists a unique family $\{V_m^s\}_{m\in\mathbb{Z}}$ of admissible (s,γ)-sets such that $0 \in V_m^s$ and $f_m(V_m^s) = V_{m+1}^s$.

The following theorem substantially strengthen the above result by claiming that (s,γ) and (u,γ)-sets are indeed smooth manifolds.

THEOREM 8.5 (Katok and Mendoza [139]). *Let $\{f_m\}_{m\in\mathbb{Z}}$ be a nonuniformly hyperbolic sequence of C^1 local diffeomorphisms defined on the whole $\mathbb{R}^n$. Given $\gamma > 0$ and a sufficiently small $\sigma > 0$, there exist a unique family $\{V_m^s\}_{m\in\mathbb{Z}}$ of C^1 admissible (s,γ)-manifolds and a unique family $\{V_m^u\}_{m\in\mathbb{Z}}$ of C^1 admissible (u,γ)-manifolds such that*:

1. $0 \in V_m^s \cap V_m^u$;
2. $f_m(V_m^s) = V_{m+1}^s$ *and* $f_m(V_m^u) = V_{m+1}^u$;
3. $T_0 V_m^s = E_m^s$ *and* $T_0 V_m^u = E_m^u$;

4. *if* $(x, y) \in V_m^s$ *then*

$$\|f_m(x, y)\| \leqslant (1+\gamma)(\lambda+\sigma_m)\|(x, y)\|$$

and if $(x, y) \in V_m^u$ *then*

$$\|f_m(x, y)\| \geqslant \big(\mu/(1+\gamma) - \sigma_m\big)\|(x, y)\|,$$

where $0 < \lambda < 1 < \mu$;

5. *for every* $(1+\gamma)(\lambda+\sigma) < \nu < \mu/(1+\gamma) - \sigma$ *and* $(x, y) \in \mathbb{R}^n$, *if there exists* $C > 0$ *such that*

$$\|\mathcal{F}_{m+k} \circ \mathcal{F}_m^{-1}(x, y)\| \leqslant C\nu^k\|(x, y)\|$$

for every $k \geqslant 0$ *then* $(x, y) \in V_m^s$, *and if there exists* $C > 0$ *such that*

$$\|\mathcal{F}_{m+k} \circ \mathcal{F}_m^{-1}(x, y)\| \leqslant C\nu^k\|(x, y)\|$$

for every $k \leqslant 0$ *then* $(x, y) \in V_m^u$.

Notice that an admissible (s, γ)-manifold (respectively, (u, γ)-manifold) is also an admissible (s, γ')-manifold (respectively, (u, γ')-manifold) for every $\gamma' > \gamma$. Therefore, the uniqueness property in Theorem 8.5 implies that both families $\{V_m^s\}_{m\in\mathbb{Z}}$ and $\{V_m^u\}_{m\in\mathbb{Z}}$ are independent of γ. These families are called, respectively, *family of invariant s-manifolds* and *family of invariant u-manifolds*. They can be characterized as follows.

PROPOSITION 8.6. *For each* $\gamma \in (0, \sqrt{\mu'/\lambda'} - 1)$ *and each sufficiently small* $\sigma > 0$:

1. *if*

$$\nu \in \big((1+\gamma)(\lambda+\sigma), \mu/(1+\gamma) - \sigma\big)$$

then

$$V_m^s = \left\{(x, y) \in \mathbb{R}^n\colon \overline{\lim_{k\to+\infty}} \frac{1}{k}\log\|\mathcal{F}_{m+k} \circ \mathcal{F}_m^{-1}(x, y)\| < \log\nu\right\}$$

and

$$V_m^u = \left\{(x, y) \in \mathbb{R}^n\colon \overline{\lim_{k\to-\infty}} \frac{1}{|k|}\log\|\mathcal{F}_{m+k} \circ \mathcal{F}_m^{-1}(x, y)\| < -\log\nu\right\};$$

2. *if* $(1+\gamma)(\lambda+\sigma) < 1 < \mu/(1+\gamma) - \sigma$ *then*

$$V_m^s = \left\{(x,y)\in\mathbb{R}^n\colon \varlimsup_{k\to+\infty}\frac{1}{k}\log\left\|\mathcal{F}_{m+k}\circ\mathcal{F}_m^{-1}(x,y)\right\|<0\right\}$$
$$=\left\{(x,y)\in\mathbb{R}^n\colon \sup_{k\geqslant 0}\left\|\mathcal{F}_{m+k}\circ\mathcal{F}_m^{-1}(x,y)\right\|<\infty\right\}$$
$$=\left\{(x,y)\in\mathbb{R}^n\colon \mathcal{F}_{m+k}\circ\mathcal{F}_m^{-1}(x,y)\to 0 \text{ as } k\to+\infty\right\}$$

and

$$V_m^u = \left\{(x,y)\in\mathbb{R}^n\colon \varlimsup_{k\to-\infty}\frac{1}{|k|}\log\left\|\mathcal{F}_{m+k}\circ\mathcal{F}_m^{-1}(x,y)\right\|<0\right\}$$
$$=\left\{(x,y)\in\mathbb{R}^n\colon \sup_{k\leqslant 0}\left\|\mathcal{F}_{m+k}\circ\mathcal{F}_m^{-1}(x,y)\right\|<\infty\right\}$$
$$=\left\{(x,y)\in\mathbb{R}^n\colon \mathcal{F}_{m+k}\circ\mathcal{F}_m^{-1}(x,y)\to 0 \text{ as } k\to-\infty\right\}.$$

The following result provides some additional information on higher differentiability of (s,γ)- and (u,γ)-manifolds.

THEOREM 8.7. *Let $\mathcal{F}$ be a sequence of C^r local diffeomorphisms, for some $r>0$. Then the unique family $\{V_m^u\}_{m\in\mathbb{Z}}$ of admissible (u,γ)-sets given by Theorem* 8.3 *is composed of C^r manifolds.*

8.3. *Hadamard–Perron Theorem: Perron's method*

We describe a version of Perron's approach to the proof of the Stable Manifold Theorem which originated in [197] and allows one to construct stable (and unstable) invariant manifolds along a single nonuniformly partially hyperbolic trajectory in the broad sense.[5]

Let f be a $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold M and $x\in M$. Assume that f is nonuniformly partially hyperbolic in the broad sense on the set $\Lambda=\{f^n(x)\}_{n\in\mathbb{Z}}$ (see Section 6.2). We obtain the local stable manifold in the form

$$V(x)=\exp_x\left\{\big(x,\psi(x)\big)\colon x\in B_1(r)\right\}, \tag{8.3}$$

where $\psi: B_1(r)\to E_2(x)$ is a smooth map, satisfying $\psi(0)=0$ and $d\psi(0)=0$, $E_1(x)$, $E_2(x)$ are invariant distributions in the tangent space (see (6.1)), and $B_1(r)\in E_1(x)$ is the ball of radius r centered at the origin. The number $r=r(x)$ is called the *size* of the local stable manifold.

We now describe how to construct the function ψ. Fix $x\in M$ and consider the map

$$\tilde{f}_x=\exp_{f(x)}^{-1}\circ f\circ\exp_x: B_1(r)\times B_2(r)\to T_{f(x)}M,$$

[5] In [197], the system is assumed to preserve a hyperbolic smooth measure. However, the proof does not use this assumption and readily extends to the case of a single nonuniformly partially hyperbolic trajectory in the broad sense. This was observed in [215].

which is well defined if r is sufficiently small. Here $B_2(r)$ is the ball of radius r in $E_1(x)$ centered at the origin. The map $\tilde{f}$ can be written in the following form:

$$\tilde{f}_x(v_1, v_2) = \big(A_x v_1 + g_{1x}(v_1, v_2),\, B_x v_2 + g_{2x}(v_1, v_2)\big),$$

where $v_1 \in E_1(x)$ and $v_2 \in E_2(x)$. Furthermore,

$$A_x : E_1(x) \to E_2\big(f(x)\big) \quad \text{and} \quad B_x : E_2(x) \to E_2\big(f(x)\big)$$

are linear maps. The map A_x is a contraction and the map B_x is an expansion. Since f is of class $C^{1+\alpha}$ we also have for $g = (g_1, g_2)$,

$$\big\|g_x(v)\big\| \leqslant C_1 \|v\|^{1+\alpha} \tag{8.4}$$

and

$$\big\|dg_x(v) - dg_x(w)\big\| \leqslant C_1 \|v - w\|^{\alpha}, \tag{8.5}$$

where $C_1 > 0$ is constant (which may depend on x).

In other words the map $\tilde{f}_x$ can be viewed as a small perturbation of the linear map $(v_1, v_2) \mapsto (A_x v_1, B_x v_2)$ by the map $g_x(v_1, v_2)$ satisfying conditions (8.4) and (8.5) in a small neighborhood U_x of the point x.

Note that size of U_x depends on x and may decay along the trajectory of x with subexponential rate (see (6.6)). This requires a substantial modification of the classical Perron's approach.

Proceeding further with Perron's approach we identify each of the tangent spaces $T_{f^m(x)}M$ with $\mathbb{R}^p = \mathbb{R}^k \times \mathbb{R}^{p-k}$ (recall that $p = \dim M$ and $1 \leqslant k < p$) via an isomorphism τ_m such that $\tau_m(E_1(x)) = \mathbb{R}^k$ and $\tau_m(E_2(x)) = \mathbb{R}^{p-k}$. The map $\tilde{F}_m = \tau_{m+1} \circ F_m \circ {\tau_m}^{-1}$ is of the form

$$\tilde{F}_m(v_1, v_2) = \big(A_m v_1 + g_{1m}(v_1, v_2),\, B_m v_2 + g_{2m}(v_1, v_2)\big), \tag{8.6}$$

where $A_m : \mathbb{R}^k \to \mathbb{R}^k$ and $B_m : \mathbb{R}^{p-k} \to \mathbb{R}^{p-k}$ are linear maps, and $g : \mathbb{R}^n \to \mathbb{R}^k$ is a nonlinear map defined for each $v_1 \in B_1(r_0) \subset \mathbb{R}^k$ and $v_2 \in B_2(r_0) \subset \mathbb{R}^{p-k}$. With respect to the Lyapunov inner product these maps satisfy:

$$\|A_m\|' \leqslant \lambda', \quad \big(\big\|B_m^{-1}\big\|'\big)^{-1} \geqslant \mu', \quad \text{where } 0 < \lambda' < \min\{1, \mu'\}, \tag{8.7}$$

and

$$g_m(0) = 0, \qquad dg_m(0) = 0,$$
$$\big\|dg_m(v) - dg_m(w)\big\|' \leqslant C_2 \gamma^{-m} \|v - w\|'^{\alpha},$$

where

$$\lambda'^{\alpha} < \gamma < 1, \quad 0 < \alpha \leqslant 1, \quad C > 0$$

(see (8.5)). We now state a general version of the Stable Manifold Theorem.

THEOREM 8.8 (Pesin [197]). *Let κ be any number satisfying*

$$\lambda' < \kappa < \min\{\mu', \gamma^{1/\alpha}\}. \tag{8.8}$$

There exist $D > 0$ and $r_0 > r > 0$, and a map $\psi : B_1(r) \to \mathbb{R}^{p-k}$ such that:

1. *ψ is of class $C^{1+\alpha}$ and $\psi(0) = 0$, $d\psi(0) = 0$;*
2. *$\|d\psi(v) - d\psi(w)\|' \leqslant D\|v - w\|'^{\alpha}$ for any $v, w \in B_1(r)$;*
3. *if $m \geqslant 0$ and $v \in B_1(r)$ then*

$$\left(\prod_{i=0}^{m-1} \tilde{F}_i\right)\big(v, \psi(v)\big) \in B_1(r) \times B_2(r),$$

$$\left\|\left(\prod_{i=0}^{m-1} \tilde{F}_i\right)\big(v, \psi(v)\big)\right\|' \leqslant D\kappa^m \big\|\big(v, \psi(v)\big)\big\|',$$

   where $\prod_{i=0}^{m-1} \tilde{F}_i$ denotes the composition $\tilde{F}_{m-1} \circ \cdots \circ \tilde{F}_0$ (with the convention that $\prod_{i=0}^{-1} \tilde{F}_i = \mathrm{Id}$);
4. *given $v \in B_1(r)$ and $w \in B_2(r)$, if there is a number $K > 0$ such that*

$$\left(\prod_{i=0}^{m-1} \tilde{F}_i\right)(v, w) \in B_1(r) \times B_2(r), \qquad \left\|\left(\prod_{i=0}^{m-1} \tilde{F}_i\right)(v, w)\right\|' \leqslant K\kappa^m$$

   for every $m \geqslant 0$, then $w = \psi(v)$;
5. *the numbers D and r depend only on the numbers λ', μ', γ, α, κ, and C.*

We outline the proof of the theorem. Consider the linear space Γ_κ of sequences of vectors $z = \{z(m) \in \mathbb{R}^p\}_{m \in \mathbb{N}}$ satisfying

$$\|z\|_\kappa = \sup_{m \geqslant 0} \big(\kappa^{-m} \big\|z(m)\big\|'\big) < \infty.$$

Γ_κ is a Banach space with the norm $\|z\|_\kappa$. Given $r > 0$, set

$$W = \big\{z \in \Gamma_\kappa \colon z(m) \in B_1(r) \times B_2(r) \text{ for every } m \in \mathbb{N}\big\}.$$

Since $0 < \kappa < 1$ the set W is open. Consider the map $\Phi_\kappa : B_1(r_0) \times W \to \Gamma_\kappa$ given by

$$\Phi_\kappa(y, z)(0) = \left(y, -\sum_{k=0}^{\infty}\left(\prod_{i=0}^{k} B_i\right)^{-1} g_{2k}\big(z(k)\big)\right),$$

and for $m > 0$,

$$\Phi_\kappa(y,z)(m) = -z(m) + \left(\left(\prod_{i=0}^{m-1} A_i\right)y, 0\right) + \left(\sum_{n=0}^{m-1}\left(\prod_{i=n+1}^{m-1} A_i\right)g_{1n}\big(z(n)\big), - \sum_{n=0}^{\infty}\left(\prod_{i=0}^{n} B_{i+m}\right)^{-1} g_{2n+m}\big(z(n+m)\big)\right).$$

Using conditions (8.7)–(8.8) one can show that the map Φ_κ is well defined, continuously differentiable over y and z and $\Phi_\kappa(0,0) = (0,0)$. Moreover, Φ_κ is of class C^1 with partial derivatives given by

$$\partial_y \Phi_\kappa(y,z)(m) = \mathcal{A}_\kappa(z) - \mathrm{Id},$$

where

$$\big(\mathcal{A}_\kappa(z)\big)t(m) = \left(\sum_{n=0}^{m-1}\left(\prod_{i=n+1}^{m-1} A_i\right)dg_{1n}\big(z(n)\big)t(n), - \sum_{n=0}^{\infty}\left(\prod_{i=0}^{n} B_{i+m}\right)^{-1} dg_{2n+m}\big(z(n+m)\big)t(m+n)\right).$$

Furthermore,

$$\big\|\mathcal{A}_\kappa(z_1) - \mathcal{A}_\kappa(z_2)\big\| \leqslant C\|z_1 - z_2\|_\kappa^\alpha, \tag{8.9}$$

where $C > 0$ is a constant. We have, in particular, that $\partial_z \Phi_\kappa(y,0) = -\mathrm{Id}$ and the map $\partial_z \Phi_\kappa(y,z)$ is continuous. Therefore, the map Φ_κ satisfies the conditions of the Implicit Function Theorem, and hence, there exist a number $r \leqslant r_0$ and a map $\varphi : B_1(r) \to W$ of class C^1 with

$$\varphi(0) = 0 \quad \text{and} \quad \Phi_\kappa\big(y, \varphi(y)\big) = 0. \tag{8.10}$$

Note that the derivatives $\partial_y \Phi_\kappa$ and $\partial_z \Phi_\kappa$ are Hölder continuous. It is clear for the former and follows for the latter in view of (8.9):

$$\begin{aligned}\big\|\partial_z \Phi_\kappa(y_1,z_1) - \partial_z \Phi_\kappa(y_2,z_2)\big\| &\leqslant \big\|\partial_z \Phi_\kappa(y_1,z_1) - \partial_z \Phi_\kappa(y_1,z_2)\big\| + \big\|\partial_z \Phi_\kappa(y_1,z_2) - \partial_z \Phi_\kappa(y_2,z_2)\big\| \\ &= 2\big\|\mathcal{A}_\kappa(z_1) - \mathcal{A}_\kappa(z_2)\big\| \leqslant CM\|z_1 - z_2\|_\kappa^\alpha.\end{aligned}$$

There is a special version of the Implicit Function Theorem for maps with Hölder continuous derivatives (see [35]) which enables one to obtain an explicit estimate of the number r and to show that it depends only on λ', μ', γ, α, κ, and C.

We now describe some properties of the map φ. Differentiating the second equality in (8.10) with respect to y we obtain

$$d\varphi(y) = -\big[\partial_z \Phi_\kappa\big(y, \varphi(y)\big)\big]^{-1} \partial_y \Phi_\kappa\big(y, \varphi(y)\big).$$

Setting $y = 0$ in this equality yields

$$d\varphi(0)(m) = \left(\prod_{i=0}^{m-1} A_i, 0\right).$$

One can write the vector $\varphi(y)(m)$ in the form

$$\varphi(y)(m) = \big(\varphi_1(y)(m), \varphi_2(y)(m)\big),$$

where $\varphi_1(y)(m) \in \mathbb{R}^k$ and $\varphi_2(y)(m) \in \mathbb{R}^{p-k}$. It follows from (8.10) that if $m \geqslant 0$ then

$$\varphi_1(y)(m) = \left(\prod_{i=0}^{m-1} A_i\right) y + \sum_{n=0}^{m-1} \left(\prod_{i=n+1}^{m-1} A_i\right) g_{1n}\big(\varphi(y)(n)\big) \tag{8.11}$$

and

$$\varphi_2(y)(m) = -\sum_{n=0}^{\infty} \left(\prod_{i=0}^{n} B_{i+m}\right)^{-1} g_{2n+m}\big(\varphi(y)(n+m)\big). \tag{8.12}$$

These equalities imply that

$$\varphi_1(y)(m+1) = A_m \varphi_1(y)(m) + g_{1m}\big(\varphi_1(y)(m), \varphi_2(y)(m)\big),$$
$$\varphi_2(y)(m+1) = B_m \varphi_2(y)(m) + g_{2m}\big(\varphi_1(y)(m), \varphi_2(y)(m)\big).$$

Indeed, iterating the first equality "forward" one easily obtains (8.11). Rewriting the second equality in the form

$$\varphi_2(y)(m) = B_m^{-1} \varphi_2(y)(m+1) - B_m^{-1} g_{2m}\big(\varphi_1(y)(m), \varphi_2(y)(m)\big)$$

and iterating it "backward" yields (8.12).

Thus, we obtain that the function $\varphi(y)$ is invariant under the family of maps $\tilde{F}_m$, i.e.,

$$\tilde{F}_m\big(\varphi(y)(m)\big) = \varphi(y)(m+1).$$

The desired map ψ^s is now defined by $\psi(v) = \varphi_2(v)(0)$ for each $v \in B^s(r)$.

Applying the above result to a diffeomorphism f which is nonuniformly partially hyperbolic in the broad sense along the trajectory of a point $x \in M$ we obtain the following version of the Stable Manifold Theorem.

THEOREM 8.9. *There exists a local stable manifold $V(x)$ such that $x \in V(x)$, $T_x V(x) = E_1(x)$, and for $y \in V(x)$ and $n \geqslant 0$,*

$$\rho\big(f^n(x), f^n(y)\big) \leqslant T(x)\lambda^n e^{\varepsilon n}\rho(x, y), \tag{8.13}$$

where $T : \Lambda \to (0, \infty)$ is a Borel function satisfying

$$T\big(f^m(x)\big) \leqslant T(x)e^{10\varepsilon|m|}, \quad m \in \mathbb{Z}. \tag{8.14}$$

In [208], Pugh constructed an explicit example of a nonuniformly completely hyperbolic diffeomorphism of a 4-dimensional manifold of class C^1 (and not of class $C^{1+\alpha}$ for any $\alpha > 0$) for which the statement of Theorem 8.9 fails. More precisely, there exists no manifold tangent to $E_1(x)$ such that (8.13) holds on some open neighborhood of x. This example illustrates that the assumption $\alpha > 0$ in Theorem 8.9 is crucial. Barreira and Valls [38] have shown that there is a class of C^1 vector fields that are not $C^{1+\alpha}$ for any $\alpha > 0$ whose nonuniformly hyperbolic trajectories possess stable manifolds.[6]

One can obtain a more refined information about smoothness of local stable manifolds. More precisely, let f be a diffeomorphism of class $C^{p+\alpha}$, with $p \geqslant 1$ and $0 < \alpha \leqslant 1$. Assume that f is nonuniformly partially hyperbolic in the broad sense along a trajectory of a point $x \in M$. Then the local stable manifold $V(x)$ is of class C^p; in particular, if f is of class C^p for some $p \geqslant 2$, then $V(x)$ is of class C^{p-1} (and even of class $C^{p-1+\alpha}$ for any $0 < \alpha < 1$). These results are immediate consequences of the following version of Theorem 8.8.

THEOREM 8.10 (Pesin [197]). *Assume that the conditions of Theorem* 8.8 *hold. In addition, assume that*:

1. *g_m are of class C^p for some $p \geqslant 2$;*
2. *there exists $K > 0$ such that for $\ell = 1, \ldots, p$,*

$$\sup_{z \in B}\big\|d^\ell g_m(z)\big\|' \leqslant K\gamma^{-m}, \qquad \sup_{z \in B}\big\|d^\ell h_m(z)\big\|' \leqslant K\gamma^{-m},$$

   where $B = B_1(r_0) \times B_2(r_0)$ (see (8.6));
3. *for $z_1, z_2 \in B$ and some $\alpha \in (0, 1)$,*

$$\big\|d^p g_m(z_1) - d^p g_m(z_2)\big\|' \leqslant K\gamma^{-m}\big(\|z_1 - z_2\|'\big)^\alpha.$$

If $\psi(u)$ is the map constructed in Theorem 8.8, *then there exists a number $N > 0$, which depends only on the numbers λ', μ', γ, α, κ, and K, such that*:

[6] A similar statement holds for diffeomorphisms, see L. Barreira and C. Valls, *Existence of stable manifolds for nonuniformly hyperbolic C^1 dynamics*, Discrete Contin. Dynam. Systems, to appear.

1. *ψ is of class $C^{p+\alpha}$;*
2. *$\sup_{u\in B_1(r)} \|d^\ell \psi(u)\|' \leqslant N$ for $\ell = 1, \dots, p$.*

In [209], Pugh and Shub strengthened the above result and showed that in fact, if f is of class C^p for some $p \geqslant 2$, then $V(x)$ is also of class C^p.

In the case of diffeomorphisms which are nonuniformly partially hyperbolic, in particular, nonuniformly completely hyperbolic, there is a symmetry between the objects marked by the index "s" and those marked by the index "u". Namely, when the time direction is reversed the statements concerning objects with index "s" become the statements about the corresponding objects with index "u". In these cases we shall denote the local stable manifold at x by $V^s(x)$. We can also construct the local unstable manifolds.

THEOREM 8.11 (Unstable Manifold Theorem). *Let f be a $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold M which is nonuniformly partially hyperbolic along the trajectory of a point $x \in M$. Then there exists a local unstable manifold $V^u(x)$ such that $x \in V^u(x)$, $T_x V^u(x) = E^u(x)$, and if $y \in V^u(x)$ and $n \leqslant 0$ then*

$$\rho\big(f^n(x), f^n(y)\big) \leqslant T(x)\mu^n e^{\varepsilon|n|}\rho(x, y),$$

where $T : \Lambda \to (0, \infty)$ is a Borel function satisfying (8.14).

Stable Manifold Theorem 8.9 was first established by Pesin in [197]. His proof is built upon classical work of Perron. Katok and Strelcyn [142] extended Stable Manifold Theorem to smooth maps with singularities (see Section 18). They essentially followed Pesin's approach. Ruelle [215] obtained another proof of Theorem 8.9, based on his study of perturbations of the matrix products in the Multiplicative Ergodic Theorem 5.5. Fathi, Herman, and Yoccoz [94] provided a detailed exposition of Theorem 8.9 which essentially follows the approaches of Pesin and Ruelle. Pugh and Shub [209] proved Stable Manifold Theorem for nonuniformly partially hyperbolic systems using graph transform techniques.

On another direction, Liu and Qian [166] established a version of Theorem 8.9 for random maps (see the article by Kifer and Liu [8] in this volume). One can extend the Stable Manifold Theorem 8.9 to infinite-dimensional spaces. Ruelle [216] proved this theorem for Hilbert spaces, closely following his approach in [215], and Mañé [173] considered Banach spaces (under certain compactness assumptions on the dynamics).

8.4. *Stable Manifold Theorem for flows*

Let φ_t be a smooth flow on a compact smooth Riemannian manifold M. The following is an analog of Theorem 8.9 for flows.

THEOREM 8.12. *Assume that φ_t is nonuniformly hyperbolic along a trajectory $\varphi_t(x)$. Then there exists a local stable manifold $V^s(x)$ satisfying*:

(a) $x \in V^s(x)$,

(b) $T_x V^s(x) = E^s(x)$,

(c) *if* $y \in V^s(x)$ *and* $t > 0$ *then*

$$\rho\big(\varphi_t(x), \varphi_t(y)\big) \leqslant T(x)\lambda^t e^{\varepsilon t}\rho(x, y),$$

where $T : \Lambda \to (0, \infty)$ *is a Borel function such that for* $s \in \mathbb{R}$,

$$T\big(\varphi_s(x)\big) \leqslant T(x)e^{10\varepsilon|s|}.$$

The proof of Theorem 8.12 can be obtained by applying Theorem 8.9 to the diffeomorphism $f = \varphi_1$ (that is nonuniformly partially hyperbolic). We call $V^s(x)$ a *local stable manifold* at x.

By reversing the time one can construct a *local unstable manifold* $V^u(x)$ at x. It has the properties similar to those of the stable manifold.

8.5. *Continuity and sizes of local manifolds*

Recall that the size of the local stable manifold $V(x)$ at a point $x \in \Lambda$ (with Λ as in Section 8.3) is the number $r = r(x)$ that is determined by Theorem 8.8 and such that (8.3) holds. It follows from statement 5 of Theorem 8.8 that the sizes of the local stable manifold at a point x and any point $y = f^m(x)$ along the trajectory of x are related by

$$r\big(f^m(x)\big) \geqslant Ke^{-\varepsilon|m|}r(x), \tag{8.15}$$

where $K > 0$ is a constant.

Assume now that f is nonuniformly partially hyperbolic in the broad sense on an invariant set Λ, and let ν be an f-invariant ergodic Borel measure with $\nu(\Lambda) = 1$. For all sufficiently large ℓ the regular set Λ^ℓ has positive measure. Therefore, the trajectory of almost every point visits Λ^ℓ infinitely many times. It follows that for typical points x the function $r(f^m(x))$ is an oscillating function of m which is of the same order as $r(x)$ for many values of m. Nevertheless, for some integers m the value $r(f^m(x))$ may become as small as it is allowed by (8.15). Let us emphasize that the rate with which the sizes of the local stable manifolds $V(f^m(x))$ decreases as $m \to +\infty$ is smaller than the rate with which the trajectories $\{f^m(x)\}$ and $\{f^m(y)\}$, $y \in V(x)$ approach each other.

It follows from statement 5 of Theorem 8.8 that the sizes of local manifolds are bounded from below on any regular set Λ^ℓ, i.e., there exists a number $r_\ell > 0$ that depends only on ℓ such that

$$r(x) \geqslant r_\ell \quad \text{for } x \in \Lambda^\ell. \tag{8.16}$$

Local stable manifolds depend uniformly continuously on $x \in \Lambda^\ell$ in the C^1 topology, i.e., if $x_n \in \Lambda^\ell$ is a sequence of points converging to x then $d_{C^1}(V(x_n), V(x)) \to 0$ as $n \to \infty$. Furthermore, by the Hölder continuity of stable distributions, local stable manifolds depend Hölder continuously on $x \in \Lambda^\ell$. More precisely, for every $\ell \geqslant 1$, $x \in \Lambda^\ell$, and points $z_1, z_2 \in V(x)$,

$$d\big(T_{z_1}V(x), T_{z_2}V(x)\big) \leqslant C\rho(z_1, z_2)^\alpha,$$

where $C > 0$ is a constant depending only on ℓ.

In the case when f is nonuniformly partially hyperbolic on an invariant subset Λ we have, for almost every $x \in \Lambda$, the local stable and unstable manifolds. Their sizes vary along the trajectory according to (8.15) and are bounded below by (8.16) on any regular set Λ^ℓ.

Finally, if f is nonuniformly completely hyperbolic on an invariant subset Λ then continuity of local stable and unstable manifolds on a regular set Λ^ℓ implies that there exists a number $\delta_\ell > 0$ such that for every $x \in \Lambda^\ell$ and $y \in \Lambda^\ell \cap B(x, \delta_\ell)$ the intersection $V^s(x) \cap V^u(y)$ is nonempty and consists of a single point which depends continuously (and in fact, Hölder continuously) on x and y.

8.6. *Graph transform property*

There is a version of the Stable Manifold Theorem known as Graph Transform Property (usually referred to as Inclination Lemma or λ-Lemma).

Consider a $C^{1+\alpha}$ diffeomorphism f which is nonuniformly partially hyperbolic in the broad sense along the trajectory of a point $x \in M$. Choose numbers r_0, b_0, and c_0 and for every $m \geqslant 0$, set

$$r_m = r_0 e^{-\varepsilon m}, \quad b_m = b_0 \mu^{-m} e^{\varepsilon m}, \quad c_m = c_0 e^{-\varepsilon m}.$$

Consider the class Ψ of $C^{1+\alpha}$ functions on $\{(m, v)\colon\ m \geqslant 0,\ v \in B_1(r_m)\}$ with values $\psi(m, v) \in E_2(f^{-m}(x))$ (where $B_1(r_m)$ is the ball in $E_1(f^{-m}(x))$ centered at 0 of radius r_m) satisfying the following conditions:

$$\big\|\psi(m, 0)\big\| \leqslant b_m, \qquad \max_{v \in B_1(r_m)} \big\|d\psi(m, v)\big\| \leqslant c_m.$$

THEOREM 8.13. *There are positive constants r_0, b_0, and c_0 such that for every $\psi \in \Psi$ one can find a function $\tilde{\psi} \in \Psi$ for which*

$$F_m^{-1}\big(\big\{\big(v, \psi(m, v)\big)\colon\ v \in B_1(r_m)\big\}\big) \supset \big\{\big(v, \tilde{\psi}(m+1, v)\big)\colon\ v \in B_1(r_{m+1})\big\}$$

for all $m \geqslant 0$.

8.7. *Regular neighborhoods*

Let $f : M \to M$ be a $C^{1+\alpha}$ diffeomorphism of a compact smooth n-dimensional Riemannian manifold M which is nonuniformly completely hyperbolic on an invariant set Λ. Viewing df as a linear cocycle over f we shall use the theory of linear extensions of cocycles (see Section 4) to construct a special coordinate system for every regular point $x \in \Lambda$. Applying the Reduction Theorem 5.10, given $\varepsilon > 0$ and a regular point $x \in M$, there exists a linear transformation $C_\varepsilon(x) : \mathbb{R}^n \to T_x M$ such that:

1. the matrix

$$A_\varepsilon(x) = C_\varepsilon(fx)^{-1} \circ d_x f \circ C_\varepsilon(x)$$

has the Lyapunov block form (5.12) (see Theorem 5.10);

2. $\{C_\varepsilon(f^m(x))\}_{m\in\mathbb{Z}}$ is a tempered sequence of linear transformations.

For every regular point $x \in M$ there is a neighborhood $N(x)$ of x such that f acts in $N(x)$ very much like the linear map $A_\varepsilon(x)$ in a neighborhood of the origin.

Denote by Λ the set of regular points for f and by $B(0,r)$ the standard Euclidean r-ball in $\mathbb{R}^n$ centered at the origin.

THEOREM 8.14 (Katok and Mendoza [139]). *For every $\varepsilon > 0$ the following properties hold*:

1. *there exists a tempered function $q : \Lambda \to (0,1]$ and a collection of embeddings $\Psi_x : B(0,q(x)) \to M$ for each $x \in \Lambda$ such that $\Psi_x(0) = x$ and $e^{-\varepsilon} < q(fx)/q(x) < e^{\varepsilon}$; these embeddings satisfy $\Psi_x = \exp_x \circ C_\varepsilon(x)$, where $C_\varepsilon(x)$ is the Lyapunov change of coordinates*;
2. *if $f_x \stackrel{\text{def}}{=} \Psi_{fx}^{-1} \circ f \circ \Psi_x : B(0,q(x)) \to \mathbb{R}^n$, then $d_0 f_x$ has the Lyapunov block form* (5.12);
3. *the C^1 distance $d_{C^1}(f_x, d_0 f_x) < \varepsilon$ in $B(0,q(x))$*;
4. *there exist a constant $K > 0$ and a measurable function $A : \Lambda \to \mathbb{R}$ such that for every $y, z \in B(0,q(x))$,*

$$K^{-1}\rho(\Psi_x y, \Psi_x z) \leqslant \|y - z\| \leqslant A(x)\rho(\Psi_x y, \Psi_x z)$$

with $e^{-\varepsilon} < A(fx)/A(x) < e^{\varepsilon}$.

We note that for each $x \in \Lambda$ there exists a constant $B(x) \geqslant 1$ such that for every $y, z \in B(0,q(x))$,

$$B(x)^{-1}\rho(\Psi_x y, \Psi_x z) \leqslant \rho'_x(\exp_x y, \exp_x z) \leqslant B(x)\rho(\Psi_x y, \Psi_x z),$$

where $\rho'_x(\cdot,\cdot)$ is the distance on $\exp_x B(0,q(x))$ with respect to the Lyapunov metric $\|\cdot\|'_x$. By Lusin's Theorem, given $\delta > 0$ there exists a set of measure at least $1 - \delta$ where $x \mapsto B(x)$ as well as $x \mapsto A(x)$ in Theorem 8.14 are bounded.

For each regular point $x \in \Lambda$ the set

$$R(x) \stackrel{\text{def}}{=} \Psi_x\big(B\big(0,q(x)\big)\big)$$

is called a *regular neighborhood* of x or a *Lyapunov chart* at x.

We stress that the existence of regular neighborhoods uses the fact that f is of class $C^{1+\alpha}$ in an essential way.

9. Global manifold theory

Let $f: M \to M$ be a $C^{1+\alpha}$ diffeomorphism of a smooth compact Riemannian manifold M which is nonuniformly partially hyperbolic in the broad sense on an invariant set $\Lambda \subset M$. Starting with local stable manifolds we will construct global stable manifolds for f.

In the case of uniformly partially hyperbolic systems (in the broad sense) global manifolds are integral manifolds of the stable distribution E_1. The latter is, in general, continuous but not smooth and hence, the classical Frobenius method fails. Instead, one can *glue* local manifolds to obtain leaves of the foliation.

In the case of nonuniformly hyperbolic systems (in the broad sense) the stable distribution E_1 may not even be continuous but measurable. The resulting "foliation" is measurable in a sense but has smooth leaves.

9.1. *Global stable and unstable manifolds*

Given a point $x \in \Lambda$, the *global stable manifold* is given by

$$W(x) = \bigcup_{n=0}^{\infty} f^{-n}\big(V\big(f^n(x)\big)\big). \tag{9.1}$$

This is a finite-dimensional immersed smooth submanifold of class $C^{r+\alpha}$ if f is of class $C^{r+\alpha}$. It has the following properties which are immediate consequences of the Stable Manifold Theorem 8.8.

THEOREM 9.1. *If $x, y \in \Lambda$, then*:

1. $W(x) \cap W(y) = \emptyset$ *if* $y \notin W(x)$;
2. $W(x) = W(y)$ *if* $y \in W(x)$;
3. $f(W(x)) = W(f(x))$;
4. $W(x)$ *is characterized as follows*:

$$W(x) = \left\{ y \in M\colon \overline{\lim_{n\to\infty}} \frac{1}{n} \log \rho\big(f^n(x), f^n(y)\big) < \log \lambda \right\}$$

(*see Section* 6.2 *for the definition of* λ).

Note that local stable manifolds are not uniquely defined. Indeed, one can choose a "smaller" submanifold containing x and lying inside $V(x)$, and view it as a "new" local manifold at x. However, such variations in the choice of local manifolds do not effect the global stable manifolds in the following sense. Fix $x \in \Lambda$. Consider its trajectory $f^n(x)$. For each $n \geqslant 0$, choose a ball $B_n \subset V(f^n(x))$ centered at $f^n(x)$ of radius $r_n > 0$.

THEOREM 9.2. *Assume that* $r_{n+1} > r_n e^{-\varepsilon n}$. *Then*

$$W(x) = \bigcup_{n=0}^{\infty} f^{-n}(B_n).$$

We give another useful characterization of global stable manifolds in the case when the diffeomorphism f possesses an invariant measure μ. Given $\ell > 1$, consider the regular set Λ^ℓ. For $x \in \Lambda^\ell$, denote by $n_i(x) > 0$ the successive moments of time for which $f^{n_i(x)}(x) \in \Lambda^\ell$. For almost every $x \in \Lambda^\ell$ the sequence $\{n_i(x)\}$ is unbounded.

THEOREM 9.3 (Pesin [198]). *For almost every* $x \in \Lambda^\ell$,

$$W(x) = \bigcup_{n=0}^{\infty} f^{-n_i(x)}\big(V\big(f^{n_i(x)}(x)\big)\big).$$

We recall that a partition W of M is called a *foliation of M with smooth leaves* if there exist $\delta > 0$, $q > 0$, and $k \in \mathbb{N}$ such that for each $x \in M$,

1. the element $W(x)$ of the partition W containing x is a smooth k-dimensional immersed submanifold; it is called the (*global*) *leaf* of the foliation at x; the connected component of the intersection $W(x) \cap B(x, \delta)$ that contains x is called the *local leaf* at x and is denoted by $V(x)$;
2. there exists a continuous map $\varphi_x : B(x, q) \to C^1(D, M)$ (where $D \subset \mathbb{R}^k$ is the unit ball) such that for every $y \in B(x, q)$ the manifold $V(y)$ is the image of the map $\varphi_x(y) : D \to M$.

The function $\Phi_x(y, z) = \varphi_x(y)(z)$ is called the *foliation coordinate chart*. This function is continuous and has continuous derivative $\frac{\partial \Phi_x}{\partial z}$.

In this section we deal only with foliations with smooth leaves and simply call them foliations. One can extend the notion of foliation to compact subsets of M (see [118] for more details).

In view of Theorem 9.1 global stable manifolds form a partition of Λ. When f is *uniformly* (partially) hyperbolic on Λ (which is compact), this partition is a foliation. When f is *nonuniformly* (partially) hyperbolic this partition is a "measurable" foliation in a certain sense (note that the partition by global manifolds may *not* be a measurable partition). We shall not discuss measurable foliations in this section (see Section 11.3 where we consider a very special class of such partitions).

Assume now that f is nonuniformly hyperbolic in the narrow sense on a set Λ. For every $x \in \Lambda$ we define the *global stable manifold* $W^s(x)$ as well as *global unstable manifold* by

$$W^u(x) = \bigcup_{n=0}^{\infty} f^n\big(V^u\big(f^{-n}(x)\big)\big).$$

This is a finite-dimensional immersed smooth submanifold (of class $C^{r+\alpha}$ if f is of class $C^{r+\alpha}$) invariant under f.

THEOREM 9.4 (Pesin [198]). *If x, $y \in \Lambda$, then*:

1. $W^u(x) \cap W^u(y) = \emptyset$ *if* $y \notin W^u(x)$;
2. $W^u(x) = W^u(y)$ *if* $y \in W^u(x)$;
3. $W^u(x)$ *is characterized as follows*:

$$W^u(x) = \left\{ y \in M\colon \varlimsup_{n\to\infty} \frac{1}{n} \log \rho\big(f^{-n}(x),\, f^{-n}(y)\big) < -\log \mu \right\}$$

(*see Section* 6.2 *for the definition of* μ).

We describe global manifolds for nonuniformly hyperbolic flows. Let φ_t be a smooth flow on M which is nonuniformly partially hyperbolic on an invariant set Λ. For every $x \in \Lambda$ we define the *global stable manifold* at x by

$$W^s(x) = \bigcup_{t>0} \varphi_{-t}\big(V^s\big(\varphi_t(x)\big)\big). \tag{9.2}$$

This is a finite-dimensional immersed smooth submanifold of class $C^{r+\alpha}$ if φ_t is of class $C^{r+\alpha}$. It satisfies statements 1–3 of Theorem 9.1. Furthermore, for every $y \in W^s(x)$ we have $\rho(\varphi_t(x), \varphi_t(y)) \to 0$ as $t \to +\infty$ with an exponential rate.

We also define the *global weakly stable manifold* at x by

$$W^{sc}(x) = \bigcup_{t\in\mathbb{R}} W^s\big(\varphi_t(x)\big).$$

It follows from (9.2) that

$$W^{sc}(x) = \bigcup_{t\in\mathbb{R}} \varphi_t\big(W^s(x)\big).$$

Furthermore, for every $x \in \Lambda$ define the *global unstable manifold* at x by

$$W^u(x) = \bigcup_{t>0} \varphi_t\big(V^u\big(\varphi_{-t}(x)\big)\big).$$

These are finite-dimensional immersed smooth submanifolds of class $C^{r+\alpha}$ if φ_t is of class $C^{r+\alpha}$. They satisfy statements 1–3 of Theorem 9.1.

We also define the *global weakly unstable manifold* at x by

$$W^{uc}(x) = \bigcup_{t\in\mathbb{R}} W^u\big(\varphi_t(x)\big).$$

It follows from (9.2) that

$$W^{sc}(x) = \bigcup_{t\in\mathbb{R}} \varphi_t\big(W^s(x)\big), \qquad W^{uc}(x) = \bigcup_{t\in\mathbb{R}} \varphi_t\big(W^u(x)\big).$$

Global (weakly) stable and unstable manifolds form partitions of the set Λ.

9.2. *Filtrations of stable manifolds*

Given a point $x \in \Lambda$, consider the Oseledets decomposition at x,

$$T_x M = \bigoplus_{j=1}^{p(x)} E_j(x).$$

Set $s(x) = \max\{j\colon\ \chi_j(x) < 0\}$ and for $i = 1, \dots, s(x)$,

$$F_i(x) = \bigoplus_{j=1}^{i} E_j(x).$$

The Stable Manifold Theorem 8.9 applies to the distribution $F_i(x)$ and provides a $C^{1+\alpha}$ local stable manifold $V_i(x)$. It is characterized as follows: there exists $r(x) > 0$ such that

$$V_i(x) = \left\{ y \in B\big(x, r(x)\big)\colon \varlimsup_{n\to+\infty} \frac{1}{n} \log d\big(f^n(x), f^n(y)\big) < \chi_i(x) \right\}.$$

Local stable manifolds form the *filtration of local stable manifolds* at x:

$$x \in V_1(x) \subset V_2(x) \subset \cdots \subset V_{s(x)}(x). \tag{9.3}$$

We define the *ith global stable manifold* at x by

$$W_i(x) = \bigcup_{n=0}^{\infty} f^{-n}\big(V_i\big(f^n(x)\big)\big).$$

It is a finite-dimensional immersed smooth submanifold of class $C^{r+\alpha}$ if f is of class $C^{r+\alpha}$. It does not depend on the particular choice of local stable manifolds in the sense of Theorem 9.2 and has the following properties which are immediate corollaries of the Stable Manifold Theorem 8.8.

THEOREM 9.5. *If x, $y \in \Lambda$, then*:

1. $W_i(x) \cap W_i(y) = \emptyset$ *if* $y \notin W_i(x)$;
2. $W_i(x) = W_i(y)$ *if* $y \in W_i(x)$;
3. $f(W_i(x)) = W_i(f(x))$;
4. $W_i(x)$ *is characterized by*

$$W_i(x) = \left\{ y \in M\colon \varlimsup_{n\to+\infty} \frac{1}{n} \log d\big(f^n(x), f^n(y)\big) < \chi_i(x) \right\}.$$

For each $x \in \Lambda$ we have the *filtration of global stable manifolds*

$$x \in W_1(x) \subset W_2(x) \subset \cdots \subset W_{s(x)}(x).$$

Consider the case when f is a nonuniformly partially hyperbolic diffeomorphism on an f-invariant set Λ. In a similar way, let $u(x) = \min\{j\colon \chi_j(x) > 0\}$ and for $i = u(x), \dots, p(x)$,

$$G_i(x) = \bigoplus_{j=i}^{p(x)} E_j(x).$$

The Unstable Manifold Theorem 8.11 applies to the distribution $G_i(x)$ and provides a $C^{1+\alpha}$ local manifold $V_i(x)$. It is characterized as follows: there exists $r(x) > 0$ such that

$$V_i(x) = \left\{ y \in B\big(x, r(x)\big)\colon \varlimsup_{n\to-\infty} \frac{1}{|n|} \log d\big(f^n(x), f^n(y)\big) < -\chi_i(x) \right\}.$$

We obtain the *filtration of local unstable manifolds* at x:

$$x \in V_{u(x)}(x) \subset V_{u(x)+1}(x) \subset \cdots \subset V_{p(x)}(x).$$

Finally, we have $V^s(x) = V_{s(x)}(x)$ and $V^u(x) = V_{u(x)}(x)$.[7]

We define the *ith global unstable manifold* at x by

$$W_i(x) = \bigcup_{n=0}^{\infty} f^n\big(V_i\big(f^{-n}(x)\big)\big).$$

It is a finite-dimensional immersed smooth submanifold of class $C^{r+\alpha}$ if f is of class $C^{r+\alpha}$. It does not depend on the particular choice of local unstable manifolds in the sense of Theorem 9.2 and is characterized as follows:

$$W_i(x) = \left\{ y \in M\colon \varlimsup_{n\to-\infty} \frac{1}{|n|} \log d\big(f^n(x), f^n(y)\big) < -\chi_i(x) \right\}.$$

For each $x \in \Lambda$ we have the *filtration of global unstable manifolds*

$$x \in W_{u(x)}(x) \subset W_{u(x)+1}(x) \subset \cdots \subset W_{p(x)}(x).$$

Finally, consider a diffeomorphism f which is a nonuniformly completely hyperbolic on an f-invariant set Λ.

Given $r \in (0, r(x))$ we denote by $B_i(x, r) \subset V_i(x)$ the ball centered at x of radius r with respect to the induced metric on $V_i(x)$.

By Theorem 8.14 there exists a special Lyapunov chart at x associated with the Oseledets decomposition at x:

[7] This notation is a bit awkward as the superscripts s and u stand for the words "stable" and "unstable", while $s(x)$ and $u(x)$ are numbers. It may get even more confusing since the functions $s(x)$ and $u(x)$, being measurable and invariant, are constant almost everywhere with respect to any invariant measure and the constant value is often denoted by s and u. We hope the reader will excuse us for such an abuse of notation.

1. there exists a local diffeomorphism $\varphi_x : U_x \to \mathbb{R}^n$ with the property that the spaces $\mathbb{E}_i = \varphi_x(\exp_x E_i(x))$ form an orthogonal decomposition of $\mathbb{R}^n$;
2. the subspaces $\mathbb{F}_k = \varphi_x(\exp_x F_k(x))$ and $\mathbb{G}_k = \varphi_x(\exp_x G_k(x))$ are independent of x;
3. if $i = 1, \ldots, p(x)$ and $v \in E_i(x)$ then

$$e^{\lambda_i(x)-\tau} \big\|\varphi_x(\exp_x v)\big\| \leqslant \big\|\varphi_{f(x)}(\exp_{f(x)} d_x f v)\big\| \leqslant e^{\lambda_i(x)+\tau} \big\|\varphi_x(\exp_x v)\big\|;$$

4. there is a constant K and a tempered function $A : \Lambda \to \mathbb{R}$ such that if $y, z \in U_x$ then

$$K\|\varphi_x y - \varphi_x z\| \leqslant d(y,z) \leqslant A(x)\|\varphi_x y - \varphi_x z\|;$$

5. there exists $\tilde{r}(x) \in (0, r(x))$ such that $B_i(x, \tilde{r}(x)) \subset V_i(x) \cap U_x$ for every $x \in \Lambda$ and $i = 1, \ldots, k(x)$ with $\lambda_i(x) \neq 0$. Moreover, for $1 \leqslant i \leqslant s(x)$, the manifolds $\varphi_x(V_i(x))$ are graphs of smooth functions $\psi_i : \mathbb{F}_i \to \mathbb{F}_{i+1}$ and for $u(x) \leqslant i \leqslant p(x)$, of smooth functions $\psi_i : \mathbb{G}_i \to \mathbb{G}_{i-1}$; the first derivatives of ψ_i are bounded by $1/3$.

It follows that for $1 \leqslant i \leqslant s(x)$,

$$f\big(V_i(x) \cap U_x\big) \subset V_i\big(f(x)\big) \cap U_{f(x)}$$

and for $u(x) \leqslant i \leqslant p(x)$,

$$f^{-1}\big(V_i(x) \cap U_x\big) \subset V_i\big(f^{-1}(x)\big) \cap U_{f^{-1}(x)}.$$

9.3. *Lipschitz property of intermediate stable manifolds*

Local manifold $V_k(y)$ in (9.3) depends Lipschitz continuously on $y \in V_{k+1}(x) \cap \Lambda^\ell$ for every $k < s(x)$. In order to state this result explicitly we shall first introduce the holonomy maps associated with families of local stable manifolds. Fix $\ell \geqslant 1$ and $x \in \Lambda^\ell$. Given transversals $T^1, T^2 \subset V_{k+1}(x)$ to the family of local stable manifolds

$$\mathcal{L}_k(x) = \big\{V_k(w) \colon\ w \in \Lambda^\ell \cap B(x,r)\big\},$$

we define the *holonomy map*

$$\pi_k : Q^\ell(x) \cap T^1 \to Q^\ell(x) \cap T^2$$

using the relation

$$\pi_k(y) = T^2 \cap V_k(w), \quad \text{where } y = T^1 \cap V_k(w) \text{ and } w \in Q^\ell(x) \cap B(x,r).$$

THEOREM 9.6 (Barreira, Pesin and Schmeling [36]). *Given $\ell \geqslant 1$, $x \in \Lambda^\ell$, and transversals $T^1, T^2 \subset V_{k+1}(x)$ to the family $\mathcal{L}_k(x)$, the holonomy map π_k is Lipschitz continuous with Lipschitz constant depending only on ℓ.*

The set Λ can be decomposed into sets Λ_β in which the numbers $k(x)$, $\dim E_i(x)$, and $\lambda_i(x)$ are constant for each i. For every ergodic measure μ invariant under f there exists a unique β for which the set Λ_β has full μ-measure. From now on we restrict our consideration to a subset $\Lambda_\beta \subset \Lambda$ and set $k(x) = k$, $s(x) = s$, $u(x) = u$, and $\lambda_i(x) = \lambda_i$ for each i and $x \in \Lambda_\beta$.

Given $\ell > 0$, consider the set $\Lambda'_{\beta\ell}$ defined by

$$\left\{x \in \Lambda_\beta\colon\ \rho(x) > \frac{1}{\ell},\ A(x) < \ell,\ \angle\left(E_i(x), \bigoplus_{j\neq i} E_j(x)\right) > \frac{1}{\ell},\ i = 1, \dots, k\right\}.$$

Let $\Lambda_{\beta\ell}$ be the closure of $\Lambda'_{\beta\ell}$. For each $x \in \Lambda'_{\beta\ell}$ there exists an invariant decomposition $T_xM = \bigoplus_{i=1}^{p(x)} E_i(x)$, filtration of local stable manifolds $V_i(x)$ and Lyapunov chart (U_x, φ_x) at x (see the previous section). In particular, the functions $\rho(x)$ and $A(x)$ can be extended to $\Lambda'_{\beta\ell}$ such that $\rho(x) > 1/\ell$, $A(x) < \ell$, and $\angle(E_i(x), \bigoplus_{j\neq i} E_j(x)) > 1/\ell$ for $i = 1, \dots, k$. The set $\Lambda_{\beta\ell}$ is compact and $\Lambda_{\beta\ell} \subset \Lambda_{\beta(\ell+1)}$, $\Lambda_\beta = \bigcup_{\ell>0} \Lambda_{\beta\ell} \pmod 0$.

Let us fix $c > 0$, $\ell > 0$, $x \in \Lambda_{\beta\ell}$, and $y' \in \Lambda_{\beta\ell} \cap B_{i+1}(x, c/\ell)$. For each $i < s$, consider two local smooth manifolds T_x and $T_{y'}$ in $V_{i+1}(x)$, containing x and y', respectively, and transverse to $V_i(z)$ for all $z \in \Lambda_{\beta\ell} \cap B_{i+1}(x, c/\ell)$. The holonomy map

$$\pi_i = \pi_i(T_x, T_{y'})\colon T_x \cap \Lambda_{\beta\ell} \cap B_{i+1}(x, c/\ell) \to T_{y'}$$

is given by

$$\pi_i(x') = V_i(x') \cap T_{y'}$$

with $x' \in T_x$. This map is well defined if c is sufficiently small (c may depend on ℓ but does not depend on x and y).

THEOREM 9.7. *Let f be a $C^{1+\alpha}$ diffeomorphism. For each $\ell > 0$, $i < s$, $x \in \Lambda_{\beta\ell}$, and $y' \in \Lambda_{\beta\ell} \cap B^i(x, c/\ell)$ the holonomy map $\pi_i(T_x, T_{y'})$ is Lipschitz continuous with the Lipschitz constant depending only on β and ℓ.*

10. Absolute continuity

Let f be a $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold M. We describe one of the most crucial properties of local stable and unstable manifolds which is known as *absolute continuity*.

Consider a foliation with smooth leaves W of M (see Section 9.2). Fix $x \in M$ and let ξ be the partition of the ball $B(x, q)$ by local manifolds $V(y)$, $y \in B(x, q)$.

The absolute continuity property addresses the following question:

> If $E \subset B(x, q)$ is a Borel set of positive volume, can the intersection $E \cap V(y)$ have zero Lebesgue measure (with respect to the Riemannian volume on $V(y)$) for almost every $y \in E$?

If the foliation W is indeed, smooth then due to the Fubini theorem, the intersection $E \cap V(y)$ has positive measure for almost all $y \in B(x, q)$. If the foliation is only continuous the absolute continuity property may not hold. A simple example which illustrates this paradoxical phenomenon was constructed by Katok (see below). A continuous but not absolutely continuous foliation does not satisfy the conditions of the Fubini theorem—a set of full Lebesgue measure may meet almost every leaf of the foliation at a single point—the phenomenon known as "Fubini's nightmare". Such pathological foliations appears generically in the stable ergodicity theory (see Section 13.8).

A celebrated result by Anosov claims that the stable and unstable invariant foliations for Anosov diffeomorphisms are absolutely continuous. We stress that generically these foliations are not smooth and therefore, the absolute continuity property is not at all trivial and requires a deep study of the structure of these foliations.

In [21], Anosov and Sinai suggested an approach to absolute continuity which is based on the study of the holonomy maps associated with the foliation. To explain this, consider a foliation W. Given x, choose two transversals T^1 and T^2 to the family of local manifolds $V(y)$, $y \in B(x, q)$. The holonomy map associates to a point $z \in T^1$ the point $w = V(z) \cap T^2$. This map is a homeomorphism. If it is absolutely continuous (see the definition below) for all points x and transversals T^1 and T^2 then the absolute continuity property follows.

For nonuniformly hyperbolic diffeomorphisms the study of absolute continuity is technically much more complicated due to the fact that the global stable and unstable manifolds may not form foliations (they may not even exist for some points in M) and the sizes of local manifolds may vary wildly from point to point. In order to overcome this difficulty one should define and study the holonomy maps associated with local stable (or unstable) manifolds on regular sets.

10.1. *Absolute continuity of stable manifolds*

Let Λ be the set of nonuniformly partially hyperbolic points in the broad sense for f so that conditions (6.2)–(6.5) hold. Let also $\{\Lambda^\ell\colon\ \ell \geqslant 1\}$ be the associated collection of regular sets. We assume that Λ is nonempty. Without loss of generality we may assume that each set Λ^ℓ is compact. We have $\Lambda^\ell \subset \Lambda^{\ell+1}$ for every ℓ. Furthermore, the stable subspaces $E_1(x)$ depend continuously on $x \in \Lambda^\ell$ and their sizes are bounded away from zero by a number r_ℓ (see (8.16)).

Fix $x \in \Lambda^\ell$, a number r, $0 < r \leqslant r_\ell$, and set

$$Q^\ell(x) = \bigcup_{w \in \Lambda^\ell \cap B(x,r)} V(w), \tag{10.1}$$

where $B(x, r)$ is the ball at x of radius r. Consider the family of local stable manifolds

$$\mathcal{L}(x) = \big\{V(w)\colon\ w \in \Lambda^\ell \cap B(x, r)\big\}$$

and a local open submanifold T which is *uniformly* transverse to it. For sufficiently small r we can chose T such that the set $\exp_x^{-1} T$ is the graph of a smooth map $\psi : B_2(q) \subset$

$E_2(x) \to E_1(x)$ (for some $q > 0$) with sufficiently small C^1 norm. In this case T intersects each local stable manifold $V(w) \in \mathcal{L}(x)$ and this intersection is transverse. We will consider local open submanifolds constructed only in this way and call them *transversals to the family* $\mathcal{L}(x)$. We also say that the map ψ *represents* T.

Let T^1 and T^2 be two transversals to the family $\mathcal{L}(x)$. We define the *holonomy map*

$$\pi : Q^\ell(x) \cap T^1 \to Q^\ell(x) \cap T^2$$

by setting

$$\pi(y) = T^2 \cap V(w), \quad \text{if } y = T^1 \cap V(w) \text{ and } w \in Q^\ell(x) \cap B(x, r).$$

The holonomy map π is a homeomorphism onto its image. It depends on x, ℓ, T^1, and T^2. Set

$$\Delta\big(T^1, T^2\big) = \big\|\psi^1 - \psi^2\big\|_{C^1}, \tag{10.2}$$

where the maps ψ^1 and ψ^2 represent T^1 and T^2, respectively.

Given a smooth submanifold W in M, we denote by ν_W the Riemannian volume on W induced by the restriction of the Riemannian metric to W. We denote by $\mathrm{Jac}(\pi)(y)$ the Jacobian of the holonomy map π at the point $y \in Q^\ell(x) \cap T^1$ specified by the measures ν_{T^1} and ν_{T^2}.

THEOREM 10.1 (Absolute Continuity). *Given $\ell \geqslant 1$, $x \in \Lambda^\ell$, and transversals T^1 and T^2 to the family $\mathcal{L}(x)$, the holonomy map π is absolutely continuous* (*with respect to the measures ν_{T^1} and ν_{T^2}*) *and the Jacobian* $\mathrm{Jac}(\pi)$ *is bounded from above and bounded away from zero.*

REMARK 10.2.

(1) One can obtain an explicit formula for the Jacobian. Namely, for every $y \in Q^\ell(x) \cap T^1$,

$$\mathrm{Jac}(\pi)(y) = \prod_{k=0}^{\infty} \frac{\mathrm{Jac}(d_{f^k(\pi(y))} f^{-1} | T_{f^k(\pi(y))} f^k(T^2))}{\mathrm{Jac}(d_{f^k(y)} f^{-1} | T_{f^k(y)} f^k(T^1))}$$

(in particular, the infinite product on the right-hand side converges).

(2) In the case when f is nonuniformly hyperbolic on Λ, one can show that the Jacobian $\mathrm{Jac}(\pi)$ satisfies

$$\big|\mathrm{Jac}(\pi) - 1\big| \leqslant C\,\Delta\big(T^1, T^2\big), \tag{10.3}$$

where $C > 0$ is a constant and $\Delta(T^1, T^2)$ is given by (10.2).

(3) If the holonomy map π is absolutely continuous then the foliation W has the absolute continuity property (see Theorem 11.1). However, the absolute continuity property of the foliation W does not necessarily imply that the holonomy map π is absolutely continuous.

The first basic proof of the Absolute Continuity theorem for nonuniformly partially hyperbolic diffeomorphisms (in the broad sense) was obtained by Pesin in [197]. A more conceptual and lucid proof (but for a less general case of nonuniform complete hyperbolicity) can be found in [35]. A somewhat different approach to absolute continuity was suggested by Pugh and Shub (see [209]).

Let us outline the main idea of the proof following the line of Pesin's argument. To estimate the Jacobian $\mathrm{Jac}(\pi)$ choose a small open set $A \subset T^1$ and let $B = \pi(A) \subset T^2$. We need to compare the measures $\nu_{T^1}(A \cap \Lambda^\ell)$ and $\nu_{T^2}(B \cap \Lambda^\ell)$. Consider the images $f^m(A)$ and $f^m(B)$, $m > 0$, which are smooth submanifolds of M. When m increases the sets $A \cap \Lambda^\ell$ and $B \cap \Lambda^\ell$ may get stretched and/or shrunk in the "unstable" direction E_2. This may occur with at most an exponential uniform rate γ with some $\lambda < \gamma < \min\{1, \mu\}$. On the other hand, the distance between the sets $f^m(A \cap \Lambda^\ell)$ and $f^m(B \cap \Lambda^\ell)$ gets exponentially small with a uniform rate λ' where $\lambda < \lambda' < \gamma$.

We then cover the set $f^m(A \cap \Lambda^\ell)$ and $f^m(B \cap \Lambda^\ell)$ by specially chosen open sets whose sizes are of order γ^m such that the multiplicity of these covers is finite and depends only on the dimension of T^1. More precisely, given a point $w \in \Lambda^\ell \cap B(x, r)$, let $y_i = V(w) \cap T^i$, $i = 1, 2$. Fix a number $q > 0$. In view of Theorem 8.13 there exists an open neighborhood $T^i_m(w, q) \subset T^i_m$ of the point $f^m(y_i)$ such that

$$T^i_m(w, q) = \exp_{w_m}\big\{\big(\psi^i_m(v), v\big)\colon\ v \in B_2(q_m)\big\},$$

where the map $\psi^i_m : B_2(q_m) \to E_1(f^m(w))$ represents $T^i_m(w, q)$ and $B_2(q_m) \subset E_1(f^m(w))$ is the ball centered at zero of radius $q_m = q\gamma^m$. If $q = q(m)$ is sufficiently small then for any $w \in \Lambda^\ell \cap B(x, r)$ and $k = 0, \dots, m$ we have that

$$f^{-1}\big(T^i_k(w, q)\big) \subset T^i_{k-1}(w, q), \quad i = 1, 2.$$

We now compare the measures $\nu_{T^1_m}|T^1_m(w, q)$ and $\nu_{T^2_m}|T^2_m(w, q)$ for sufficiently large m.

LEMMA 10.3. *There exists $C_1 > 0$ such that the following holds: for any $m > 0$ there exists $q_0 = q_0(m) > 0$ such that for any $0 < q \leqslant q_0$ we have*

$$C_1^{-1} \leqslant \frac{\nu_{T^1_m}(T^1_m(w, q))}{\nu_{T^2_m}(T^2_m(w, 2q))} \leqslant C_1.$$

LEMMA 10.4. *For any sufficiently large $m > 0$ there are points $w_j \in \Lambda^\ell \cap B(x, r)$, $j = 1, \dots, p = p(m)$ and a number $q = q(m) > 0$ such that the sets $W^1_m(w_j, q)$ form an open cover of the set $f^m(Q^\ell(x) \cap T^1)$ (see* (10.1)*) of finite multiplicity which depends only on the dimension of T^1.*

For sufficiently large m the sets $T^2_m(w, 2q)$ cover the set $f^m(B \cap \Lambda^\ell)$. It follows from Lemmas 10.3 and 10.4 that the ratio of the measures of the sets $f^m(A \cap \Lambda^\ell)$ and $f^m(B \cap \Lambda^\ell)$ is bounded.

To return back to the measure $\nu_{T^1}(A \cap \Lambda^\ell)$ we use the well-known relation

$$\nu_{T^1}\big(A \cap \Lambda^\ell\big) = \int_{f^m(A\cap\Lambda^\ell)} \operatorname{Jac}\big(df^{-m}|T_y f^m\big(T^1\big)\big)\, d\nu_{f^m(T^1)}(y).$$

Similar relation holds for the measures $\nu_{T^2}(B \cap \Lambda^\ell)$ and $\nu_{f^m(T^2)}(f^m(B \cap \Lambda^\ell))$. It remains to estimate the ratio of the Jacobians of the pullbacks $df^{-m}|T_y f^m(T^1)$ and $df^{-m}|T_{\pi(y)} f^m(T^2)$ for $y \in f^m(A \cap \Lambda^\ell)$. To do this choose a point $z \in f^{-m}(T_m^i(w,q))$ and set $z_m = f^m(z)$ and

$$D^i(z,m) = \operatorname{Jac}\big(d_{z_m} f^{-m}|T_{z_m} T_m^i(w,q)\big).$$

LEMMA 10.5. *There exist $C_2 > 0$ and $m_1(\ell) > 0$ such that for every $w \in \Lambda^\ell \cap B(x,r)$ and $m \geqslant m_1(\ell)$ one can find $q = q(m)$ such that*

$$C_2^{-1} \leqslant \left|\frac{D^2(y_m^2,m)}{D^1(y_m^1,m)}\right| \leqslant C_2,$$

and for $z \in f^{-m}(T_m^1(w,q))$,

$$C_2^{-1} \leqslant \left|\frac{D^1(z_m,m)}{D^1(y_m^1,m)}\right| \leqslant C_2.$$

This result allows one to compare the measures of the preimages under f^{-m} of $T_m^1(w,q)$ and $T_m^2(w,q)$. More precisely, the following statement holds.

LEMMA 10.6. *There exist $C_3 > 0$ and $m_2(\ell) > 0$ such that if $w \in \Lambda^\ell \cap B(x,r)$ and $m \geqslant m_2(\ell)$, then one can find $q = q(m)$ such that*

$$C_3^{-1} \leqslant \frac{\nu_{T^1}(f^{-m}(T_m^1(w,q)))}{\nu_{T^2}(f^{-m}(T_m^2(w,q)))} \leqslant C_3.$$

10.2. *Nonabsolutely continuous foliation*

We describe an example due to Katok of a nonabsolutely continuous foliation (another version of this example can be found in [185]; see also Section 6.2 of the Chapter "Partially hyperbolic dynamical systems" by B. Hasselblatt and Ya. Pesin in this volume [6]). Consider a hyperbolic automorphism A of the torus $\mathbb{T}^2$ and let $\{f_t\colon\ t \in S^1\}$ be a family of diffeomorphisms preserving the area m and satisfying the following conditions:

1. f_t is a small perturbation of A for every $t \in S^1$;
2. f_t depends smoothly on t;
3. the function $h(t) = h_m(f_t)$ is strictly monotone in a small neighborhood of $t = 0$ (here $h_m(f_t)$ is the metric entropy of the diffeomorphism f_t).

Note that for any family f_t the entropy is given by

$$h(t) = \int_{\mathbb{T}^2} \log \|d_x f_t | E^u_t(x)\| \, dm(x),$$

where $E^u_t(x)$ denotes the unstable subspace of f_t at the point x (see Section 14). Hence, one can modify A in a small neighborhood such that $h(t)$ is strictly monotone.

We introduce the diffeomorphism $F : \mathbb{T}^2 \times S^1 \to \mathbb{T}^2 \times S^1$ by $F(x,t) = (f_t(x), t)$. Since f_t is sufficiently close to A, they are conjugate via a Hölder homeomorphism g_t, i.e., $f_t = g_t \circ A \circ g_t^{-1}$. Given $x \in \mathbb{T}^2$, consider the set

$$H(x) = \big\{\big(g_t(x), t\big) \colon t \in S^1\big\}.$$

It is diffeomorphic to the circle S^1 and the collection of these sets forms an F-invariant foliation H of $\mathbb{T}^2 \times S^1 = \mathbb{T}^3$ with $F(H(x)) = H(A(x))$. Note that $H(x)$ depends Hölder continuously on x. However, the holonomy maps associated with the foliation H are not absolutely continuous. To see this consider the holonomy map

$$\pi_{t_1,t_2} : \mathbb{T}^2 \times \{t_1\} \to \mathbb{T}^2 \times \{t_2\}.$$

We have that

$$\pi_{0,t}(x,0) = \big(g_t(x), t\big) \quad \text{and} \quad F\big(\pi_{0,t}(x,0)\big) = \pi_{0,t}\big(A(x), 0\big).$$

If the map $\pi_{0,t}$ (with t being fixed) were absolutely continuous the measure $(\pi_{0,t})_* m$ would be absolutely continuous with respect to m. Note that each map f_t is ergodic (it is conjugate to the ergodic map A) and hence, m is the only absolutely continuous f_t-invariant probability measure. Thus, $(\pi_{0,t})_* m = m$. In particular, $h(t) = h(0)$. Since the entropy function $h(t)$ is strictly monotone in a small neighborhood of $t = 0$, the map g_t is not absolutely continuous for small t and so is the map $\pi_{0,t}$.

This example is a particular case of a more general situation of partially hyperbolic systems with nonintegrable central foliations, see [117].

11. Smooth invariant measures

In this section we deal with dynamical systems on compact manifolds which preserve smooth measures and are nonuniformly hyperbolic on some invariant subsets of positive measure (in particularly, on the whole manifold). We will present a sufficiently complete description of ergodic properties of the system. Note that most complete results (for example, on ergodicity, K-property and Bernoulli property) can be obtained when the system is completely hyperbolic. However, some results (for example, on Pinsker partition) hold true if only partial hyperbolicity in the broad sense is assumed. One of the main technical tools to in the study is the absolute continuity property of local stable and unstable invariant manifolds established in the previous section.

11.1. *Absolute continuity and smooth measures*

We begin with a more detailed description of absolute continuity of local stable and unstable manifolds for diffeomorphisms with respect to smooth measures.

Let f be a $C^{1+\alpha}$ diffeomorphism of a smooth compact Riemannian manifold M without boundary and let ν be a *smooth measure*, i.e., a probability measure which is equivalent to the Riemannian volume m. Let also Λ be the set of nonuniformly partially hyperbolic points in the broad sense for f. We assume that $\nu(\Lambda) = 1$.

Consider a regular set Λ^ℓ of positive measure. For every $x \in \Lambda^\ell$ we have the filtration of stable subspaces at x:

$$0 \in F_1(x) \subset F_2(x) \subset \cdots \subset F_{s(x)}(x)$$

and the corresponding filtration of local stable manifolds at x:

$$x \in V_1(x) \subset V_2(x) \subset \cdots \subset V_{s(x)}(x)$$

(see Section 9.2). Since $V_k(x)$ depends continuously on $x \in \Lambda^\ell$, without loss of generality we may assume that $s(x) = s$, $\dim V_k(x) = d_k$ for every $x \in \Lambda^\ell$ and $1 \leqslant s \leqslant p$. Fix $x \in \Lambda^\ell$ and consider the family of local stable manifolds

$$\mathcal{L}_k^\ell(x) = \big\{V_k(y)\colon\ y \in B(x,r) \cap \Lambda^\ell\big\}.$$

For $y \in B(x,r) \cap \Lambda^\ell$, denote by $m^k(y)$ the Riemannian volume on $V_k(w)$ induced by the Riemannian metric on M. Consider the set

$$P_k^\ell(x,r) = \bigcup_{y \in B(x,r) \cap \Lambda^\ell} V_k(w)$$

and its partition ξ^k by local manifolds $V_k(y)$. Denote by $\nu^k(y)$ the conditional measure on $V_k(y)$ generated by the partition ξ^k and the measure ν. The factor space $P^\ell(x,r)/\xi^k$ can be identified with the subset

$$A_k(x) = \big\{w \in T\colon \text{ there is } y \in \Lambda^\ell \cap B(x,r) \text{ such that } w = T \cap V_k(y)\big\},$$

where T is a transverse to the family $\mathcal{L}_k^\ell$.

THEOREM 11.1. *The following statements hold*:

1. *for ν-almost every $y \in \Lambda^\ell \cap B(x,r)$, the measures $\nu^k(y)$ and $m^k(y)$ are equivalent, i.e.,*

   $$d\nu^k(y)(z) = \kappa_k(y,z)\, dm^k(y)(z),$$

   where $\kappa_k(y,z)$, $z \in V_k(y)$ is the density function;

2. $$\kappa_k(y,z) = \prod_{i=0}^{\infty} \frac{\mathrm{Jac}(df|F_k(f^i(z)))}{\mathrm{Jac}(df|F_k(f^i(y)))};$$

3. *the function $\kappa_k(y,z)$ is Hölder continuous*;
4. *there is $C = C(\ell) > 0$ such that*

$$C^{-1} dm^k(y)(z) \leqslant d\nu^k(y)(z) \leqslant C dm^k(y)(z);$$

5. $m^k(x)(V^k(x) \setminus \Lambda) = 0$ *for ν-almost every $x \in \Lambda$.*

We now consider the case when Λ is a nonuniformly completely hyperbolic set for f. The above results apply to the families of local stable and unstable manifolds. For $y \in B(x,r) \cap \Lambda^\ell$ let $m^s(y)$ and $m^u(y)$ be the Riemannian volumes on $V^s(y)$ and $V^u(y)$, respectively. Let also ξ^s and ξ^u be the partitions of $B(x,r)$ by local stable and unstable manifolds, and $\nu^s(y)$ (respectively, $\nu^u(y)$) the conditional measures on $V^s(y)$ (respectively, $V^u(y)$) generated by ν and the partitions ξ^s (respectively, ξ^u). Finally, let $\hat{\nu}^s$ (respectively, $\hat{\nu}^u$) be the factor measures.

THEOREM 11.2. *The following statements hold*:

1. *for ν-almost every $y \in \Lambda^\ell \cap B(x,r)$ the measures $\nu^s(y)$ and $m^s(y)$ are equivalent; moreover, $d\nu^s(y)(z) = \kappa(y,z) dm^s(y)(z)$ where*

$$\kappa(y,z) = \prod_{i=0}^{\infty} \frac{\mathrm{Jac}(df|E^s(f^i(z)))}{\mathrm{Jac}(df|E^s(f^i(y)))};$$

2. *the factor measures $\hat{\nu}^s$ is equivalent to the measure $m^u(x)|A_k(x)$*;
3. $m^s(x)(V^s(x) \setminus \Lambda) = 0$ *for ν-almost every $x \in \Lambda$*;
4. *similar statements hold for the family of local unstable manifolds.*

11.2. *Ergodic components*

The following statement is one of the main results of smooth ergodic theory. It describes the decomposition of a hyperbolic smooth invariant measure into its ergodic components.

THEOREM 11.3 (Pesin [198]). *Let f be a $C^{1+\alpha}$ diffeomorphism of a smooth compact Riemannian manifold M and ν an f-invariant smooth (completely) hyperbolic measure on M. There exist invariant sets $\Lambda_0, \Lambda_1, \ldots$ such that*:

1. $\bigcup_{i \geqslant 0} \Lambda_i = \Lambda$, *and* $\Lambda_i \cap \Lambda_j = \emptyset$ *whenever* $i \neq j$;
2. $\nu(\Lambda_0) = 0$, *and* $\nu(\Lambda_i) > 0$ *for each* $i \geqslant 1$;
3. $f|\Lambda_i$ *is ergodic for each* $i \geqslant 1$.

The proof of this theorem exploits a simple yet deep argument due to Hopf [121]. Consider the regular sets Λ^ℓ of positive measure and let $x \in \Lambda^\ell$ be a Lebesgue point. For each $r > 0$ set

$$P^\ell(x,r) = \bigcup_{y \in \Lambda^\ell \cap B(x,r)} V^s(y).$$

Clearly, $P^\ell(x,r)$ has positive measure. It turns out that for a sufficiently small $r = r(\ell)$ the set

$$Q(x) = \bigcup_{n \in \mathbb{Z}} f^n\big(P^\ell(x,r)\big)$$

is an ergodic component, i.e., the map $f|Q(x)$ is ergodic. Indeed, given an f-invariant continuous function φ, consider the functions

$$\bar\varphi(x) = \lim_{n\to\infty} \frac{1}{2n+1} \sum_{k=-n}^{n} \varphi\big(f^k(x)\big),$$

$$\varphi^+(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \varphi\big(f^k(x)\big) \quad \text{and} \quad \varphi^-(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \varphi\big(f^{-k}(x)\big)$$

which are well defined for ν-almost every point x. We also have that $\bar\varphi(x) = \varphi^+(x) = \varphi^-(x)$ outside a subset $N \subset M$ of zero measure.

Since $\rho(f^n(z), f^n(w)) \to 0$ as $n \to \infty$ and φ is continuous, we obtain

$$\bar\varphi(z) = \varphi^+(z) = \varphi^+(w) = \bar\varphi(w).$$

Notice that the continuous functions are dense in $L^1(M,\nu)$ and hence, the functions of the form $\bar\varphi$ are dense in the set of f-invariant Borel functions.

It remains to show that the function $\bar\varphi(z)$ is constant almost everywhere. By Theorem 11.2 there exists a point $y \in (\Lambda^\ell \cap B(x,r)) \setminus N$ such that $m^u(y)(V^u(y) \cap N) = 0$ (recall that $\nu^s(y)$ and $\nu^u(y)$ are, respectively, the measures induced on $V^s(y)$ and $V^u(y)$ by the Riemannian volume). Let

$$P^s = \bigcup V^s(w),$$

where the union is taken over all points $w \in \Lambda^\ell \cap B(x, r_\ell)$ for which, respectively, $V^s(w) \cap V^u(y) \in N$. By absolute continuity property, we have $\nu(P^s) = 0$.

Let $z_1, z_2 \in P^\ell(x,r) \setminus (P^s \cup N)$. There are points $w_i \in \Lambda^\ell \cap B(x,r)$ such that $z_i \in V^s(w_i)$ for $i = 1, 2$. Note that the intersection $V^s(w_i) \cap V^u(y)$ is nonempty and consists of a single point y_i, $i = 1, 2$. We have that

$$\bar\varphi(z)(z_1) = \bar\varphi(z)(y_1) = \bar\varphi(z)(y_2) = \bar\varphi(z)(z_2)$$

and the ergodicity of $f|Q(x)$ follows.

Since almost every point $x \in \Lambda$ is a Lebesgue point of Λ^ℓ for some ℓ, the invariant sets $Q(x)$ cover the set $\Lambda \pmod 0$ and there is at most countable many such sets. We denote them by $Q_1, Q_2, \ldots$. We have $\nu(Q_i) > 0$ for each $i \geqslant 1$, and the set $\Lambda_0 = \Lambda \setminus \bigcup_{i \geqslant 1} Q_i$ has zero measure. Since $f|Q_i$ is ergodic $Q_i \cap Q_j = \emptyset \pmod 0$ whenever $i \neq j$. If we set $\Lambda_n = Q_n \setminus \bigcup_{i=1}^{n-1} Q_i$ then $\Lambda_i \cap \Lambda_j = \emptyset$ and $\nu(Q_i) = \nu(\Lambda_i) > 0$.

We describe an example of a diffeomorphism with nonzero Lyapunov exponents that has more than one ergodic component. Consider the diffeomorphism $G_{\mathbb{T}^2}$ of the torus $\mathbb{T}^2$ constructed in Section 2.2. This map is ergodic. The punched torus $\mathbb{T}^2 \setminus \{0\}$ is C^∞-diffeomorphic to the manifold $\mathbb{T}^2 \setminus \bar{U}$, where U is a small open disk around 0 and $\bar{U}$ denotes its closure. Therefore, we obtain a C^∞ diffeomorphism $F_{\mathbb{T}^2}$ of the manifold $\mathbb{T}^2 \setminus U$ with $F_{\mathbb{T}^2}|\partial U = \mathrm{Id}$. We have that $F_{\mathbb{T}^2}$ preserves a smooth measure, has nonzero Lyapunov exponents, and is ergodic.

Let $(\tilde{M}, \tilde{F}_{\mathbb{T}^2})$ be a copy of $(M, F_{\mathbb{T}^2})$. By gluing the manifolds M and $\tilde{M}$ along ∂U we obtain a smooth compact manifold $\mathcal{M}$ without boundary and a diffeomorphism $\mathcal{F}$ of $\mathcal{M}$ which preserves a smooth measure and has nonzero Lyapunov exponents almost everywhere. However, the map $\mathcal{F}$ is not ergodic and has two ergodic components of positive measure (M and $\tilde{M}$).

Similarly, one can obtain a diffeomorphism with nonzero Lyapunov exponents with n ergodic components of positive measure for an arbitrary n. However, it does not seem feasible to push this construction further and obtain a diffeomorphism with nonzero Lyapunov exponents with countably many ergodic components of positive measure. Such an example was constructed by Dolgopyat, Hu and Pesin in [87] using a different approach. It illustrates that Theorem 11.3 cannot be improved.

EXAMPLE 11.4. There exists a volume-preserving C^∞ diffeomorphism f of the three-dimensional torus $\mathbb{T}^3$ with nonzero Lyapunov exponents almost everywhere and countably many ergodic components which are open $(\mathrm{mod}\, 0)$.

The construction starts with a linear hyperbolic automorphism $A : \mathbb{T}^2 \to \mathbb{T}^2$ which has at least two fixed points p and p'. The desired map f is obtained as a perturbation of the map $F = A \times \mathrm{Id}$ of the three-dimensional torus $\mathbb{T}^3 = \mathbb{T}^2 \times S^1$. More precisely, consider a countable collection of intervals $\{I_n\}_{n=1}^\infty$ on the circle S^1, where

$$I_{2n} = \big[(n+2)^{-1}, (n+1)^{-1}\big], \qquad I_{2n-1} = \big[1 - (n+1)^{-1}, 1 - (n+2)^{-1}\big].$$

Clearly, $\bigcup_{n=1}^\infty I_n = (0,1)$ and $\operatorname{int} I_n$ are pairwise disjoint.

The main result in [87] states that for any $k \geqslant 2$ and $\delta > 0$, there exists a map g of the three-dimensional manifold $M = \mathbb{T}^2 \times I$ such that:

1. g is a C^∞ volume-preserving diffeomorphism of M;
2. $\|F - g\|_{C^k} \leqslant \delta$;
3. for all $0 \leqslant m < \infty$, $D^m g|\mathbb{T}^2 \times \{z\} = D^m F|\mathbb{T}^2 \times \{z\}$ for $z = 0$ and 1;
4. g is ergodic with respect to the Riemannian volume and has nonzero Lyapunov exponents almost everywhere.

Applying this result, for each n, one can construct a C^∞ volume-preserving ergodic diffeomorphism $f_n : \mathbb{T}^2 \times [0,1] \to \mathbb{T}^2 \times [0,1]$ satisfying

1. $\|F - f_n\|_{C^n} \leqslant e^{-n^2}$;
2. $D^m f_n | \mathbb{T}^2 \times \{z\} = D^m F | \mathbb{T}^2 \times \{z\}$ for $z = 0$ or 1 and all $0 \leqslant m < \infty$;
3. f_n has nonzero Lyapunov exponents μ-almost everywhere.

Let $L_n : I_n \to [0, 1]$ be the affine map and $\pi_n = (\mathrm{Id}, L_n) : \mathbb{T}^2 \times I_n \to \mathbb{T}^2 \times [0, 1]$. The desired map f is given by $f | \mathbb{T}^2 \times I_n = \pi_n^{-1} \circ f_n \circ \pi_n$ for all n and $f | \mathbb{T}^2 \times \{0\} = F | \mathbb{T}^2 \times \{0\}$. Note that for every $n > 0$ and $0 \leqslant m \leqslant n$,

$$\begin{aligned} \left\| D^m F | \mathbb{T}^2 \times I_n - \pi_n^{-1} \circ D^m f_n \circ \pi_n \right\|_{C^n} &\leqslant \left\| \pi_n^{-1} \circ (D^m F - D^m f_n) \circ \pi_n \right\|_{C^n} \\ &\leqslant e^{-n^2} \cdot (n+1)^n \to 0 \end{aligned}$$

as $n \to \infty$. It follows that f is C^∞ on M and it has the required properties.

In the following section we describe a result (see Theorem 11.9) which provides some additional conditions guaranteeing that the number of ergodic component in Theorem 11.3 is finite. Roughly speaking one should require that:

(1) the global stable (or unstable) foliation extends to a continuous foliation of the manifold and
(2) the Lyapunov exponents $\chi_i(x)$ are away from zero uniformly over x.

We now consider the case of a smooth flow φ_t on a compact manifold M preserving a smooth hyperbolic measure ν. We also assume that ν vanishes on the set of fixed points of φ_t.

Since the time-one map of the flow is nonuniformly partially hyperbolic we conclude that the families of local stable and unstable manifolds possess the absolute continuity property. This is a key fact which allows one to study the ergodic properties of nonuniformly hyperbolic flows.

THEOREM 11.5 (Pesin [198]). *There exist invariant sets $\Lambda_0, \Lambda_1, \ldots$ such that*

1. $\bigcup_{i \geqslant 0} \Lambda_i = \Lambda$, *and* $\Lambda_i \cap \Lambda_j = \emptyset$ *whenever* $i \neq j$;
2. $\nu(\Lambda_0) = 0$, *and* $\nu(\Lambda_i) > 0$ *for each* $i \geqslant 1$;
3. $\varphi_t | \Lambda_i$ *is ergodic for each* $i \geqslant 1$.

Using the flow described in Section 2.6 one can construct an example of a flow with nonzero Lyapunov exponents which has an arbitrary finite number of ergodic components.

11.3. *Local ergodicity*

Consider a $C^{1+\alpha}$ diffeomorphism of a compact manifold M preserving a smooth hyperbolic measure. In this section we discuss the *local ergodicity problem*—under what conditions ergodic components are open (up to a set of measure zero).

In this connection the following two problems are of interest:

PROBLEM 11.6. Is there a volume-preserving diffeomorphism which has nonzero Lyapunov exponents almost everywhere such that some (or even all) of its ergodic components with positive measure are not open (mod 0)?

PROBLEM 11.7. Is there a volume-preserving diffeomorphism which has nonzero Lyapunov exponents on an open (mod 0) and dense set U such that U has positive but not full measure? Is there a volume preserving diffeomorphism with the above property such that $f|U$ is ergodic?

The main obstacles for local ergodicity are the following:

1. the stable and unstable distributions are measurable but not necessarily continuous;
2. the global stable (or unstable) leaves may not form a foliation;
3. the unstable leaves may not *expand* under the action of f^n (note that they are defined as being exponentially *contracting* under f^{-n}, so that they are determined by the *negative* semitrajectory); the same is true for stable leaves with respect to the action of f^{-n}.

There are three different ways to obtain sufficient conditions for local ergodicity. Each of them is based on requirements which eliminate one or more of the above mentioned obstacles.

1. The first one is due to Pesin [198]. It requires a special structure of the global stable or unstable manifolds and is used to establish local ergodicity of geodesic flows (see Section 17).
2. The second one is due to Katok and Burns [136]. Its main advantage is that it relies on requirements on the local behavior of the system.
3. The third one is due to Liverani and Wojtkowski [169]. It deals with symplectic dynamical systems and is an adaptation of the Sinai method (that was developed for billiard dynamical systems; see [233]) to nonuniformly hyperbolic dynamical systems (both smooth and smooth with singularities; see Section 18).

1. We first describe the approach in [198]. Roughly speaking it requires that the stable (or unstable) leaves form a foliation of a measurable subset of full measure in M. First, we extend the notion of foliation of M with smooth leaves, introduced in Section 9.2, to foliation of a measurable subset.

Given a subset $X \subset M$, we call a partition ξ of X a *(δ, q)-foliation of X with smooth leaves* or simply a *(δ, q)-foliation of X* if there exist continuous functions $\delta : X \to (0, \infty)$ and $q : X \to (0, \infty)$ and an integer $k > 0$ such that for each $x \in X$:

1. there exists a smooth immersed k-dimensional submanifold $W(x)$ containing x for which $\xi(x) = W(x) \cap X$ where $\xi(x)$ is the element of the partition ξ containing x; the manifold $W(x)$ is called the (*global*) *leaf* of the (δ, q)-foliation at x; the connected component of the intersection $W(x) \cap B(x, \delta(x))$ that contains x is called the *local leaf* at x and is denoted by $V(x)$;
2. there exists a continuous map $\varphi_x : B(x, q(x)) \to C^1(D, M)$ ($D \subset \mathbb{R}^k$ is the open unit ball) such that for every $y \in X \cap B(x, q(x))$ the manifold $V(y)$ is the image of the map $\varphi_x(y) : D \to M$.

For every $x \in X$ and $y \in B(x, q(x))$ we set $U(y) = \varphi(y)(D)$ and we call it the *local leaf* of the (δ, q)-foliation at y. Note that $U(y) = V(y)$ for $y \in X$.

The following result establishes the local ergodicity property in the case when the stable (or unstable) foliation for f extends to a continuous foliation of M with smooth leaves.

THEOREM 11.8 (Pesin [198]). *Let f be a $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold M preserving a smooth measure ν and nonuniformly hyperbolic on an invariant set Λ. Assume that $\nu(\Lambda) > 0$ and that there exists a (δ, q)-foliation W of Λ such that $W(x) = W^s(x)$ for every $x \in \Lambda$ (where $W^s(x)$ is the global stable manifold at x; see Section 8). Then every ergodic component of f of positive measure is open* (mod 0) *in Λ (with respect to the induced topology).*

This theorem provides a way to establish the ergodicity of the map $f|\Lambda$. Namely, under the conditions of Theorem 11.8 every ergodic component of f of positive measure that lies in Λ is open (mod 0), hence, the set Λ is open (mod 0) and, if $f|\Lambda$ is topologically transitive, then $f|\Lambda$ is ergodic.

In general, a diffeomorphism f preserving a smooth hyperbolic measure may have countably many ergodic components which are open (mod 0) (see Example 11.4). We describe a criterion which guarantees that the number of open (mod 0) ergodic components is finite.

THEOREM 11.9. *Let f be a $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold M preserving a smooth measure ν and nonuniformly hyperbolic on an invariant set Λ. Assume that $\nu(\Lambda) > 0$ and that there exists a continuous foliation W of M with smooth leaves such that $W(x) = W^s(x)$ for every $x \in \Lambda$. Assume, in addition, that there exists a number $a > 0$ such that for almost every $x \in M$,*

$$\left|\chi_i(x)\right| > a. \tag{11.1}$$

Then $f|\Lambda$ has at most finitely many ergodic components of positive measure.

To see this observe that assumption (11.1) allows one to apply Proposition 13.16 and find a number $r > 0$ with the following property: for almost every $x \in \Lambda$ there is $n = n(x)$ such that the size of a local unstable manifold $V^u(f^n(x))$ is at least r. Let x be a density point of Λ. Consider the set

$$P(x, r) = \bigcup_{y \in V^u(f^n(x))} B^s(y, r),$$

where $B^s(y, r)$ is the ball in $W^s(y)$ centered at y of radius r. This set is contained in an ergodic component. It is also open and contains a ball of radius $\varepsilon > 0$ which does not depend on x. Thus, every ergodic component contains a ball of radius ε.

For a general diffeomorphism preserving a smooth hyperbolic measure, one should not expect the unstable (and stable) leaves to form a (δ, q)-foliation for some functions $\delta(x)$ and $q(x)$. In order to explain why this can happen consider a local unstable manifold $V^u(x)$ passing through a point $x \in \Lambda$. For a typical x and sufficiently large ℓ, the set $V^u(x) \cap \Lambda^\ell$ has positive Riemannian volume (as a subset of the smooth manifold $V^u(x)$) but is, in general, a Cantor-like set. When the local manifold is moved forward a given time n one should expect a sufficiently small neighborhood of the set $V^u(x) \cap \Lambda^\ell$ to expand. Other pieces of the local manifold (corresponding to bigger values of ℓ) will also expand

but with smaller rates. As a result the global leaf $W^u(x)$ (defined by (9.1)) may bend "uncontrollably"—the phenomenon that is yet to be observed but is thought to be "real" and even "typical" in some sense. As a result the map $x \mapsto \varphi_x$ in the definition of a (δ, q)-foliation may not be, indeed, continuous.

Furthermore, the global manifold $W^u(x)$ may be "bounded", i.e., it may not admit an embedding of an arbitrarily large ball in $\mathbb{R}^k$ (where $k = \dim W^u(x)$). This phenomenon is yet to be observed too.

The local continuity of the global unstable leaves often comes up in the following setting. Using some additional information on the system one can build an invariant foliation whose leaves contain local unstable leaves. This alone may not yet guarantee that global unstable leaves form a foliation. However, one often may find that the local unstable leaves expand in a "controllable" and somewhat uniform way when they are moved forward. We will see below that this guarantees the desired properties of unstable leaves. Such a situation occurs, for example, for geodesic flows on compact Riemannian manifolds of nonpositive curvature (see Section 17.1).

We now state a formal criterion for local ergodicity.

THEOREM 11.10 (Pesin [198]). *Let f be a $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold, preserving a smooth hyperbolic measure ν, and nonuniformly hyperbolic on an invariant set Λ of full measure. Let also W be a (δ, q)-foliation of Λ with the following properties*:

1. *$W(x) \supset V^s(x)$ for every $x \in \Lambda$;*
2. *there exists a number $\delta_0 > 0$ and a measurable function $n(x)$ on Λ such that for almost every $x \in \Lambda$ and any $n \geqslant n(x)$,*

$$f^{-n}\big(V^s(x)\big) \supset B_W\big(f^{-n}(x), \delta_0\big).$$

Then every ergodic component of f of positive measure is open (mod 0).

In the case of one-dimensional (δ, q)-foliations the second condition of Theorem 11.10 holds automatically and hence, can be omitted.

THEOREM 11.11 (Pesin [198]). *Let W be a one-dimensional (δ, q)-foliation of Λ, satisfying the following property*: *$W(x) \supset V^s(x)$ for every $x \in \Lambda$. Then every ergodic component of f of positive measure is open* (mod 0). *Moreover, $W^s(x) = W(x)$ for almost every $x \in \Lambda$.*

One can readily extend Theorems 11.10 and 11.11 to the case when the set Λ is open (mod 0) and has positive (not necessarily full) measure as well as to dynamical systems with continuous time.

THEOREM 11.12. *Let f be a $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold preserving a smooth measure ν and nonuniformly hyperbolic on an invariant set Λ. Assume that Λ is open* (mod 0) *and has positive measure. Let also W be a (δ, q)-foliation of Λ which satisfies properties* 1 *and* 2 *in Theorem* 11.10. *Then every ergodic component of $f|\Lambda$ of positive measure is open* (mod 0).

THEOREM 11.13. *Let φ_t be a smooth flow of a compact smooth Riemannian manifold preserving a smooth measure ν and nonuniformly hyperbolic on an invariant set Λ. Assume that Λ is open* (mod 0) *and has positive measure. Let also W be a (δ, q)-foliation of Λ with the following properties*:

1. *$W(x) \supset V^s(x)$ for every $x \in \Lambda$;*
2. *there exists a number $\delta_0 > 0$ and a measurable function $t(x)$ on Λ such that for almost every $x \in \Lambda$ and any $t \geqslant t(x)$,*

$$\varphi_{-t}\big(V^s(x)\big) \supset B_W\big(\varphi_{-t}(x), \delta_0\big).$$

Then every ergodic component of the flow $\varphi_t|\Lambda$ of positive measure is open (mod 0).

2. We now describe the approach in [136] to study the local ergodicity. A continuous function $Q : TM \to \mathbb{R}$ is called an *infinitesimal eventually strict Lyapunov function* for f over a set $U \subset M$ if:

1. for each $x \in U$ the function $Q_x = Q|T_xM$ is homogeneous of degree one, and takes on both positive and negative values;
2. there exist continuous distributions $D^s_x \subset C^s(x)$ and $D^u_x \subset C^u(x)$ such that $T_xM = D^s_x \oplus D^u_x$ for all $x \in U$, where

$$C^s(x) = Q^{-1}((-\infty, 0)) \cup \{0\} \quad \text{and} \quad C^u(x) = Q^{-1}((0, \infty)) \cup \{0\};$$

3. for every $x \in U$, $n \in \mathbb{N}$, $f^n(x) \in U$, and $v \in T_xM$,

$$Q_{f^n(x)}(d_x f^n v) \geqslant Q_x(v);$$

4. for ν-almost every $x \in U$ there exist $k = k(x)$, $\ell = \ell(x) \in \mathbb{N}$ such that $f^k(x) \in U$, $f^{-\ell}(x) \in U$, and for $v \in T_xM \setminus \{0\}$,

$$Q_{f^k(x)}\big(d_x f^k v\big) > Q_x(v) \quad \text{and} \quad Q_{f^{-\ell}(x)}\big(d_x f^{-\ell} v\big) < Q_x(v).$$

A function Q is called an *infinitesimal eventually uniform Lyapunov function* for f over a set $U \subset M$ if it satisfies conditions 1–3 and the following condition: there exists $\varepsilon > 0$ such that for ν-almost every $x \in M$ one can find $k = k(x)$, $\ell = \ell(x) \in \mathbb{N}$ for which $f^k(x) \in U$, $f^{-\ell}(x) \in U$, and if $v \in T_xM \setminus \{0\}$ then

$$Q_{f^k(x)}\big(d_x f^k v\big) > Q_x(v) + \varepsilon\|v\|$$

and

$$Q_{f^{-\ell}(x)}\big(d_x f^{-\ell} v\big) < Q_x(v) - \varepsilon\|v\|.$$

The following result gives a criterion for local ergodicity in terms of infinitesimal Lyapunov functions.

THEOREM 11.14 (Katok and Burns [136]). *The following properties hold*:

1. *If f possesses an infinitesimal eventually strict Lyapunov function Q over an open set $U \subset M$, then almost every ergodic component of f on the set $\bigcup_{n\in\mathbb{Z}} f^n(U)$ is open* (mod 0).
2. *If f possesses an infinitesimal eventually uniform Lyapunov function Q over an open set $U \subset M$, then every connected component of the set $\bigcup_{n\in\mathbb{Z}} f^n(U)$ belongs to one ergodic component of f. Moreover, if U is connected then $f|U$ is a Bernoulli transformation.*

This theorem was first proved by Burns and Gerber [65] for flows in dimension 3.

We sketch the proof of this theorem. When Q is an infinitesimal eventually strict Lyapunov function, given a compact set $K \subset U$, one can use the uniform continuity of $x \mapsto Q_x$ on the set K, and requirement 3 in the definition of Lyapunov function to show that the size of the stable and unstable manifolds on K is uniformly bounded away from zero. Furthermore, using requirement 4 one can show that for ν-almost every point $z \in M$ there exist $\theta = \theta(z) > 0$ and a neighborhood N of z such that for ν-almost every $x \in N$ and $y \in V^u(x) \cap N$ the tangent space $T_y V^u(x)$ is in the θ-interior of $C^u(y)$. A similar statement holds for stable manifolds.

Together with requirement 2 this implies that the stable and unstable manifolds have almost everywhere a "uniform" product structure; namely, for almost every $x \in U$ there exist a neighborhood $N(x)$ of x and $\delta > 0$ such that:

1. $V^s(y)$ and $V^u(y)$ have size at least δ for almost every $y \in N(x)$;
2. $V^s(y) \cap V^u(z) \neq \emptyset$ for $(\nu \times \nu)$-almost every $(y, z) \in N(x) \times N(x)$.

The proof of statement 1 follows now by applying the Hopf argument.

When Q is an infinitesimal eventually uniform Lyapunov function, the function $\theta(z)$ is uniformly bounded away from zero. This can be used to establish that for every x (and not only almost every x) there exists a neighborhood $N(x)$ of x and $\delta > 0$ with the above properties. A similar argument now yields the first claim in statement 2. The last claim is an immediate consequence of Theorem 11.19.

3. Finally we outline the approach in [169] to study the local ergodicity in the symplectic case. This approach is built upon a method which was developed by Sinai [233] in his pioneering work on billiard systems. It has been later improved by Sinai and Chernov [234] and by Krámli, Simányi and Szász [152] who considered semidispersing billiards.

Let M be a smooth compact symplectic manifold of dimension $2d$ with the symplectic form ω. Let also $f : M \to M$ be a symplectomorphism (i.e., a diffeomorphism of M which preserves the symplectic structure).

Fix $x \in M$. A subspace $V \subset T_x M$ is called *Lagrangian* if V is a maximal subspace on which ω vanishes (it has dimension d). Given two transverse Lagrangian subspaces V_1 and V_2 define the *sector* between them by

$$C = C(V_1, V_2) = \big\{v \in T_x M\colon\ \omega(v_1, v_2) \geqslant 0 \text{ for } v = v_1 + v_2,\ v_i \in V_i,\ i = 1, 2\big\}.$$

Define the quadratic form associated with an ordered pair of transverse Lagrangian subspaces V_1 and V_2 by

$$Q(v) = Q(V_1, V_2, v) = \omega(v_1, v_2) \quad \text{for } v = v_1 + v_2,\ v_i \in V_i,\ i = 1, 2.$$

Using this quadratic form we can write the cone $C(V_1, V_2)$ in the form

$$C(V_1, V_2) = \big\{v \in T_x M\colon\ Q(v) \geqslant 0\big\}.$$

We define the interior of the cone by

$$\operatorname{int} C(V_1, V_2) = \big\{v \in T_x M\colon\ Q(v) > 0\big\}.$$

We assume that two continuous subbundles of transverse Lagrangian subspaces are chosen in an open (not necessarily dense) subset $U \subset M$. We denote them by $\{V_1(x)\}_{x\in U}$ and $\{V_2(x)\}_{x\in U}$, respectively. For $x \in U$ let $C(x) = C(V_1(x), V_2(x))$ and $C'(x) = C(V_2(x), V_1(x))$.

If $x \in U$ and $f^n(x) \in U$ let us define

$$\sigma(d_x f^n) = \inf_{v\in \operatorname{int} C(x)} \sqrt{\frac{Q(V_1(x), V_2(x), d_x f^n v)}{Q(V_1(x), V_2(x), v)}}.$$

THEOREM 11.15. *Assume that the following conditions hold*:

1. *Monotonicity condition*: *if $x \in U$ and $f^k(x) \in U$ for $k \geqslant 0$ then*

$$d_x f^k C(x) \subset C\big(f^k(x)\big);$$

2. *Strict monotonicity condition*: *for almost every point $x \in U$ there exist $n > 0$ and $m < 0$ such that $f^n(x)$, $f^m(x) \in U$ and*

$$d_x f^n C(x) \subset \operatorname{int} C\big(f^n(x)\big) \cup \{0\}, \quad d_x f^m C'(x) \subset \operatorname{int} C'\big(f^m(x)\big) \cup \{0\}. \quad (11.2)$$

Then for any $n \geqslant 1$ and any point $x \in U$ such that $f^n(x) \in U$ and $\sigma(d_x f^n) > 1$ there is a neighborhood of x which is contained in one ergodic component of f.

It follows from this theorem that if U is connected and every point in it is strictly monotone (i.e., (11.2) holds) then $\bigcup_{k\in\mathbb{Z}} f^k(U)$ belongs to one ergodic component of f. This is a symplectic version of Theorem 11.14. We observe that Theorem 11.15 is a particular case of a more general result by Liverani and Wojtkowski for smooth dynamical systems with singularities (see Section 18).

11.4. *Pinsker partition, K-property and Bernoulli property*

In the ergodic theory there is a hierarchy of ergodic properties of which ergodicity (or the description of ergodic components) is the first and weakest one. Among the stronger properties are (weak and strong) mixing, K-property (including the description of the Pinsker

or π-partition) and the strongest among them—the Bernoulli property (or the description of Bernoulli components). The latter means essentially that the system is isomorphic in the measure-theoretical sense to the classical Bernoulli scheme.

We shall see that dynamical systems with nonzero Lyapunov exponents (nonuniformly hyperbolic systems) have all of these properties with respect to smooth invariant measures.

Let $f: M \to M$ be a $C^{1+\alpha}$ diffeomorphism of a smooth compact Riemannian manifold M preserving a smooth measure ν. Assume that f is nonuniformly partially hyperbolic in the broad sense on an invariant set Λ of positive measure. For every $x \in \Lambda$ we have that

$$\chi_1(x) < \cdots < \chi_{s(x)}(x) < 0 \leqslant \chi_{s(x)+1}(x) < \cdots < \chi_{p(x)}(x),$$

where $\chi_i(x)$, $i = 1, \ldots, p(x)$, are the distinct values of the Lyapunov exponent at x each with multiplicity $k_i(x)$. We also have the filtration of local (stable) manifolds at x,

$$x \in V_1(x) \subset V_2(x) \subset \cdots \subset V_{s(x)}(x), \tag{11.3}$$

as well as the filtration of global (stable) manifolds at x,

$$x \in W_1(x) \subset W_2(x) \subset \cdots \subset W_{s(x)}(x) \tag{11.4}$$

(see Section 9.2). Fix $j > 0$ and $m > 0$ and consider the sets

$$\Lambda_{j,m} = \{x \in \Lambda\colon \dim W_j(x) = m\}, \qquad \hat{\Lambda}_{j,m} = \bigcup_{x \in \Lambda_{j,m}} W_j(x). \tag{11.5}$$

For some j and m we have that $\nu(\Lambda_{j,m}) > 0$. Note that $W_j(x) \subset \Lambda_{j,m} \pmod 0$ for almost every $x \in \Lambda_{j,m}$. Hence, $\hat{\Lambda}_{j,m} = \Lambda_{j,m} \pmod 0$.

Consider the partition W_j of $\hat{\Lambda}_{j,m}$ by global manifolds $W_j(x)$. In general, this partition is not measurable. However, one can construct a special subpartition of W_j which we call *pseudo* π-partition for $f|\Lambda_{j,m}$—when f is nonuniformly *completely* hyperbolic on the set Λ this partition is the π-partition for $f|\Lambda_{j,m}$, i.e., the maximal partition with zero entropy.

We denote the measurable hull of a partition ξ by $\mathcal{H}(\xi)$ and we use the notation ε for the partition by points.

THEOREM 11.16. *There exists a measurable partition $\eta = \eta_{j,m}$ of $\hat{\Lambda}_{j,m}$ with the following properties*:

1. *for almost every $x \in \Lambda_{j,m}$ the set $C_\eta(x)$ is an open* $(\mathrm{mod}\, 0)$ *subset of $W_j(x)$*;
2. $f\eta \geqslant \eta$;
3. $\eta^+ = \bigvee_{i=0}^{\infty} f^i \eta = \varepsilon$;
4. $\bigwedge_{i>-\infty}^{0} f^i \eta = \mathcal{H}(W_j)$;
5. *if f is nonuniformly completely hyperbolic on Λ then $\mathcal{H}(W_j) = \pi(f|\Lambda_{j,m})$.*

Sinai [232, Theorem 5.2] proved this theorem for a class of dynamical systems with *transverse foliation*. Pesin [198] adapted this approach for nonuniformly hyperbolic dynamical systems.

We stress that the measurable hull $\mathcal{H}(W_j)$ does not depend on j (see statement 1 of Theorem 11.17); this is a manifestation of the Lipschitz property of intermediate stable manifolds (see Theorem 9.6). One can estimate the entropy of f with respect to η from below (see Theorem 12.11).

In order to construct the partition η, given $\ell > 1$, consider the regular set Λ^ℓ. For a sufficiently small $r = r(\ell) > 0$ and $x \in \Lambda^\ell$, set

$$P_j^\ell(x) = \bigcup_{y \in \Lambda^\ell \cap B(x,r)} V_j(y), \qquad Q(x) = \bigcup_{n=-\infty}^{\infty} f^n\big(P_j^\ell(x)\big). \tag{11.6}$$

It suffices to construct the partition η on the set $Q(x)$. Consider the partition $\tilde{\xi}$ of $P_j^\ell(x)$ by local manifolds $V_j(y)$, $y \in \Lambda^\ell \cap B(x,r)$. Adding the element $Q(x) \setminus P_j^\ell(x)$ we obtain a partition of $Q(x)$ which we denote by ξ. The partition

$$\eta = \xi^- = \bigvee_{i \leqslant 0} f^i \xi$$

has the desired properties.

In [160], Ledrappier and Young constructed a special countable partition of M of finite entropy which is a refinement of the partition η. We describe this partition in Section 16.3.

An important manifestation of Theorem 11.16 is the establishment of the K-property of a $C^{1+\alpha}$ diffeomorphism f which preserves a smooth measure ν and is nonuniformly completely hyperbolic on an invariant set Λ of positive measure. By Theorem 11.3 the set Λ can be decomposed into ergodic components Λ_i, $i = 1, 2, \ldots$, of positive measure. Fix i and denote by η_j the measurable partition of Λ_i associated with the foliation W_j, see Theorem 11.16.

THEOREM 11.17 (Pesin [198]). *The following properties hold*:

1. $\mathcal{H}(W_{j_1}|\hat{\Lambda}_i) = \mathcal{H}(W_{j_2}|\hat{\Lambda}_i) = \pi(f|\Lambda_i)$ *for any* $1 \leqslant j_1 < j_2 \leqslant s$ *or* $s+1 \leqslant j_1 < j_2 \leqslant p$;
2. *the* π*-partition of* $f|\Lambda_i$ *is finite and consists of* n_i *elements* Λ_i^k, $k = 1, \ldots, n_i$, *such that* $f(\Lambda_i^k) = \Lambda_i^{k+1}$, $k = 1, \ldots, n_i - 1$ *and* $f(\Lambda_i^n) = \Lambda_i^1$;
3. $f^{n_i}|\Lambda_i^k$ *is a* K*-automorphism.*

We now discuss the Bernoulli property. There are examples in general ergodic theory of systems which have K-property but fail to be Bernoulli. This cannot happen for smooth systems with nonzero exponents: Bernoulli property holds automatically as long as the system has the K-property (indeed, the mixing property is already sufficient).

THEOREM 11.18. *Let* f *be a* $C^{1+\alpha}$ *diffeomorphism of a smooth compact Riemannian manifold* M *preserving a smooth hyperbolic measure* ν. *Assume that* f *is weakly mixing with respect to* ν. *Then* f *is a Bernoulli automorphism.*

Ornstein and Weiss [191] established the Bernoulli property for geodesic flows on compact manifolds of negative curvature. Pesin [198] used a substantially more general version of their approach to prove Theorem 11.18. The proof exploits the characterization of a Bernoulli map in terms of *very weakly Bernoulli partitions* (see [191]). More precisely, there is a finite measurable partition α of the manifold M whose elements have piecewise smooth boundaries and arbitrarily small diameter. Indeed, one can construct a sequence of such partitions $\alpha_1 \leqslant \alpha_2 \leqslant \cdots$ such that $\alpha_n \to \varepsilon$. The proof goes to show that each partition α_n is very weakly Bernoulli and the result follows. An important technical tool of the proof is the refined estimate (10.3) of the Jacobian of the holonomy map.

Combining Theorems 11.17 and 11.18 we obtain the following Spectral Decomposition Theorem for systems with nonzero Lyapunov exponents preserving smooth measures.

THEOREM 11.19. *For each $i \geqslant 1$ the following properties hold*:

1. *Λ_i is a disjoint union of sets Λ_i^j, for $j = 1, \ldots, n_i$, which are cyclically permuted by f, i.e., $f(\Lambda_i^j) = \Lambda_i^{j+1}$ for $j = 1, \ldots, n_i - 1$, and $f(\Lambda_i^{n_i}) = \Lambda_i^1$;*
2. *$f^{n_i}|\Lambda_i^j$ is a Bernoulli automorphism for each j.*

We consider the case of dynamical systems with continuous time. Let φ_t be a C^2 flow on a smooth compact Riemannian manifold M preserving a smooth measure ν and nonuniformly hyperbolic on M. By Theorem 11.5, M can be decomposed into ergodic components Λ_i, $i = 1, 2, \ldots$, of positive measure. Applying Theorem 11.16 to the nonuniformly partially hyperbolic diffeomorphism $\varphi_1|\Lambda_i$ we obtain the following result.

THEOREM 11.20 (Pesin [198]). *There exists a partition $\eta = \eta_i$ of Λ_i for which*:

1. *for almost every $x \in \Lambda_i$ the element $C_\eta(x)$ is an open* (mod 0) *subset of $W^s(x)$;*
2. *$\varphi_1\eta \geqslant \eta$;*
3. *$\bigvee_{i=0}^{\infty} \varphi_i \eta = \varepsilon$;*
4. *$\bigwedge_{i>-\infty}^{0} \varphi_i \eta = \mathcal{H}(W^s) = \pi(\varphi_i|\Lambda_i)$.*

The following result establishes the K-property of the flow φ_t on the set Λ_i. For simplicity we will drop the index i. We remind the reader that a flow φ_t is a *K-flow* if and only if the diffeomorphism φ_t is a K-automorphism for every $t \in \mathbb{R}$.

THEOREM 11.21 (Pesin [198]). *Assume that the flow $\varphi_t|\Lambda$ has continuous spectrum. Then it is a Bernoulli flow and in particular, a K-flow.*

The following result is an immediate consequence of this theorem.

COROLLARY 11.22. *Let φ_t be a smooth flow on a compact smooth Riemannian manifold M preserving a smooth measure ν. Assume that ν is hyperbolic and that φ_t is mixing with respect to ν. Then φ_t is a Bernoulli flow.*

12. Metric entropy

A crucial idea in Smooth Ergodic Theory is that sufficient instability of trajectories yields rich ergodic properties of the system. The entropy formula is in a sense a "quantitative manifestation" of this idea and is yet another pearl of Smooth Ergodic Theory. It expresses the Kolmogorov–Sinai entropy $h_\nu(f)$ of a diffeomorphism, preserving a smooth hyperbolic measure, in terms of the values of the Lyapunov exponent.

12.1. *Margulis–Ruelle inequality*

Let f be a C^1 diffeomorphism of a compact smooth manifold M. The following very general result provides an upper bound for the entropy of f with respect to any Borel invariant probability measure ν.

THEOREM 12.1 (Margulis–Ruelle inequality). *The following estimate holds*:

$$h_\nu(f) \leqslant \int_M \Sigma_+ \, d\nu(x), \tag{12.1}$$

where

$$\Sigma_+ = \sum_{i:\ \chi_i(x)>0} k_i(x)\chi_i(x).$$

In the case of volume-preserving diffeomorphisms this estimate was obtained by Margulis (unpublished). The inequality in the general case was established by Ruelle in [212] (see also [35] and [175]).

We sketch the proof of the theorem. By decomposing ν into its ergodic components we may assume without loss of generality that ν is ergodic. Then $s(x) = s$ and $k_i(x) = k_i$, $\chi_i(x) = \chi_i$ are constant ν-almost everywhere for each $1 \leqslant i \leqslant s$. Fix $m > 0$. Since M is compact, there exists $t_m > 0$ such that for every $0 < t \leqslant t_m$, $y \in M$, and $x \in B(y,t)$ we have

$$\frac{1}{2} d_x f^m\big(\exp_x^{-1} B(y,t)\big) \subset \exp_{f^m x}^{-1} f^m\big(B(y,t)\big) \subset 2 d_x f^m\big(\exp_x^{-1} B(y,t)\big),$$

where for a set $A \subset T_z M$ and $z \in M$, we write $\alpha A = \{\alpha v\colon\ v \in A\}$.

There is a special partition of the manifold M which is described in the following statement.

LEMMA 12.2. *Given $\varepsilon > 0$, there is a partition ξ of M such that*:

1. $\operatorname{diam}\xi \leqslant t_m/10$ *and* $h_\nu(f^m, \xi) \geqslant h_\nu(f^m) - \varepsilon$;
2. *for every element $C \in \xi$ there exist balls $B(x,r)$ and $B(x,r')$, such that $r < 2r' \leqslant t_m/20$ and $B(x,r') \subset C \subset B(x,r)$*;

3. *there exists* $0 < r < t_m/20$ *such that if* $C \in \xi$ *then* $C \subset B(y, r)$ *for some* $y \in M$, *and if* $x \in C$ *then*

$$\frac{1}{2} d_x f^m \big(\exp_x^{-1} B(y, r)\big) \subset \exp_{f^m x}^{-1} f^m C \subset 2 d_x f^m \big(\exp_x^{-1} B(y, r)\big).$$

To construct such a partition, given $\alpha > 0$, consider a maximal α-separated set Γ, i.e., a finite set of points for which $d(x, y) > \alpha$ whenever $x, y \in \Gamma$. For $x \in \Gamma$ set

$$\mathcal{D}_\Gamma(x) = \big\{y \in M\colon\ d(y, x) \leqslant d(y, z) \text{ for all } z \in \Gamma \setminus \{x\}\big\}.$$

Obviously, $B(x, \alpha/2) \subset \mathcal{D}_\Gamma(x) \subset B(x, \alpha)$. Note that the sets $\mathcal{D}_\Gamma(x)$ corresponding to different points $x \in \Gamma$ intersect only along their boundaries, i.e., at a finite number of submanifolds of codimension greater than zero. Since ν is a Borel measure, if necessary, we can move the boundaries slightly so that they have zero measure. Thus, we obtain a partition ξ with $\operatorname{diam} \xi \leqslant \alpha$ which can be chosen to satisfy

$$h_\nu(f^m, \xi) > h_\nu(f^m) - \varepsilon \quad \text{and} \quad \operatorname{diam} \xi < t_m/10.$$

This guarantees the properties in the lemma.

Continuing with the proof of the theorem observe that

$$\begin{aligned} h_\nu(f^m, \xi) &= \lim_{k \to \infty} H_\nu\big(\xi \,|\, f^m \xi \vee \cdots \vee f^{km} \xi\big) \\ &\leqslant H_\nu(\xi \,|\, f^m \xi) = \sum_{D \in f^m \xi} \nu(D) H(\xi \,|\, D) \\ &\leqslant \sum_{D \in f^m \xi} \nu(D) \log \operatorname{card}\{C \in \xi\colon\ C \cap D \neq \emptyset\}, \end{aligned} \tag{12.2}$$

where $H(\xi|D)$ is the entropy of ξ with respect to the conditional measure on D induced by ν. The following is a uniform exponential estimate for the number of elements $C \in \xi$ which have nonempty intersection with a given element $D \in f^m \xi$.

LEMMA 12.3. *There exists a constant* $K_1 > 0$ *such that for* $D \in f^m \xi$,

$$\operatorname{card}\{C \in \xi\colon\ D \cap C = \emptyset\} \leqslant K_1 \sup\big\{\|d_x f\|^{mn}\colon\ x \in M\big\},$$

where $n = \dim M$.

This can be shown by estimating the volume of each element C and using property 2 of the partition ξ.

We also have an exponential bound for the number of those sets $D \in f^m \xi$ which contain regular points. Namely, given $\varepsilon > 0$, let $R_m = R_m(\varepsilon)$ be the set of forward regular points $x \in M$ which satisfy the following condition: for $k > m$ and $v \in T_x M$,

$$e^{k(\chi(x,v)-\varepsilon)} \|v\| \leqslant \big\|d_x f^k v\big\| \leqslant e^{k(\chi(x,v)+\varepsilon)} \|v\|.$$

LEMMA 12.4. *If $D \in f^m\xi$ has nonempty intersection with R_m then there exists a constant $K_2 > 0$ such that*

$$\operatorname{card}\{C \in \xi\colon\ D \cap C \neq \emptyset\} \leqslant K_2 e^{\varepsilon m} \prod_{i\colon\ \chi_i > 0} e^{m(\chi_i+\varepsilon)k_i}.$$

To establish the inequality note that

$$\operatorname{card}\{C \in \xi\colon\ D \cap C \neq \emptyset\} \leqslant \operatorname{vol}(B)(\operatorname{diam}\xi)^{-n},$$

where $\operatorname{vol}(B)$ denotes the volume of

$$B = \big\{y \in M\colon\ d\big(y, \exp_{f^m(x)}\big(d_x f^m\big(\exp_x^{-1} B'\big)\big)\big) < \operatorname{diam}\xi\big\}$$

and $B' = B(x, 2\operatorname{diam} C')$ for some $C' \in \xi$ such that $C' \cap R_m \neq \emptyset$ and $f^m(C') = D$, and some $x \in C' \cap f^{-m}(R_m)$. Up to a bounded factor, $\operatorname{vol}(B)$ is bounded by the product of the lengths of the axes of the ellipsoid $d_x f^m(\exp_x^{-1} B')$. Those of them that correspond to nonpositive exponents are at most subexponentially large. The remaining ones are of size at most $e^{m(\chi_i+\varepsilon)}$, up to a bounded factor, for all sufficiently large m. Thus,

$$\begin{aligned}\operatorname{vol}(B_1) &\leqslant K e^{m\varepsilon}(\operatorname{diam} B)^n \prod_{i\colon\ \chi_i>0} e^{m(\chi_i+\varepsilon)k_i}\\ &\leqslant K e^{m\varepsilon}(2\operatorname{diam}\xi)^n \prod_{i\colon\ \chi_i>0} e^{m(\chi_i+\varepsilon)k_i},\end{aligned}$$

for some constant $K > 0$. The lemma follows.

By Lemmas 12.3 and 12.4 and (12.2), we obtain

$$\begin{aligned}mh_\nu(f) - \varepsilon = h_\nu(f^m) - \varepsilon &\leqslant h_\nu(f^m, \xi)\\ &\leqslant \sum_{D\cap R_m \neq \emptyset} \nu(D)\left(\log K_2 + \varepsilon m + m \sum_{i\colon\ \chi_i>0} (\chi_i+\varepsilon)k_i\right)\\ &\quad + \sum_{D\cap R_m = \emptyset} \nu(D)\big(\log K_1 + nm \log\sup\big\{\|d_x f\|\colon\ x \in M\big\}\big)\\ &\leqslant \log K_2 + \varepsilon m + m \sum_{i\colon\ \chi_i>0} (\chi_i+\varepsilon)k_i\\ &\quad + \big(\log K_1 + nm \log\sup\big\{\|d_x f\|\colon\ x \in M\big\}\big)\nu(M \setminus R_m).\end{aligned}$$

By the Multiplicative Ergodic Theorem 5.5, we have $\bigcup_{m\geqslant 0} R_m(\varepsilon) = M \pmod 0$ for every sufficiently small ε. It follows that

$$h_\nu(f) \leqslant \varepsilon + \sum_{i\colon\ \chi_i>0} (\chi_i+\varepsilon)k_i.$$

Letting $\varepsilon \to 0$ we obtain the desired upper bound.

As an immediate consequence of Theorem 12.1 we obtain an upper bound for the topological entropy $h(f)$ of a diffeomorphism f. Namely,

$$h(f) = \sup_{\nu} h_{\nu}(f) \leqslant \sup_{\nu} \int_M \Sigma_+ \, d\nu, \tag{12.3}$$

where the suprema are taken over all f-invariant Borel probability measures on M.

In general inequalities (12.1) and (12.3) can be strict. In fact, as the following example shows, there are C^∞ diffeomorphisms for which $h(f) < \inf_\nu \int_M \Sigma_+ \, d\nu$, and hence, $h_\nu(f) < \int_M \Sigma_+ \, d\nu$ for *any* invariant measure ν.

EXAMPLE 12.5 (Figure-Eight; Bowen and Katok (see [135])). Let f be a diffeomorphism of the two-dimensional sphere S^2 with three repelling fixed points p_1, p_2, p_3 and one saddle fixed point q. Suppose that the stable and unstable manifolds of the point q form two loops γ_1, γ_2 that divide S^2 into three regions A_1, A_2, and A_3. For $i = 1, 2, 3$, we have $p_i \in A_i$ and any point in $A_i \setminus \{p_i\}$ tends, respectively, to γ_1, γ_2, and $\gamma_1 \cup \gamma_2$. Thus, any f-invariant *finite* measure ν is supported on the finite set $\{p_1, p_2, p_3, q\}$. Therefore, $h_\nu(f) = 0$ while $\int_M \Sigma_+ \, d\nu > c > 0$ for some c independent of ν. In addition, we have

$$h(f) = \sup_{\nu} h_{\nu}(f) < \inf_{\nu} \int_M \Sigma_+ \, d\nu,$$

where the supremum and infimum are taken over all f-invariant Borel probability measures on S^2.

EXAMPLE 12.6 (Two-dimensional horseshoes). Let Λ be a basic set (i.e., a locally maximal hyperbolic set), of a topologically transitive Axiom A surface diffeomorphism of class C^1. McCluskey and Manning [182] showed that for every $x \in \Lambda$ the Hausdorff dimension of the set $W^u(x) \cap \Lambda$ is the unique root s of Bowen's equation

$$P\big(-s \log \|df|E^u\|\big) = 0,$$

where P is the topological pressure on $f|\Lambda$. In particular, s is independent of x.

Assume that $s < 1$. Since $s \mapsto P(-s \log \|df|E^u\|)$ is decreasing, we obtain

$$P\big(-\log \|df|E^u\|\big) < 0.$$

By the Variational Principle for the topological pressure, for every f-invariant measure ν,

$$h_\nu(f) < \int_\Lambda \log \big\| d_x f|E^u(x) \big\| \, d\nu(x) = \int_\Lambda \Sigma_+ \, d\nu$$

(we use here Birkhoff's Ergodic Theorem and the fact that $\dim E^u = 1$).

Note that $h_\nu(f^{-1}) = h_\nu(f)$ and the Lyapunov exponents of f^{-1} are those of f taken with opposite sign. Therefore, it follows from Theorem 12.1 that

$$h_\nu(f) \leqslant -\int_M \sum_{i:\ \chi_i(x)<0} \chi_i(x) k_i(x)\, d\nu(x).$$

Set

$$a = \int_M \sum_{i:\ \chi_i(x)>0} \chi_i(x) k_i(x)\, d\nu(x)$$

and

$$b = -\int_M \sum_{i:\ \chi_i(x)<0} \chi_i(x) k_i(x)\, d\nu(x).$$

In Example 12.5 one can choose the eigenvalues of df at the critical points, and the measure ν to guarantee any of the relations: $a < b$ or $a = b$ or $a > b$. One can also show that if ν is the Riemannian volume on M, then $a = b$.

An important manifestation of Margulis–Ruelle's inequality is that positivity of topological entropy implies the existence of at least one nonzero Lyapunov exponent.

COROLLARY 12.7. *If the topological entropy of a C^1 diffeomorphism f of a compact manifold is positive, then there exists an ergodic f-invariant measure with at least one positive and one negative Lyapunov exponent.*

For surface diffeomorphisms, Corollary 12.7 means that any diffeomorphism with positive topological entropy possesses an ergodic invariant measure whose Lyapunov exponents are all nonzero.

Let us point out that the positivity of topological entropy can sometimes be determined using pure topological information. For example, theorems of Manning [177], Misiurewicz and Przytycki [186,187], and Yomdin [253,254] relate the topological entropy to the action of the diffeomorphism on the homology groups (see also [133]); see Section 15.5.

Other immediate consequences of Theorem 12.1 are as follows.

COROLLARY 12.8. *Let ν be a measure which is invariant under a C^1 diffeomorphism f of a compact manifold. If $h_\nu(f) > 0$ then ν has at least one positive and one negative Lyapunov exponent.*

For surface diffeomorphisms, Corollary 12.8 implies that if $h_\nu(f) > 0$ then the Lyapunov exponents of ν are all nonzero, i.e., ν is hyperbolic (see Sections 10.1 and 15).

COROLLARY 12.9. *We have*

$$\begin{aligned} h(f) &\leqslant \dim M \times \inf_{m \geqslant 1} \frac{1}{m} \log^+ \sup_{x \in M} \|d_x f^m\| \\ &= \dim M \times \lim_{m \to \infty} \frac{1}{m} \log^+ \sup_{x \in M} \|d_x f^m\|. \end{aligned}$$

12.2. *The entropy formula*

Let $f : M \to M$ be a $C^{1+\alpha}$ diffeomorphism, $\alpha > 0$ and ν an f-invariant measure which is absolutely continuous with respect to the Riemannian volume. The main result of this section is the *Pesin entropy formula* which expresses the entropy of f with respect to ν via its Lyapunov exponents. It was first proved by Pesin in [198]. The proof relies on properties of the unstable foliation and in particular, absolute continuity. Another proof of the entropy formula was obtained by Mañé in [172] (see also [175]). It does not involve directly the existence of stable and unstable foliations but instead uses some subtle properties of the action of the differential df with respect to the Lyapunov exponents in the presence of a smooth invariant measure.

THEOREM 12.10 (Pesin [198]). *The following formula holds true*:

$$h_\nu(f) = \int_M \Sigma_+ \, d\nu. \tag{12.4}$$

In view of the Margulis–Ruelle inequality we only need to establish the lower bound

$$h_\nu(f) \geqslant \int_M \sum_{i:\ \chi_i(x)>0} k_i(x)\chi_i(x)\, d\nu(x),$$

or equivalently (by replacing f by f^{-1} and using Theorem 5.5)

$$h_\nu(f) \geqslant -\int_M \sum_{i:\ \chi_i(x)<0} k_i(x)\chi_i(x)\, d\nu(x).$$

This inequality is a corollary of a more general result which we now state.

Let $f : M \to M$ be a $C^{1+\alpha}$ diffeomorphism of a smooth compact Riemannian manifold M preserving a smooth measure ν and nonuniformly partially hyperbolic in the broad sense on an invariant set Λ of positive measure. For every $x \in \Lambda$ we have that

$$\chi_1(x) < \cdots < \chi_{s(x)}(x) < 0 \leqslant \chi_{s(x)+1}(x) < \cdots < \chi_{p(x)}(x),$$

where $\chi_i(x)$, $i = 1, \ldots, p(x)$, are the distinct values of the Lyapunov exponent at x each with multiplicity $k_i(x)$. We also have the filtration of local (stable) manifolds (11.3) as well as the filtration of global (stable) manifolds (11.4) at x. Given $j > 0$ and $m > 0$, consider the sets (11.5). Note that $\nu(\Lambda_{j,m}) > 0$ for some j and m and $W_j(x) \subset \Lambda_{j,m} \pmod 0$ for almost every $x \in \Lambda_{j,m}$. Hence, $\hat{\Lambda}_{j,m} = \Lambda_{j,m} \pmod 0$.

Consider the partition $\eta = \eta_{j,m}$ of $\hat{\Lambda}_{j,m}$ constructed in Theorem 11.16.

THEOREM 12.11 (Pesin [199]). *The entropy of f with respect to η admits the following estimate from below*:

$$h_\nu(f,\eta) \geqslant -\int_{\Lambda_{j,m}} \sum_{i=1}^{j} k_i(z)\chi_i(z)\,d\nu(z).$$

We shall sketch the proof of the theorem. Given $\ell > 0$ consider the regular set Λ^ℓ. For sufficiently small $r = r(\ell)$ and $x \in \Lambda^\ell$ consider also the sets $P^{\ell,j}(x)$ and $Q(x)$ defined by (11.6). Let $\tilde{\nu}$ be the measure on $Q(x)$ given for any measurable subset $A \subset Q(x)$ by $\tilde{\nu}(A) = \nu(A)(\nu(Q(x))^{-1}$. It suffices to show that

$$h\big(f|Q(x),\eta\big) \geqslant -\int_{Q(x)} \sum_{i=1}^{j} k_i(z)\chi_i(z)\,d\tilde{\nu}(z). \tag{12.5}$$

Consider the function

$$g(z) = \prod_{i=1}^{j} \exp\big(\chi_i(z)\big)^{k_i(z)}.$$

Given $\varepsilon > 0$, let $Q_p = \{z \in Q(x)\colon\ p\varepsilon < g(z) \leqslant (p+1)\varepsilon\}$. It suffices to show the inequality (12.5) for the restriction $\bar{f} = f|Q_p$ and the measure $\bar{\nu}$ defined by $\bar{\nu}(A) = \nu(A)(\nu(Q_p))^{-1}$ for any measurable subset $A \subset Q_p$.

Set $J_n(z) = \mathrm{Jac}(df^n|T_zW^j(z))$. It follows from the Multiplicative Ergodic Theorem 5.5 that there exists a positive Borel function $T(z,\varepsilon)$, $z \in Q_p$ and $\varepsilon > 0$ such that for $n > 0$,

$$J_n(z) \leqslant T(z,\varepsilon)g(z)\exp(\varepsilon n).$$

Set for $t \geqslant 0$,

$$Q_p^t = \big\{z \in Q_p\colon\ T(z,\varepsilon) \leqslant t\big\}.$$

We have that for any $\alpha > 0$ and all sufficiently large t,

$$\bar{\nu}\big(Q_p^t\big) \geqslant 1-\alpha. \tag{12.6}$$

It follows from Theorem 10.1 that there exists $C_1 = C_1(t) > 0$ such that for any $z \in Q_p^t$ and $n > 0$,

$$\nu^j(z)\big(f^n\big(C_\eta(z)\big)\big) \leqslant C_1 J^n(z). \tag{12.7}$$

Denote by $B_\eta(z,r)$ the ball in $C_\eta(z)$ centered at z of radius r.

LEMMA 12.12. *For any $\beta > 0$ there exists $q = q(t)$ and a subset $A^t \subset Q_p^t$ such that*:

1. $\bar{\nu}(Q_p^t \setminus A^t) \leqslant \beta$;
2. *for any* $z \in A^t$ *the element* $C_\eta(z)$ *contains the ball* $B_\eta(z, q)$.

Denote by $\nu_\eta(z)$ the conditional measure on the element $C_\eta(z)$ of the partition η generated by the measure ν. For every $z \in A^t$,

$$C_4^{-1} \leqslant \frac{d\nu_\eta(z)}{dm^j(z)} \leqslant C_4, \tag{12.8}$$

where $C_4 = C_4(t) > 0$ is a constant independent of z. For any $n > 0$,

$$h_{\bar{\nu}}(f) = \frac{1}{n} h_{\bar{\nu}}(f^n) \geqslant \frac{1}{n} H(f^n\eta|\eta).$$

We use here the fact that

$$\eta = \eta^- = \bigvee_{i \leqslant 0} f^i \eta$$

(see Theorem 11.16). It follows from (12.6)–(12.8) that for every $x \in A^t$ and $n > 0$,

$$\begin{aligned} H\big(\bar{f}^n\eta|C_\eta(z)\big) &= -\int_{C_\eta(x)} \frac{\nu_\eta(C_\eta(x) \cap C_{f^n\eta}(z))}{\nu_\eta(C_\eta(y))}\, d\nu_\eta(y) \\ &\geqslant -\log\big[C_4^2 C_1 t\big((p+1)\varepsilon\big)^n e^{\varepsilon n} V\big(B_\eta\big(z, q(t)\big)\big)^{-1}\big] = I_n, \end{aligned} \tag{12.9}$$

where $V(B_\eta(z, q(t)))$ is the Riemannian volume of the ball $B_\eta(z, q(t))$. We have that $V(B_\eta(z, q(t))) \geqslant C_5 q^m(t)$ where $C_5 > 0$ is a constant. It follows that

$$\begin{aligned} I_n &\geqslant -\log\big(C_4^2 C_1 t\big)\big(C_5 q^m(t)\big)^{-1} - n\big(\log\big((p+1)\varepsilon\big) + \varepsilon\big) \\ &\geqslant C_6 - n\big(\log g(z) + \varepsilon\big). \end{aligned} \tag{12.10}$$

By (12.6) and statement 1 of Lemma 12.12, we obtain that $\bar{\nu}(Q_p \setminus A^t) \leqslant \alpha + \beta$. Therefore, integrating inequality (12.9) over the elements $C_\eta(x)$ and taking (12.10) into account we conclude that

$$\begin{aligned} \frac{1}{n} H\big(\bar{f}^n\eta|\eta\big) &\geqslant \frac{1}{n} I_n \bar{\nu}\big(A^t\big) \geqslant \frac{1}{n} \int_{Q_p} I_n\, d\bar{\nu}(1 - \alpha\beta) \\ &\geqslant \int_{Q_p} \sum_{i=1}^{j} k_i(z)\chi_i(z)\, d\bar{\nu}(z) - \gamma, \end{aligned}$$

where γ can be made arbitrary small if ε, α, and β are chosen sufficiently small and n sufficiently large. The desired result follows.

In the two-dimensional case the assumption that $f \in C^{1+\alpha}$ can be relaxed for a residual set of diffeomorphisms.

THEOREM 12.13 (Tahzibi [237]). *Let M be a compact smooth surface. There exists a residual subset $\mathcal{G}$ in the space* $\mathrm{Diff}^1(M,m)$ *of C^1 volume preserving diffeomorphisms of M such that every $f \in \mathcal{G}$ satisfies the entropy formula* (12.4). *Moreover, $\mathcal{G}$ contains all volume-preserving diffeomorphisms of class $C^{1+\alpha}$.*

The main idea of the proof is the following. In the two-dimensional case a volume-preserving diffeomorphism f has at most one positive Lyapunov exponents $\chi^+(x)$ almost everywhere. For $f \in \mathrm{Diff}^1(M,m)$ set $L(f) = \int_M \chi^+(x)\,d\mu$. One can show that the set of continuity points of the functions $L(f)$ and $h_m(f)$ is residual in the C^1 topology. Let f be a continuity point. One obtains the entropy formula for f by approximating f by a sequence f_n of $C^{1+\alpha}$ diffeomorphisms for which the entropy formula (12.4) holds.

Ledrappier and Strelcyn [142] extended the entropy formula to SRB-measures invariant under $C^{1+\alpha}$ diffeomorphisms (see Section 14) and Ledrappier and Young [160] obtained a general version of the entropy formula for arbitrary C^2 diffeomorphisms (see Section 16.1).

13. Genericity of systems with nonzero exponents

13.1. *Existence of diffeomorphisms with nonzero exponents*

Presence of an Anosov diffeomorphism f on a compact Riemannian manifold M imposes strong conditions on the topology of the manifold. For example, M should admit two foliations with smooth leaves (invariant under f). Anosov diffeomorphisms are only known to exist on multi-dimensional tori or more generally on factors of nilpotent Lie groups. On the contrary nonuniform hyperbolicity imposes no restrictions on the topology of M.

THEOREM 13.1 (Dolgopyat and Pesin [88]). *Given a compact smooth Riemannian manifold $M \neq S^1$ there exists a C^∞ volume-preserving Bernoulli diffeomorphism f of M with nonzero Lyapunov exponents almost everywhere.*

Let us comment on the proof of this theorem.

1. Katok [134] proved this theorem in the two-dimensional case. His argument goes as follows. Consider the diffeomorphism G_{S^2} of the sphere, constructed in Section 2.3. It has four *singularity* points $p_i = \zeta(x_i)$. Let ξ be a C^∞ map which blows up the point p_4. Consider the map $G_{D^2} = \xi \circ G_{S^2} \circ \xi^{-1}$ of the closed unit disk D^2. It is a C^∞ diffeomorphism which preserves the area, has the Bernoulli property and nonzero Lyapunov exponents almost everywhere.

The disk D^2 can be embedded into any surface. This is a corollary of a more general statement (see [134]).

PROPOSITION 13.2. *Given a p-dimensional compact C^∞ manifold M and a smooth measure μ on M, there exists a continuous map $h : D^p \to M$ (D^p is the unit ball in $\mathbb{R}^p$) such that*

1. *the restriction $h|\operatorname{int} D^p$ is a diffeomorphic embedding*;

2. $h(D^p) = M$;
3. $\mu(M \setminus h(D^p)) = 0$;
4. $h_* m = \mu$ *where* m *is the volume in* $\mathbb{R}^p$.

Note that G_{D^2} is identity on the boundary ∂D^2. Moreover, one can choose the function ψ in the construction of maps $G_{\mathbb{T}^2}$ and G_{S^2} such that the map G_{D^2} is "sufficiently flat" near the boundary of the disk.

More precisely, let $\rho = \{\rho_n\}$ be a sequence of nonnegative real-valued continuous functions on D^p which are strictly positive inside the disc. Let $C^\infty_\rho(D^p)$ be the set of all C^∞ functions on D^p satisfying the following condition: for any $n \geqslant 0$ there exists a sequence of numbers $\varepsilon_n > 0$ such that for all $(x_1, \dots, x_p) \in D^p$ for which $x_1^2 + \cdots + x_p^2 \geqslant (1 - \varepsilon_n)^2$ we have

$$\left| \frac{\partial^n h(x_1, \dots, x_p)}{\partial^{i_1} x_1 \dots \partial^{i_p} x_p} \right| < \rho_n(x_1, \dots, x_p),$$

where $i_1, \dots, i_p$, are nonnegative integers and $i_1 + \cdots + i_p = n$.

Any diffeomorphism G of the disc D^p can be written in the form $G(x_1, \dots, x_p) = (G_1(x_1, \dots, x_p), \dots, G_p(x_1, \dots, x_p))$. Set

$$\mathrm{Diff}^\infty_\rho(D^p) = \left\{ g \in \mathrm{Diff}^\infty(D^p) \colon\ G_i(x_1, \dots, x_p) - x_i \in C^\infty_\rho(D^p),\ i = 1, \dots, p \right\}.$$

PROPOSITION 13.3 (Katok [134]). *Given a compact* C^∞ *Riemannian manifold* M *there exists a sequence of functions* ρ *such that for any* $G \in \mathrm{Diff}^\infty_\rho(D^p)$ *the map* g *defined as* $g(x) = h(G(h^{-1}(x)))$ *for* $x \in h(\mathrm{int}\, D^p)$ *and* $g(x) = x$ *otherwise, is a* C^∞ *diffeomorphism of* M (*the map* h *is from Proposition* 13.2).

The function ψ can be chosen so that $G_{D^2} \in \mathrm{Diff}^\infty_\rho(D^2)$ and hence, the map f, defined as $f(x) = h(G_{D^2}(h^{-1}(x)))$ for $x \in h(\mathrm{int}\, D^p)$ and $f(x) = x$ otherwise, has all the desired properties: it preserves area, has nonzero Lyapunov exponents and is a Bernoulli map.

2. For any smooth compact Riemannian manifold M of dimension $p = \dim M \geqslant 5$, Brin [58] constructed a C^∞ volume-preserving Bernoulli diffeomorphism which has all but one nonzero Lyapunov exponents. His construction goes as follows.

Let A be a volume-preserving hyperbolic automorphism of the torus $\mathbb{T}^{p-3}$ and φ_t the suspension flow over A with the roof function

$$H(x) = H_0 + \varepsilon \tilde{H}(x),$$

where H_0 is a constant and the function $\tilde{H}(x)$ is such that $|\tilde{H}(x)| \leqslant 1$. The flow φ_t is an Anosov flow on the phase space Y^{p-2} which is diffeomorphic to the product $\mathbb{T}^{p-3} \times [0, 1]$, where the tori $\mathbb{T}^{p-3} \times \{0\}$ and $\mathbb{T}^{p-3} \times \{1\}$ are identified by the action of A. Consider the skew product map R of the manifold $N = D^2 \times Y^{n-2}$ given by

$$R(z) = R(x, y) = \left(G_{D^2}(x), \varphi_{\alpha(x)}(y)\right), \quad z = (x, y), \tag{13.1}$$

where $\alpha : D^2 \to \mathbb{R}$ is a nonnegative C^∞ function which is equal to zero in a small neighborhood U of the *singularity set* $\{q_1, q_2, q_3\} \cap \partial D^2$ and is strictly positive otherwise. The map R is of class C^∞ and preserves volume. One can choose the function $\tilde{H}(x)$ such that R is a Bernoulli diffeomorphism which has all but one nonzero Lyapunov exponents (the zero exponent corresponds to the direction of the flow φ_t).

Brin proved that there exists a smooth embedding of the manifold Y^{p-2} into $\mathbb{R}^p$. It follows that there is a smooth embedding $\chi_1 : D^2 \times Y^{p-2} \to D^p$ which is a diffeomorphism except for the boundary $\partial D^2 \times Y^{p-2}$. Using Proposition 13.2 one can find a smooth embedding $\chi : D^p \to M$ which is a diffeomorphism except for the boundary ∂D^p. Since the map R is identity on the boundary $\partial D^2 \times Y^{p-2}$ the map $h = (\chi_1 \circ \chi) \circ R \circ (\chi_1 \circ \chi)^{-1}$ has all the desired properties.

3. Dolgopyat and Pesin [88] constructed the required map P as a small perturbation of the map R (defined by (13.1)). The diffeomorphism P can be found in the form $P = \varphi \circ R$ where $\varphi(x, y) = (x, \varphi_x(y))$ and $\varphi_x : Y^{p-2} \to Y^{p-2}$, $x \in N$, is a family of volume preserving C^∞ diffeomorphisms satisfying $d_{C^1}(\varphi_x, \mathrm{Id}) \leqslant \varepsilon$. To construct such a family fix a sufficiently small number $\gamma > 0$, any point $y_0 \in Y^{p-2}$, and a point $x_0 \in D^2$ such that

$$G^j_{D^2}\big(B(x_0, \gamma)\big) \cap B(x_0, \gamma) = \emptyset, \quad -N < j < N,\ j \neq 0,$$
$$G^j_{D^2}\big(B(x_0, \gamma)\big) \cap \partial D^2 = \emptyset, \quad -N < j < N.$$

Set $\Delta = B(x_0, \gamma) \times B(q_0, \gamma)$ and choose a coordinate system $\{\xi_1, \xi_2, \eta_1, \dots, \eta_{p-2}\}$ in Δ such that $x = (\xi_1, \xi_2)$, $y = (\eta_1, \dots, \eta_{p-2})$, $dm = dx\,dy$ (recall that m is the volume) and

$$E^c_{\varphi_t}(y_0) = \frac{\partial}{\partial \eta_1}, \qquad E^s_{\varphi_t}(y_0) = \left(\frac{\partial}{\partial \eta_2}, \dots, \frac{\partial}{\partial \eta_k}\right),$$
$$E^u_{\varphi_t}(y_0) = \left(\frac{\partial}{\partial \eta_{k+1}}, \dots, \frac{\partial}{\partial \eta_{p-2}}\right)$$

for some k, $2 \leqslant k < p-2$. Let $\psi(t)$ be a C^∞ function with compact support. Set

$$\tau = \frac{1}{\gamma^2}\big(\|\xi_1\|^2 + \|\xi_2\|^2 + \|\eta_1\|^2 + \cdots + \|\eta_{p-2}\|^2\big)$$

and define

$$\varphi_x^{-1}(y) = \big(\xi_1, \xi_2, \eta_1 \cos\big(\varepsilon\psi(\tau)\big) + \eta_2 \sin\big(\varepsilon\psi(\tau)\big),$$
$$-\eta_1 \sin\big(\varepsilon\psi(\tau)\big) + \eta_2 \cos\big(\varepsilon\psi(\tau)\big), \eta_3, \dots, \eta_{p-2}\big).$$

The family φ_x determines the map φ so that the map $P = \varphi \circ R$ is a volume-preserving Bernoulli diffeomorphism with nonzero Lyapunov exponents.

4. We discuss the case $\dim M = 3$. Consider the manifold $N = D^2 \times S^1$ and the skew product map R,

$$R(z) = R(x, y) = \big(G_{D^2}(x), R_{\alpha(x)}(y)\big), \quad z = (x, y), \tag{13.2}$$

where $R_{\alpha(x)}$ is the rotation by the angle $\alpha(x)$ and $\alpha : D^2 \to \mathbb{R}$ is a nonnegative C^∞ function which is equal to zero in a small neighborhood of the singularity set $\{q_1, q_2, q_3\} \cap \partial D^2$ and is strictly positive otherwise.

We define a perturbation P of R in the form $P = \varphi \circ R$. Consider a coordinate system $\xi = \{\xi_1, \xi_2, \xi_3\}$ in a small neighborhood of a point $z_0 \in N$ such that $dm = d\xi$ and

$$E^c_R(z_0) = \frac{\partial}{\partial \xi_1}, \quad E^s_R(z_0) = \frac{\partial}{\partial \xi_2}, \quad E^u_R(z_0) = \frac{\partial}{\partial \xi_3}.$$

Let $\psi(t)$ be a C^∞ function with compact support. Set $\tau = \|\xi\|^2/\gamma^2$ and define

$$\begin{aligned} \varphi^{-1}(\xi) = \big(&\xi_1 \cos\big(\varepsilon\psi(\tau)\big) + \xi_2 \sin\big(\varepsilon\psi(\tau)\big), \\ &-\xi_1 \sin\big(\varepsilon\psi(\tau)\big) + \xi_2 \cos\big(\varepsilon\psi(\tau)\big), \xi_3\big). \end{aligned} \tag{13.3}$$

One can choose the function $\alpha(x)$ and the point z_0 such that the map P has all the desired properties.

5. We now proceed with the case $\dim M = 4$. Consider the manifold $N = D^2 \times \mathbb{T}^2$ and the skew product map R defined by (13.2) where $R_{\alpha(x)}$ is the translation by the vector $\alpha(x)$ and the function $\alpha(x)$ is chosen as above. Consider a perturbation P of R in the form $P = \varphi \circ R$ and choose the map φ as above to ensure that

$$\int_N \big[\chi^c_1(z, P) + \chi^c_2(z, P)\big]\, dz < 0,$$

where $\chi^c_1(z, P) \geqslant \chi^c_2(z, P)$ are the Lyapunov exponents of P along the central subspace $E^c_P(z)$. One can further perturb the map P in the C^1 topology to a map $\bar{P}$ to guarantee that

$$\int_N \big[\chi^c_1(z, \bar{P}) + \chi^c_2(z, \bar{P})\big]\, dz < 0, \qquad \int_N \big[\chi^c_1(z, \bar{P}) - \chi^c_2(z, \bar{P})\big]\, dz \leqslant \varepsilon,$$

where $\chi^c_1(z, \bar{P}) \geqslant \chi^c_2(z, \bar{P})$ are the Lyapunov exponents of $\bar{P}$ along the central subspace $E^c_{\bar{P}}(z)$ and $\varepsilon > 0$ is sufficiently small. This can be done using the approach described in the proof of Theorem 13.8 (this is one of the reasons why $\bar{P}$ is close to P in the C^1 topology only). The map $\bar{P}$ has all the desired properties.

13.2. *Existence of flows with nonzero exponents*

In [126], Hu, Pesin and Talitskaya established a continuous time version of Theorem 13.1.

THEOREM 13.4. *Given a compact smooth Riemannian manifold M of* $\dim M \geqslant 3$, *there exists a C^∞ volume-preserving Bernoulli flow φ_t such that at m-almost every point $x \in M$ it has nonzero Lyapunov exponent except for the exponent in the direction of the flow.*

We sketch the proof of this theorem. Assume first that $\dim M \geqslant 5$. Consider the map

$$R = G_{D^2} \times A : D^2 \times \mathbb{T}^{p-3} \to D^2 \times \mathbb{T}^{p-3},$$

where $p = \dim M$, G_{D^2} is the above constructed diffeomorphism of the two-dimensional disk with nonzero Lyapunov exponents and A is a linear automorphism of the torus $\mathbb{T}^{p-3}$.

Consider further the suspension flow g_t over R with the roof function $H = 1$ and the suspension manifold $K = D^2 \times \mathbb{T}^{p-3} \times [0,1]/\sim$, where $\sim$ is the identification $(x, y, 1) = (G_{D^2}(x), A(y), 0)$. Denote by Z the vector field of the suspension flow.

Finally, consider the suspension flow h_t over A with the roof function $H = 1$ and the suspension manifold $L = \mathbb{T}^{p-3} \times [0,1]/\sim$, where $\sim$ is the identification $(y, 1) = (Ay, 0)$. Let $N = D^2 \times \mathbb{T}^{p-3} \times [0,1]/\sim$, where $\sim$ is the identification $(x, y, 1) = (x, Ay, 0)$ for any $x \in D^2$, $y \in \mathbb{T}^{p-3}$.

The proof goes by showing that there exists a volume-preserving C^∞ diffeomorphism $F : K \to N$ so that the vector field $Y = dFZ$ is divergence free and

$$Y(x, y, t) = \big(Y_1(x, y, t), 0, 1\big).$$

Choose a C^∞ function $a : D^2 \to [0,1]$ which vanishes on the boundary ∂D^2 with all its partial derivatives of any order, strictly positive otherwise and $a(x) = 1$ outside small neighborhood of the boundary. Define the vector field V on N by

$$V(x, y, t) = \big(Y_1(x, y, t), 0, a(x)\big).$$

The flow on K corresponding to the vector field $dF^{-1}VF$ is volume-preserving, has nonzero Lyapunov exponents (except for the exponent in the flow direction) and is Bernoulli. The manifold K can be embedded into M and this embedding carries over the flow into a flow on M with all the desired properties.

13.3. *Genericity conjecture*

Little is known about genericity of systems with nonzero Lyapunov exponents. On any manifold M of dimension $\dim M \geqslant 2$ and for sufficiently large r there are open sets of volume-preserving C^r diffeomorphisms of M which possess positive measure sets with all of the exponents to be zero: these sets consist of codimension one invariant tori on which the system is conjugate to a Diophantine translation (see [71,116,251,252]).

In this regard the following conjecture is of a great interest in the field.

CONJECTURE 13.5. *Let f be a $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold M preserving a smooth measure μ. Assume that f has nonzero Lyapunov exponents almost everywhere. Then there exists a neighborhood $\mathcal{U}$ of f in* $\mathrm{Diff}^{1+\alpha}(M, \mu)$ *and*

a G_δ-set $\mathcal{A} \subset \mathcal{U}$ such that any diffeomorphism $g \in \mathcal{A}$ has nonzero Lyapunov exponents on a set A_g of positive measure.

13.4. C^1-genericity for maps

We stress that the assumption on the regularity of f (i.e., f is of class $C^{1+\alpha}$) is crucial: in the C^1 topology one should expect quite a different behavior. Let us describe some relevant results. We first consider the case of a compact surface M.

THEOREM 13.6 (Bochi [46]). *There exists a residual subset $\mathcal{U}$ in the space of area preserving C^1 diffeomorphisms such that any $f \in \mathcal{U}$ is either Anosov or has zero Lyapunov exponents almost everywhere.*

This theorem was first announced by Mañé around 1983. Although the proof was never published a sketch of it appeared in [174] (see also [176] for a symplectic version of this result). A version of Theorem 13.6 for manifolds of higher dimension was obtained by Bochi and Viana in [49].

Let f be a volume-preserving ergodic C^1 diffeomorphism of a compact smooth Riemannian manifold M and x a Lyapunov regular point for f. Consider the Oseledets decomposition (6.7) along the orbit of x and two subspaces $E_i(x)$ and $E_j(x)$ corresponding to two distinct values of the Lyapunov exponent, $\chi_i > \chi_j$ (since f is ergodic these values do not depend on x). Given a point y in the orbit of x there is $m = m(y, i, j) \geqslant 1$ such that

$$\|df^m|E_i(y)\| \cdot \|df^m|E_j(y)\| \leqslant \frac{1}{2}.$$

Let $m(y) = \max_{i,j} m(y, i, j)$. We say that the Oseledets decomposition has the *dominated property* if $m(y)$ does not depend on y. In other words, the fact that df^n eventually expends $E_i(y)$ more than $E_j(y)$ can be observed in finite time *uniformly over the orbit of* f. The dominated property implies that the angles between the Oseledets subspaces are bounded away from zero along the orbit.

THEOREM 13.7 (Bochi and Viana [47–49]). *Let M be a compact smooth Riemannian manifold. There exists a residual subset $\mathcal{U}$ in the space of volume-preserving C^1 diffeomorphisms such that for any $f \in \mathcal{U}$ and almost every $x \in M$ the Oseledets decomposition is either dominated along the orbit of x or is trivial, i.e., all Lyapunov exponents at x are zero.*

This theorem is a corollary of the following result that provides necessary conditions for continuity of Lyapunov exponents $\chi_i(f, x)$ over f. For $j = 1, \ldots, p-1$ define

$$\mathrm{LE}_j(f) = \int_M \big[\chi_1(f, x) + \cdots + \chi_j(f, x)\big]\, dm(x).$$

It is well known that the function

$$f \in \mathrm{Diff}^1(M, m) \to \mathrm{LE}_j(f)$$

is upper-semicontinuous.

THEOREM 13.8 (Bochi and Viana [49]). *Let $f_0 \in \mathrm{Diff}^1(M, m)$ be such that the map*

$$f \in \mathrm{Diff}^1(M, m) \to \big(\mathrm{LE}_1(f), \ldots, \mathrm{LE}_{p-1}(f)\big) \in \mathbb{R}^{p-1}$$

is continuous at $f = f_0$. Then for almost every $x \in M$ the Oseledets decomposition is either dominated along the orbit of x or is trivial.

The main idea of the proof can be described as follows (we borrow this description from [49]). If the Oseledets decomposition is neither dominated nor trivial over a set of orbits of positive volume then for some i and arbitrary large m there exist infinitely many iterates $y_j = f^{n_j}(x)$ for which

$$\big\|df^m|E_i^-(y)\big\|\,\big\|\big(df^m|E_j^+(y)\big)^{-1}\big\| > \frac{1}{2}, \tag{13.4}$$

where

$$E_i^+(y) = E_1(y) \oplus \cdots \oplus E_i(y)$$

and

$$E_i^-(y) = E_{i+1}(y) \oplus \cdots \oplus E_{p(y)}(y).$$

Applying a small perturbation one can move a vector originally in $E_i^+(y)$ to $E_i^-(y)$ thus "blending" different expansion rates.

More precisely, fix $\varepsilon > 0$, sufficiently large m and a point $x \in M$. For n much bigger than m choose an iterate $y = f^\ell(x)$ with $\ell \approx \frac{n}{2}$ as in (13.4). By composing df with small rotations near the first m iterates of y one can cause the orbit of some $df_x^\ell v \in E_i^+(y)$ to move to $E_i^-(z)$. This creates a perturbation $g = f \circ h$ which preserves the orbit segment $\{x, \ldots, f^n(x)\}$ and is such that $dg_x^s v \in E_i^+$ during the first ℓ iterates and $dg_x^s v \in E_i^-$ during the last $n - \ell - m \approx \frac{n}{2}$ iterates. We wish to conclude that dg_x^n lost some expansion if compared to df_x^n. To this end we compare the kth exterior products of these linear maps with $k = \dim E_i^+$. We have

$$\big\|\wedge^p(dg_x^n)\big\| \leqslant \exp\left(n\left(\chi_1 + \cdots + \chi_{k-1} + \frac{1}{2}(\chi_k + \chi_{k+1})\right)\right),$$

where the Lyapunov exponents are computed at (f, x). Notice that $\chi_{k+1} = \hat{\lambda}_{i+1}$ is strictly smaller than $\chi_k = \hat{\lambda}_i$. This local procedure is then repeated for a positive volume set of points $x \in M$. Using the fact that

$$\mathrm{LE}_k(g) = \inf_n \frac{1}{n} \int_M \log \left\| \wedge^p (dg_x^n) \right\| dm$$

one can show that $\mathrm{LE}_k(g)$ drops under such arbitrary small perturbations contradicting continuity.

For the above construction to work one should arrange various intermediate perturbations around each $f^s(y)$ not to interfere with each other nor with other iterates of x in the time interval $\{0, \dots, n\}$. One can achieve this by rescaling the perturbation $g = f \circ h$ near each $f^s(y)$ if necessary to ensure that its support is contained in a sufficiently small neighborhood of the point. In a local coordinate w around $f^s(y)$ rescaling corresponds to replacing $h(w)$ by $rh(w/r)$ for some small $r > 0$. This does not affect the value of the derivative at $f^s(y)$ nor the C^1 norm of the perturbation and thus it can be made close to f in the C^1 topology. It is not clear whether the argument can be modified to work in C^q with $q > 1$.

One can establish a version of Theorem 13.7 in the symplectic case.

THEOREM 13.9 (Bochi and Viana [49]). *Let M be a compact smooth Riemannian manifold. There exists a residual subset $\mathcal{U}$ in the space of C^1 symplectic diffeomorphisms such that every $f \in \mathcal{U}$ is either Anosov or has at least two zero Lyapunov exponents at almost every $x \in M$.*

13.5. *C^0-genericity for cocycles*

We now describe a version of Theorem 13.7 for linear cocycles.

Let $S \subset GL(n, \mathbb{R})$ be an embedded submanifold (with or without boundary). We say that S is *accessible* if it acts transitively on the projective space $\mathbb{R}P^{n-1}$. More precisely, for any $C > 0$, $\varepsilon > 0$ there are $m > 0$ and $\alpha > 0$ with the following property: given $\xi, \eta \in \mathbb{R}P^{n-1}$ with $\angle(\xi, \eta) \leqslant \alpha$ and any $A_0, \dots, A_{m-1} \in S$ with $\|A_i^{\pm 1}\| \leqslant C$ one can find $\tilde{A}_0, \dots, \tilde{A}_{m-1} \in S$ such that $\|A_i - \tilde{A}_i\| \leqslant \varepsilon$ and

$$\tilde{A}_0, \dots, \tilde{A}_{m-1}(\xi) = A_0, \dots, A_{m-1}(\eta).$$

Let X be a compact Hausdorff space and $f : X \to X$ a homeomorphism preserving a Borel probability measure μ. Let also $\mathcal{A} : X \times \mathbb{Z} \to GL(n, \mathbb{R})$ be the cocycle over f generated by a function $A : X \to GL(n, \mathbb{R})$.

THEOREM 13.10 (Bochi and Viana [49]). *For any accessible set $S \subset GL(n, \mathbb{R})$ there exists a residual set $\mathcal{R} \subset C(X, S)$ such that for every $A \in \mathcal{R}$ and almost every $x \in X$ either all Lyapunov exponents of the cocycle $\mathcal{A}$, generated by A, are equal to each other or the Oseledets decomposition for $\mathcal{A}$ (see* (5.1)*) is dominated.*

This result applies to cocycles associated with Schrödinger operators. In this case $X = S^1$, $f : S^1 \to S^1$ is an irrational rotation, $f(x) = x + \alpha$, and the generator $A : S^1 \to SL(2, \mathbb{R})$ is given by

$$A(\theta) = \begin{pmatrix} E - V(\theta) & -1 \\ 1 & 0 \end{pmatrix},$$

where $E \in \mathbb{R}$ is the total energy and $V : S^1 \to \mathbb{R}$ is the potential energy. The cocycle generated by A is a point of discontinuity for the Lyapunov exponents, as functions of $V \in C^0(S^1, \mathbb{R})$, if and only if the exponents are nonzero and E lies in the spectrum of the associated Schrödinger operator (E lies in the complement of the spectrum if and only if the cocycle is uniformly hyperbolic which for cocycles with values in $SL(2, \mathbb{R})$ is equivalent to domination; see also Ruelle [214], Bourgain [55] and Bourgain and Jitomirskaya [56]).

For $V \in C^r(S^1, \mathbb{R})$ with $r = \omega, \infty$, Avila and Krikorian [25] proved the following result on nonuniform hyperbolicity for Schrödinger cocycles: *if α satisfies the recurrent Diophantine condition* (*i.e, there are infinitely many $n > 0$ for which the nth image of α under the Gauss map satisfies the Diophantine condition with fixed constant and power*) *then for almost every E the Schrödinger cocycle either has nonzero Lyapunov exponents or is C^r-equivalent to a constant cocycle*.

For some C^1-genericity results on positivity of the maximal Lyapunov exponents see Sections 7.3.3 and 7.3.5.

13.6. *L^p-genericity for cocycles*

Let (X, μ) be a probability space and $f : X \to X$ a measure preserving automorphism. Consider the cocycle $\mathcal{A} : X \times \mathbb{Z} \to GL(n, \mathbb{R})$ over f generated by a measurable function $A : X \to GL(n, \mathbb{R})$. We endow the space $\mathcal{G}$ of these functions with a special L^p-like topology. Set for $1 \leqslant p < \infty$,

$$\|A\|_p = \left(\int_X \|A(x)\|^p \, d\mu(x) \right)^{1/p}$$

and

$$\|A\|_\infty = \operatorname*{ess\,sup}_{x \in X} \|A\|.$$

We have $0 \leqslant \|A\|_p \leqslant \infty$. For $A, B \in \mathcal{G}$ let

$$\tau_p(A, B) = \|A - B\|_p + \left\|A^{-1} - B^{-1}\right\|_p$$

and

$$\rho_p(A, B) = \frac{\tau_p(A, B)}{1 + \tau_p(A, B)}.$$

Here we agree that $\|A - B\|_p = \infty$ or $\|A^{-1} - B^{-1}\|_p = \infty$ if and only if $\rho_p(A, B) = 1$. One can check that ρ_p is a metric on $\mathcal{G}$ and that the space $(\mathcal{G}, \rho_p)$ is complete.

Assume that f is ergodic. Following Arbieto and Bochi [22] we denote by $\mathcal{G}_{IC} \subset \mathcal{G}$ the subset of all maps A satisfying the integrability condition (5.3) and by $\mathcal{G}_{OPS}$ the subset of all those $A \in \mathcal{G}_{IC}$ which have one-point spectrum, i.e., for which the Lyapunov spectrum of the cocycle $\mathcal{A}$ consists of a single point. It turns out that the "one-point spectrum property" is typical in the following sense (see Arbieto and Bochi [22]; an earlier but weaker result is obtained by Arnold and Cong in [24]).

THEOREM 13.11 (Arbieto and Bochi [22]). *Assume that f is ergodic. Then $\mathcal{G}_{OPS}$ is a residual subset of $\mathcal{G}_{IC}$ in the L^p topology for any $1 \leqslant p \leqslant \infty$.*

The proof of this result is based upon the study of the functions $\Lambda_k : \mathcal{G}_{IC} \to \mathbb{R}$, $k = 1, \ldots, n$, given by

$$\Lambda_k(A) = \int_X \big(\chi_1(A, x) + \cdots + \chi_k(A, x)\big)\, d\mu(x).$$

THEOREM 13.12 (Arbieto and Bochi [22]). *The following statements hold*:

1. *the function Λ_k is upper-semicontinuous (i.e., for any $A \in \mathcal{G}_{IC}$ and $\varepsilon > 0$ there exists $\delta > 0$ such that $\Lambda_k(B) < \Lambda_k(A) + \varepsilon$ for any $B \in \mathcal{G}_{IC}$ with $\rho_p(A, B) < \delta$)*;
2. *the function Λ_n is continuous*;
3. *if f is ergodic then Λ_k is continuous at $A \in \mathcal{G}_{IC}$ if and only if $A \in \mathcal{G}_{OPS}$.*

For some other results on genericity of cocycles with low differentiability see [53].

13.7. *Mixed hyperbolicity*

We consider the situation of *mixed* hyperbolicity, i.e., hyperbolicity is uniform throughout the manifold in some but not all directions. More precisely, we assume that f is partially hyperbolic, i.e., the tangent bundle TM is split into three df-invariant continuous subbundles

$$TM = E^s \oplus E^c \oplus E^u. \tag{13.5}$$

The differential df contracts uniformly over $x \in M$ along the *strongly stable* subspace $E^s(x)$, it expands uniformly along the *strongly unstable* subspace $E^u(x)$, and it can act either as nonuniform contraction or expansion with weaker rates along the *central* direction $E^c(x)$. More precisely, there exist numbers

$$0 < \lambda_s < \lambda_c' \leqslant 1 \leqslant \lambda_c'' < \lambda_u$$

such that for every $x \in M$,

$$\lambda_c'\|v\| \leqslant \begin{cases} \|d_x f(v)\| \leqslant \lambda_s\|v\|, & v \in E^s(x), \\ \|d_x f(v)\| \leqslant \lambda_c''\|v\|, & v \in E^c(x), \\ \lambda_u\|v\| \leqslant \|d_x f(v)\|, & v \in E^u(x). \end{cases}$$

We say that a partially hyperbolic diffeomorphism preserving a smooth measure μ has *negative central exponents* on a set A of positive measure if $\chi(x, v) < 0$ for every $x \in A$ and every nonzero $v \in E^c(x)$. The definition of *positive central exponents* is analogous. Partially hyperbolic systems with negative (positive) central exponents as explained above were introduced by Burns, Dolgopyat and Pesin [64] in connection to stable ergodicity of partially hyperbolic systems (see below). Their work is based upon earlier results of Alves, Bonatti and Viana [19] who studied SRB-measures for partially hyperbolic systems for which the tangent bundle is split into two invariant subbundles, one uniformly contracting and the other nonuniformly expanding (see Section 14.3).

For $x \in M$ one can construct local stable manifolds $V^s(x)$ and local unstable manifolds $V^u(x)$ and their sizes are bounded away from zero uniformly over $x \in M$. In addition, for $x \in A$ one can construct *local weakly stable manifolds* $V^{sc}(x)$ whose size varies with x and may be arbitrary close to zero.

THEOREM 13.13 (Burns, Dolgopyat and Pesin [64]). *Let f be a C^2 diffeomorphism of a compact smooth Riemannian manifold M preserving a smooth measure μ. Assume that there exists an invariant subset $A \subset M$ with $\mu(A) > 0$ on which f has negative central exponents. Then every ergodic component of $f|A$ is open* (mod 0) *and so is the set A.*

To see this let us take a density point $x \in A$ and consider the sets

$$P(x) = \bigcup_{y \in V^{sc}} V^u(y), \qquad Q(x) = \bigcup_{n \in \mathbb{Z}} f^n\big(P(x)\big). \tag{13.6}$$

$P(x)$ is open and so is $Q(x)$. Using absolute continuity of local unstable manifolds and repeating argument in the proof of Theorem 11.3 we obtain that $f|Q(x)$ is ergodic.

In general, one should not expect the set A to be of full measure nor the map $f|A$ to be ergodic. We introduce a sufficiently strong condition which guarantees this.

We call two points $p, q \in M$ *accessible* if there are points $p = z_0, z_1, \dots, z_{\ell-1}, z_\ell = q, z_i \in M$, such that $z_i \in V^u(z_{i-1})$ or $z_i \in V^s(z_{i-1})$ for $i = 1, \dots, \ell$. The collection of points $z_0, z_1, \dots, z_\ell$ is called a *us-path* connecting p and q. Accessibility is an equivalence relation. The diffeomorphism f is said to have the *accessibility property* if any two points $p, q \in M$ are accessible and to have the *essential accessibility property* if the partition into accessibility classes is trivial (i.e., a measurable union of equivalence classes must have zero or full measure).

A crucial manifestation of the essential accessibility property is that the orbit of almost every point $x \in M$ is dense in M. This implies the following result.

THEOREM 13.14 (Burns, Dolgopyat and Pesin [64]). *Let f be a C^2 partially hyperbolic diffeomorphism of a compact smooth Riemannian manifold M preserving a smooth mea-*

sure μ. Assume that f has negative central exponents on an invariant set A of positive measure and is essentially accessible. Then f has negative central exponents on the whole of M, the set A has full measure, f has nonzero Lyapunov exponents almost everywhere, and f is ergodic.

Accessibility plays a crucial role in stable ergodicity theory. A C^2 diffeomorphism f preserving a Borel measure μ is called *stably ergodic* if any C^2 diffeomorphism g which is sufficiently close to f in the C^1 topology, which preserves μ, is ergodic. Volume-preserving Anosov diffeomorphisms are stably ergodic.

THEOREM 13.15 (Burns, Dolgopyat and Pesin [64]). *Under the assumption of Theorem* 13.14, *f is stably ergodic.*

One can show that indeed, f is stably Bernoulli, i.e., any C^2 diffeomorphism g which is sufficiently close to f in the C^1 topology, which preserves μ, is Bernoulli.

The proof of Theorem 13.15 is based upon some delicate properties of Lyapunov exponents for systems with mixed hyperbolicity which are of interest by themselves.

1. Since the map f is ergodic the values of the Lyapunov exponents are constant almost everywhere. Therefore,

$$\chi(x, v) \leqslant a < 0 \tag{13.7}$$

uniformly over x and $v \in E^c(x)$. It follows that

$$\int_M \log \|df|E^c(x)\| \, d\mu \leqslant a < 0.$$

Since the splitting (13.5) depends continually on the perturbation g of f we obtain that

$$\int_M \log \|df|E^c_g(x)\| \, d\mu \leqslant \frac{a}{2} < 0$$

(we assume that g is sufficiently close to f and preserves the measure μ). This, in turn, implies that g has negative central exponents on a set A_g of positive μ-measure.

2. Condition (13.7) allows one to estimate the sizes of global weakly stable manifolds along a typical trajectory of g.

PROPOSITION 13.16. *Under the assumption* (13.7) *there is a number $r > 0$ such that for any C^2 diffeomorphism g which is sufficiently close to f in the C^1 topology and for any $x \in A_g$ one can find $n \geqslant 0$ such that the size of the global manifold $W^{sc}(g^{-n}(x))$ is at least r.*

The proof of this statement uses the notion of σ-hyperbolic times which is of interest by itself and provides a convenient technical tool in studying the behavior of local manifolds

along trajectories. It was introduced by Alves in [17] (see also [19]) but some basic ideas behind this notion go back to the work of Pliss [205] and Mañé [175]. Given a partially hyperbolic diffeomorphism f and a number $0 < \sigma < 1$, we call the number n a *σ-hyperbolic time for f at x* if for every $0 \leqslant j \leqslant n$,

$$\prod_{k=1}^{j} \left\| df | E_f^c\big(f^{k-n}(x)\big) \right\| \leqslant \sigma^j. \tag{13.8}$$

It is shown by Alves, Bonatti and Viana in [19] that if f satisfies (13.7) then any point $x \in A_f$ has infinitely many hyperbolic times. The proof of this statement is based on a remarkable result known as Pliss lemma. Although technical this lemma provides an important observation related to nonuniform hyperbolicity.

LEMMA 13.17 (Pliss [205]; see also [175, Chapter IV.11]). *Let $H \geqslant c_2 > c_1 > 0$ and $\zeta = (c_2 - c_1)/(H - c_1)$. Given real numbers $a_1, \ldots, a_N$ satisfying*

$$\sum_{j=1}^{N} a_j \geqslant c_2 N \quad \text{and} \quad a_j \leqslant H \quad \text{for all } 1 \leqslant j \leqslant N,$$

there are $\ell > \zeta N$ and $1 < n_1 < \cdots < n_\ell \leqslant N$ such that

$$\sum_{j=n+1}^{n_j} a_j \geqslant c_1(N_i - n) \quad \text{for each } 0 \leqslant n < n_i,\ i = 1, \ldots, \ell.$$

Alves and Araújo [18] estimated the frequency of σ-hyperbolic times. More precisely, given $\theta > 0$ and $x \in M$ we say that the frequency of σ-hyperbolic times $n_1 < n_2 < \cdots < n_\ell$ at x exceeds θ if for large n we have $n_\ell \leqslant n$ and $\ell \geqslant \theta n$. We also introduce the function h on M which is defined almost everywhere and assigns to $x \in M$ its first σ-hyperbolic time.

THEOREM 13.18. *If for some $\sigma \in (0, 1)$ the function h is Lebesgue integrable then there are $\hat{\sigma} > 0$ and $\theta > 0$ such that almost every $x \in M$ has frequency of hyperbolic times bigger than θ.*

We return to the proof of the proposition. As we saw the map g also satisfies (13.7). Applying (13.8) to g we obtain that there is a number $r > 0$ such that for any σ-hyperbolic time n and $0 \leqslant j \leqslant n$,

$$\operatorname{diam}\big(g^j\big(B^{sc}\big(g^{-n}(x), r\big)\big)\big) \leqslant \sigma^j,$$

where $B^{sc}(y, r)$ is the ball in the global manifold $W^{sc}(y)$ centered at y of radius r. Since $\sigma < 1$ and the hyperbolic time n can be arbitrary large this ensures that $g^n(B^{sc}(g^{-n}(x), r))$ lies in the local manifold $V^{sc}(x)$ and hence, $B^{sc}(g^{-n}(x), r)$ is contained in the global manifold $W^{sc}(g^{-n}(x))$.

3. The perturbation g possesses the *ε-accessibility property* where $\varepsilon = d_{C^1}(f, g)$. This means that for any two points $p, q \in M$ there exists a us-path connecting p with the ball centered at q of radius ε. Although ε-accessibility is weaker then accessibility it still allows one to establish ergodicity of the perturbation g. Indeed, choosing a density point $x \in A_g$ and a number n such that the size of $W^{sc}(g^{-n}(x))$ is at least r we obtain that the set $P(x)$ (see (13.6)) contains a ball of radius $r \geqslant 2\varepsilon$.

13.8. *Open sets of diffeomorphisms with nonzero Lyapunov exponents*

It is shown in [51] that any partially hyperbolic diffeomorphism f_0 with one-dimensional central direction preserving a smooth measure μ can be slightly perturbed such that the new map f is partially hyperbolic, preserves μ and has negative central exponents (hence, the results of the previous section apply to f). This result was first obtained by Shub and Wilkinson [229] in the particular case when f_0 is the direct product of a hyperbolic automorphism of two-torus and the identity map of the circle. The perturbation that remove zero exponents can be arranged in the form (13.3). The proof in the general case is a modification of the argument in [229] (see also [34] and [84]).

One can use this observation to obtain an open set of non-Anosov diffeomorphisms with nonzero Lyapunov exponents on multi-dimensional tori. Consider a diffeomorphism $f_0 = A \times \mathrm{Id}$ of the torus $\mathbb{T}^p = \mathbb{T}^{p-1} \times S^1$, $p \geqslant 3$, where A is a linear hyperbolic automorphism of $\mathbb{T}^{p-1}$. It is partially hyperbolic and preserves volume. Let f be a small C^2 perturbation of f_0 preserving volume and having negative central exponents. One can arrange the perturbation f to have the accessibility property. Then any volume-preserving C^2 diffeomorphism g which is sufficiently close to f is ergodic and has nonzero Lyapunov exponents almost everywhere.

Note that g is partially hyperbolic and the central distribution E^c is one dimensional. By a result in [118] this distribution is integrable and the leaves W^c of the corresponding foliation are smooth closed curves which are diffeomorphic to circles. The foliation W^c is continuous (indeed, it is Hölder continuous) but is not absolutely continuous (see [229]). Moreover, there exists a set E of full measure and an integer $k > 1$ such that E intersects almost every leaf $W^c(x)$ at exactly k points (see [219]; the example in Section 10.2 is of this type).

14. SRB-measures

We shall consider hyperbolic invariant measures which are not smooth. This includes, in particular, dissipative systems for which the support of such measures is attracting invariant sets. A general hyperbolic measure may not have "nice" ergodic properties: its ergodic components may be of zero measure and it may have zero metric entropy. There is, however, an important class of hyperbolic measures known as SRB-measures (after Sinai, Ruelle and Bowen). They appear naturally in applications due to the following observation.

Let f be a diffeomorphism of a smooth Riemannian p-dimensional manifold M. An open set $U \subset M$ is called a *trapping region* if $\overline{f(U)} \subset U$ (where $\bar{A}$ denotes the closure of the set A). The closed f-invariant set

$$\Lambda = \bigcap_{n \geqslant 0} f^n(U)$$

is an *attractor* for f so that f is dissipative in U.

Consider the evolution of the Riemannian volume m under f, i.e., the sequence of measures

$$\nu_n = \frac{1}{n} \sum_{k=0}^{n-1} f_*^k m, \tag{14.1}$$

where the measure $f_*^k m$ is defined by $f_*^k m(A) = m(f^{-1}(A))$ for any Borel set $A \subset F^k(U)$. Any limit measure ν of this sequence is supported on Λ. If indeed, the sequence (14.1) *converges* the limit measure ν is the *physical* (or *natural*) measure on Λ. The latter plays an important role in applications and is defined by the following property: for any continuous function φ on M, called *observable*, and m-almost every point $x \in U$,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi\big(f^k(x)\big) = \int_M \varphi \, d\nu. \tag{14.2}$$

We call ν an *SRB-measure* if there is a set $B = B(\nu) \subset U$ of positive Lebesgue measure such that for any continuous observable φ the identity (14.2) holds for $x \in B$ (in this case Λ is a *Milnor attractor*, see [184]). The set $B(\nu)$ is the *basin of attraction* of ν.

Assume that for m-almost every point $x \in U$ the Lyapunov exponents $\chi_i(x)$, $i = 1, \ldots, p$, are not equal to zero. More precisely, there is a number $1 \leqslant k(x) < p$ such that $\chi_i(x) < 0$ for $i = 1, \ldots, k(x)$ and $\chi_i(x) > 0$ for $i = k(x) + 1, \ldots, p$. It is not known whether *under this assumption the measure ν is hyperbolic*.

We stress that a physical measure need not be an SRB-measure as Example 12.5 of the figure-eight attractor shows. In the following sections we give another (equivalent) definition of SRB-measures in the case when these measures are hyperbolic. We also discuss their ergodic properties, and present some examples of systems with SRB-measures (for a somewhat less elaborated account of SRB-measures see [1]). In uniformly hyperbolic dynamics SRB-measures are examples of more general Gibbs measures (see the recent excellent survey on this topic by Ruelle [218]). It is an open problem to extend the theory of Gibbs measures to nonuniformly hyperbolic dynamical systems.

14.1. *Definition and ergodic properties of SRB-measures*

Let f be a $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold M and ν a hyperbolic invariant measure for f. Denote by $\Lambda = \Lambda_\nu$ the set of points with nonzero

Lyapunov exponents. We have that $\nu(\Lambda) = 1$. Fix a regular set Λ^ℓ of positive measure, a point $x \in \Lambda^\ell$, and a number $0 < r < r_\ell$ (see (8.16)). Set

$$R^\ell(x, r) = \bigcup_{y \in \Lambda^\ell \cap B(x,r)} V^u(y)$$

and denote by $\xi^\ell(x)$ the partition of $R^\ell(x, r)$ by local unstable manifolds $V^u(y)$, $y \in \Lambda^\ell \cap B(x, r)$.

A hyperbolic measure ν is called an *SRB-measure* if for every $\ell > 0$ and almost every $x \in \Lambda^\ell$, $y \in \Lambda^\ell \cap B(x, r)$, the conditional measure $\nu^u(y)$, generated by ν and partition $\xi^\ell(x)$ on $V^u(y)$, is absolutely continuous with respect to the Riemannian volume $m^u(y)$ on $V^u(y)$.

There is a measurable density function $\kappa(y, z)$, $z \in V^u(y)$, such that $d\nu^u(y)(z) = \kappa(y, z)\, dm^u(y)(z)$. The following result gives a description of the density function $\kappa(y, z)$.

THEOREM 14.1. *For any $y \in \Lambda^\ell \cap B(x, r)$ and $z \in V^u(y)$,*

$$\kappa(y, z) = \prod_{i=1}^{\infty} \frac{J(df^{-1}|E^u(f^{-i}(z)))}{J(df^{-1}|E^u(f^{-i}(y)))}.$$

The density function $\kappa(y, z)$ is Hölder continuous and strictly positive.

SRB-measures have ergodic properties similar to those of smooth measures. The proofs of the corresponding results use Theorem 14.1 and are modifications of arguments in the case of smooth measures (those proofs in the latter case use only absolute continuity of local unstable manifolds), see Ledrappier [156].

THEOREM 14.2. *There exist invariant sets $\Lambda_0, \Lambda_1, \ldots$ such that*:

1. $\bigcup_{i \geqslant 0} \Lambda_i = \Lambda$, *and* $\Lambda_i \cap \Lambda_j = \emptyset$ *whenever* $i \neq j$;
2. $\nu(\Lambda_0) = 0$, *and* $\nu(\Lambda_i) > 0$ *for each* $i \geqslant 1$;
3. $f|\Lambda_i$ *is ergodic for each* $i \geqslant 1$.

THEOREM 14.3. *There exists a measurable partition η of Λ with the following properties*:

1. *for almost every $x \in \Lambda$ the element $C_\eta(x)$ is open* (mod 0) *subset of $W^u(x)$*;
2. $f\eta \geqslant \eta$;
3. $\eta^+ = \bigvee_{i=0}^{\infty} f^i \eta = \varepsilon$;
4. $\bigwedge_{i>-\infty}^{0} f^i \eta = \mathcal{H}(W^u) = \pi(f|\Lambda) = \nu$ (*the trivial partition of Λ*);
5. *for each $i = 1, 2, \ldots$ the π-partition of $f|\Lambda_i$ is finite and consists of n_i elements Λ_i^k, $k = 1, \ldots, n_i$, such that $f(\Lambda_i^k) = \Lambda_i^{k+1}$, $k = 1, \ldots, n_i - 1$, and $f(\Lambda_i^n) = \Lambda_i^1$.*

THEOREM 14.4.

1. $f^{n_i}|\Lambda_i^k$ *is a Bernoulli automorphism.*
2. *If the map $f|\Lambda$ is mixing then it is a Bernoulli automorphism.*

3. (*Ledrappier and Strelcyn* [158]) *The entropy of f is*

$$h_\nu(f) = h_\nu(f,\eta) = \int_M \sum_{i:\ \chi_i(x)>0} k_i(x)\chi_i(x)\,d\nu(x).$$

Let ν be an SRB-measure for a $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold M and Λ the set of points with nonzero Lyapunov exponents. We have that $\nu(\Lambda) = 1$ and $V^u(x) \subset \Lambda \pmod 0$ for almost every $x \in \Lambda$. In view of the absolute continuity of local stable manifolds we obtain that the set $\bigcup_{x\in\Lambda} V^s(x)$ has positive volume. As an immediate corollary of this observation we have the following result.

THEOREM 14.5. *A $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold possesses at most countably many ergodic SRB-measures. The basin of attraction of every SRB-measure has positive volume.*

14.2. *Characterization of SRB-measures*

It turns out that the entropy formula (see statement 3 of Theorem 14.4) completely characterizes SRB-measures.

THEOREM 14.6. *For a Borel measure ν invariant under a C^2 diffeomorphism, the entropy formula holds if and only if ν is an SRB-measure.*

This characterization was first established by Ledrappier [156] for systems with nonzero Lyapunov exponents and in the general case by Ledrappier and Young (see [159]; see also Section 16.1 for a discussion of this result). It is also shown in [159] that the Radon–Nikodym derivatives $d\nu^u(x)/dm^u(x)$ are strictly positive functions which are C^1 along unstable manifolds.

Qian and Zhu [210] extended the notion of SRB-measures to C^2 endomorphism via their inverse limits. They also established the entropy formula and the same characterization of SRB-measures as in the above theorem.

14.3. *Existence of SRB-measures I: Some general results*

We describe here results on existence of SRB-measures in some general situations.

1. A topologically transitive Anosov diffeomorphism f possesses an ergodic SRB-measure: it is the limit of the sequence of measures (14.1). This result extends to uniformly hyperbolic attractors, i.e., attractors which are hyperbolic sets. For "almost Anosov" diffeomorphisms Hu [123] found conditions which guarantee existence of SRB-measures, while Hu and Young [125] described examples of such maps with no finite SRB-measures (see the articles [4, Section 3.6] and [1, Section 3] for relevant definitions and details).

2. More generally, consider a partially hyperbolic attractor Λ, i.e., an attractor such that $f|\Lambda$ is partially hyperbolic (see Section 9 in the chapter "Partially hyperbolic dynamical systems" by B. Hasselblatt and Ya. Pesin in this volume [6]). Observe that $W^u(x) \subset \Lambda$ for every $x \in \Lambda$.

Let ν be an invariant Borel probability measure supported on Λ. Given a point $x \in \Lambda$, and a small number $r > 0$, set

$$R(x,r) = \bigcup_{y \in \Lambda \cap B(x,r)} V^u(y).$$

Denote by $\xi(x)$ the partition of $R(x,r)$ by $V^u(y)$, $y \in \Lambda \cap B(x,r)$. Following [204] we call ν a *u-measure* if for almost every $x \in \Lambda$ and $y \in \Lambda \cap B(x,r)$, the conditional measure $\nu^u(y)$, generated by ν and partition $\xi(x)$ on $V^u(y)$, is absolutely continuous with respect to $m^u(y)$.

THEOREM 14.7 (Pesin and Sinai [204]). *Any limit measure of the sequence of measures* (14.1) *is a u-measure on* Λ.

Since partially hyperbolic attractors may not admit Markov partitions, the proof of this theorem exploits quite a different approach than the one used to establish existence of SRB-measures for classical hyperbolic attractors (see Section 19 where this approach is outlined).

In general, the sequence of measures (14.1) may not converge and some strong conditions are required to guarantee convergence.

THEOREM 14.8 (Bonatti and Viana [52]). *Assume that*:

1. *every leaf of the foliation* W^u *is everywhere dense in* Λ;
2. *there exists a limit measure* ν *for the sequence of measures* (14.1) *with respect to which* f *has negative central exponents*.

Then the sequence of measures (14.1) *converges and the limit measure is the unique u-measure on* Λ. *It is an SRB-measure.*

Every SRB-measure on Λ is a u-measure. The converse statement is not true in general but it is true in the following two cases:

(a) Λ is a (completely) hyperbolic attractor;

(b) f has negative central exponents.

Here are two results in this direction.

THEOREM 14.9 (Alves, Bonatti and Viana [19]). *Assume that* f *is nonuniformly expanding along the center-unstable direction, i.e.,*

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{j=1}^{n} \log \left\| df^{-1} | E^{cu}_{f^j(x)} \right\| < 0 \tag{14.3}$$

for all x in a positive Lebesgue measure set $A \subset M$. Then f has an ergodic SRB-measure supported in $\bigcap_{j=0}^{\infty} f^j(M)$. Moreover, if the limit in (14.3) *is bounded away from zero then A is contained* (mod 0) *in the union of the basins of finitely many SRB-measures.*

THEOREM 14.10 (Burns, Dolgopyat and Pesin [64]). *Let ν be a u-measure on Λ. Assume that there exists an invariant subset $A \subset \Lambda$ with $\mu(A) > 0$ on which f has negative central exponents. Assume also that for every $x \in \Lambda$ the global unstable manifold $W^u(x)$ is dense in Λ. Then ν is the only u-measure for f and f has negative central exponents at ν-almost every $x \in \Lambda$; hence, (f, ν) is ergodic, ν is an SRB-measure and its basin contains the topological basin of Λ* (mod 0).

3. The following general statement links convergence of the sequence of measures (14.1) to the existence of SRB-measures.

THEOREM 14.11 (Tsujii [241]). *Let f be a $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold M and $A \subset M$ a set of positive volume such that for every $x \in A$ the sequence of measures*

$$\frac{1}{n} \sum_{i=0}^{n-1} \delta_{f^k(x)}$$

converges weakly to an ergodic hyperbolic measure ν_x. If the Lyapunov exponents at x coincide with those of ν_x then ν_x is an SRB-measure for Lebesgue almost every $x \in A$.

4. In [259], Young suggested an axiomatic approach for constructing SRB-measures. It is built upon her work on tower constructions for nonuniformly hyperbolic systems and presents the system as a Markov extension (see Appendix A). This approach is a basis to establish existence of SRB-measures for Hénon-type attractors as well as existence of absolutely continuous invariant measures for some piecewise hyperbolic maps and logistic maps.

Let f be a $C^{1+\alpha}$ diffeomorphism of a compact smooth Riemannian manifold M.

An embedded disk $\gamma \subset M$ is called an *unstable disk* if for any $x, y \in \gamma$ the distance $\rho(f^{-n}(x), f^{-n}(y)) \to 0$ exponentially fast as $n \to \infty$; it is called a *stable disk* if for any $x, y \in \gamma$ the distance $\rho(f^n(x), f^n(y)) \to 0$ exponentially fast as $n \to \infty$.

We say that a set Λ has a *hyperbolic product structure* if there exist a continuous family of unstable disks $\Gamma^u = \{\gamma^u\}$ and a continuous family of stable disks $\Gamma^s = \{\gamma^s\}$ such that

1. $\dim \gamma^u + \dim \gamma^s = \dim M$;
2. the γ^u-disks are transversal to the γ^s-disks with the angles between them bounded away from zero;
3. each γ^u-disk meets each γ^s-disk at exactly one point;
4. $\Lambda = (\bigcup \gamma^u) \cap (\bigcup \gamma^s)$.

We impose some conditions on the map f (see conditions (P1)–(P5) below) which guarantee the existence of an SRB-measure for f. Roughly speaking they mean that there exists a set Λ with a hyperbolic product structure and a return map f^R from Λ to itself such

that f is a Markov extension over f^R (see Appendix A). More precisely, we assume the following.

(P1) There exists $\Lambda \subset M$ with a hyperbolic product structure and such that $\mu_{\gamma^u}\{\gamma^u \cap \Lambda\} > 0$ for every $\gamma^u \in \Gamma^u$.

(P2) There are pairwise disjoint subsets $\Lambda_1, \Lambda_2, \ldots \subset \Lambda$ with the properties that

1. each Λ_i has a hyperbolic product structure and its defining families can be chosen to be Γ^u and $\Gamma_i^s \subset \Gamma^s$; we call Λ_i an s-subset; similarly, one defines u-subsets;
2. on each γ^u-disk, $\mu_{\gamma^u}\{(\Lambda \setminus \bigcup \Lambda_i) \cap \gamma^u\} = 0$;
3. there exists $R_i \geqslant 0$ such that $f^{R_i}(\Lambda_i)$ is a u-subset of Λ; moreover, for all $x \in \Lambda_i$ we require that $f^{R_i}(\gamma^s(x)) \subset \gamma^s(f^{R_i}(x))$ and $f^{R_i}(\gamma^u(x)) \supset \gamma^u(f^{R_i}(x))$;
4. for each n, there are at most finitely many i's with $R_i = n$;
5. $\min R_i \geqslant R_0$ for some $R_0 > 0$ depending only on f.

Condition (P2) means that the set Λ has the structure of a "horseshoe", however, infinitely many branches returning at variable times.

In order to state remaining conditions (P3)–(P5) we assume that there is a function $s_0(x, y)$—a *separation time* of the points x and y—which satisfy:

(i) $s_0(x, y) \geqslant 0$ and depends only on the γ^s-disks containing the two points;

(ii) the maximum number of orbits starting from Λ that are pairwise separated before time n is finite for each n;

(iii) for $x, y \in \Lambda$, $s_0(x, y) \geqslant R_i + s_0(f^{R_i}(x), f^{R_i}(y))$; in particular, $s_0(x, y) \geqslant R_i$;

(iv) for $x \in \Lambda_i$, $y \in \Lambda_j$, $i \neq j$ but $R_i = R_j$, we have $s_0(x, y) < R_i - 1$.

Conditions (iii) and (iv) describe the relations between $s_0(x, y)$ and returns to Λ, namely, that points in the same Λ_i must not separate before they return, while points in distinct Λ_i's must first separate if they are to return simultaneously.

We assume that there exist $C > 0$ and $\alpha < 1$ such that for all $x, y \in \Lambda$ the following conditions hold:

(P3) *contraction along γ^s-disks*: $\rho(f^n(x), f^n(y)) \leqslant C\alpha^n$ for all $n \geqslant 0$ and $y \in \gamma^s(x)$;

(P4) *backward contraction and distortion along γ^u*: for $y \in \gamma^u(x)$ and $0 \leqslant k \leqslant n < s_0(x, y)$, we have

(a) $\rho(f^n(x), f^n(y)) \leqslant C\alpha^{s_0(x,y)-n}$;

(b) $$\log \prod_{i=k}^{n} \frac{\det df^u(f^i(x))}{\det df^u(f^i(y))} \leqslant C\alpha^{s_0(x,y)-n};$$

(P5) *convergence of $d(f^i|\gamma^u)$ and absolute continuity of Γ^s*:

(a) for $y \in \Gamma^s(x)$ and $n \geqslant 0$,

$$\log \prod_{i=k}^{\infty} \frac{\det df^u(f^i(x))}{\det df^u(f^i(y))} \leqslant C\alpha^n;$$

(b) for $\gamma, \gamma' \in \Gamma^u$ define $\Theta : \gamma \cap \Lambda \to \gamma' \cap \Lambda$ by $\Theta(x) = \gamma^s(x) \cap \gamma'$. Then Θ is absolutely continuous and

$$\frac{d(\Theta_*^{-1}\mu_{\gamma'})}{d\mu_\gamma}(x) = \prod_{i=0}^{\infty} \frac{\det df^u(f^i(x))}{\det df^u(f^i(\Theta(x)))}.$$

In [259], Young showed that a map f satisfying conditions (P1)–(P5) admits a Markov extension (see Appendix A). As an important corollary one has the following result.

THEOREM 14.12 (Young [259]). *Assume that for some* $\gamma \in \Gamma^u$,

$$\int_{\gamma \cap \Lambda} R \, d\mu_\gamma < \infty.$$

Then f *admits an SRB-measure.*

14.4. *Existence of SRB-measures II: Hénon attractors*

Constructing SRB-measures for nonuniformly hyperbolic dissipative systems is a challenging problem and few examples have been successfully studied.

In Section 19 we will discuss existence of SRB-measures for uniformly hyperbolic dissipative maps with singularities possessing generalized hyperbolic attractors. The behavior of trajectories in these systems is essentially nonuniformly hyperbolic.

An example of nonuniformly hyperbolic dissipative systems possessing SRB-measures is the Hénon map. It was introduced by Hénon in 1977 (see [113]) as a simplified model for the Poincaré first return time map of the Lorenz system of ordinary differential equations. The Hénon family is given by

$$H_{a,b}(x, y) = \big(1 - ax^2 + by, x\big).$$

Hénon carried out numerical studies of this family and suggested the presence of a "chaotic" attractor for parameter values near $a = 1.4$ and $b = 0.3$. Observe that for $b = 0$ the family $H_{a,b}$ reduces to the logistic family Q_a. By continuity, given $a \in (0, 2)$, there is a rectangle in the plane which is mapped by $H_{a,b}$ into itself. It follows that $H_{a,b}$ has an attractor provided b is sufficiently small. This attractor is called the *Hénon attractor*.

In the seminal paper [40], Benedicks and Carleson, treating $H_{a,b}$ as small perturbations of Q_a, developed highly sophisticated techniques to describe the dynamics near the attractor. Building on this analysis, Benedicks and Young [41] established existence of SRB-measures for the Hénon attractors and described their ergodic properties.

THEOREM 14.13. *There exist* $\varepsilon > 0$ *and* $b_0 > 0$ *such that for every* $0 < b \leqslant b_0$ *one can find a set* $\Delta_b \in (2 - \varepsilon, 2)$ *of positive Lebesgue measure with the property that for each* $a \in \Delta_b$ *the map* $H_{a,b}$ *admits a unique SRB-measure* $\nu_{a,b}$.

In [42], Benedicks and Young studied ergodic properties of the measure $\nu_{a,b}$ showing that besides being Bernoulli this measure has exponential decay of correlations and satisfies a central limit theorem. More precisely, they proved the following result.

Let f be a transformation of a Lebesgue space X preserving a probability measure ν and $\mathcal{L}$ be a class of functions on X. We say that f has *exponential decay of correlations* for functions in $\mathcal{L}$ if there is a number $\tau < 1$ such that for every pair of functions $\varphi, \psi \in \mathcal{L}$ there is a constant $C = C(\varphi, \psi) > 0$ such that for all $n \geqslant 0$,

$$\left| \int \varphi(\psi \circ f^n)\, d\nu - \int \varphi\, d\nu \int \psi\, d\nu \right| \leqslant C\tau^n$$

(see Appendix A for more information on decay of correlations). Further, we say that f satisfies a *central limit theorem* for φ with $\int \varphi\, d\nu = 0$ if for some $\sigma > 0$ and all $t \in \mathbb{R}$,

$$\nu\left\{ \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} \varphi \circ f^i < t \right\} \to \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{t} e^{-u^2/2\sigma^2}\, du$$

as $n \to \infty$.

THEOREM 14.14 [42]. *With respect to $\nu_{a,b}$ the map $H_{a,b}$*

1. *has exponential decay of correlations for Hölder continuous functions* (*the rate of decay may depend on the Hölder exponent*);
2. *satisfies the central limit theorem for Hölder continuous functions with zero mean*; *the standard deviation σ is strictly positive if and only if $\varphi \neq \psi \circ f - \psi$ for some $\psi \in L^2(\nu)$.*

In [246], Wang and Young introduced a 2-parameter family of maps of the plane to which the above results extend. This family is defined as follows.

Let $A = S^1 \times [-1, 1]$ and a 2-parameter family $T_{a,b} : A \to A$, $a \in [a_0, a_1]$, $b \in [0, b_1]$, be constructed via the following four steps.

Step I. Let $f : S^1 \to S^1$ satisfies the *Misiurewicz conditions*: if $C = \{x:\ f'(x) = 0\}$ then

1. $f'' \neq 0$ for all $x \in C$;
2. f has negative Schwarzian derivative on $S^1 \setminus C$;
3. $f^n(x) \neq x$ and $|(f^n)'(x)| \leqslant 1$ for any $x \in S^1$ and $n \in \mathbb{Z}$;
4. $\inf_{n \geqslant 0} d(f^n(x), C) > 0$ for all $x \in C$.

Observe that for $p \in S^1$ with $\inf_{n \geqslant 0} d(f^n(p), C) > 0$, and any g sufficiently close to f in the C^2 topology there is a unique point $p(g)$ having the same symbolic dynamics with respect to g as p does with respect to f. If f_a is a 1-parameter family through f with f_a sufficiently close to f in the C^2 topology for all a we let $p(a) = p(f_a)$. For $x \in C$ we denote by $x(a)$ the corresponding critical point of f_a.

Step II. Let $f_a : S^1 \to S^1$ be a 1-parameter family for which $f = f_{a^*}$ for some $a^* \in [a_0, a_1]$ with f as in Step I. We assume that the following *transversality condition* holds: for every $x \in C$ and $p = f(x)$,

$$\frac{d}{dx} f_a\big(x(a)\big) \neq \frac{d}{da} p(a) \quad \text{at } a = a^*.$$

Step III. Let $f_{a,b}: S^1 \times \{0\} \to A$ be a 2-parameter family which is an extension of the 1-parameter family in Step II, i.e., $f_{a,0} = f_a$ and $f_{a,b}$ is an embedding for $b > 0$.

Step IV. Let $T_{a,b}: A \to A$ be an extension of $f_{a,b}$ in such a way that $T_{a,0} \subset S^1 \to A$ and $T_{a,b}$ maps A diffeomorphically onto its image for $b > 0$. Assume also that the following *nondegeneracy condition* holds:

$$\partial_y T_{a^*,0}(x, 0) \neq 0 \quad \text{whenever } f'_{a^*}(x) = 0.$$

For a version of this construction in higher dimensions see [247].

On another direction, Mora and Viana [188] modified Benedicks and Carleson's approach in a way which allowed them to treat Hénon-like maps using some techniques from the general bifurcation theory such as homoclinic tangencies. Later Viana [244] extended results from [188] to higher dimensions; see [170] for a more detailed account of these results and further references.

15. Hyperbolic measures I: Topological properties

One can extend some techniques widely used in the theory of locally maximal hyperbolic sets to measures with nonzero exponents. These tools are not only important for applications but they provide a crucial nontrivial geometric structure to measures with nonzero exponents. In particular, one can close recurrent orbits, shadow pseudo-orbits, construct almost Markov covers, and determine the cohomology class of Hölder cocycles by periodic data.

Let f be a $C^{1+\alpha}$ diffeomorphism of a compact Riemannian manifold M, for some $\alpha > 0$, and ν an f-invariant hyperbolic probability measure.

15.1. *Closing and shadowing lemmas*

We address the following two fundamental problems:

1. Given a recurrent point x is it possible to find a nearby periodic point y which follows the orbit of x during the period of time that the points in the orbit of x return very close to x?
2. Given a sequence of points $\{x_n\}$ with the property that the image of x_n is very close to x_{n+1} for every n (such a sequence $\{x_n\}$ is called a *pseudo-orbit*), is it possible to find a point x such that $f^n(x)$ is close to x_n for every i? In other words, if a sequence of points $\{x_n\}$ resembles an orbit can one find a real orbit that shadows (or closely follows) the pseudo-orbit?

This sort of problems are known respectively as *closing problem* and *shadowing problem*, while the corresponding properties are called the *closing property* and the *shadowing property*.

The following result by Katok [135] establishes the closing property for nonuniformly hyperbolic diffeomorphisms.

THEOREM 15.1. *For every $\ell > 0$ and $\eta > 0$ there exists $\delta = \delta(\ell, \eta) > 0$ with the following property: if $x \in \Lambda^\ell$ and $f^m(x) \in \Lambda^\ell$ with $d(f^m(x), x) < \delta$, then there exists $z = z(x)$ such that*

1. *z is a hyperbolic periodic point for f with $f^m(z) = z$;*
2. *for $i = 0, \dots, m$,*

$$d\big(f^i(z), f^i(x)\big) \leqslant \eta A_\ell \max\big\{e^{\varepsilon i}, e^{\varepsilon(m-i)}\big\},$$

where A_ℓ is a constant depending only on ℓ.

An immediate corollary of this result is the existence of periodic orbits in a regular set Λ^ℓ of a nonuniformly hyperbolic diffeomorphism. In fact, a stronger result holds. Denote by $\mathrm{Per}_h(f)$ the set of hyperbolic periodic points for f.

THEOREM 15.2 (Katok [135]). *We have* $\operatorname{supp} \nu \subset \overline{\mathrm{Per}_h(f)}$.

The proof of Theorem 15.2 is an application of Theorem 15.1. Fix $x_0 \in \operatorname{supp} \mu$, $\alpha > 0$ and $\ell \geqslant 1$ such that $\mu(B(x_0, \alpha/2) \cap \Lambda^\ell) > 0$. Choose $\delta > 0$ according to Theorem 15.1 and such that $\eta A_\ell < \alpha/2$ and a set $B \subset B(x_0, \alpha/2) \cap \Lambda^\ell$ of positive measure and diameter at most δ. By the Poincaré Recurrence Theorem, for μ-almost every $x \in B$ there exists a positive integer $n(x)$ such that $f^{n(x)}(x) \in B$ and hence, $d(f^{n(x)}(x), x) < \delta$. By Theorem 15.1, there exists a hyperbolic periodic point z of period $n(x)$ such that $d(x, z) < \alpha/2$, and thus $d(x_0, z) < d(x_0, x) + d(x, z) < \alpha$.

A further application of Theorem 15.1 is the following statement.

COROLLARY 15.3. *For an ergodic measure ν, if all the Lyapunov exponents of f are negative* (*respectively, all are positive*) *on a set of full ν-measure, then* $\operatorname{supp} \nu$ *is an attracting* (*respectively, repelling*) *periodic orbit.*

See also Corollaries 15.7, 15.15, and 15.16 below for related results.

We now present an analog of the shadowing lemma for nonuniformly hyperbolic diffeomorphisms. Given $a \in \mathbb{Z} \cup \{-\infty\}$ and $b \in \mathbb{Z} \cup \{\infty\}$, a sequence $\{x_n\}_{a<n<b}$ is called an *ε-orbit* or *ε-pseudo-orbit* for f if $d(x_{n+1}, x_n) < \varepsilon$ for all $a < n < b$. It is *δ-shadowed* by the orbit of x if $d(x_n, f^n(x)) < \delta$ for all $a < n < b$.

Given $\ell > 0$, denote by

$$\tilde{\Lambda}^\ell = \bigcup_{x \in \Lambda^\ell} R(x),$$

where $R(x)$ is a regular neighborhood of x (see Section 8.7).

THEOREM 15.4 (Katok and Mendoza [139]). *For every sufficiently small $\alpha > 0$ there exists $\beta = \beta(\alpha, \ell)$ such that given a β-pseudo-orbit $\{x_m\} \subset \tilde{\mathcal{R}}^\ell$, there exists $y \in M$ such that its orbit α-shadows $\{x_m\}$.*

The following result is a nonuniformly hyperbolic version of the famous Livshitz theorem that determines the cohomology class of Hölder cocycles by periodic data.

THEOREM 15.5 (Katok and Mendoza [139]). *Let $\varphi: M \to \mathbb{R}$ be a Hölder continuous function such that for each periodic point p with $f^m(p) = p$ we have $\sum_{i=0}^{m-1} \varphi(f^i(p)) = 0$. Then there exists a Borel measurable function h such that for ν-almost every x,*

$$\varphi(x) = h\big(f(x)\big) - h(x).$$

15.2. *Continuous measures and transverse homoclinic points*

In the neighborhood of any transverse homoclinic point there exists a hyperbolic horseshoe, that is, a (uniformly) hyperbolic invariant set obtained by a horseshoe-like construction (see, for example, [137, Theorem 6.5.5]). This phenomenon persists under small perturbations. It turns out that transverse homoclinic points are present whenever the diffeomorphism possesses hyperbolic continuous measures.

THEOREM 15.6 (Katok [135]). *Let ν be a continuous and nonatomic Borel invariant measure. Then*

1. $\operatorname{supp} \nu$ *is contained in the closure of the set of hyperbolic periodic points that have transverse homoclinic points*;
2. *if ν is ergodic, then* $\operatorname{supp} \nu$ *is contained in the closure of the set of transverse homoclinic points of exactly one hyperbolic periodic point.*

Let $P_m(f)$ be the number of periodic points of f of period m.

COROLLARY 15.7. *Let ν be a continuous and nonatomic Borel invariant measure. Then f has a compact f-invariant set $\Lambda \subset M$ such that*

1. $$\overline{\lim_{m\to\infty}} \frac{1}{m} \log P_m(f) \geqslant h(f|\Lambda) > 0; \tag{15.1}$$

2. *Λ is a horseshoe for f, i.e., Λ is a (uniformly) hyperbolic set for f and $f|\Lambda$ is topologically conjugate to a topological Markov chain.*

In particular, $h(f) > 0$ whenever there exists a continuous nonatomic hyperbolic invariant measure. One can strengthen Theorem 15.6 and obtain a Spectral Decomposition Theorem for hyperbolic measures.

THEOREM 15.8 (Katok and Mendoza [139]). *For each $\ell > 0$, the Pesin set Λ^ℓ can be decomposed into finitely many closed f-invariant sets Λ_i such that for each i there exists $x_i \in M$ with $\Lambda_i \subset \overline{\{f^n(x_i)\colon n \in \mathbb{Z}\}}$.*

Set now

$$\chi(x) = \min\{|\chi_i(x)|\colon 1 \leqslant i \leqslant s\},$$

where $\chi_i(x)$ are the values of the Lyapunov exponent at x. If ν is an ergodic hyperbolic measure, then $\chi(x) = \chi_\nu$, where χ_ν is a nonzero constant.

THEOREM 15.9 (Katok and Mendoza [139]). *Let ν be ergodic. If $x \in \operatorname{supp}\nu$, then for any $\rho > 0$, any neighborhoods V of x and W of $\operatorname{supp}\nu$, and any continuous functions $\varphi_1, \dots, \varphi_k$, there exists a hyperbolic periodic point $z \in V$ such that*:

1. *the orbit of z is contained in W*;
2. *$\chi(z) \geqslant \chi_\nu - \rho$*;
3. *if $m(z)$ is the period of z, then for $i = 1, \dots, k$,*

$$\left| \frac{1}{m(z)} \sum_{k=0}^{m(z)-1} \varphi_i\big(f^k(x)\big) - \int_M \varphi_i \, d\nu \right| < \rho.$$

Theorem 15.9 has the following consequence.

COROLLARY 15.10. *If $\{f_n\}_{n\geqslant 1}$ is a sequence of $C^{1+\alpha}$ diffeomorphisms converging to f in the C^1 topology, then for each $n \geqslant 1$, f_n has a hyperbolic invariant probability measure ν_n such that $\{\nu_n\}_{n\geqslant 1}$ converges weakly to ν. Furthermore, ν_n may be chosen such that* $\operatorname{supp}\nu_n \subset \operatorname{Per}_h(f_n)$ *for each $n \geqslant 1$.*

An application of Corollary 15.10 to a constant sequence of diffeomorphisms yields the following result.

COROLLARY 15.11. *For a $C^{1+\alpha}$ diffeomorphism $f : M \to M$ of a compact smooth manifold one of the following mutually exclusive alternatives holds*:

1. *the measures supported on hyperbolic periodic points are weakly dense in the set of hyperbolic measures*;
2. *there are no hyperbolic measures* (*and hence, there are no hyperbolic periodic points*).

Corollaries 15.10 and 15.11 suggest a "weak stability" of hyperbolic measures.

The estimate (15.1) can be strengthened in the following way to become somewhat a multiplicative estimate.

THEOREM 15.12 (Ugarcovici [242]). *Assume that $h_\nu(f) > 0$. If ν is not a locally maximal ergodic measure in the class of f-invariant ergodic measures then there exist multiplicatively enough periodic orbits which are equidistributed with respect to ν. In other words, for any $r > 0$ and any collection of continuous functions $\varphi_1, \dots, \varphi_k$ there exist a sequence*

$n_m \to \infty$ and sets $P_{n_m} = P_{n_m}(r, \varphi_1, \dots, \varphi_k)$ of periodic orbits of period n_m such that for any $z \in P_{n_m}$,

$$\left| \frac{1}{n_m} \sum_{i=1}^{n_m} \varphi_i\big(f^l(z)\big) - \int \varphi_i \, d\nu \right| < r$$

and

$$\overline{\lim_{m\to\infty}} \frac{\operatorname{card} P_{n_m}}{e^{n_m h_\nu(f)}} \geqslant 1.$$

15.3. *Entropy, horseshoes, and periodic points*

Recall that a set Λ is a horseshoe for a diffeomorphism f if there exist s, k and sets $\Lambda_0, \dots, \Lambda_{k-1}$ such that $\Lambda = \Lambda_0 \cup \dots \cup \Lambda_{k-1}$, $f^k(\Lambda_i) = \Lambda_i$, $f(\Lambda_i) = \Lambda_{i+1 \bmod k}$, and $f^k|\Lambda_0$ is conjugate to a full shift in s symbols. For a horseshoe Λ we set

$$\chi(\Lambda) = \inf\{\chi_\nu \colon \operatorname{supp} \nu \text{ is a periodic orbit on } \Lambda\}.$$

THEOREM 15.13 (Katok and Mendoza [139]). *Assume that ν is ergodic and $h_\nu(f) > 0$. Then for any $\varepsilon > 0$ and any continuous functions $\varphi_1, \dots, \varphi_k$ on M, there exists a hyperbolic horseshoe Λ such that*:

1. $h(f|\Lambda) > h_\nu(f) - \varepsilon$;
2. *Λ is contained in an ε-neighborhood of* $\operatorname{supp} \nu$;
3. $\chi(\Lambda) > \chi_\nu - \varepsilon$;
4. *there exists a measure ν_0 supported on Λ such that for $i = 1, \dots, k$,*

$$\left| \int_M \varphi_i \, d\nu_0 - \int_M \varphi_i \, d\nu \right| < \delta.$$

We outline the proof of this result. Given $\ell \geqslant 1$, let ζ be a finite measurable partition of M refining the partition $\{\Lambda^\ell, M \setminus \Lambda^\ell\}$. Fix $r > 0$. For each $m \geqslant 1$, let Λ_m^ℓ be the set of points $x \in \Lambda^\ell$ such that $f^q(x) \in \zeta(x)$ for some $q \in [m, (1+r)m]$, and

$$\left| \frac{1}{s} \sum_{j=0}^{s-1} \varphi_i\big(f^j(x)\big) - \int_M \varphi \, d\nu \right| < \frac{r}{2}$$

for $s \geqslant m$ and $i = 1, \dots, k$. Using Birkhoff's Ergodic Theorem, one can show that $\nu(\Lambda_m^\ell) \to \nu(\Lambda^\ell)$ as $m \to \infty$. From now on we choose m such that $\nu(\Lambda_m^\ell) > \nu(\Lambda^\ell) - r$. Given $\delta > 0$, there exists a cover $\{R(x_1), \dots, R(x_t)\}$ of Λ^ℓ by closed rectangles (with $x_i \in \Lambda_m^\ell$) and numbers $\lambda \in (0, 1)$, satisfying $e^{-\chi_\nu - \delta} < \lambda < e^{-\chi_\nu + \delta}$, and $\gamma > 0$ such that

1. $\Lambda^\ell \subset \bigcup_{i=1}^t B(x_i, \delta)$, with $B(x_i, \delta) \subset \operatorname{int} R(x_i)$ for each i;
2. $\operatorname{diam} R(x_i) < r$ for each i;

3. if $x \in \Lambda^\ell \cap B(x_i, \delta)$ and $f^m(x) \in \Lambda^\ell \cap B(x_j, \delta)$ for some $m > 0$, then the connected component $\mathcal{C}(R(x_i) \cap f^{-m}(R(x_j)), x)$ of $R(x_i) \cap f^{-m}(R(x_j))$ containing x is an admissible (s, γ)-rectangle in $R(x_i)$ and $f^m(\mathcal{C}(R(x_i) \cap f^{-m}(R(x_j)), x))$ is an admissible (u, γ)-rectangle in $R(x_j)$;
4. for $k = 0, \dots, m$,

$$\operatorname{diam} f^k\big(\mathcal{C}\big(R(x_i) \cap f^{-m}\big(R(x_j)\big), x\big)\big) \leqslant 3 \operatorname{diam} R(x_i) \max\big\{\lambda^k, \lambda^{m-k}\big\}.$$

Here an *admissible* (s, γ)*-rectangle* is the set of points

$$\big\{(v, u) \in [-h, h]^2 \colon u = \theta\psi_1(v) + (1-\theta)\psi_2(v),\ 0 \leqslant \theta \leqslant 1\big\},$$

where ψ_1 and ψ_2 are two (s, γ)-curves (for some $h \leqslant 1$ and some appropriate parametrization in each Lyapunov chart; see Section 8.2). The definition of (u, γ)-rectangles is analogous. The cover can be easily obtained from the behavior of (s, γ)- and (u, γ)-curves under iteration by f, and by using Theorem 15.1 to establish the last property.

Let $E_m \subset \Lambda_m^\ell$ be an (m, ε)-separated set of maximal cardinality. By the Brin–Katok formula for the metric entropy, there exist infinitely many m such that $\operatorname{card} E_m \geqslant e^{m(h_\nu(f)-r)}$. For each $q \in [m, (1+r)m]$, let $V_q = \{x \in E_m \colon f^q(x) \in \zeta(x)\}$ and let n be the value of q that maximizes $\operatorname{card} V_q$. Since $e^{mr} > mr$ we have $\operatorname{card} V_n \geqslant e^{m(h_\nu(f)-3r)}$. Consider now the value j for which $\operatorname{card}(V_n \cap R(x_j))$ is maximal. Then

$$\operatorname{card}\big(V_n \cap R(x_j)\big) \geqslant \frac{1}{t} \operatorname{card} V_n \geqslant \frac{1}{t} e^{m(h_\nu(f)-3r)}. \tag{15.2}$$

Each point $x \in V_n \cap R(x_j)$ returns to the rectangle $R(x_j)$ in n iterations, and thus $\mathcal{C}(R(x_j) \cap f^n(R(x_j)), f^n(x))$ is an admissible (u, γ)-rectangle in $R(x_j)$ and $f^{-n}(\mathcal{C}(R(x_j) \cap f^n(R(x_j)), f^n(x)))$ is an admissible (s, γ)-rectangle in $R(x_j)$. This follows from the fact that $d(x_j, x) < \delta$ and $d(f^n(x), x_j) < \delta$, and from property 2 of the cover. If $y \in \mathcal{C}(R(x_j) \cap f^{-n}(R(x_j)), x)$ then by the last property of the cover, $d(f^i(x), f^i(y)) \leqslant 3r$ for $i = 0, \dots, n$. This implies that given a point $y \in \mathcal{C}(R(x_j) \cap f^{-n}(R(x_j)), x) \setminus \{x\}$, we must have $y \notin V_n$; otherwise it would contradict the separability of V_n. Hence, there exist $\operatorname{card} V_n$ disjoint admissible (s, γ)-rectangles mapped by f^n onto $\operatorname{card} V_n$ admissible (u, γ)-rectangles.

Let

$$\Lambda(m) = \bigcup_{l \in \mathbb{Z}} f^{nl}\Bigg(\bigcup_{x \in V_n \cap R(x_j)} \mathcal{C}\big(R(x_j) \cap f^{-n}\big(R(x_j)\big), x\big)\Bigg).$$

The map $f^n|\Lambda(m)$ is conjugate to the full shift on $\operatorname{card}(V_n \cap R(x_j))$ symbols. Now observe that for each $y \in \Lambda(m)$ its orbit remains in the union of the regular neighborhoods $R(x_j), \dots, R(f^n(x_j))$, and thus $f^n|\Lambda(m)$ is a hyperbolic horseshoe.

The entropy of $f^n|\Lambda(m)$ equals $\log \operatorname{card}(V_n \cap R(x_j))$. By (15.2),

$$h\big(f|\Lambda(m)\big) = \frac{1}{n} \log \operatorname{card}\big(V_n \cap R(x_j)\big) \geqslant \frac{1}{n} \log \frac{1}{t} e^{m(h_\nu(f)-3r)}.$$

Since $m/n > 1/(1+r)$, we obtain the desired properties.

The following are immediate consequences of Theorem 15.13.

COROLLARY 15.14. *Assume that ν is ergodic and $h_\nu(f) > 0$. There exists a sequence of f-invariant measures ν_n supported on hyperbolic horseshoes Λ_n such that*:

1. $\nu_n \to \nu$ *in the weak* topology*;
2. *if $h_\nu(f) > 0$ then $h_{\nu_n}(f) \to h_\nu(f)$.*

COROLLARY 15.15. *Assume that ν is ergodic and $h_\nu(f) > 0$. Given $\varepsilon > 0$,*

$$h_\nu(f) \leqslant \overline{\lim_{m\to\infty}} \frac{1}{m} \log^+ \operatorname{card}\bigl\{x \in M\colon\ f^m(x) = x \text{ and } \chi(x) \geqslant \chi(\nu) - \varepsilon\bigr\}.$$

In particular,

$$h(f) \leqslant \overline{\lim_{m\to\infty}} \frac{1}{m} \log^+ P_m(f).$$

In the two-dimensional case, any measure with positive entropy is hyperbolic (see Corollary 12.8). Therefore, Corollary 15.15 implies the following relation between periodic points and topological entropy.

COROLLARY 15.16. *For any $C^{1+\alpha}$ diffeomorphism f of a two-dimensional manifold,*

$$h(f) \leqslant \overline{\lim_{m\to\infty}} \frac{1}{m} \log^+ P_m(f). \tag{15.3}$$

In the multi-dimensional case, the inequality (15.3) does not hold for arbitrary diffeomorphisms.

The following result shows that hyperbolic measures persist under C^1 perturbations. This is a consequence of the structural stability of hyperbolic measures.

COROLLARY 15.17. *Assume that ν is ergodic and $h_\nu(f) > 0$. Given $C^{1+\alpha}$ diffeomorphisms f_n for each $n \geqslant 1$ such that f_n converges to f in the C^1 topology, there exist f_n-invariant ergodic measures ν_n satisfying the following properties*:

1. $\nu_n \to \nu$ *in the weak topology*;
2. $h_{\nu_n}(f_n) \to h_\nu(f)$;
3. $\chi_{\nu_n} \to \chi_\nu$.

15.4. *Continuity properties of entropy*

It follows from Theorem 15.13 that

$$h(f) = \sup\bigl\{h(f|\Lambda)\colon\ \Lambda \text{ is a hyperbolic horseshoe}\bigr\}.$$

The following two results describe continuity-like properties of topological and metric entropies on the space of diffeomorphisms. The first result deals with diffeomorphisms of class $C^{1+\alpha}$ and follows from Corollary 15.17 and the structural stability of horseshoes.

THEOREM 15.18. *The topological entropy on the space of $C^{1+\alpha}$ diffeomorphisms of a given surface is lower-semicontinuous.*

The second result deals with C^∞ diffeomorphisms.

THEOREM 15.19. *For a C^∞ map $f : M \to M$ of a compact manifold*

1. *the map $\mu \mapsto h_\mu(f)$ is upper-semicontinuous on the space of f-invariant probability measures on M;*
2. *the map $f \mapsto h(f)$ is upper-semicontinuous.*

The first statement is due to Newhouse [190] and the second one was established independently by Newhouse [190] and Yomdin [253]. We refer to [190] for references in the case of interval maps. It follows from Theorems 15.18 and 15.19 that the topological entropy is continuous for C^∞ diffeomorphisms of a given compact surface.

15.5. *Yomdin-type estimates and the entropy conjecture*

In [227, §V], Shub conjectured that for any C^1 map $f : M \to M$ of a compact manifold,

$$h(f) \geqslant \log \sigma(f_*), \tag{15.4}$$

where $f_* : H_*(M, \mathbb{R}) \to H_*(M, \mathbb{R})$ is the linear map induced by f on the total homology of M,

$$H_*(M, \mathbb{R}) = \bigoplus_{i=0}^{\dim M} H_i(M, \mathbb{R})$$

and

$$\sigma(f_*) = \lim_{n\to\infty} \|f_*^n\|^{1/n} = \max\{\sigma(f_{*i}) \colon i = 0, \ldots, \dim M\}$$

is the spectral radius of f_*. This is referred to as the *entropy conjecture*. For a $C^{1+\alpha}$ diffeomorphism f one could use (15.4) if available to establish positivity of the topological entropy and hence, existence of a measure with some positive Lyapunov exponents and the associated nontrivial stochastic behavior (see Section 15.2). We give here an account of the results in the direction of the conjecture (see also the survey by Katok [133] for the status of the conjecture prior to 1986).

In the case of the first homology $f_{*1} : H_1(M, \mathbb{R}) \to H_1(M, \mathbb{R})$ we have the following result for arbitrary continuous maps.

THEOREM 15.20 (Manning [177]). *If f is a continuous map of a smooth compact manifold then $h(f) \geqslant \log \sigma(f_{*1})$.*

There exists a stronger version of Theorem 15.20 due to Katok [133] with the number $\log \sigma(f_{*1})$ replaced by the so-called algebraic entropy of the action induced by f on the (not necessarily commutative) fundamental group $\pi_1(M)$. It follows from Theorem 15.20 and Poincaré duality that the entropy conjecture holds for any homeomorphism of a manifold M with $\dim M \leqslant 3$ (see [177]).

In the case of the top homology group the following result holds (recall that $f_{*\dim M}$ is the same as multiplication by the degree $\deg f$).

THEOREM 15.21 (Misiurewicz and Przytycki [187]). *If f is a C^1 map of a compact smooth manifold, then $h(f) \geqslant \log|\deg f|$.*

In particular, this implies that the entropy conjecture holds for any smooth map of a sphere (in any dimension) and any smooth map of a compact manifold with dimension at most 2.

On some manifolds the entropy conjecture turns out to hold for arbitrary continuous maps.

THEOREM 15.22 (Misiurewicz and Przytycki [186]). *The entropy conjecture holds for any continuous map of a torus* (*in any dimension*).

Since any Anosov automorphism of the torus is topologically conjugate to an algebraic automorphism (see [137, Theorem 18.6.1]), we conclude that if f is an Anosov diffeomorphism of a torus, then $h(f) = \log \sigma(f_*)$.

Shub formulated the entropy conjecture in connection with the problem of defining the simplest diffeomorphisms in each isotopy class of diffeomorphisms. From this point of view, it is important to discuss the entropy conjecture for example for structurally stable diffeomorphisms. In [228], Shub and Sullivan described an open and dense subset (in the C^0-topology) of the set of structurally stable diffeomorphisms for which the entropy conjecture holds. Later Shub and Williams obtained a more general result which does no require the nonwandering set to have zero dimension.

THEOREM 15.23 (Shub and Williams [230]). *The entropy conjecture holds for any axiom A no-cycles diffeomorphism.*

More recently Yomdin established the C^∞ version of the entropy conjecture with an approach using semialgebraic geometry.

THEOREM 15.24 (Yomdin [253,254]). *The entropy conjecture holds for any C^∞ map of a compact manifold.*

Yomdin also proved more generally that for a C^k map $f: M \to M$ of a compact manifold, with $1 \leqslant k \leqslant \infty$, and $j = 0, \dots, \dim M$,

$$h(f) + \frac{j}{k} \lim_{n\to\infty} \frac{1}{n} \log \max_{x\in M} \|d_x f^n\| \geqslant v_j(f) \geqslant \log \sigma(f_{*j}).$$

Here $v_j(f)$ is the exponential growth rate of j-volumes,

$$v_j(f) = \sup \overline{\lim_{n\to\infty}} \frac{1}{n} \log \operatorname{vol}\bigl(f^n(A)\bigr),$$

where the supremum is taken over all submanifolds $A \subset M$ of dimension j, and the volume is counted with multiplicities. Newhouse proved earlier in [189] that $h(f) \leqslant \max_j v_j(f)$ for a $C^{1+\alpha}$ map of a compact manifold. In particular, we have $h(f) = \max_j v_j(f) = \log \sigma(f_*)$ for a C^∞ map.

16. Hyperbolic measures II: Entropy and dimension

16.1. *Entropy formula*

We describe results of Ledrappier and Young [159,160] including the general formula for the entropy of a diffeomorphism. Let f be a C^2 diffeomorphism of a compact smooth Riemannian manifold M preserving a Borel measure on M. For a regular point $x \in M$ and $i = 1, \dots, u(x) = \max\{i\colon \lambda_i(x) > 0\}$, consider the ith-unstable global manifold $W_i(x)$ of f at x (see Section 9.2). We introduce the notion of the entropy "along" the W_i-foliation.

For $n > 0$, and $\varepsilon > 0$ set

$$V_i(x, n, \varepsilon) = \bigl\{y \in W_i(x)\colon \rho_{W_i}\bigl(f^k(x), f^k(y)\bigr) < \varepsilon \text{ for } 0 \leqslant k < n\bigr\}.$$

Consider a measurable partition ξ of M. We say that ξ is *subordinate* to the W_i-foliation if for ν-almost every $x \in M$ we have $\xi(x) \subset W_i(x)$ and $\xi(x)$ contains an open neighborhood of x in the topology of $W_i(x)$. Let $\{\nu_i(x)\}$ be the system of conditional measures associated with ξ. Define

$$\underline{h}_i(x, \xi) = \lim_{\varepsilon\to 0} \underline{\lim_{n\to\infty}} -\frac{1}{n} \log \nu_i(x)\bigl(V_i(x, n, \varepsilon)\bigr),$$

$$\bar{h}_i(x, \xi) = \lim_{\varepsilon\to 0} \overline{\lim_{n\to\infty}} -\frac{1}{n} \log \nu_i(x)\bigl(V_i(x, n, \varepsilon)\bigr).$$

THEOREM 16.1 [160]. *The following properties hold*:

1. $h_i(x) := \underline{h}_i(x, \xi) = \bar{h}_i(x, \xi)$ *for* ν*-almost every* $x \in M$*, independently of the choice of the partition* ξ;
2. $\int_M h_{u(x)}(x)\, d\nu(x) = h_\nu(f)$.

The number $h_i(x)$ is called the *local entropy* of f at x *along the* W_i*-foliation*.

We also consider the pointwise dimension of conditional measures "along" the W^i-foliation. Let $B_i(x,r)$ be the ball in $W_i(x)$ centered at x of radius ε and ξ a measurable partition subordinate to W_i. For a regular point $x \in M$ define

$$\underline{d}^i_\nu(x,\xi) = \varliminf_{r\to 0} \frac{\log \nu_i(x)(B_i(x,r))}{\log r}, \qquad \bar{d}^i_\nu(x,\xi) = \varlimsup_{r\to 0} \frac{\log \nu_i(x)(B_i(x,r))}{\log r}.$$

THEOREM 16.2 [160]. *The following properties hold*:

1. $d^i_\nu(x) := \underline{d}^i_\nu(x,\xi) = \bar{d}^i_\nu(x,\xi)$ *for* ν*-almost every* $x \in M$*, independently of the choice of the partition* ξ;
2. $0 \leqslant d^i_\nu(x) - d^{i-1}_\nu(x) \leqslant \dim E_i(x)$ *for* $2 \leqslant i \leqslant u(x)$.

The number $d^i_\nu(x)$ is called the *pointwise dimension* of ν *along the* W_i*-foliation*. The number $d^i_\nu(x) - d^{i-1}_\nu(x)$ can be interpreted as a "transverse dimension" of ν on the quotient W_i / W_{i-1} (recall that each leaf of W_i is foliated by leaves of W_{i-1}).

In the particular cases $i = s$ and $i = u$ the quantities

$$d^s_\nu(x) \stackrel{\text{def}}{=} \lim_{r\to 0} \frac{\log \nu^s_x(B^s(x,r))}{\log r}, \qquad d^u_\nu(x) \stackrel{\text{def}}{=} \lim_{r\to 0} \frac{\log \nu^u_x(B^u(x,r))}{\log r}$$

are called *stable* and *unstable local* (*pointwise*) *dimensions* of ν. They are well defined for almost every $x \in M$ and are constant almost everywhere; we denote these constants by d^s_ν and d^u_ν. Set also $d^0_\nu(x) = 0$.

THEOREM 16.3 [160]. *The metric entropy of a* C^2 *diffeomorphism* f *is expressed by the following formula*:

$$h_\nu(f) = \int_M \sum_{i=1}^{u(x)} \lambda_i(x)\big(d^i_\nu(x) - d^{i-1}_\nu(x)\big)\, d\nu(x).$$

The proof goes by showing that for ν-almost every $x \in M$ and $i = 2, \dots, u(x)$,

$$h_1(x) = \lambda_1(x) d^1_\nu(x), \qquad h_i(x) - h_{i-1}(x) = \lambda_i(x)\big(d^i_\nu(x) - d^{i-1}_\nu(x)\big).$$

To prove this Ledrappier and Young constructed a special countable partition $\mathcal{P}$ of M of finite entropy related to the Pinsker partition (see Theorem 11.16). Given integers $k, \ell \in \mathbb{N}$ we also consider the partition $\mathcal{P}^l_k = \bigvee_{n=-k}^{l} f^{-n}\mathcal{P}$.

THEOREM 16.4 [159,160]. *Let* ν *be ergodic. Given* $0 < \varepsilon < 1$*, there exists a set* $\Gamma \subset M$ *of measure* $\nu(\Gamma) > 1 - \varepsilon/2$*, an integer* $n_0 \geqslant 1$*, and a number* $C > 1$ *such that for every* $x \in \Gamma$ *and any integer* $n \geqslant n_0$*, the following statements hold*:

1. *for all integers* $k, l \geqslant 1$,

$$C^{-1} e^{-(l+k)h-(l+k)\varepsilon} \leqslant \nu\big(\mathcal{P}^l_k(x)\big) \leqslant C e^{-(l+k)h+(l+k)\varepsilon}, \tag{16.1}$$

$$C^{-1}e^{-kh-k\varepsilon} \leqslant \nu_x^s\big(\mathcal{P}_k^0(x)\big) \leqslant Ce^{-kh+k\varepsilon}, \tag{16.2}$$

$$C^{-1}e^{-lh-l\varepsilon} \leqslant \nu_x^u\big(\mathcal{P}_0^l(x)\big) \leqslant Ce^{-lh+l\varepsilon}, \tag{16.3}$$

where $h = h_\nu(f)$;

2. $$\xi^s(x) \cap \bigcap_{n\geqslant 0} \mathcal{P}_0^n(x) \supset B^s(x, e^{-n_0}),$$

$$\xi^u(x) \cap \bigcap_{n\geqslant 0} \mathcal{P}_n^0(x) \supset B^u(x, e^{-n_0}); \tag{16.4}$$

3. $$e^{-d^s n-n\varepsilon} \leqslant \nu_x^s\big(B^s(x, e^{-n})\big) \leqslant e^{-d^s n+n\varepsilon}, \tag{16.5}$$

$$e^{-d^u n-n\varepsilon} \leqslant \nu_x^u\big(B^u(x, e^{-n})\big) \leqslant e^{-d^u n+n\varepsilon}; \tag{16.6}$$

4. $$\mathcal{P}_{an}^{an}(x) \subset B(x, e^{-n}) \subset \mathcal{P}(x), \tag{16.7}$$

$$\mathcal{P}_{an}^0(x) \cap \xi^s(x) \subset B^s(x, e^{-n}) \subset \mathcal{P}(x) \cap \xi^s(x), \tag{16.8}$$

$$\mathcal{P}_0^{an}(x) \cap \xi^u(x) \subset B^u(x, e^{-n}) \subset \mathcal{P}(x) \cap \xi^u(x), \tag{16.9}$$

where a *is the integer part of* $2(1+\varepsilon)\max\{\lambda_1, -\lambda_p, 1\}$;

5. *if* $Q_n(x)$ *is defined by*

$$Q_n(x) = \bigcup \mathcal{P}_{an}^{an}(y)$$

where the union is taken over $y \in \Gamma$ *for which*

$$\mathcal{P}_0^{an}(y) \cap B^u(x, 2e^{-n}) \neq \emptyset \quad \textit{and} \quad \mathcal{P}_{an}^0(y) \cap B^s(x, 2e^{-n}) \neq \emptyset;$$

then

$$B(x, e^{-n}) \cap \Gamma \subset Q_n(x) \subset B(x, 4e^{-n}), \tag{16.10}$$

$$B^s(x, e^{-n}) \cap \Gamma \subset Q_n(x) \cap \xi^s(x) \subset B^s(x, 4e^{-n}), \tag{16.11}$$

$$B^u(x, e^{-n}) \cap \Gamma \subset Q_n(x) \cap \xi^u(x) \subset B^u(x, 4e^{-n}). \tag{16.12}$$

We outline the construction of the partition $\mathcal{P}$, and discuss its relation to the Pinsker partition (compare with Theorem 11.16). We proceed in a manner similar to that in Section 11.4. Consider a regular set Λ^ℓ with $\nu(\Lambda^\ell) > 0$. For a sufficiently small $r = r(\ell) > 0$ and $x \in \Lambda^\ell$, set

$$P^\ell(x) = \bigcup_{y\in\Lambda^\ell\cap B(x,r)} V^u(y), \qquad Q(x) = \bigcup_{n=-\infty}^{\infty} f^n\big(P^\ell(x)\big).$$

Since ν is ergodic the set $Q(x)$ has full ν-measure. Let ξ be the partition of $Q(x)$ by local unstable manifolds $V^u(y)$, $y \in \Lambda^\ell \cap B(x, r)$, and the element $Q(x) \setminus P^\ell(x)$. Then

$\xi^+ = \bigvee_{i\geqslant 0} f^i\xi$ is the Pinsker partition subordinate to the partition into global unstable manifolds.

Let now $\Lambda \subset M$ be the set of regular points, and $\Psi_x : B(0, q(x)) \to M$ a family of Lyapunov charts for $x \in \Lambda$ (see Theorem 8.14). Fix $\delta > 0$ and consider a partition $\mathcal{P}$ of finite entropy satisfying:

1. $\mathcal{P}$ is "adapted" to the Lyapunov charts in the sense that the elements of the partition $\mathcal{P}^+ = \bigvee_{n=0}^{\infty} f^n\mathcal{P}$ satisfy for each $x \in \Lambda$,
$$\mathcal{P}^+(x) \subset \Psi_x\big(\big\{y \in B\big(0, q(x)\big)\colon \big\|\big(\Psi^{-1}_{f^{-n}(x)} \circ f^{-n} \circ \Psi_x\big)(y)\big\| \leqslant \delta q\big(f^{-n}(x)\big)\big\}\big);$$
2. $h_\nu(f, \mathcal{P}) \geqslant h_\nu(f) - \varepsilon$;
3. the partition $\mathcal{P}$ refines $\{E, M \setminus E\}$ for some measurable set E of positive measure such that there exists a transversal T to W^u with the following property: if an element $C \in \xi^+$ intersects E, then T intersects C in exactly one point.

Ledrappier and Young [159] have shown that a certain partition constructed by Mañé in [172] possesses property 1. Property 3 is related to the construction of "transverse metrics" to ξ^+. Namely, consider the partitions $\eta_1 = \xi^+ \vee \mathcal{P}^+$ and $\eta_2 = \mathcal{P}^+$. One can construct a special metric on $\eta_2(x)/\eta_1$ for every $x \in \bigcup_{n\geqslant 0} f^n(E)$.

One can obtain the inclusions (16.4) from the fact that the partition $\mathcal{P}$ is adapted to the Lyapunov charts. Since the Lyapunov exponents at almost every point are constant, (16.7), (16.8), and (16.9) follow from (16.4) and an appropriate choice of a. The inequalities (16.5) and (16.6) are easy consequences of existence of the stable and unstable pointwise dimensions d^s_ν and d^u_ν (see Theorem 16.7). The inclusions (16.10) are based upon the continuous dependence of stable and unstable manifolds in the $C^{1+\alpha}$ topology on the base point in each regular set. The inclusions in (16.11) and (16.12) follow readily from (16.10).

Property (16.1) is an immediate corollary of Shannon–McMillan–Breiman's Theorem applied to the partition $\mathcal{P}$. Properties (16.2) and (16.3) follow from "leaf-wise" versions of this theorem. More precisely, Ledrappier and Young have shown (see [160, Lemma 9.3.1] and [159, Proposition 5.1]) that for ν-almost every x,

$$\lim_{n\to\infty} -\frac{1}{n} \log \nu^u_x\big(\mathcal{Q}^n_0(x)\big) = h_\nu(f),$$

where $\mathcal{Q}$ is any partition of finite entropy. Since $\mathcal{P}^n_0(x) \supset (\xi^+ \cap \mathcal{P})^n_0(x)$, we conclude that

$$\overline{\lim_{n\to\infty}} -\frac{1}{n} \log \nu^u_x\big(\mathcal{P}^n_0(x)\big) \leqslant h_\nu(f) \tag{16.13}$$

for ν-almost every x. Moreover, using the fact that $\mathcal{P}$ is adapted to the Lyapunov charts one can show that the partition $\mathcal{P}$ additionally possesses the property that given $\delta > 0$ there exists $n_0 \geqslant 0$ such that $\mathcal{P}^n_0(x) \cap \xi^u(x) \subset V_{u(x)}(x, n, \delta)$ for ν-almost every x and every $n \geqslant n_0$ (see [160, Lemma 9.3.3]). It follows from Theorem 16.1 that for ν-almost every x,

$$\begin{aligned}\overline{\lim_{n\to\infty}} -\frac{1}{n} \log \nu^u_x\big(\mathcal{P}^n_0(x)\big) &\geqslant \lim_{\delta\to 0} \overline{\lim_{n\to\infty}} -\frac{1}{n} \log \nu^u_x\big(V_{u(x)}(x, n, \delta)\big)\\ &= h_\nu(f).\end{aligned} \tag{16.14}$$

Putting together (16.13) and (16.14) we obtain (16.3). A similar argument can be used to obtain (16.2).

Note that the Margulis–Ruelle inequality is an immediate corollary of Theorem 16.3 and so is the fact that any measure ν with absolutely continuous conditional measures on unstable manifolds satisfies Pesin's entropy formula.

16.2. *Dimension of measures. Local dimension*

For a Borel measure ν on a complete metric space X define the *Hausdorff dimension* $\dim_H \nu$, and *lower* and *upper box dimensions*, $\underline{\dim}_B \nu$, and $\overline{\dim}_B \nu$ by

$$\begin{aligned}
\dim_H \nu &= \inf\{\dim_H Z\colon\ \nu(Z) = 1\},\\
\underline{\dim}_B \nu &= \lim_{\delta\to 0} \inf\{\underline{\dim}_B Z\colon\ \nu(Z) \geqslant 1-\delta\},\\
\overline{\dim}_B \nu &= \lim_{\delta\to 0} \inf\{\overline{\dim}_B Z\colon\ \nu(Z) \geqslant 1-\delta\},
\end{aligned}$$

where $\dim_H Z$, $\underline{\dim}_B Z$ and $\overline{\dim}_B Z$ are respectively the Hausdorff dimension, lower and upper box dimensions of the set Z. It follows from the definition that

$$\dim_H \nu \leqslant \underline{\dim}_B \nu \leqslant \overline{\dim}_B \nu.$$

Another important characteristic of dimension type of ν is its *information dimension*. Given a partition ξ of X, define the *entropy of ξ with respect to ν* by

$$H_\nu(\xi) = -\sum_{C_\xi} \nu(C_\xi)\log \nu(C_\xi),$$

where C_ξ is an element of the partition ξ. Given a number $\varepsilon > 0$, set

$$H_\nu(\varepsilon) = \inf\{H_\nu(\xi)\colon\ \operatorname{diam}\xi \leqslant \varepsilon\},$$

where $\operatorname{diam}\xi = \max \operatorname{diam} C_\xi$. We define the *lower* and *upper information dimensions of ν* by

$$\underline{I}(\nu) = \underline{\lim}_{\varepsilon\to 0} \frac{H_\nu(\varepsilon)}{\log(1/\varepsilon)}, \qquad \bar{I}(\nu) = \overline{\lim}_{\varepsilon\to 0} \frac{H_\nu(\varepsilon)}{\log(1/\varepsilon)}.$$

Young established a powerful criterion that guarantees the coincidence of the Hausdorff dimension and lower and upper box dimensions of measures as well as their lower and upper information dimensions. Define the *local* (*pointwise*) *dimension* of ν by

$$d_\nu(x) = \lim_{r\to 0} \frac{\log \nu(B(x,r))}{\log r}, \tag{16.15}$$

where $B(x,r)$ is the ball centered at x of radius r (provided the limit exists). It was introduced by Young in [255] and characterizes the local geometrical structure of ν with respect to the metric in X. If the limit in (16.15) does not exist we consider the lower and upper limits and introduce respectively *the lower* and *upper local* (*pointwise*) *dimensions* of ν at x and we denote them by $\underline{d}_\nu(x)$ and $\bar{d}_\nu(x)$.

THEOREM 16.5 (Young [255]). *Let X be a compact metric space of finite topological dimension and ν a Borel probability measure on X. Assume that*

$$\underline{d}_\nu(x) = \bar{d}_\nu(x) = d_\nu \tag{16.16}$$

for ν-almost every $x \in X$. Then

$$\dim_H \nu = \underline{\dim}_B \nu = \overline{\dim}_B \nu = \underline{I}(\nu) = \bar{I}(\nu) = d_\nu.$$

A measure ν satisfying (16.16) is called *exact dimensional*.

We will discuss the problem of existence of the limit in (16.15) for hyperbolic invariant measures. This problem is often referred to as the Eckmann–Ruelle conjecture. Its affirmative solution was obtained by Barreira, Pesin and Schmeling in [36].

THEOREM 16.6. *Let f be a $C^{1+\alpha}$ diffeomorphism of a smooth Riemannian manifold M without boundary, and ν an f-invariant compactly supported hyperbolic ergodic Borel probability measure. Then ν is exact dimensional and*

$$d_\nu = d_\nu^s + d_\nu^u.$$

In general, when the measure ν is not ergodic the stable and unstable local dimensions as well as the local dimension itself depend on the point x. In this case one can prove that for ν-almost every $x \in M$,

$$d_\nu(x) = d_\nu^s(x) + d_\nu^u(x).$$

Let us comment on the proof of Theorem 16.6. The upper bound for the pointwise dimension of *any* Borel f-invariant measure ν was obtained by Ledrappier and Young in [160].

THEOREM 16.7. *Let f be a C^2 diffeomorphism of M. For ν-almost every $x \in M$,*

$$\bar{d}_\nu \leqslant d_\nu^s + d_\nu^u + \dim E^c(x).$$

In the case when the measure ν is hyperbolic (i.e., $\dim E^c(x) = 0$ for ν-almost every $x \in M$) this result can be extended to $C^{1+\alpha}$ diffeomorphisms (not necessarily C^2). Barreira, Pesin and Schmeling [36] have shown that

$$\bar{d}_\nu \leqslant d_\nu^s + d_\nu^u.$$

The lower bound for the pointwise dimension, $\underline{d}_\nu \geqslant d^s_\nu + d^u_\nu$, is an immediate corollary of Theorem 16.9.

Young proved Theorem 16.6 for surface diffeomorphisms.

THEOREM 16.8 (Young [255]). *Let f be a $C^{1+\alpha}$ diffeomorphism of a smooth compact surface M and ν a hyperbolic ergodic measure with Lyapunov exponents $\lambda^1_\nu > 0 > \lambda^2_\nu$. Then*

$$\underline{d}_\nu = \bar{d}_\nu = h_\nu(f)\left(\frac{1}{\lambda^1_\nu} - \frac{1}{\lambda^2_\nu}\right).$$

Let us point out that neither of the assumptions of Theorem 16.6 can be omitted. Pesin and Weiss presented an example of a Hölder homeomorphism with Hölder constant arbitrarily close to 1 whose ergodic measure of maximal entropy is not exact dimensional (see [203]). Ledrappier and Misiurewicz [157] constructed an example of a smooth map of a circle preserving an ergodic measure with zero Lyapunov exponent which is not exact dimensional. Kalinin and Sadovskaya [131] strengthened this result by showing that for a residual set of circle diffeomorphisms with irrational rotation number the unique invariant measure has lower pointwise dimension 0 and upper pointwise dimension 1 for almost every point in S^1.

16.3. *Local product structure of hyperbolic measures*

The following principle result establishes a crucial property of hyperbolic measures: these measures have asymptotically "almost" local product structure.

THEOREM 16.9 (Barreira, Pesin and Schmeling [36]). *Let f be a $C^{1+\alpha}$ diffeomorphism of a smooth Riemannian manifold M without boundary, and ν an f-invariant compactly supported hyperbolic ergodic Borel probability measure. Then for every $\delta > 0$ there exist a set $\Lambda \subset M$ with $\nu(\Lambda) > 1 - \delta$ such that for every $x \in \Lambda$ and every sufficiently small r (depending on x), we have*

$$r^\delta \nu^s_x\big(B^s(x,r)\big)\nu^u_x\big(B^u(x,r)\big) \leqslant \nu\big(B(x,r)\big) \leqslant r^{-\delta}\nu^s_x\big(B^s(x,r)\big)\nu^u_x\big(B^u(x,r)\big).$$

The proof of Theorem 16.9 uses the crucial Markov property of the special countable partition $\mathcal{P}$ of M constructed in Theorem 16.4.

THEOREM 16.10 (Barreira, Pesin and Schmeling [36]). *For every $x \in \Gamma$ and $n \geqslant n_0$,*

$$\mathcal{P}^{an}_{an}(x) \cap \xi^s(x) = \mathcal{P}^0_{an}(x) \cap \xi^s(x);$$
$$\mathcal{P}^{an}_{an}(x) \cap \xi^u(x) = \mathcal{P}^{an}_0(x) \cap \xi^u(x).$$

Note that any SRB-measure possesses a stronger property of local product structure and so does any Gibbs measure on a locally maximal hyperbolic set.

We emphasize that Theorem 16.9 is not trivial even for measures supported on locally maximal uniformly hyperbolic sets. In this situation the stable and unstable foliations need not be Lipschitz (in fact, they are "generically" not Lipschitz), and in general, the measure need not have a local product structure despite the fact that the set itself does.

Let us illustrate Theorems 16.6 and 16.9 by considering the full shift σ on the space Σ_p of two-sided infinite sequences of numbers in $\{1, \dots, p\}$. This space is endowed with the usual "symbolic" metric d_β, for each fixed number $\beta > 1$, defined as follows:

$$d_\beta\big(\omega^1, \omega^2\big) = \sum_{i \in \mathbb{Z}} \beta^{-|i|} \big|\omega_i^1 - \omega_i^2\big|,$$

where $\omega^1 = (\omega_i^1)$ and $\omega^2 = (\omega_i^2)$.

Let ν be a σ-invariant ergodic measure on Σ_p. By Shannon–McMillan–Breiman's Theorem, for ν-almost every $\omega \in \Sigma_p$,

$$\lim_{n\to\infty} -\frac{1}{2n+1} \log \nu\big(C_n(\omega)\big) = h_\nu(\sigma), \tag{16.17}$$

were $C_n(\omega)$ is the cylinder at ω of "size" n. Since $C_n(\omega)$ is the ball in the symbolic metric centered at ω of radius β^n, the quantity in the right-hand side in (16.17) is the local dimension of ν at ω. Thus, Shannon–McMillan–Breiman's Theorem claims that the local dimension of ν is almost everywhere constant and that the common value is the measure-theoretical entropy of ν.

Further, fix $\omega = (\omega_i) \in \Sigma_p$. The cylinder $C_n(\omega)$ can be identified with the direct product $C_n^+(\omega) \times C_n^-(\omega)$ where

$$C_n^+(\omega) = \big\{\bar\omega = (\bar\omega_i)\colon\ \bar\omega_i = \omega_i \text{ for } i = 0, \dots, n\big\}$$

and

$$C_n^-(\omega) = \big\{\bar\omega = (\bar\omega_i)\colon\ \bar\omega_i = \omega_i \text{ for } i = -n, \dots, 0\big\}$$

are the "positive" and "negative" cylinders at ω of "size" n. Define measures

$$\nu_n^+(\omega) = \nu | C_n^+(\omega) \quad \text{and} \quad \nu_n^-(\omega) = \nu | C_n^-(\omega).$$

It follows from Theorem 16.9 that for every $\delta > 0$ there exist a set $\Lambda \subset \Sigma_p$ with $\nu(\Lambda) > 1 - \delta$ and an integer $m \geqslant 1$ such that for every $\omega \in \Lambda$ and every sufficiently large n (depending on ω), we have

$$\beta^{-\delta|n|} \nu_{n+m}^+(\omega) \times \nu_{n+m}^-(\omega) \leqslant \nu | C_n(\omega) \leqslant \beta^{\delta|n|} \nu_{n-m}^+(\omega) \times \nu_{n-m}^-(\omega).$$

17. Geodesic flows on manifolds without conjugate points

For a long time geodesic flows have played an important stimulating role in developing the hyperbolic theory. Already in the beginning of the 20th century Hadamard and Morse,

while studying the statistics of geodesics on surfaces of negative curvature, observed that the local instability of trajectories is the prime reason for the geodesic flow to be ergodic and topologically transitive. The further study of geodesic flows has led researchers to introduce different classes of hyperbolic dynamical systems (Anosov systems, uniformly partially hyperbolic systems, and nonuniformly hyperbolic systems). On the other hand, geodesic flows always were one of the main areas for applying new advanced methods of the hyperbolic theory of dynamical systems. This in particular, has led to some new interesting results in differential and Riemannian geometry.

17.1. *Ergodic properties of geodesic flows*

Consider the geodesic flow g_t on a compact smooth Riemannian p-dimensional manifold M without conjugate points. The flow preserves the Liouville measure μ on the tangent bundle. Let the set Δ be given by (2.10). We assume that Δ is of positive Liouville measure. By Theorem 2.3 the geodesic flow is nonuniformly hyperbolic on Δ and hence, the results of Section 11.2 apply and show that ergodic components of $g_t|\Delta$ are of positive Liouville measure (see Theorem 11.3). Indeed, under some mild geometric assumptions the geodesic flow on Δ is ergodic. To see this we will first observe that every ergodic component of positive measure is open (mod 0) and then will use a remarkable result by Eberlein on topological transitivity of geodesic flows.

To establish local ergodicity of $g_t|\Delta$ we shall describe two invariant foliations (known as the stable and unstable horospherical foliations) of SM, W^- and W^+, such that $W^s(x) = W^-(x)$ and $W^u(x) = W^+(x)$ for almost every $x \in \Delta$.

We denote by H the universal Riemannian cover of M, i.e., a simply connected p-dimensional complete Riemannian manifold for which $M = H/\Gamma$ where Γ is a discrete subgroup of the group of isometries of H, isomorphic to $\pi_1(M)$. According to the Hadamard–Cartan theorem, any two points $x, y \in H$ are joined by a single geodesic which we denote by γ_{xy}. For any $x \in H$, the exponential map $\exp_x : \mathbb{R}^p \to H$ is a diffeomorphism. Hence, the map

$$\varphi_x(y) = \exp_x\left(\frac{y}{1 - \|y\|}\right) \tag{17.1}$$

is a homeomorphism of the open p-dimensional unit disk D onto H.

Two geodesics $\gamma_1(t)$ and $\gamma_2(t)$ in H are said to be *asymptotic* if

$$\sup_{t>0} \rho\big(\gamma_1(t), \gamma_2(t)\big) < \infty.$$

The asymptoticity is an equivalence relation, and the equivalence class $\gamma(\infty)$ corresponding to a geodesic γ is called a *point at infinity*. The set of these classes is denoted by $H(\infty)$ and is called the *ideal boundary* of H. Using (17.1) one can extend the topology of the space H to $\bar{H} = H \cup H(\infty)$ so that $\bar{H}$ becomes a compact space.

The map φ_x can be extended to a homeomorphism (still denoted by φ_x) of the closed p-dimensional disk $\bar{D} = D \cup S^{p-1}$ onto $\bar{H}$ by the equality

$$\varphi_x(y) = \gamma_y(+\infty), \quad y \in S^{p-1}.$$

In particular, φ_x maps S^{p-1} homeomorphically onto $H(\infty)$.

For any two distinct points x and y on the ideal boundary there is a geodesic which joins them. This geodesic is uniquely defined if the Riemannian metric is of strictly negative curvature (i.e., if inequality (2.7) is strict). Otherwise, there may exist a pair of distinct points x, $y \in H(\infty)$ which can be joined by more than one geodesic. If the manifold has no focal points there exists a geodesic isometric embedding into H of an infinite strip of zero curvature which consists of geodesics joining x and y. This statement is know as the *flat strip theorem*.

The fundamental group $\pi_1(M)$ of the manifold M acts on the universal cover H by isometries. This action can be extended to the ideal boundary $H(\infty)$. Namely, if $p = \gamma_v(+\infty) \in H(\infty)$ and $\zeta \in \pi_1(M)$, then $\zeta(p)$ is the equivalent class of geodesics which are asymptotic to the geodesic $\zeta(\gamma_v(t))$.

We now describe the invariant foliations for the geodesic flow.

Fix a point $x \in H$ and a vector $v \in SH$. Consider a sequence of vectors $v_n \in SH$, $v_n \to v$, a sequence of points $x_n \in H$, $x_n \to x$ and a sequence of numbers $t_n \to \infty$. Denote by γ_n the geodesic joining the points x_n and $\gamma_{v_n}(t_n)$. Since the sequence of vectors $\dot{\gamma}_n(0)$ is compact the sequence of geodesics has a limit geodesic. Following [200] we say that the manifold M satisfies the *asymptoticity axiom* if for any choice of $x_n, x \in H$, $v_n, v \in SH$, $x_n \to x$, $v_n \to v$ and $t_n \to \infty$ any limit geodesic of the sequence of geodesic γ_n is asymptotic to the geodesic γ.

If the manifold M satisfies the asymptoticity axiom then the sequence γ_n, indeed, converges to γ. Moreover, given a geodesic γ and a point $x \in H$, there exists a unique geodesic γ' passing through x and asymptotic to γ.

PROPOSITION 17.1 (Pesin [200]). *If the manifold M has no focal points then it satisfies the asymptoticity axiom.*

We consider the distributions E^- and E^+ introduced by (2.8) and (2.9).

PROPOSITION 17.2 (Pesin [200]). *Assume that the manifold M satisfies the asymptoticity axiom. Then the distributions E^- and E^+ are integrable. Their integral manifolds form continuous foliations of SM with smooth leaves. These foliations are invariant under the geodesic flow.*

Denote by W^- and W^+ the foliations of SM corresponding to the invariant distributions E^- and E^+. These foliations can be lifted from SM to SH and we denote these lifts by $\tilde{W}^-$ and $\tilde{W}^+$, respectively.

Given $x \in H$ and $p \in H(\infty)$, set

$$L(x, p) = \pi\left(\tilde{W}^-(v)\right),$$

where $x = \pi(v)$ and $p = \gamma_v(\infty)$. The set $L(x, p)$ is called the *horosphere* through x centered at p.

We summarize the properties of the foliations and horospheres in the following statement.

PROPOSITION 17.3. *The following statements hold*:

1. *for any $x \in H$ and $p \in H(\infty)$ there exists a unique horosphere $L(x, p)$ centered at p which passes through x; it is a limit in the C^1 topology of spheres $S^p(\gamma(t), t)$ as $t \to +\infty$ where γ is the unique geodesic joining x and p;*
2. *the leaf $W^-(v)$ is the framing of the horosphere $L(x, p)$ ($x = \pi(v)$ and $p = \gamma_v(+\infty)$) by orthonormal vectors which have the same direction as the vector v (i.e., they are "inside" the horosphere). The leaf $W^+(v)$ is the framing of the horosphere $L(x, p)$ ($x = \pi(v)$ and $p = \gamma_v(-\infty) = \gamma_{-v}(+\infty)$) by orthonormal vectors which have the same direction as the vector v (i.e., they are "outside" the horosphere);*
3. *for every $\zeta \in \pi_1(M)$,*

$$\zeta\big(L(x, p)\big) = L\big(\zeta(x), \zeta(p)\big),$$
$$d_v\zeta\, \tilde{W}^-(v) = \tilde{W}^-(d_v\zeta\, v), \qquad d_v\zeta\, \tilde{W}^+(v) = \tilde{W}^+(d_v\zeta\, v);$$

4. *for every $v, w \in SH$, for which $\gamma_v(+\infty) = \gamma_w(+\infty) = p$, the geodesic $\gamma_w(t)$ intersects the horosphere $L(\pi(v), p)$ at some point.*

THEOREM 17.4 (Pesin [200]). *Assume that the manifold M satisfies the asymptoticity axiom. Assume also that the set Δ has positive Liouville measure. Then for almost every $v \in SM$ we have that $W^-(v) = W^s(v)$ and $W^+(v) = W^u(v)$.*

By Theorem 11.8 we conclude that ergodic components of $g_t|\Delta$ of positive measure are open (mod 0). In particular, the set Δ is open (mod 0). See Theorem 2.5 that gives sufficient conditions for the set Δ to be of positive Liouville measure.

We describe the topological transitivity of geodesic flows. Following Eberlein [89] we say that the manifold M satisfies the *uniform visibility axiom* if for any $\varepsilon > 0$ there exists $R = R(\varepsilon) > 0$ such that from each point $x \in H$ any geodesic segment γ with $d(x, \gamma) \geqslant R$ is visible at an angle less than ε.

PROPOSITION 17.5 (Eberlein [89]). *The following statements hold*:

1. *if a manifold satisfies the uniform visibility axiom with respect to a Riemannian metric then it satisfies this axiom with respect to any other Riemannian metric with no conjugate points;*
2. *a manifold of nonpositive curvature satisfies the uniform visibility axiom if its universal cover does not admit an isometric geodesic embedding of $\mathbb{R}^2$;*
3. *a compact two-dimensional manifold M of genus $\geqslant 1$ satisfies the uniform visibility axiom;*
4. *if M satisfies the uniform visibility axiom the it also satisfies the asymptoticity axiom.*

THEOREM 17.6 (Eberlein [89]). *Assume that the compact manifold M satisfies the uniform visibility axiom. Then the geodesic flow g_t is topologically transitive.*

The following principal result is an immediate corollary of Theorems 17.4 and 17.6.

THEOREM 17.7. *Let M be a compact smooth Riemannian manifold without focal points satisfying the uniform visibility axiom. If the set Δ has positive Liouville measure then it is open* (mod 0) *and is everywhere dense. The geodesic flow $g_t|\Delta$ is nonuniformly hyperbolic and ergodic. Indeed, $g_t|\Delta$ is Bernoulli.*

The Bernoulli property follows from Theorem 11.21 and from a result by Arnold (see [20, §23]) implying that the geodesic flow has continuous spectrum.

It is an open problem whether the set Δ has full Liouville measure. Brin and Burago have proved this under the additional assumption that the set of negative curvature in M has finitely many connected components. The same result was obtained by Hertz who used different methods. None of these results is published.

Further results on ergodic and topological properties of geodesic flows on manifolds of nonpositive curvature were obtained by Knieper [147–149]. His celebrated result establishes existence and uniqueness of the measure of maximal entropy thus extending the classical result by Margulis to nonpositively curved manifolds. Knieper also obtained multiplicative asymptotic bounds for the growth of volume of spheres (and hence, also that of balls) and the number of periodic orbits. For a detailed account of this work see the chapter [9].

In [109,110], Gunesch strengthened Knieper's results and obtained precise asymptotic formulae for the growth of volume of spheres and the number of homotopy classes of periodic orbits for the geodesic flow on rank 1 manifolds of nonpositive curvature. This extends results by Margulis to the nonuniformly hyperbolic case.

Let M be a compact Riemannian manifold of nonpositive curvature. Given a tangent vector $v \in SM$, $\operatorname{rank}(v)$ is the dimension of the space of parallel Jacobi fields along the geodesic γ_v. The minimum of $\operatorname{rank}(v)$ over all $v \in SM$ is called the *rank* of M, $\operatorname{rank}(M)$. If $\operatorname{rank}(v) = \operatorname{rank}(M)$ the geodesic γ_v and the corresponding vector v are called *regular*. It is easy to see that $1 \leqslant \operatorname{rank}(M) \leqslant \dim M$.

The following result describes the fundamental rank rigidity for nonpositively curved manifolds. It was obtained independently by Ballmann [32] (see also [33]) and by Burns and Spatzier [66] (see also [91]).

THEOREM 17.8. *Let M be a compact smooth Riemannian manifold of nonpositive curvature with irreducible universal cover H. Then the manifold has either rank* 1 *or H is a symmetric space of higher rank.*

In other words, the universal cover of a nonpositively curved manifold can be represented as a product of Euclidean, symmetric, and rank 1 spaces.

THEOREM 17.9 (see [109,110]).

1. *Given $x \in H$, let $b_r(x) = \mathrm{vol}(B(x,r))$ be the Riemannian volume of the ball centered at x of radius r in the universal cover H of M. Then*

$$b_r(x) \sim c(x)e^{hr},$$

where $c(x)$ is a continuous function on M and $h = h(g^t)$ is the topological entropy of the geodesic flow.

2. *Let $P(t)$ be the number of homotopy classes of periodic orbits of length at most t. Then*

$$P(t) \sim \frac{1}{ht} e^{ht}.$$

Observe that unlike in the case of negatively curved manifolds, for nonpositively curved manifolds there may be uncountably many periodic geodesics homotopic to a given one but they all have the same length.

17.2. *Entropy of geodesic flows*

For $v \in SM$ let $v^\perp$ be the set of vectors $w \in SM$ which are orthogonal to v. Consider the linear map $S_v : v^\perp \to v^\perp$ defined by the equality: $S_v w = K_\xi(w)$, where $\xi(w)$ is the vector in $E^-(v)$ such that $d\pi\xi(w) = w$.

THEOREM 17.10. *For a Riemannian metric of class C^4 of nonpositive curvature S_v is a linear self-adjoint operator of the second quadratic form for the horosphere $L(\pi(v), \gamma_v(+\infty))$ at the point $\pi(v)$ (which is a submanifold in H of class C^2).*

Denote by $\{e_i(v)\}$, $i = 1, \dots, p-1$, the orthonormal basis in $v^\perp$ consisting of eigenvectors of S_v. Let $K_i(v)$ be the corresponding eigenvalues. The numbers $K_i(v)$ are called the *principal curvatures* and the directions determined by the vectors $e_i(v)$ the *directions of principal curvatures* for the limit sphere at $\pi(v)$.

THEOREM 17.11 (Pesin [201], Freire and Mañé [100]).

1. *The entropy of the geodesic flow is*

$$h_\mu(g^1) = -\int_M \sum_{i=1}^{p-1} K_i(v)\, d\mu(v) = -\int_M \mathrm{tr}\, S_v\, d\mu(v),$$

where μ is the Liouville measure and $\mathrm{tr}\, S_v$ denotes the trace of S_v.

2. *Let ν be a g^t-invariant probability measure. Then*

$$h_\mu(g^1) \leqslant -\int_M \mathrm{tr}\, S_v\, d\nu(v).$$

Statement 1 of Theorem 17.11 is analogous to a result of [234] for dispersing billiards.

For the topological entropy $h(g^t)$ of the geodesic flow on manifolds without conjugate points Freire and Mañé [100] established the following formula.

THEOREM 17.12.

$$h(g^t) = \lim_{r\to\infty} \frac{\log \operatorname{vol}(B(x,r))}{r},$$

where $x \in H$ is a point in the universal cover of M (the limit exists and does not depend on x).

18. Dynamical systems with singularities: The conservative case

18.1. *General systems with singularities*

In this and the following sections we shall discuss how the core results of smooth ergodic theory can be extended to dynamical systems with singularities (where the map or its differential are discontinuous). We consider two cases: the conservative one when the system preserves volume and the dissipative one when the system possesses an attractor. The main example in the first case is billiards while the main example in the second case is the Lorenz attractor. In both cases the system is uniformly hyperbolic either on the whole phase space or in an neighborhood of the attractor. However, the presence of singularities may effect the behavior of trajectories in a crucial way so that along some trajectories Lyapunov exponents are zero. We shall describe some general conditions on the singularity set which guarantee that Lyapunov exponents along the "majority" of trajectories are nonzero and methods of nonuniform hyperbolicity theory apply.

Let M be a compact smooth Riemannian manifold and $S \subset M$ a closed set. Following Katok and Strelcyn [142] we call a map $f : M \setminus S \to f(M \setminus S)$ a *map with singularities* and the set S the *singularity set* for f if the following conditions hold:

(A1) f is a C^2 diffeomorphism;

(A2) there exist constants $C_1 > 0$ and $a \geqslant 0$ such that

$$\left\|d_x^2 f\right\| \leqslant C_1 \rho(x,S)^{-a}, \quad x \in M \setminus S,$$
$$\left\|d_x^2 f^{-1}\right\| \leqslant C_1 \rho(x,S^-)^{-a}, \quad x \in f(M \setminus S),$$

where $S^- = \{y \in M$: there are $z \in S$ and $z_n \in M \setminus S$ such that $z_n \to z$, $f(z_n) \to y\}$ is the *singularity set* for f^{-1}.

Let μ be a probability measure on M invariant under f. We assume that

(A3) $\displaystyle\int_M \log^+ \|df\|\, d\mu < \infty$ and $\displaystyle\int_M \log^+ \left\|df^{-1}\right\| d\mu < \infty;$

(A4) for every $\varepsilon > 0$ there exist constants $C_2 > 0$ and $b \in (0, 1]$ such that

$$\mu\big(\{x \in M\colon\ \rho(x, S) < \varepsilon\}\big) \leqslant C_2\varepsilon^b.$$

Condition (A2) means that the derivative of f may grow with a "moderate" polynomial rate near the singularity set and condition (A4) implies that $\mu(S) = 0$, i.e., the singularity set is "small".

Conditions (A1)–(A4) constitute the basis of the Katok–Strelcyn theory and allow one to extend results of smooth ergodic theory to smooth systems with singularities. In particular, at every Lyapunov regular point with nonzero Lyapunov exponents one can construct local stable and unstable manifolds, establish the crucial absolute continuity property, describe ergodic properties of the map with respect to a smooth hyperbolic invariant measure and obtain the entropy formula.

We shall now proceed with a formal description. Set $N^+ = \{x \in M\colon\ f^n(x) \notin S$ for all $n \geqslant 0\}$ and let $N = \bigcap_{n\geqslant 0} f^n(N^+)$. For each $\alpha \in (0, 1)$ and $\gamma > 0$ set

$$\Omega^{\alpha,\gamma} = \big\{x \in N\colon\ \rho\big(f^n(x), A\big) \geqslant \gamma\alpha^{|n|} \text{ for every } n \in \mathbb{Z}\big\}.$$

PROPOSITION 18.1. *We have that* $\mu(N) = 0$ *and* $\mu(\Omega^{\alpha,\gamma}) \to 1$ *as* $\gamma \to 0$.

To see this note that

$$N \setminus \Omega^{\alpha,\gamma} = \big\{x \in N\colon\ \rho\big(f^n(x), A\big) < \gamma\alpha^{|n|} \text{ for some } n \in \mathbb{Z}\big\}$$

and hence, by (A4),

$$\begin{aligned}\mu(N \setminus \Omega^{\alpha,\gamma}) &\leqslant \sum_{n\in\mathbb{Z}} \mu\big(\{x \in N\colon\ \rho\big(f^n(x), A\big) < \gamma\alpha^{|n|}\}\big)\\ &\leqslant \sum_{n\in\mathbb{Z}} \mu\big(\{x \in X\colon\ \rho(x, A) < \gamma\alpha^{|n|}\}\big)\\ &\leqslant \sum_{n\in\mathbb{Z}} C_2\gamma^a\alpha^{a|n|} \leqslant \frac{2C_2\gamma^a}{1-\alpha^a}.\end{aligned}$$

Let $\Lambda \subset M \setminus N$ be the set of points with nonzero Lyapunov exponents. We assume that $\mu(\Lambda) > 0$ and we consider the collection of regular sets Λ^ℓ, $\ell \geqslant 1$, for f (see Section 4.5). Denote by

$$\Lambda^{\ell,\alpha,\gamma} = \Lambda^\ell \cap \Omega^{\alpha,\gamma}.$$

From now on we fix a sufficiently large $\ell > 0$, $\alpha = \alpha(\ell)$ and $\gamma = \gamma(\ell)$ such that the set $A^\ell = \Lambda^{\ell,\alpha(\ell),\gamma(\ell)}$ has positive measure.

As an immediate corollary of Proposition 18.1 we obtain the following statement.

THEOREM 18.2 (Stable manifold theorem). *Let f be a diffeomorphism with singularities satisfying conditions* (A1)–(A4). *Then for every $x \in A^\ell$ there exists a local stable manifold $V^s(x)$ such that $x \in V^s(x)$, $T_x V^s(x) = E^s(x)$, and for $y \in V^s(x)$ and $n \geqslant 0$,*

$$\rho\big(f^n(x), f^n(y)\big) \leqslant T(x)\lambda^n e^{\varepsilon n}\rho(x, y),$$

where $T : X \to (0, \infty)$ is a Borel function satisfying

$$T\big(f^m(x)\big) \leqslant T(x)e^{10\varepsilon|m|}, \quad m \in \mathbb{Z}.$$

Furthermore, for every $x \in A^\ell$ there exists a local unstable manifold $V^u(x)$ which have similar properties.

When a point moves under f its stable (and/or unstable) manifold may "meet" the singularity set and be cut by it into several pieces whose sizes, in general, may be *uncontrollably* small. It is condition (A4) that allows one to control this process for almost every trajectory, see Proposition 18.1. In particular, the size of the local manifolds $V^s(x)$ and $V^u(x)$ may depend on the point x and may deteriorate along the trajectory but only with subexponential rate. Moreover, local manifolds satisfy the absolute continuity property, see Section 10.1. This provides the basis for obtaining a complete description of ergodic and topological properties of the system including:

(1) descriptions of ergodic and K-components (see Theorems 11.3 and 11.17),
(2) the entropy formula[8] (see Theorems 12.1 and 12.10), and
(3) density of periodic points (see Theorem 15.2).

Liverani and Wojtkowski [169] designed a method which allows one to study local ergodicity of smooth systems with singularities. The systems to which this method applies are defined axiomatically by a number of conditions. They include some assumptions on the singularity set, existence of invariant cone families which are monotone and strictly monotone (see Section 11.3), and an adaptation of the Sinai–Chernov Ansatz for billiards (see [75]).

Other results on local ergodicity of smooth systems with singularities were obtained by Chernov [72], Markarian [178], and Vaienti [243] (for some particular map).

In [250], Wojtkowski and Liverani introduced a special class of dynamical systems with singularities—conformally Hamiltonian systems with collisions. They are determined by a nondegenerate 2-form Θ and a function H (called Hamiltonian). The form does not have to be closed but $d\Theta = \gamma \wedge \Theta$ for some closed 1-form γ. This condition guarantees that, at least locally, the form Θ can be multiplied by a nonzero function to give a bona fide symplectic structure (such a structure is called *conformally symplectic*). Examples of systems with conformally symplectic structure include the Gaussian isokinetic dynamics and the Nosé–Hoover dynamics. The main result in [250] claims that the Lyapunov spectrum of the corresponding conformally Hamiltonian flow is symmetric. This recovers and generalizes results by Benettin, Galgani, Giorgilli and Strelcyn [43], Dettmann and Morriss

[8] To establish the upper bound for the entropy, one needs to assume, in addition to (A2), that $\|d_x f\| \leqslant C_1\rho(x, S)^{-a}$ and $\|d_x f^{-1}\| \leqslant C_1\rho(x, S^-)^{-a}$; see [142] for details.

[79,80], Garrido and Galavotti [103], Dellago, Posch and Hoover [78], and Bonetto, Galavotti and Garrido [54].

18.2. *Billiards*

We consider billiards in the plane which form a special class of maps with singularities. Let Q be a compact connected subset of $\mathbb{R}^2$ such that ∂Q consists of a finite number of curves of class C^3. The *billiard flow* in Q is generated by the free motion of a particle in the interior of Q, with unit speed and elastic reflections at the boundary (reflection is elastic if the angle of reflection equals the angle of incidence). The flow acts on the unit tangent bundle SQ but is not well defined in the corners of ∂Q. It can be shown that the billiard flow preserves the Liouville measure on SQ. We refer to [77] for more details.

Consider the set $X \subset SQ$ consisting of the unit vectors in SQ at the boundary ∂Q and pointing inside Q. The *billiard map* on Q is defined as the first return map $f: X \to X$ induced by the billiard flow. Given $(q, v) \in X$, its image $f(q, v)$ is the point $(q', v') \in X$, where q' and v' are the position and velocity of the particle with initial condition (q, v) immediately after the next reflection at ∂Q. We introduce the coordinates (s, θ) for a point $(q, v) \in X$ where s is the length of ∂Q up to q measured with respect to a given point in ∂Q and $\theta \in [-\pi/2, \pi/2]$ is the angle that the vector v makes with the inward normal of ∂Q at q. We endow X with the Riemannian metric $ds^2 + d\theta^2$. The billiard map preserves the measure $d\nu = (2c)^{-1} \cos\theta \, ds \, d\theta$ where c is the length of ∂Q.

The billiard map, in general, is not well defined everywhere in X. Let Z be the set of corners of ∂Q, i.e., the points where ∂Q is not of class C^1, and let $q' \neq q$ be the first point of ∂Q where the particle with initial condition $(q, v) \in X$ hits ∂Q. Then f is not defined on the set

$$S^+ = \big\{(q, v) \in X\colon\ q' \in Z \text{ or the segment } \overline{qq'} \text{ is tangent to } \partial Q \text{ at } q'\big\}.$$

Thus, f is not defined at (q, v) if the particle with initial condition (q, v) either hits a corner of ∂Q or reflects at ∂Q with a null angle. Define $R: X \to X$ by $R(s, \theta) = (s, \pi - \theta)$ for $(s, \theta) \in X$. The map f^{-1} is not defined on the set $S^- = RS^+$. One can show that the sets S^+ and S^- consist of a finite number of curves of class C^2 that intersect only at their endpoints (see [142]). They are called respectively the *singularity sets* for f and f^{-1}. For each $n > 0$, the sets where f^n and f^{-n} are not defined are, respectively,

$$S_n^+ = S^+ \cup f^{-1}S^+ \cup \cdots \cup f^{-n+1}S^+,$$
$$S_n^- = S^- \cup fS^- \cup \cdots \cup f^{n-1}S^-.$$

Let

$$S_\infty^+ = \bigcup_{n>0} S_n^+, \qquad S_\infty^- = \bigcup_{n>0} S_n^- \quad \text{and} \quad S_\infty = S_\infty^- \cap S_\infty^+.$$

The points in S_∞^+ (respectively, S_∞^-) hit a corner of ∂Q after a finite number of iterations of f (respectively, f^{-1}). The points in S_∞ hit a corner of ∂Q after a finite number of

iterations of f and after a finite number of iterations of f^{-1}, and thus they have orbits of finite length. Clearly, $\nu(S_n^+) = \nu(S_n^-) = 0$ for every $n > 0$.

According to the former observations, there exists an integer $m > 0$ such that the sets $X \setminus S^+$ and $X \setminus S^-$ consist both of a finite number of open connected sets, say $X_1^+, \dots, X_m^+$ and $X_1^-, \dots, X_m^-$, such that $f : X_i^+ \to X_i^-$ is a C^2 diffeomorphism for $i = 1, \dots, m$. Therefore, the billiard map f is a map with singularities on X (in this case $S = S^+$). Following Katok and Strelcyn [142] we will describe a sufficiently large class of billiards which satisfy conditions (A1)–(A7).

THEOREM 18.3. *Let f be a billiard map on Q. If ∂Q is piecewise C^2, has finite length, and has a uniformly bounded curvature, then condition* (A3) *holds with respect to the measure ν.*

For example, any billiard whose boundary is a union of a finite number of closed arcs and closed curves of class C^2 satisfies the hypotheses of Theorem 18.3.

To establish condition (A4) we consider the class P_k, $k \geqslant 2$, of billiards whose boundary is a union of a finite number of intervals and strictly convex or strictly concave C^k curves.

THEOREM 18.4. *Let f be a billiard map of class P_k, $k \geqslant 2$. Then the singularity set for f is a union of a finite number of closed curves of class C^{k-1} of finite length and of a finite number of isolated points. In particular, f satisfies condition* (A4).

We now discuss condition (A2). Let γ be a strictly convex smooth curve in the plane. For each $p \in \gamma$, let ℓ_1 be the oriented tangent line to γ at p (one-sided tangent line if p is an endpoint of γ), and let ℓ_2 be the line through p orthogonal to ℓ_1. Orienting the line ℓ_2 in a suitable way, one can assume that in a neighborhood of p, with respect to the orthogonal coordinate system given by ℓ_1 and ℓ_2, the curve γ is the graph of a smooth strictly convex function (that we also denote by γ). We consider the class Π_k, $k \geqslant 2$, of billiards in P_k for which there exists $C > 0$ such that all strictly convex pieces of ∂Q satisfy

$$\frac{(s-t)\gamma'(s) - \gamma(s) + \gamma(t)}{\gamma(s) - \gamma(t) - (s-t)\gamma'(t)} \geqslant C$$

for every $s \neq t$ in a neighborhood of zero (at the endpoints we consider appropriate one-sided neighborhoods of zero). It is shown in [142] that the class Π_k includes the billiards in P_k for which the following holds: for each γ of class C^{m+2} as above, $\gamma^{(i)}(0) = 0$ for $2 \leqslant i \leqslant m-1$, and $\gamma^{(m)}(0) \neq 0$.

THEOREM 18.5 [142]. *Any billiard map $f \in \Pi_k$, $k \geqslant 3$, satisfies condition* (A2).

It follows from Theorems 18.3–18.5 that billiard maps of class Π_k, $k \geqslant 3$, satisfy conditions (A1)–(A4). Thus, the results of the previous section apply. Particular cases include dispersing (Sinai's) billiards and some semidispersing billiards. Recall that a curve Γ in the boundary ∂Q of a billiard is *dispersing*, *focusing*, or *flat* if Γ is, respectively, strictly concave outward (with respect to Q), strictly convex outward, or Γ is a straight segment.

We denote by Γ_-, Γ_+, and Γ_0 the unions of the curves forming ∂Q that are, respectively, dispersing, focusing, and flat. A billiard is called *dispersing* if $\Gamma_+ = \Gamma_0 = \emptyset$, and *semidispersing* if $\Gamma_+ = \emptyset$, $\Gamma_- \neq \emptyset$.

We refer to the survey [63] for more details. See also the collection of surveys in [236].

19. Hyperbolic attractors with singularities

In this section we consider dissipative hyperbolic dynamical systems with singularities. They possess attractors and act uniformly hyperbolic in their vicinity. However, due to singularities the behavior of trajectories is effectively nonuniformly hyperbolic. We call these attractors generalized hyperbolic attractors. They were introduced by Pesin in [202]. Examples include Lorenz attractor, Lozi attractor and Belykh attractor. We describe a construction SRB-measures for these systems.

19.1. *Definitions and local properties*

Let M be a smooth Riemannian manifold, $K \subset M$ an open bounded connected set and $N \subset K$ a closed set. Let also $f : K \setminus N \to K$ be a map satisfying the following conditions:

(H1) f is a C^2 diffeomorphism from the open set $K \setminus N$ onto its image;

(H2) there exist constants $C > 0$ and $\alpha \geqslant 0$ such that

$$\|d_x f\| \leqslant C\rho(x, N^+)^{-\alpha}, \quad \|d_x^2 f\| \leqslant C\rho(x, N^+)^{-\alpha}, \quad x \in K \setminus N,$$
$$\|d_x f^{-1}\| \leqslant C\rho(x, N^-)^{-\alpha}, \quad \|d_x^2 f^{-1}\| \leqslant C\rho(x, N^-)^{-\alpha}, \quad x \in f(K \setminus N),$$

where $N^+ = N \cup \partial K$ is the *singularity set* for f and $N^- = \{y \in K\colon$ there are $z \in N^+$ and $z_n \in K \setminus N^+$ such that $z_n \to z$, $f(z_n) \to y\}$ is the *singularity set* for f^{-1}.

Set

$$K^+ = \left\{x \in K\colon\ f^n(x) \notin N^+ \text{ for all } n \in \mathbb{N}\right\}$$

and

$$D = \bigcap_{n \in \mathbb{N}} f^n(K^+).$$

The set $\Lambda = \bar{D}$ is called the *attractor*. We have that

$$D = \Lambda \setminus \bigcup_{n \in \mathbb{Z}} f^n(N^+)$$

and that the maps f and f^{-1} are defined on D, with $f(D) = D$.

Let us fix $\varepsilon > 0$ and set for $\ell \geqslant 1$,

$$D^+_{\varepsilon,\ell} = \{z \in \Lambda\colon\ \rho(f^n(z), N^+) \geqslant \ell^{-1}e^{-\varepsilon n} \text{ for } n \geqslant 0\},$$
$$D^-_{\varepsilon,\ell} = \{z \in \Lambda\colon\ \rho(f^{-n}(z), N^-) \geqslant \ell^{-1}e^{-\varepsilon n} \text{ for } n \geqslant 0\},$$
$$D^{\pm}_{\varepsilon} = \bigcup_{\ell \geqslant 1} D^{\pm}_{\varepsilon,\ell}, \qquad D^0_\varepsilon = D^+ \cap D^-_\varepsilon.$$

The set D^0_ε is f- and f^{-1}-invariant and $D^0_\varepsilon \subset D$ for every ε. This set is an "essential part" of the attractor and in general, may be empty. Even if it is not it may not support any f-invariant Borel finite measure. We say that Λ is *observable* if

(H3) for all sufficiently small ε the set D^0_ε supports an f-invariant Borel finite measure.

We shall provide some conditions that ensure that Λ is observable. Given $A \subset \Lambda$, write $f^{-1}(A) = \{z \in \Lambda \setminus N^+\colon\ f(z) \in A\}$. We denote by $U(\varepsilon, N^+)$ the open ε-neighborhood (in K) of N^+, by $\mathcal{M}_f$ the family of f-invariant Borel probability measures on Λ and by $\varphi(z) = \rho(z, N^+)$.

PROPOSITION 19.1. *The set Λ is observable if one of the following conditions holds*:

1. *there exists $\mu \in \mathcal{M}_f$ such that $\mu(D) > 0$ and $\int_\Lambda |\log\varphi|\, d\mu < \infty$;*
2. *there exist $C > 0$, $q > 0$ such that for any $\varepsilon > 0$ and $n \in \mathbb{N}$,*

$$\nu\big(f^{-n}\big(U(\varepsilon, N^+) \cap f^n(K^+)\big)\big) \leqslant C\varepsilon^q \tag{19.1}$$

(*here ν is the Riemannian volume in K*).

Let us stress that condition (19.1) is similar to condition (A4) for conservative systems with singularities.

Denote by $C(x, \alpha, P)$ the cone at $x \in M$ ($\alpha > 0$ is a real number and P is a linear subspace of T_xM), composed of all vectors $v \in T_xM$ for which

$$\angle(v, P) = \min_{w \in P} \angle(v, w) \leqslant \alpha.$$

We say that Λ is a *generalized hyperbolic attractor* if there exist $C > 0$, $\lambda \in (0, 1)$, a function $\alpha(z)$, and two fields of subspaces $P^s(z)$, $P^u(z) \subset T_zM$, $\dim P^s(z) = q$, $\dim P^u(z) = p - q$ ($p = \dim M$) for $z \in K \setminus N^+$ such that the cones $C^s(z) = C^s(z, \alpha(z), P^s(z))$ and $C^u(z) = C(z, \alpha(z), P^u(z))$ satisfy the following conditions:

(H4) the angle between $C^s(x)$ and $C^u(x)$ is uniformly bounded away from zero over $x \in M \setminus S$; in particular, $C^s(x) \cap C^u(x) = \emptyset$;

(H5) $df\big(C^u(x)\big) \subset C^u\big(f(x)\big)$ for any $x \in M \setminus S$,
$df^{-1}\big(C^s(x)\big) \subset C^s\big(f^{-1}(x)\big)$ for any $x \in f(M \setminus S)$;

(H6) for any $n > 0$,

$$\|df^n v\| \geqslant C\lambda^{-n}\|v\| \quad \text{for any } x \in N^+,\ v \in C^u(x),$$
$$\|df^{-n} v\| \geqslant C\lambda^{-n}\|v\| \quad \text{for any } x \in f^n(N^+),\ v \in C^s(x).$$

Given $z \in D$, the subspaces

$$E^s(z) = \bigcap_{n\geqslant 0} df^{-n} C^s\big(f^n(z)\big), \qquad E^u(z) = \bigcap_{n\geqslant 0} df^n C^u\big(f^{-n}(z)\big)$$

satisfy

(E1) $T_z M = E^s(z) \oplus E^u(z)$, $E^s(z) \cap E^u(z) = \{0\}$;
(E2) the angle between $E^s(z)$ and $E^u(z)$ is uniformly bounded away from zero;
(E3) for each $n \geqslant 0$,

$$\|df^n v\| \leqslant C\lambda^n\|v\|, \quad v \in E^s(z),$$
$$\|df^{-n} v\| \geqslant C^{-1}\lambda^{-n}\|v\|, \quad v \in E^u(z).$$

The subspaces $E^s(z)$ and $E^u(z)$ determine a uniform hyperbolic structure for f on the set D. One can construct local stable and unstable manifolds $V^s(z)$, $V^u(z)$, at every point $z \in D^0_\varepsilon$; in fact, local stable (respectively, unstable) manifolds can be constructed for every $z \in D^+_\varepsilon$ (respectively, for every $z \in D^-_\varepsilon$). Since f has singularities the "size" of local manifolds is a measurable (not continuous) function on D^0_ε, despite the fact that the hyperbolic structure on D is uniform. The size can deteriorate along trajectories but with subexponential rate; it is uniform over the points in $D^0_{\varepsilon,\ell}$.

To simplify our notations we drop the subscript ε from in $D^\pm_\varepsilon$, $D^\pm_{\varepsilon,\ell}$, $D^0_{\varepsilon,\ell}$, etc.

PROPOSITION 19.2. *$V^u(z) \subset D^-$ for any $z \in D^-$.*

Let $A \subset \Lambda$. Define

$$\hat f(A) = f(A \setminus N^+), \qquad \hat f^{-1}(A) = \hat f^{-1}(A \setminus N^-).$$

The sets $\hat f^n(A)$ and $\hat f^{-n}(A)$ for $n > 1$ are defined in the same way. Given $z \in D^0$ we set

$$W^s(z) = \bigcup_{n\geqslant 0} \hat f^{-n}\big(V^s\big(f^n(z)\big)\big), \qquad W^u(z) = \bigcup_{n\geqslant 0} \hat f^n\big(V^u\big(f^{-n}(z)\big)\big).$$

The set $W^s(z)$ is a smooth embedded, but possibly not connected, submanifold in K. It is called the *global stable manifold* at z. If $y \in W^s(z)$ then all images $f^n(y)$, $n \geqslant 0$, are well defined. Similar statements hold for $W^u(z)$, the *global unstable manifold* at z.

For $y \in W^s(z)$ denote by $B^s(y,r)$ the ball in $W^s(z)$ of radius r centered at y (we restrict ourselves to a connected component of $W^s(z)$). Fix $r > 0$ and take $y \in W^s(z)$, $w \in B^s(y,r)$, $n \geqslant 0$ (respectively, $y \in W^u(z)$, $w \in B^u(y,r)$, $n \leqslant 0$). We have

$$\rho^s\big(f^n(y), f^n(w)\big) \leqslant C\mu^n \rho^s(y,w)$$

and, respectively,

$$\rho^u\big(f^{-n}(y), f^{-n}(w)\big) \leqslant C\mu^n \rho^u(y,w),$$

where $C = C(r) > 0$ is a constant, ρ^s and ρ^u are respectively the distances induced by ρ on $W^s(z)$ and $W^u(z)$.

19.2. *SRB-measures: Existence and ergodic properties*

We outline a construction of SRB-measures for diffeomorphisms with generalized hyperbolic attractors. Denote by $J^u(z)$ the Jacobian of the map $df|E^u(z)$ at a point $z \in D^0$. Fix $\ell > 0$, $z \in D^0_\ell$, $y \in W^u(z)$, and $n > 0$, and set

$$\kappa_n(z,y) = \prod_{j=0}^{n-1} \frac{J^u(f^{-j}(z))}{J^u(f^{-j}(y))}.$$

PROPOSITION 19.3. *The following properties hold*:

1. *For any $\ell \geqslant 1$ and $z \in D^0_\ell$, $y \in W^u(z)$ there exists the limit*

$$\kappa(z,y) = \lim_{n\to\infty} \kappa_n(z,y) > 0.$$

Moreover, there is $r^1_\ell > 0$ such that for any $\varepsilon > 0$, $r \in (0, r^1_\ell)$ one can find $N = N(\varepsilon, r)$ such that for any $n \geqslant N$,

$$\max_{z\in D^0_\ell} \max_{y\in \bar{B}^u(z,r)} \big|\kappa_n(z,y) - \kappa(z,y)\big| \leqslant \varepsilon.$$

2. *The function $\kappa(z,y)$ is continuous on D^0_ℓ.*
3. *For any $z \in D^0_\ell$ and $y_1, y_2 \in W^u(z)$,*

$$\kappa(z,y_1)\kappa(y_1,y_2) = \kappa(z,y_2).$$

Fix $\ell \geqslant 1$, $z \in D^0_\ell$ and let $B(z,r)$ be a ball in K centered at z of radius r. Define a *rectangle* at z by

$$\Pi = \Pi(z,r) = \bigcup_{y\in B(z,r)\cap D^0_\ell} B^u\big([y,z],r\big),$$

where $[y, z] = V^u(y) \cap V^s(z)$. Consider the partition $\xi = \xi(\Pi)$ of $\Pi(z, r)$ by the sets $C_\xi(y) = B^u([y, z], r)$, $y \in B(z, r) \cap D^0_\ell$. This partition is continuous and measurable with respect to any Borel measure μ on Λ.

Fix $z \in D^0_\ell$ and a rectangle $\Pi = \Pi(z, r)$ at z. Assume that $\mu(\Pi) > 0$ and denote by $\mu_\xi(y)$, $y \in B(z, r) \cap D^0_\ell$ the family of conditional measures on the sets $C_\xi(y)$. We say that μ is an *SRB-measure* if for any $\ell \geqslant 0$, $z \in D^0_\ell$, and $\Pi = \Pi(z, r)$ with $\mu(\Pi) > 0$,

$$d\mu_\xi(y') = r(y)\kappa\big([z, y], y'\big)\, d\nu^u(y').$$

Here ν^u is the Riemannian volume on $W^u(z)$ induced by the Riemannian metric, $y \in B(z, r) \cap D^0_\ell$, $y' \in B^u([z, y], r)$ and $r(y)$ is the "normalizing factor",

$$r(y) = \left(\int_{B^u([y,z], r^3_\ell)} \kappa\big([z, y], y'\big)\, d\nu^u(y') \right)^{-1}.$$

Denote by $\mathcal{M}'_f$ the family of measures $\mu \in \mathcal{M}_f$ for which $\mu(D^0) = 1$ and by $\mathcal{M}^u_f$ the family of SRB-measures in $\mathcal{M}'_f$. Any $\mu \in \mathcal{M}'_f$ is a measure with nonzero Lyapunov exponents $\chi^1(x), \ldots, \chi^p(x)$ and if μ is ergodic the function $\chi^i(x)$ are constant μ-almost everywhere. We denote the corresponding values by χ^i_μ and assume that

$$\chi^1_\mu \geqslant \cdots \geqslant \chi^q_\mu > 0 > \chi^{q+1}_\mu \geqslant \cdots \geqslant \chi^p_\mu.$$

Fix $z \in D^0_\ell$, $r > 0$ and set

$$U_0 = B^u(z, r), \quad \tilde{U}_0 = U_0, \quad \tilde{U}_n = f(U_{n-1}), \quad U_n = \tilde{U}_n \setminus N^+,$$

and

$$c_0 = 1, \qquad c_n = \left(\prod_{k=0}^{n-1} J^u\big(f^k(z)\big) \right)^{-1}.$$

We define measures $\tilde{\nu}_n$ on U_n by

$$d\tilde{\nu}_n(y) = c_n \kappa\big(f^n(z), y\big)\, d\nu^u(y), \quad n \geqslant 0,$$

and measures ν_n on Λ by

$$\nu_n(A) = \tilde{\nu}_n(A \cap U_n), \quad n \geqslant 0, \tag{19.2}$$

for each Borel set $A \subset \Lambda$.

We say that the attractor Λ satisfies condition (H7) if there exist a point $z \in D^0$ and constants $C > 0$, $t > 0$, $\varepsilon_0 > 0$ such that for any $0 < \varepsilon \leqslant \varepsilon_0$ and $n \geqslant 0$,

$$\nu^u\big(V^u(z) \cap f^{-n}\big(U(\varepsilon, N^+)\big)\big) \leqslant C\varepsilon^t.$$

If Λ satisfies condition (H7) then $\nu_n(A) = \nu_0(f^{-n}A))$ for any $n > 0$ and any Borel set $A \subset \Lambda$.

THEOREM 19.4 (Pesin [202]). *Assume that Λ is a generalized hyperbolic attractor satisfying condition* (H7). *Then there exists a measure $\mu \in \mathcal{M}_f^u$ supported on D^0 which satisfies conditions* 1 *and* 2 *of Proposition* 19.1.

We outline the proof of the theorem. Let $z \in D^0$ be the point mentioned in condition (H7) and ν_k the sequence of measures on Λ as in (19.2). Consider the sequence of measures on Λ defined by

$$\mu_n = \frac{1}{n}\sum_{k=0}^{n-1} \nu_k. \tag{19.3}$$

First, using condition (H7), one can show that for any $\gamma > 0$ there exists $\ell_0 > 0$ such that $\mu_n(D_\ell^0) \geqslant 1 - \gamma$ for any $n > 0$ and $\ell \geqslant \ell_0$. It follows that some limit measure μ for the sequence of measures μ_n is supported on D^0. Next, one can prove that μ is f-invariant and an SRB-measure on Λ.

From now on we assume that Λ is a generalized hyperbolic attractor satisfying condition (H7) and that $\mu \in \mathcal{M}_f^u$, $\mu(D^0) = 1$. We will describe the ergodic properties of μ.

PROPOSITION 19.5. *For μ-almost every $z \in D^0$,*

$$\mu^u\big(D^0 \cap V^u(z)\big) = 1. \tag{19.4}$$

Fix $z \in D^0$ for which (19.4) holds and choose ℓ such that $\nu^u(D_\ell^0 \cap V^u(z)) > 0$. Let W be a smooth submanifold in a small neighborhood of $V^u(z)$ of the form

$$W = \big\{\exp_z\big(w, \varphi(w)\big)\colon\ w \in I \subset E^u(z)\big\},$$

where I is an open subset and $\varphi : I \to E^s(z)$ is a diffeomorphism. W has the same dimension as $V^u(z)$ and is transverse to $V^s(y)$ for all $y \in D_\ell^0 \cap V^u(z)$. Consider the map $p : D_\ell^0 \cap V^u(z) \to W$ where $p(y)$ is the point of intersection of $V^s(y)$ and W. We denote by ν_W the measure on W induced by the Riemannian metric on W (considered as a submanifold of M). One can prove the following result using arguments in the proof of Theorem 10.1.

PROPOSITION 19.6. *The measure $p_*\nu^u$ is absolutely continuous with respect to ν_W.*

Fix $z \in D^0$ and for each $\ell > 0$ set

$$Q(\ell, z) = \bigcup_{y \in D_\ell^0 \cap V^u(z)} V^s(y) \cap \Lambda.$$

One can show that for μ-almost every $z \in \Lambda$ and any sufficiently large $\ell > 0$ we have $\mu(Q(\ell, z)) > 0$ and the set $Q = \bigcup_{n\in\mathbb{Z}} f^n(Q(\ell, z))$ is an ergodic component of positive measure for the map $f|\Lambda$. This implies the following description of the ergodic properties of the map $f|\Lambda$ with respect to the SRB-measure μ.

THEOREM 19.7 (Pesin [202]). *Let $\mu \in \mathcal{M}_f^u$. Then there exist sets $\Lambda_i \subset \Lambda$, $i = 0, 1, 2, \ldots$, such that*:

1. $\Lambda = \bigcup_{i\geqslant 0} \Lambda_i$, $\Lambda_i \cap \Lambda_j = \emptyset$ *for* $i \neq j$, $i, j = 0, 1, 2, \ldots$;
2. $\mu(\Lambda_0) = 0$, $\mu(\Lambda_i) > 0$ *for* $i > 0$;
3. *for* $i > 0$, $\Lambda_i \subset D$, $f(\Lambda_i) = \Lambda_i$, $f|\Lambda_i$ *is ergodic*;
4. *for* $i > 0$, *there exists a decomposition* $\Lambda_i = \bigcup_{j=1}^{n_i} \Lambda_i^j$, $n_i \in \mathbb{N}$, *where*
   (a) $\Lambda_i^{j_1} \cap \Lambda_i^{j_2} = \emptyset$ *for* $j_1 \neq j_2$;
   (b) $f(\Lambda_i^j) = \Lambda_i^{j+1}$ *for* $j = 1, 2, \ldots, n_i - 1$, *and* $f(\Lambda_i^{n_i}) = \Lambda_i^1$;
   (c) $f^{n_i}|\Lambda_i^1$ *is isomorphic to a Bernoulli automorphism*;
5. *the metric entropy* $h_\mu(f|\Lambda)$ *satisfies*

$$h_\mu(f|\Lambda) = \int_\Lambda \sum_{i=1}^{u(x)} \chi_i(x)\, d\mu(x),$$

   where $\chi_1(x), \chi_2(x), \ldots, \chi_{u(x)}(x)$ *is the collection of positive values of the Lyapunov exponent, counted with multiplicities*;
6. *there exists a partition* η *of* Λ *with the following properties*:
   (a) *for* μ*-almost every* $x \in \Lambda$ *the element* $C_\eta(x)$ *of the partition* η *is an open subset in* $W^u(x)$;
   (b) $f\eta \geqslant \eta$, $\bigvee_{k\geqslant 0} f^k\eta = \varepsilon$, $\bigwedge_{k\geqslant 0} f^k\eta = \nu(W^u)$, *where* $\nu(W^u)$ *is the measurable hull of the partition of* Λ *consisting of single leaves* $W^u(x)$ *if* $x \in D^0$ *and single points* $\{x\}$ *if* $x \in \Lambda \setminus D^0$;
   (c) $h(f|\Lambda, \eta) = h_\mu(f|\Lambda)$.

Set

$$W^s(\Lambda) = \bigcup_{z\in D^0} W^s(z).$$

The following is a direct consequence of Proposition 19.1 and Theorem 19.7.

THEOREM 19.8 (Pesin [202]). *Let $\mu \in \mathcal{M}_f^u$. Then for any set Λ_i with $i > 0$ as in Theorem* 19.7 *we have*:

1. *the Riemannian volume of* $W^s(\Lambda_i)$ *is positive*;
2. *there exists* $A_i \subset \Lambda$ *such that* $\mu(A_i) = \mu(\Lambda_i)$ *and for any* $z \in W^s(A_i)$ *and any continuous function* φ *on* M *there exists the limit*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi\big(f^k(z)\big) = \frac{1}{\mu(\Lambda_i)} \int_{\Lambda_i} \varphi\, d\mu.$$

Using the above results one can now describe the class of all SRB-measures on Λ.

THEOREM 19.9 (Pesin [202]). *There exists sets Λ_n, $n = 0, 1, 2, \ldots$ and measures $\mu_n \in \mathcal{M}_f^u$, $n = 1, 2, \ldots$, such that*:

1. $\Lambda = \bigcup_{n \geqslant 0} \Lambda_n$, $\Lambda_n \cap \Lambda_m = \emptyset$ *for* $n \neq m$;
2. *the Riemannian volume of* $W^s(\Lambda_n) \cap W^s(\Lambda_m)$ *is zero for* $n \neq m$, $n, m > 0$;
3. *for* $n > 0$, $\Lambda_n \subset D$, $f(\Lambda_n) = \Lambda_n$, $\mu_n(\Lambda_n) = 1$, *and* $f|\Lambda_n$ *is ergodic with respect to* μ_n;
4. *for* $n > 0$, *there exist* $k_n > 0$ *and a subset* $A_n \subset \Lambda_n$ *such that*
   (a) *the sets* $A_{n,i} = f^i(A_n)$ *are pairwise disjoint for* $i = 1, \ldots, k_n - 1$ *and* $A_{n,k_n} = A_{n,1}$, $\Lambda = \bigcup_{i=1}^{k_n - 1} A_{n,i}$;
   (b) $f^{k_n}|A_{n,1}$ *is a Bernoulli automorphism with respect to* μ_n;
5. *if* $\mu \in \mathcal{M}_f^u$, *then* $\mu = \sum_{n>0} \alpha_n \mu_n$ *with* $\alpha_n \geqslant 0$ *and* $\sum_{n>0} \alpha_n = 1$;
6. *if* ν *is a measure on* K *absolutely continuous with respect to the Riemannian volume and* $\nu_n = \nu|W^s(\Lambda_n)$, $n > 0$, *then*

$$\lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k-1} f_*^i \nu_n = \mu_n.$$

To conclude let us mention a connection between SRB-measures and condition (H7). Notice that any accumulation point of the sequence of measures in (19.3) is an SRB-measure (this essentially follows from Theorem 19.4). We describe a special property of such measures.

PROPOSITION 19.10. *If μ is the SRB-measure constructed in Theorem* 19.4 *then there exists $\varepsilon_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0)$ and any $n > 0$,*

$$\mu\big(U(\varepsilon, N^+)\big) \leqslant C\varepsilon^t, \tag{19.5}$$

where $C > 0$, $t > 0$ are constants independent of ε and n.

We have seen that condition (H7) is sufficient to prove the existence of an SRB-measure on a generalized hyperbolic attractor. We will now show that it is "almost" necessary.

PROPOSITION 19.11. *Let $\mu \in \mathcal{M}_f^u$ (μ is an SRB-measure on Λ and $\mu(D^0) = 1$) satisfy* (19.5) *for some constants C, t, ε. Then for μ-almost every point $z \in D^0$ there exists $\varepsilon(z) > 0$ such that condition* (H7) *holds with respect to z and any $\varepsilon \in (0, \varepsilon(z))$.*

19.3. *Examples*

We now consider a number of examples of maps with generalized hyperbolic attractor when M is a two-dimensional manifold. First we formulate some general assumptions which guarantee the validity of properties (H3) and (H7). Let f be a map satisfying condition (H1) and assume that:

(G1) $K = \bigcup_{i=1}^{m} K^i$, with K^i to be a closed sets, $\operatorname{int} K^i \cap \operatorname{int} K^j = \emptyset$ whenever $i \neq j$,

$$\partial K^i = \bigcup_{j=1}^{r_i} N_{ij} \cup \bigcup_{j=1}^{q_i} M_{ij},$$

where N_{ij} and M_{ij} are smooth curves, and

$$N = \bigcup_{i=1}^{m} \bigcup_{j=1}^{r_i} N_{ij}, \qquad \partial K = \bigcup_{i=1}^{m} \bigcup_{j=1}^{q_i} M_{ij};$$

(G2) f is continuous, and differentiable on each K^i, $i = 1, \dots, m$;

(G3) f possesses two families of stable and unstable cones $C^s(z)$ and $C^u(z)$, $z \in K \setminus \bigcup_{i=1}^{m} \partial K^i$;

(G4) the unstable cone $C^u(z)$ at z depends continuously on $z \in K^i$ and there exists $\alpha > 0$ such that for any $z \in N_{ij} \setminus \partial N_{ij}$, $v \in C^u(z)$, and any vector w tangent to N_{ij} we have $\angle(v, w) \geqslant \alpha$;

(G5) there exists $\tau > 0$ such that $f^k(N) \cap N = \emptyset$, $k = 0, \dots, \tau$ and $a^\tau > 2$ where

$$a = \inf_{z \in K \setminus N} \inf_{v \in C^u(z)} \|d_z f v\| > 1.$$

THEOREM 19.12 (Pesin [202]). *If f satisfies conditions* (H1) *and* (G1)–(G5), *then it also satisfies condition* (H7) *for any $z \in D^0$ and* (19.1) (*in particular, f satisfies condition* (H3)).

Assume now that f satisfies conditions (H1)–(H2), (G1)–(G2), (G4), and (instead of (G3) and (G5)) the following holds:

(G3′) $\rho(f^k(N), n) \geqslant A \exp(-\gamma k)$, $k = 1, 2, \dots$,

where $A > 0$ is a constant and $\gamma > 0$ is sufficiently small (when compared with λ; in particular, $f^k(N) \cap N = \emptyset$, $k = 1, 2, \dots$). Then f satisfies condition (H7) for any $z \in D^0$ and condition (H3).

We now describe some particular two-dimensional maps with generalized hyperbolic attractors.

Lorenz type attractors. Let $I = (-1, 1)$ and $K = I \times I$. Let also $-1 = a_0 < a_1 < \cdots < a_q < a_{q+1} = 1$. Set

$$P_i = I \times (a_i, a_{i+1}), \quad i = 0, \dots, q, \qquad \ell = I \times \{a_0, a_1, \dots, a_q, a_{q+1}\}.$$

Let $T : K \setminus \ell \to K$ be an injective map,

$$T(x, y) = \big(f(x, y), g(x, y)\big), \quad x, y \in I, \tag{19.6}$$

where the functions f and g satisfy the following conditions:

(L1) f and g are continuous on $\bar{P}_i$ and

$$\lim_{y \nearrow a_i} f(x,y) = f_i^-, \qquad \lim_{y \nearrow a_i} g(x,y) = g_i^-,$$
$$\lim_{y \searrow a_i} f(x,y) = f_i^+, \qquad \lim_{y \searrow a_i} g(x,y) = g_i^+,$$

where $f_i^\pm$ and $g_i^\pm$ do not depend on x, $i = 1, 2, \dots, q$;

(L2) f and g have continuous second derivatives on P_i and if $(x,y) \in P_i$, $i = 1, \dots, q$, then

$$df(x,y) = B_i^1 (y - a_i)^{-\nu_i^1} \big(1 + A_i^1(x,y)\big),$$
$$dg(x,y) = C_i^1 (y - a_i)^{-\nu_i^2} \big(1 + D_i^1(x,y)\big)$$

whenever $y - a_i \leqslant \gamma$, and

$$df(x,y) = B_i^2 (a_{i+1} - y)^{-\nu_i^3} \big(1 + A_i^2(x,y)\big),$$
$$dg(x,y) = C_i^2 (a_{i+1} - y)^{-\nu_i^4} \big(1 + D_i^2(x,y)\big)$$

whenever $a_{i+1} - y \leqslant \gamma$, where $\gamma > 0$ is sufficiently small, B_i^1, B_i^2, C_i^1, C_i^2 are some positive constants, $0 \leqslant \nu_i^1, \nu_i^2, \nu_i^3, \nu_i^4 < 1$, and $A_i^1(x,y)$, $A_i^2(x,y)$, $D_i^1(x,y)$, $D_i^2(x,y)$ are continuous functions, which tend to zero when $y \to a_i$ or $y \to a_{i+1}$ uniformly over x; furthermore, the norms of the second derivatives $\|f_{xx}\|$, $\|f_{xy}\|$, $\|g_{xy}\|$, and $\|g_{xx}\|$ are bounded;

(L3) we have the inequalities

$$\|f_x\| < 1, \qquad \big\|g_y^{-1}\big\| < 1,$$
$$1 - \big\|g_y^{-1}\big\| \cdot \|f_x\| > 2\sqrt{\big\|g_y^{-1}\big\| \cdot \|g_x\| \cdot \big\|g_y^{-1} f_y\big\|},$$
$$\big\|g_y^{-1}\big\| \cdot \|g_x\| < \big(1 - \|f_x\|\big)\big(1 - \big\|g_y^{-1}\big\|\big),$$

where $\|\cdot\| = \max_{i=0,\dots,q} \sup_{(x,y)\in P_i} |\cdot|$.

The class of maps satisfying (L1)–(L3) was introduced in [13]. It includes the famous geometric model of the Lorenz attractor. The latter can be described as follows.

THEOREM 19.13. *Assume that $\ell = I \times \{0\}$, $K = I \times I$, and that $T : \backslash\ell \to K$ is a map of the form* (19.6) *where the functions f and g are given by*

$$f(x,y) = \big(-B|y|^{\nu_0} + Bx \operatorname{sgn} y |y|^{\nu} + 1\big) \operatorname{sgn} y,$$
$$g(x,y) = \big((1+A)|y|^{\nu_0} - A\big) \operatorname{sgn} y.$$

If $0 < A < 1$, $0 < B < 1/2$, $\nu > 1$, and $1/(1+A) < \nu_0 < 1$, then T satisfies conditions (L1)–(L3).

The class of maps introduced above is somewhat representative.

THEOREM 19.14. *On an arbitrary smooth compact Riemannian manifold of dimension at least* 3 *there exists a vector field* X *having the following property*: *there is a smooth submanifold* S *such that the first-return map* T *induced on* S *by the flow of* X *satisfies conditions* (L1)–(L3).

We now describe the ergodic and topological properties of the maps with Lorenz type attractors.

THEOREM 19.15 (Pesin [202]). *The following properties hold*:

1. *A map* T *satisfying* (L1)–(L3) *also satisfies conditions* (H1), (H2) *and the attractor* Λ *for* T *is an observable generalized hyperbolic attractor*; *the stable* (*unstable*) *cone at each point* $z \in K$ *is the set of vectors having angle at most* $\pi/6$ *with the horizontal* (*vertical*) *line*.
2. *The stable lamination* W^s *can be extended to a continuous* C^1*-foliation in* K.
3. *Assume that one of the following condition holds*:
   (a) $v_i^j = 0$, $i = 1, \dots, q$, $j = 1, 2, 3, 4$;
   (b) $\rho(T^n(f_i^{\pm}, g_i^{\pm}), \ell) \geqslant C_i \exp(-\gamma n)$ *for any* $n \geqslant 0$, $i = 1, \dots, q$ ($C_i > 0$ *are constants independent of* n; γ *is sufficiently small*).

   Then T *satisfies conditions* (G1)–(G5) (*as well as* (G1), (G2), (G3′), *and* (G4)). *In particular, it satisfies condition* (H7) *for any* $z \in D^0$ *and* (19.1).

The existence of an SRB-measure for the classical geometric model of Lorenz attractor (when K is a square, and ℓ consists of a single interval) was shown in [62]. The proof uses Markov partitions. If the stable foliation W^s is smooth (it takes place, for example, when g does not depend on x) the existence of an SRB-measure follows from a well-known result in the theory of one-dimensional maps (one can show that Λ is isomorphic to the inverse limit of a one-dimensional piecewise expanding map for which (a_i, a_{i+1}), $i = 0, \dots, q$, are intervals of monotonicity; see [13] for details and references).

We now give an example of Lorenz type attractor for which the discontinuity set consists of countable number of intervals and the corresponding map has countable number of components of topological transitivity. Consider a one-dimensional map $g(y)$, $y \in [0, 1]$, given by

$$g(y) = \begin{cases} \frac{1}{n+2} + \frac{2}{2n+1} y & \text{if } \frac{1}{n+1} \leqslant y < \frac{2n+1}{2(n+1)}, \\ \frac{2n+1}{2(n+1)} + \frac{1}{2(n+1)} y & \text{if } \frac{2n+1}{2(n+1)} \leqslant y < \frac{1}{n}, \end{cases}$$

for $n = 1, 2, 3, \dots$. One can show that there exists a function $f(x, y)$ such that the map $T(x, y) = (f(x, y), g(y))$ satisfies conditions (L1)–(L3). However, each set $\Lambda \cap I \times [1/(n+1), 1/n]$ is a component of topological transitivity for T.

Lozi type attractors. Let $c > 0$, $I = (0, c)$, $K = I \times I$, and $0 = a_0 < a_1 < \cdots < a_q < a_{q+1} = c$. Set $\ell = \{a_0, a_1, \ldots, a_q, a_{q+1}\} \times I$ and let $T : K \to K$ be an injective continuous map

$$T(x, y) = \big(f(x, y), g(x, y)\big), \quad x, y \in I,$$

satisfying the following conditions:

Loz1. $T|(K \setminus \ell)$ is a C^2-diffeomorphism and the second derivatives of the maps T and T^{-1} are bounded from above;

Loz2. $\mathrm{Jac}(T) < 1$;

Loz3. $\inf\{(|\frac{\partial f}{\partial x}| - |\frac{\partial f}{\partial y}|) - (|\frac{\partial g}{\partial x}| + |\frac{\partial g}{\partial y}|)\} \geqslant 0$;

Loz4. $\inf\{|\frac{\partial f}{\partial x}| - |\frac{\partial f}{\partial y}|\} \stackrel{\text{def}}{=} u > 1$;

Loz5. $\sup\{(|\frac{\partial f}{\partial x}| + |\frac{\partial g}{\partial y}|)/(|\frac{\partial f}{\partial x}| - |\frac{\partial f}{\partial y}|)^2\} < 1$;

Loz6. there exists $N > 0$ such that $T^k(\ell) \cap \ell = \emptyset$ for $1 \leqslant k \leqslant N$ and $u^N > 2$.

This class of maps was introduced by Young in [256]. It includes the map

$$T(x, y) = \big(1 + by - a|x|, x\big) \tag{19.7}$$

which is obtained from the well-known Lozi map by a change of coordinates (see [161]). It is easy to verify that there exist open intervals of a and b such that (19.7) takes some square $[0, c] \times [0, c]$ into itself and satisfies Loz1–Loz6.

THEOREM 19.16 (Pesin [202]). *The following properties hold*:

1. *A map T satisfying* Loz1–Loz6 *also satisfies conditions* (H1), (G1)–(G5), *and the attractor Λ for T is an observable generalized hyperbolic attractor*; *the stable* (*respectively, unstable*) *cone at each point $z \in K$ has a vertical* (*respectively, horizontal*) *line as the center line. This map also satisfies condition* (H7) *for any $z \in D^0$ and* (19.1).
2. *The stable lamination W^s can be extended to a continuous C^0-foliation in K.*

Belykh type attractors. Let $I = [-1, 1]$, $K = I \times I$, and $\ell = \{(x, y)\colon\ y = kx\}$. Consider the map

$$T(x, y) = \begin{cases} (\lambda_1(x - 1) + 1, \lambda_2(y - 1) + 1) & \text{for } y > kx, \\ (\mu_1(x + 1) - 1, \mu_2(y + 1) - 1) & \text{for } y < kx. \end{cases}$$

In the case $\lambda_1 = \mu_1$, $\lambda_2 = \mu_2$ this map was introduced by Belykh in [39] and was the simplest model in the so-called phase synchronization theory.

THEOREM 19.17. *The following properties hold*:

1. *Assume that*

$$0 < \lambda_1 < \frac{1}{2}, \quad 0 < \mu_1 < \frac{1}{2}, \quad 1 < \lambda_2 < \frac{2}{1 - |k|}, \quad 1 < \mu_2 < \frac{2}{1 - |k|}, \quad |k| < 1.$$

Then T is a map from $K \setminus \ell$ into K satisfying conditions (H1), (G1)–(G4), *and the attractor Λ for T is a generalized hyperbolic attractor* (*the stable and unstable one-dimensional subspaces at each point $z \in D^0$ are respectively horizontal and vertical lines*; *the stable and unstable cones at each point $z \in K$ are the set of vectors having angle at most $\pi/4$ with the horizontal or vertical lines*).

2. *If $\lambda_2 > 2$ and $\lambda_2 > 2$, then T satisfies condition* (G5), *and hence, condition* (H7) *for any $z \in D^0$ and* (19.1).

Acknowledgements

We would like to thank D. Dolgopyat, A. Katok, and the referees for valuable comments. L. Barreira was partially supported by the Center for Mathematical Analysis, Geometry, and Dynamical Systems, through FCT by Program POCTI/FEDER. Ya. Pesin was partially supported by the National Science Foundation, Division of Mathematical Science. Some parts of the chapter were written when Ya. Pesin visited the Research Institute for Mathematical Science (RIMS) at Kyoto University. Ya. Pesin thanks RIMS for hospitality.

Appendix A. Decay of correlations, by Omri Sarig

A.1. *Introduction*

One way of saying that a probability preserving transformation $(X, \mathcal{B}, m, T)$ has unpredictable dynamics is to claim that the results of a 'measurement at time zero' $f(x)$ and a 'measurement at time n' $g(T^n x)$ are correlated very weakly for large n. The correlation coefficient of two random variables f_1, f_2 is defined to be $\frac{\mathrm{Cov}(f_1, f_2)}{\|f_1\|_2 \|f_2\|_2}$, where $\mathrm{Cov}(f_1, f_2) := \int f_1 f_2 - \int f_1 \int f_2$. This suggests the following definition:

DEFINITION 1. A probability preserving transformation $(X, \mathcal{B}, m, T)$ is called *strongly mixing* if for every $f, g \in L^2$, $\mathrm{Cov}(f, g \circ T^n) := \int fg \circ T^n - \int f \int g \underset{n\to\infty}{\longrightarrow} 0$.

It is natural to ask for the speed of convergence (the faster it is the less predictable the system seems to be). Unfortunately, without extra assumptions, the convergence can be arbitrarily slow: For all sequences $\varepsilon_n \downarrow 0$ and all $0 \neq g \in L^2$ s.t. $\int g = 0$, $\exists f \in L^2$ with $\mathrm{Cov}(f, g \circ T^n) \neq O(\varepsilon_n)$.[9]

We will therefore refine the question stated above and ask: *How fast does* $\mathrm{Cov}(f, g \circ T^n) \to 0$ *for f, g in a given collection of functions $\mathcal{L} \subsetneq L^2$*? The collection $\mathcal{L}$ varies from problem to problem. In practice, the challenge often reduces to the problem of identifying a class of functions $\mathcal{L}$ which is large enough to generate $\mathcal{B}$, but small enough to admit analysis.

[9] Otherwise, the functionals $\varphi_n(f) := \frac{1}{\varepsilon_n} \int fg \circ T^n$ are pointwise bounded on L^2, whence by the Banach–Steinhaus theorem uniformly bounded. But $\|\varphi_n\| = \frac{1}{\varepsilon_n}\|g\|_2 \to \infty$ (Y. Shalom).

We discuss this problem below. The literature on this subject is vast, and cannot be covered in an appendix of this size. We will therefore focus on the *methods* used to attack the problem, rather than their actual application (which is almost always highly nontrivial, but also frequently very technical). The reader is referred to Baladi's book [28] for a more detailed account and a more complete bibliography.

In what follows, $(X, \mathcal{B}, m, T)$ is a probability preserving transformation, and $\mathcal{L}$ is a collection of square integrable functions. We assume for simplicity that T is *non*invertible (the methods we describe below can be applied in invertible situations, but are easier to understand in the noninvertible setting). A key concept is:

DEFINITION 2. The *transfer operator* (or *dual operator*, or *Frobenius–Perron operator*) of T is $\hat{T} : L^1 \to L^1$ where $\hat{T} f$ is the unique L^1-function s.t.:

$$\forall g \in L^\infty, \quad \int g \cdot \hat{T} f = \int g \circ T \cdot f.$$

The definition of $\hat{T}$ is tailored to make the following statement correct: If $d\mu = f\, dm$, then $d\mu \circ T^{-1} = \hat{T} f\, dm$. Thus, $\hat{T}$ is the action of T on density functions.

It is easy to check that $\hat{T}$ is a positive operator, a contraction (i.e., $\|\hat{T} f\|_1 \leqslant \|f\|_1$) and that $\|\hat{T} f\|_1 = \|f\|_1$ for all $f \geqslant 0$. The T-invariance of m implies that $\hat{T} 1 = 1$. The relation between $\hat{T}$ and $\operatorname{Cov}(f, g \circ T^n)$ is the following identity:

$$\operatorname{Cov}(f, g \circ T^n) = \int \left[\hat{T}^n f - \int f \right] g. \tag{A.1}$$

We see that the asymptotic behavior of $\operatorname{Cov}(f, g \circ T^n)$ can be studied by analyzing the asymptotic behavior of $\hat{T}^n$ as $n \to \infty$. This is the viewpoint we adopt here.

A.2. *Spectral gap and exponential decay of correlations*

Suppose $\mathcal{L}$ is a Banach space of square integrable functions s.t. $1 \in \mathcal{L}$, $\hat{T}(\mathcal{L}) \subseteq \mathcal{L}$, and $\|\cdot\|_{\mathcal{L}} \geqslant \|\cdot\|_1$. We already mentioned that 1 is an eigenvalue of $\hat{T}$. The operator $Pf := \int f$ is a projection onto its eigenspace.

We say that $\hat{T}$ has a *spectral gap* in $\mathcal{L}$, if the spectrum of $\hat{T} - P \in \operatorname{Hom}(\mathcal{L}, \mathcal{L})$ is a proper subset of the open unit disc, or equivalently, if the $\mathcal{L}$-spectral radius of $\hat{T} - P$, which we denote by $\rho_{\mathcal{L}}$, is strictly less than one.

To see the connection with decay of correlations, note that $\hat{T}^n - P = (\hat{T} - P)^n$, because $\hat{T} P = P \hat{T}$ and $P^2 = P$. Therefore, if $\rho_{\mathcal{L}} < \lambda < 1$, $f \in \mathcal{L}$ and $g \in L^\infty$, then

$$\begin{aligned} \left| \operatorname{Cov}(f, g \circ T^n) \right| &\leqslant \int \left| g (\hat{T} - P)^n f \right| \leqslant \|g\|_\infty \left\| (\hat{T} - P)^n f \right\|_{\mathcal{L}} \\ &= O(\lambda^n) \|f\|_{\mathcal{L}} \|g\|_\infty. \end{aligned}$$

Thus, a spectral gap in $\mathcal{L}$ implies exponential decay of correlations in $\mathcal{L}$.

The question arises how to find a space $\mathcal{L}$ such that $\hat{T}:\mathcal{L}\to\mathcal{L}$ has a spectral gap. We discuss two general methods. The first establishes a spectral gap directly, and the second indirectly.

A.2.1. *Double norm inequalities* Consider the action of T on mass distributions $f\,dm$. If T is very chaotic, then its action will tend to flatten the mountains of $f\,dm$ and to fill-up its crevices. After many iterations, the irregularity of the original mass distribution disappears, and the shape of $\hat{T}^n f\,dm \equiv (f\,dm)\circ T^{-n}$ depends only on the size (total mass) of $f\,dm$, and not on its shape.

It is a deep insight of Doeblin and Fortet [81] that this phenomena is captured by the following double norm inequality, and that this inequality can be used to establish a spectral gap:

$$\|\hat{T}^n f\|_{\mathcal{L}} \leqslant \theta^n \|f\|_{\mathcal{L}} + M\|f\|_{\mathcal{C}} \quad (n\in\mathbb{N}).$$

Here $\|\cdot\|_{\mathcal{L}}$ measures regularity (Lipschitz, BV, etc.), $\|\cdot\|_{\mathcal{C}}$ measures size (L^∞, L^1, etc.), and $0<\theta<1$, $M>0$ are independent of n. We present the functional-analytic machinery in the form given by Ionescu-Tulcea and Marinescu [127] (see Hénon [112] for refinements):

THEOREM A.1 (Doeblin and Fortet, Ionescu-Tulcea and Marinescu). *Let $\mathcal{C}\supseteq\mathcal{L}$ be two Banach spaces such that $\mathcal{L}$-bounded sets are precompact in $\mathcal{C}$, and such that*

$$x_n\in\mathcal{L},\quad \sup\|x_n\|_{\mathcal{L}}<\infty,\quad \|x_n-x\|_{\mathcal{C}}\to 0$$
$$\Rightarrow\quad x\in\mathcal{L},\quad \textit{and}\quad \|x\|_{\mathcal{L}}\leqslant\sup\|x_n\|_{\mathcal{L}}.$$

Let S be a bounded linear operator on $\mathcal{L}$. If $\exists M, H>0$, $0<\theta<1$ s.t. for all $x\in\mathcal{L}$,

$$\sup_{n\geqslant 1}\|S^n x\|_{\mathcal{C}}\leqslant H\|x\|_{\mathcal{L}}\quad \textit{and}\quad \|Sx\|_{\mathcal{L}}\leqslant\theta\|x\|_{\mathcal{L}}+M\|x\|_{\mathcal{C}},$$

then $S=\sum_{i=1}^{p}\lambda_i P_i+N$ where $p<\infty$, $P_i^2=P_i$, $P_iP_j=0$ $(i\neq j)$, $P_iN=NP_i=0$, $\dim\operatorname{Im}(P_i)<\infty$, and $\|N^n\|=O(\kappa^n)$ for some $0<\kappa<1$.

In other words, the theorem gives sufficient conditions for the $\mathcal{L}$-spectrum of S to consist of a compact subset of the open unit disc, and a finite number of eigenvalues λ_i with finite multiplicity. The assumptions of the theorem clearly also imply that $|\lambda_i|\leqslant 1$ for all i.

It follows that if S has no eigenvalues of modulus one other than a simple eigenvalue $\lambda=1$, then S has a spectral gap. This is always the case for the transfer operator as soon as $\mathcal{L}\subset L^1$ and $(X,\mathcal{B},m,T)$ is exact (i.e., $\bigcap_{n=1}^{\infty}T^{-n}\mathcal{B}=\{\emptyset,X\}\bmod m$). Indeed, a theorem of M. Lin [165] says that for exact systems $\|\hat{T}^n f\|_1\underset{n\to\infty}{\longrightarrow}0$ for all $f\in L^1$ with integral zero, so there can be no nonconstant L^1-eigenfunctions with eigenvalue λ such that $|\lambda|=1$.

The key step in applying the double-norm method is the choice of Banach spaces $\mathcal{L}$ and $\mathcal{C}$: It is here that the specifics of the dynamics enter the picture. We indicate some typical choices (our list is by no means complete).

Maps with Markov partitions can be studied in terms of their symbolic dynamics using the sup norm for 'size' and the (symbolic) Hölder norm for 'regularity' (see Ruelle [213], Bowen [57] for finite partitions, and Aaronson and Denker [12] for infinite partitions). The resulting spaces depend on the Markov partition, and they therefore change when the map is perturbed. This makes the study of some stability questions difficult. In the case of Anosov diffeomorphisms, there is an alternative choice of Banach spaces due to Gouëzel and Liverani [108] and Blank, Keller and Liverani [45] which avoids symbolic dynamics, and is thus better suited to the study of such issues.

Without Markov partitions, it is not reasonable to expect the transfer operator to preserve Hölder continuity, and a different choice for $\mathcal{L}$ is needed. In one-dimensional systems, one can sometime use the choice $\mathcal{L} = BV$, $\mathcal{C} = L^1$, see Lasota and Yorke [155], Rychlik [220], Hofbauer and Keller [119], Baladi and Keller [29], Keller [143], and Baladi [28] and references therein.

The multi-dimensional non-Markovian expanding case is more intricate, because of the absence of a canonical *BV* norm, and because of the difficulty in controlling the propagation of singularities in high iterates. Various generalizations of the BV norm have been suggested in this context, see Saussol [224] and Buzzi and Keller [67], and references therein.

Skew-products, i.e., maps of the form $(x, \xi) \mapsto (Tx, f_x(\xi))$, can also be treated using double norm inequalities, at least when the transfer operator of T is well behaved. Additional conditions are required, however, to guarantee mixing: it is possible for the transfer operator of the skew product to have nontrivial eigenvalues of modulus one, even when T is mixing. We refer the reader to the works by Kowalski [151], Dolgopyat [86], Parry and Pollicott [193], Field, Melbourne and Török [98], and references therein.

A.2.2. *Cones* This method is to find a cone of functions C such that $\hat{T}(C) \subseteq C$. If $\hat{T}(C)$ is sufficiently 'small' in C, then $\operatorname{span}\{\hat{T}^n f\}$ converges exponentially fast to $\operatorname{span}\{1\}$ (the precise statements follow shortly). This convergence can then be used to derive a spectral gap on a suitable space, or to prove exponential decay of correlations in C directly.

We present the necessary machinery due to G. Birkhoff [44], and introduced to the study of decay of correlations by Liverani [167] (see also Ferrero and Schmitt [97] and Bakhtin [26,27]).

A subset C of a normed vector space V is a *cone*, if $f \in C, \lambda > 0 \Rightarrow \lambda f \in C$. A cone is called *proper* if $C \cap -C = \emptyset$, *convex* if $f, g \in C \Rightarrow f + g \in C$, and *closed* if $C \cup \{0\}$ is closed. *Hilbert's projective metric* is the following pseudo-metric on C:

$$\Theta(f, g) := \log\left(\frac{\inf\{\mu > 0\colon\ g \preccurlyeq \mu f\} \cup \{0\}}{\sup\{\lambda > 0\colon\ \lambda f \preccurlyeq g\} \cup \{\infty\}}\right), \quad \text{where } f \preccurlyeq g \Leftrightarrow g - f \in C.$$

Alternatively, $\Theta(f, g) = \log \frac{\beta^*}{\alpha^*}$ where α^*, β^* are the best constants in the inequality $\alpha^* f \preccurlyeq g \preccurlyeq \beta^* f$. Observe that $\Theta(\alpha f, \beta g) = \Theta(f, g)$ for all $\alpha, \beta > 0$: Θ measures the distance between the directions generated by f, g, not between f, g themselves.

THEOREM A.2 (G. Birkhoff). *Let C be a closed convex proper cone inside a normed vector space $(V, \|\cdot\|)$, and let $S: V \to V$ be a linear operator such that $S(C) \subseteq C$. If $\Delta := \sup\{\Theta(Sf, Sg): f, g \in C\} < \infty$, then S contracts Θ uniformly*:

$$\Theta(Sf, Sg) \leqslant \tanh\left(\frac{\Delta}{4}\right)\Theta(f, g) \quad (f, g \in C). \tag{A.2}$$

In particular, if we can find a closed convex proper cone $C \subset L^1$ which contains the constants and for which $\hat{T}(C) \subset C$ and $\Delta < \infty$, then the iteration of (A.2) gives for every $f \in L^1$, $\Theta(\hat{T}^n f, Pf) = \Theta(\hat{T}^n f, \hat{T}^n Pf) \leqslant \tanh^{n-1}(\frac{\Delta}{4})\Delta$, and this tends to zero exponentially. (Recall that $Pf = \int f$.) We see that the Θ-distance between the rays determined by $\hat{T}^n f$ and Pf tends to zero geometrically.

The next step is to estimate the L^1-distance between $\hat{T}^n f$ and Pf. In general, this step depends on the cone in a noncanonical way, and cannot be described in a general terms. If we add the assumption that all functions in C are nonnegative and that $f, g \in C$, $f \pm g \in C \Rightarrow \|f\|_1 \geqslant \|g\|_1$, then the situation simplifies considerably, because in this case (see, e.g., [167]),

$$\left\| \frac{f}{\|f\|_1} - \frac{g}{\|g\|_1} \right\|_1 \leqslant e^{\Theta(f,g)} - 1 \quad (f, g \in C).$$

Since $\|\hat{T}^n f\|_1 = \|f\|_1 = \|Pf\|_1$ whenever $f \geqslant 0$, we see that for all $f \in C$,

$$\begin{aligned} \|\hat{T}^n f - Pf\|_1 &= \|f\|_1 \left\| \frac{\hat{T}^n f}{\|\hat{T}^n f\|_1} - \frac{Pf}{\|Pf\|_1} \right\|_1 \leqslant \left(e^{\Theta(\hat{T}^n f, \hat{T}^n Pf)} - 1\right)\|f\|_1 \\ &= O(\rho^n)\|f\|_1 \end{aligned}$$

with $\rho = \tanh\frac{\Delta}{4}$. It now follows from (A.1) that $|\operatorname{Cov}(f, g \circ T^n)| = O(\rho^n)\|f\|_1\|g\|_\infty$ uniformly for $f \in C$, $g \in L^\infty$ and we proved exponential decay of correlations.

The assumption $f, g \in C$, $f \pm g \in C \Rightarrow \|f\|_1 \geqslant \|g\|_1$ is not satisfied in many dynamical situations of interest. In these cases other relations between the Banach distance and Hilbert distance occur, depending on the type of the cone that is used. We refer the reader to [167] for methods which handle this difficulty.

Finally, we mention that Birkhoff's inequality can be generalized for operators mapping one cone to another cone (see Liverani [167, Theorem 1.1]). This is important in nonuniformly expanding situations, where one is forced to consider a chain of cones $\hat{T}(C_i) \subsetneq C_{i+1}$, see Maume-Deschamps [181] for examples.

A.2.3. *Decay of correlations for flows* We now turn from discrete time to continuous time.

Let $\sigma_t : X \to X$ be a strongly mixing probability preserving semiflow on $(X, \mathcal{B}, m, T)$. The decay of correlations of σ_t is captured by the asymptotic behavior as $t \to \infty$ of the *correlation function*:[10]

$$\rho(t) := \int f \cdot g \circ \sigma_t \, d\mu - \int f \, d\mu \int g \, d\mu \quad (t > 0).$$

In order to keep the exposition as simple as possible, we assume that the semiflow is given as a *suspension* over a map $T : \Sigma \to \Sigma$ with *roof function* $r : \Sigma \to \mathbb{R}^+$:

$$X = \big\{(x, \xi) \in \Sigma \times \mathbb{R} \colon 0 \leqslant \xi < r(x)\big\},$$
$$\sigma_t(x, \xi) = (x, \xi + t) \quad \text{with the identifications } (x, \xi) \sim \big(Tx, \xi - r(x)\big),$$
$$dm(x, \xi) = \frac{1}{\int r \, d\mu}(\mu \times d\xi)\Big|_X.$$

The reader may want to think of $\Sigma \simeq \Sigma \times \{0\}$ as of a Poincaré section for the (semi)flow with section map $T : \Sigma \to \Sigma$, and first return time function $r : \Sigma \to \mathbb{R}$. This is the standard way to obtain such a representation.

The main difficulty in continuous time is that the decay of correlations of σ_t depends in a subtle way on the properties of $r : \Sigma \to \mathbb{R}_+$ and $T : \Sigma \to \Sigma$ as a *pair*. There are examples of Ruelle [217] and Pollicott [206] which show that σ_t may not have exponential decay of correlations, even when $T : \Sigma \to \Sigma$ does. In fact, they exhibit (strongly mixing) suspensions over the same section map which have exponential decay of correlations with one roof function, but not with another. In the other direction, there are examples by Kocergin [150] and Khanin and Sinai [144] of mixing suspension flows built over *non*mixing base transformations (see Fayad [96] for the decay of correlations for examples of this type).

It is only recently that Chernov [73] has identified the properties of T and r which are responsible for super-polynomial mixing for Anosov flows, and that Dolgopyat [82] has shown how to use these properties to show that the rate of mixing is in fact exponential for smooth observables, thus settling a problem that has remained open since the early days of hyperbolic dynamics.

Ruelle [217] and Pollicott [206,207] suggested to study $\rho(t)$ as $t \to \infty$ by considering the analytic properties of its Fourier transform

$$\hat{\rho}(s) := \int_{-\infty}^{\infty} e^{-ist} \rho(t) 1_{[0,\infty)}(t) \, dt = \int_0^{\infty} e^{-ist} \rho(t) \, dt,$$

and then appealing to a suitable Tauberian theorem, for example [231, IX.14]:

PROPOSITION A.3. *If $\hat{\rho}(s)$ extends analytically to a strip $\{s = x + iy \colon |y| < \varepsilon\}$ and the functions $\mathbb{R} \ni x \mapsto \hat{\rho}(x + iy)$ $(|y| < \varepsilon)$ are absolutely integrable, with uniformly bounded L^1-norm, then $|\rho(t)| = O(e^{-\varepsilon_0 t})$ for every $0 < \varepsilon_0 < \varepsilon$.*

[10] This is a standard abuse of terminology: $\rho(t)$ is the covariance, not the correlation.

To apply this method, we must first find an analytic extension of $\hat{\rho}$ to some horizontal strip, and then control the growth of this extension.

The starting point is a formula for $\hat{\rho}(s)$ in terms of the transfer operator $\hat{T}$ of T. To obtain such a formula we break $\int_0^\infty dt$ into $\int_0^{r(x)-\xi} + \sum_{n\geqslant 1}\int_{r_n(x)-\xi}^{r_{n+1}(x)-\xi}$ in accordance to the times t when the flow 'hits the roof' (here and throughout $r_n = \sum_{k=0}^{n-1} r\circ T^k$). Setting $E(s) := \int_X \int_0^{r(x)-\xi} e^{-ist} fg\circ\sigma_t\, dt\, dm$, and

$$\hat{f}_s(x) := \int_0^{r(x)} e^{-is\xi} f(x,\xi)\, d\xi, \qquad \hat{g}_s(x) := \int_0^{r(x)} e^{is\xi} g(x,\xi)\, d\xi,$$

and assuming f, g both have integral zero and $\int r\, d\mu = 1$, we obtain

$$\hat{\rho}(s) = E(s) + \sum_{n=1}^{\infty}\int_\Sigma \hat{T}^n\big(e^{isr_n}\hat{f}_s\big)\hat{g}_s\, d\mu \equiv E(s) + \sum_{n=1}^{\infty}\int_\Sigma \hat{T}_s^n(\hat{f}_s)\hat{g}_s\, d\mu,$$

where $\hat{T}_s$ is defined by $\hat{T}_s : F \mapsto \hat{T}\big(e^{isr}F\big)$.

The point of this representation is that, as long as r is bounded, $s\mapsto \hat{T}_s$ has an obvious extension to $s\in\mathbb{C}$. When $\hat{T}$ has a spectral gap, one can study the analyticity of this extension using the analytic perturbation theory of bounded linear operators (Pollicott [207]). The term $E(s)$ is of no importance, because it is an entire function of s.

The integrability conditions of Proposition A.3 turn out to be more delicate. The problem is to control the infinite sum; the term $E(s)$ can be handled in a standard way under some reasonable assumptions on g. This sum is majorized by $\|\hat{f}_s\|_\infty \|\hat{g}_s\|_{\mathcal{L}} \sum_{n\geqslant 1}\|\hat{T}_s^n\|$, so is natural to try to bound $\sum_{n\geqslant 1}\|\hat{T}_s^n\|$ in some strip $S = \{s = x+iy\colon |y| < \varepsilon\}$, at least for $|x|$ large. This amounts to considering expressions of the form

$$\hat{T}_s^n F = \hat{T}^n\big(e^{ixr_n}e^{-yr_n}F\big) \quad (s = x+iy\in S)$$

and showing that the cancellation effect of e^{ixr_n} is powerful enough to make $\hat{T}_s^n$ small. It is at this point that the counterexamples of Ruelle and Pollicott we mentioned before behave badly, and where additional structure is required.

In the case of Anosov flows, Dolgopyat was able to carry out the estimate using Chernov's 'axiom of uniform nonintegrability'.[11] We present his result in a special case, where this axiom is satisfied, and in a weaker form than that used in his paper. The reader is referred to [82] for more general statements.

THEOREM A.4 (D. Dolgopyat). *Let g^t be a geodesic flow on the unit tangent bundle of a smooth, compact, negatively curved surface M. There exist a Poincaré section Σ, a Banach space $\mathcal{L}$, and an $\varepsilon > 0$ s.t. $\sum_{n\geqslant 1}\|\hat{T}_s^n\| = O(|\operatorname{Re}(s)|^\alpha)$ for some $0<\alpha<1$ and all $s\in\{s = x+iy\colon |y|<\varepsilon\}$ with $|\operatorname{Re}(s)|$ large.*

[11] 'Nonintegrability' here refers to foliations, not functions.

We get

$$\begin{aligned}\left|\hat{\rho}(s)\right| &\leqslant \left|E(s)\right| + \|\hat{f}_s\|_\infty \|\hat{g}_s\|_{\mathcal{L}} \sum_{n=1}^{\infty} \|\hat{T}_s^n\| \\ &= \left|E(s)\right| + \|\hat{f}_s\|_\infty \|\hat{g}_s\|_{\mathcal{L}} O\big(\left|\mathrm{Re}(s)\right|^\alpha\big).\end{aligned}$$

Under certain smoothness assumptions on f, g, $\|\hat{f}_s\|_\infty$, $\|\hat{g}_s\|_{\mathcal{L}}$, $|E(s)|$ can be shown to decay fast enough so that the integrability conditions of Proposition A.3 hold. Exponential decay of correlations follows.

We end this section by mentioning the works of Pollicott [207] and Baladi and Vallée [30] for versions of Dolgopyat's estimate for semiflows over piecewise expanding maps of the interval, Dolgopyat's study of exponential and rapid mixing for generic hyperbolic flows [83,85], the paper by Stoyanov [235] for the case of open billiard flows, and the recent paper by Liverani [168] for an extension of Dolgopyat's work to contact Anosov flows.

A.3. *No spectral gap and subexponential decay of correlations*

There are examples (typically nonuniformly hyperbolic systems) where the decay of correlations is slower than exponential. Obviously, the transfer operator for these examples cannot have a spectral gap. We discuss two methods which can be used in this case.

Both methods rely on Kakutani's *induced transformation* construction, which we now review. Let $(X, \mathcal{B}, m, T)$ be a probability preserving transformation and fix $A \in \mathcal{B}$ with $m(A) \neq 0$. By Poincaré's Recurrence Theorem,

$$\varphi_A(x) := 1_A(x) \inf\{n \geqslant 1\colon\ T^n x \in A\}$$

is finite a.e., so $T_A : A \to A$ given by $T_A(x) = T^{\varphi_A(x)}(x)$ is well defined almost everywhere. The map T_A is called the *induced transformation* on A. It is known that if T preserves m, then T_A preserves the measure $m_A(E) := m(E|A)$.

Observe that one iteration of T_A corresponds to several iterations of T, so T_A is more 'chaotic' than T. As a result, $\hat{T}_A$ averages densities much faster than $\hat{T}$, and it is natural to expect it to behave better as an operator. The first method we describe applies when $\hat{T}_A$ has better spectral properties than $\hat{T}$. The second applies when it has better distortion properties.

A.3.1. *Renewal theory* This is a method for determining the asymptotic behavior of $\hat{T}^n$ when $\hat{T}$ has no spectral gap but $\hat{T}_A$ does. Define the following operators on $L^1(A) = \{f \in L^1\colon\ f \text{ is zero outside } A\}$:

$$T_n f := 1_A \hat{T}^n (f 1_A) \quad \text{and} \quad R_n f := 1_A \hat{T}^n (f 1_{[\varphi_A = n]}).$$

Now form the generating functions $T(z) := I + \sum_{n\geqslant 1} z^n T_n$, $R(z) := \sum_{n\geqslant 1} z^n R_n$. Note that $R(1) = \hat{T}_A$. The *renewal equation* is the following identity [223]:

$$T(z) = \big(I - R(z)\big)^{-1} \quad \big(|z| < 1\big).$$

The left-hand side contains information on T_n which are almost the same as $\hat{T}^n$ ($T_n f = \hat{T}^n f$ on A whenever $f \in L^1(A)$), whereas the right-hand side involves $R(z)$ which is a (singular) perturbation of $R(1) = \hat{T}_A$.

The spectral gap of $R(1)$, if it exists, allows us to analyze $R(z)$ using perturbation theory. The analytic problem we are facing is how to translate information on $R(z)$ to information on $T(z)$. If $R(z)$ were an ordinary power series with nonnegative coefficients, this problem would be covered by classical renewal theory. The following result [223] is an operator theoretic version of parts of this theory. In what follows, $\mathbb{D} = \{z \in \mathbb{C}\colon |z| < 1\}$:

THEOREM A.5 (O. Sarig). *Let T_n be bounded linear operators on a Banach space $\mathcal{L}$ such that $T(z) = I + \sum_{n\geqslant 1} z^n T_n$ converges in* $\mathrm{Hom}(\mathcal{L}, \mathcal{L})$ *for every $z \in \mathbb{D}$. Assume that*:

1. Renewal Equation: *for every $z \in \mathbb{D}$, $T(z) = (I - R(z))^{-1}$ where $R(z) = \sum_{n\geqslant 1} z^n R_n$, $R_n \in \mathrm{Hom}(\mathcal{L}, \mathcal{L})$ and $\sum \|R_n\| < \infty$.*
2. Spectral Gap: *the spectrum of $R(1)$ consists of an isolated simple eigenvalue at* 1 *and a compact subset of $\mathbb{D}$.*
3. Aperiodicity: *the spectral radius of $R(z)$ is strictly less than one for all $z \in \bar{\mathbb{D}} \setminus \{1\}$.*

Let P be the eigenprojection of $R(1)$ at 1. *If $\sum_{k>n} \|R_k\| = O(1/n^\beta)$ for some $\beta > 2$ and $PR'(1)P \neq 0$, then for all n,*

$$T_n = \frac{1}{\mu} P + \frac{1}{\mu^2} \sum_{k=n+1}^{\infty} P_n + E_n,$$

where μ is given by $PR'(1)P = \mu P$, $P_n = \sum_{\ell>n} PR_\ell P$, and $E_n \in \mathrm{Hom}(\mathcal{L}, \mathcal{L})$ satisfy $\|E_n\| = O(1/n^{\lfloor\beta\rfloor})$.

Gouëzel has relaxed some of the conditions of this theorem, and has shown how to get higher order terms in this asymptotic expansion [107].

In the special case $T_n f = 1_A \hat{T}^n(f 1_A)$, $R_n f = 1_A \hat{T}^n(f 1_{[\varphi_A = n]})$, one checks that $\mu = \frac{1}{m(A)}$, $Pf = 1_A \frac{1}{m(A)} \int_A f\,dm$, $P_n f = 1_A \frac{1}{m(A)^2} \sum_{\ell>n} m[\varphi_A > \ell] \int_A f\,dm$. The theorem then implies that if f, g are supported inside A, $g \in L^\infty$, $f \in \mathcal{L}$, then

$$g\hat{T}^n f = g \int f + g \sum_{k=n+1}^{\infty} m[\varphi_A > k] \int f + g E_n f.$$

It follows from (A.1) that if $\|\cdot\|_1 \leqslant \|\cdot\|_{\mathcal{L}}$, then

$$\mathrm{Cov}(f, g \circ T^n) = \left(\sum_{k=n+1}^{\infty} m[\varphi_A > k] \right) \int f \int g + O\big(n^{-\lfloor\beta\rfloor}\big).$$

This is often enough to determine $\mathrm{Cov}(f, g \circ T^n)$ up to asymptotic equivalence (see [223, 107] for examples). In particular, unlike the other methods we discuss here, the renewal method—when applicable—yields lower bounds, not just upper bounds for the decay of correlations.

A.3.2. *Coupling* Fix a set A, and consider two positive functions f, g such that $\|f\|_1 = \|g\|_1$. The coupling method for estimating $\|\hat{T}^n f - \hat{T}^n g\|_1$ is based on the following heuristic: Suppose $\exists \varepsilon_1 > 0$ such that $\hat{T} f = \varepsilon_1 1_A + f_1$, $\hat{T} g = \varepsilon_1 1_A + g_1$ with f_1, g_1 positive. If $\delta_1 := 1 - \frac{\|f_1\|_1}{\|f\|_1}$ and $n > 1$, then

$$\hat{T}^n f - \hat{T}^n g \equiv \hat{T}^{n-1} f_1 - \hat{T}^{n-1} g_1 \quad \text{and} \quad \|f_1\|_1 = \|g_1\|_1 = (1 - \delta_1)\|f\|_1.$$

A fraction δ_1 of the total mass was 'coupled' and cancelled out. We now iterate this procedure. If this is possible, then $\exists f_k > 0$ and ε_k, δ_k such that $\hat{T} f_k = f_{k+1} + \varepsilon_k 1_A$ and $\|f_k\|_1 = \|g_k\|_1 = \prod_{i=1}^{k}(1 - \delta_i)\|f\|_1$, where $\delta_i = 1 - \frac{\|f_i\|_1}{\|f_{i-1}\|_1}$. For all $n > N$,

$$\begin{aligned} \left\|\hat{T}^n f - \hat{T}^n g\right\|_1 &= \left\|\hat{T}^{n-N} f_N - \hat{T}^{n-N} g_N\right\|_1 \leqslant \left\|\hat{T}^{n-N} f_N\right\|_1 + \left\|\hat{T}^{n-N} g_N\right\|_1 \\ &= \|f_N\|_1 + \|g_N\|_1 = 2\prod_{i=1}^{N}(1 - \delta_i)\|f\|_1. \end{aligned}$$

If we start with $g = Pf$, we get an upper bound for $\|\hat{T}^n f - Pf\|_1$ which we can then translate using (A.1) to an upper bound for $\mathrm{Cov}(f, h \circ T^n)$ for all $h \in L^\infty$.

The upper bound that we get depends on how much we were able to 'couple' away at every stage. It is a deep insight of L.-S. Young [259,260] that this can be done very efficiently in many important nonuniformly hyperbolic systems, if the set A is such that the induced transformation T_A is a piecewise onto map with uniform bounded distortion.

We describe the class of examples which can be treated this way abstractly. The reader interested in applications to 'real' systems is referred to Bálint and Tóth [31], Markarian [180], Chernov [74], Chernov and Young [76], Young [259] for a treatment of Billiard systems; Young [260], Bruin, van Strien and Luzzatto [61], and Holland [120] for interval maps; and Benedicks and Young [42] and Buzzi and Maume-Deschamps [68] for some higher-dimensional examples.

A *L.-S. Young tower* is a nonsingular conservative transformation $(\Delta, \mathcal{B}, m, F)$ equipped with a generating measurable partition $\{\Delta_{\ell,i} : i \in \mathbb{N}, \ell = 0, \ldots, R_i - 1\}$ with the following properties:

(T1) The measure of $\Delta_{\ell,i}$ is positive and finite for every i and ℓ, and $m(\Delta_0) < \infty$ where $\Delta_0 = \biguplus_{i \geqslant 1} \Delta_{0,i}$.

(T2) g.c.d.$\{R_i : i = 1, 2, 3, \ldots\} = 1$.

(T3) If $\ell + 1 < R_i$, then $F : \Delta_{\ell,i} \to \Delta_{\ell+1,i}$ is a measurable bijection, and $m|_{\Delta_{\ell+1,i}} \circ F|_{\Delta_{\ell,i}} = m|_{\Delta_{\ell,i}}$.

(T4) If $\ell + 1 = R_i$, then $F : \Delta_{\ell,i} \to \Delta_0$ is a measurable bijection.

(T5) Let $R: \Delta_0 \to \mathbb{N}$ be the function $R|_{\Delta_{0,i}} \equiv R_i$ and set $\varphi := \log \frac{dm|_{\Delta_0}}{dm|_{\Delta_0} \circ F^R}$. φ has an a.e. version for which $\exists C > 0$, $\theta \in (0,1)$ s.t. $\forall i$ and $\forall x, y \in \Delta_{0,i}$,

$$\left| \sum_{k=0}^{R(x)-1} \varphi\big(F^k x\big) - \sum_{k=0}^{R(y)-1} \varphi\big(F^k y\big) \right| < C\theta^{s(F^R x, F^R y)},$$

where $s(x, y) = \min\{n \geqslant 0\colon\ (F^R)^n x, (F^R)^n y$ lie in distinct $\Delta_{0,i}\}$.

THEOREM A.6 (L.-S. Young). *Suppose $(\Delta, \mathcal{B}, m, F)$ is a probability preserving L.-S. Young tower with θ as above. Set $\mathcal{L} := \{f : \Delta \to \mathbb{R}\colon \sup |f(x) - f(y)|/\theta^{s(x,y)} < \infty\}$, and define $\hat{R}(x) := \inf\{n \geqslant 0\colon\ F^n(x) \in \Delta_0\}$. For every $f \in \mathcal{L}$ and $g \in L^\infty$,*

1. *if $m[\hat{R} > n] = O(n^{-\alpha})$ for some $\alpha > 0$, then $|\operatorname{Cov}(f, g \circ T^n)| = O(n^{-\alpha})$;*
2. *if $m[\hat{R} > n] = O(\rho_0^n)$ with $0 < \rho_0 < 1$, then $|\operatorname{Cov}(f, g \circ T^n)| = O(\rho^n)$ for some $0 < \rho < 1$;*
3. *if $m[\hat{R} > n] = O(\rho_0^{n^\gamma})$ with $0 < \rho_0 < 1$, $0 < \gamma_0 \leqslant 1$, then $|\operatorname{Cov}(f, g \circ T^n)| = O(\rho^{n^\gamma})$ for some $0 < \rho < 1$, $0 < \gamma < \gamma_0$.*

We remark that if $m[\hat{R} > n] \asymp n^{-\alpha}$, then the bound in (1) was shown to be optimal in a particular example by Hu [124] and in the general case using the methods of the previous subsection by Sarig [223] and Gouëzel [107].

Acknowledgements

O. Sarig thanks D. Dolgopyat, M. Pollicott and Ya. Pesin for helpful comments, and acknowledges support by NSF grant 0400687.

References

Surveys in volume 1A and this volume

[1] N. Chernov, *Invariant measures for hyperbolic dynamical systems*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier (2002), 321–407.
[2] G. Forni, *On the Lyapunov exponents of the Kontsevich–Zorich cocycle*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier (2006), 549–580.
[3] A. Furman, *Random walks on groups and random transformations*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier (2002), 931–1014.
[4] B. Hasselblatt, *Hyperbolic dynamical systems*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier (2002), 239–319.
[5] B. Hasselblatt and A. Katok, *Principal structures*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier (2002), 1–203.
[6] B. Hasselblatt and Ya. Pesin, *Partially hyperbolic dynamical systems*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier (2006), 1–55.
[7] M. Jakobson and G. Światek, *One-dimensional maps*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier (2002), 599–664.

[8] Yu. Kifer and P.-D. Liu, *Random dynamics*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier (2006), 379–499.
[9] G. Knieper, *Hyperbolic dynamics and Riemannian geometry*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier (2002), 453–545.
[10] S. Luzzatto, *Stochastic-like behaviour in nonuniformly expanding maps*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier (2006), 265–326.
[11] E. Pujals and M. Sambarino, *Homoclinic bifurcations, dominated splitting and robust transitivity*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier (2006), 327–378.

Other sources

[12] J. Aaronson and M. Denker, *Local limit theorems for partial sums of stationary sequences generated by Gibbs–Markov maps*, Stoch. Dyn. **1** (2001), 193–237.
[13] V. Afraimovich and Ya. Pesin, *Dimension of Lorenz type attractors*, Sov. Sci. Rev., Sect. C, Math. Phys. Rev. **6** (1987), 169–241.
[14] V. Alekseev, *Quasirandom dynamical systems. I. Quasirandom diffeomorphisms*, Math. USSR Sb. **5** (1) (1968), 73–128.
[15] V. Alekseev, *Quasirandom dynamical systems. II. One-dimensional nonlinear vibrations in a periodically perturbed field*, Math. USSR Sb. **6** (4) (1968), 505–560.
[16] V. Alekseev, *Quasirandom dynamical systems. III. Quasirandom vibrations of one-dimensional oscillators*, Math. USSR Sb. **7** (1) (1969), 1–43.
[17] J. Alves, *SRB measures for nonhyperbolic systems with multidimensional expansion*, Ann. Sci. École Norm. Sup. (4) **33** (2000), 1–32.
[18] J. Alves and V. Araújo, *Hyperbolic times: Frequency versus hyperbolicity*, Ergodic Theory Dynam. Systems **24** (2004), 329–346.
[19] J. Alves, C. Bonatti and M. Viana, *SRB measures for partially hyperbolic systems whose central direction is mostly expanding*, Invent. Math. **140** (2000), 351–398.
[20] D. Anosov, *Geodesic flows on closed Riemann manifolds with negative curvature*, Proc. Steklov Inst. Math. **90** (1969), 1–235.
[21] D. Anosov and Ya. Sinai, *Certain smooth ergodic systems*, Russian Math. Surveys **22** (5) (1967), 103–167.
[22] A. Arbieto and J. Bochi, *L^p-generic cocycles have one-point Lyapunov spectrum*, Stoch. Dyn. **3** (2003), 73–81.
[23] L. Arnold, *Random Dynamical Systems*, Monographs in Mathematics, Springer (1998).
[24] L. Arnold and N. Cong, *Linear cocycles with simple Lyapunov spectrum are dense in L^∞*, Ergodic Theory Dynam. Systems **19** (1999), 1389–1404.
[25] A. Avila and R. Krikorian, *Reducibility or non-uniform hyperbolicity for quasiperiodic Schrödinger cocycles*, Preprint (2004).
[26] V. Bakhtin, *Random processes generated by a hyperbolic sequence of mappings I*, Russian Acad. Sci. Izv. Math. **44** (1995), 247–279.
[27] V. Bakhtin, *Random processes generated by a hyperbolic sequence of mappings II*, Russian Acad. Sci. Izv. Math. **44** (1995), 617–627.
[28] V. Baladi, *Positive transfer operators and decay of correlations*, Adv. Ser. in Nonlinear Dynam., Vol. 16, World Scientific (2000).
[29] V. Baladi and G. Keller, *Zeta functions and transfer operators for piecewise monotonic maps of the interval*, Comm. Math. Phys. **127** (1990), 459–479.
[30] V. Baladi and B. Vallée, *Exponential decay of correlations for surface semi-flows without finite Markov partitions*, Proc. Amer. Math. Soc. **133** (2005), 865–874.
[31] P. Bálint and I. Tóth, *Correlation decay in certain soft billiards*, Comm. Math. Phys. **243** (2003), 55–91.
[32] W. Ballmann, *Nonpositively curved manifolds of higher rank*, Ann. of Math. (2) **122** (1985), 597–609.
[33] W. Ballmann, M. Brin and P. Eberlein, *Structure of manifolds of nonpositive curvature. I*, Ann. of Math. (2) **122** (1985), 171–203.

[34] A. Baraviera and C. Bonatti, *Removing zero Lyapunov exponents*, Ergodic Theory Dynam. Systems **23** (2003), 1655–1670.
[35] L. Barreira and Ya. Pesin, *Lyapunov Exponents and Smooth Ergodic Theory*, Univ. Lecture Series, Vol. 23, Amer. Math. Soc. (2002).
[36] L. Barreira, Ya. Pesin and J. Schmeling, *Dimension and product structure of hyperbolic measures*, Ann. of Math. (2) **149** (1999), 755–783.
[37] L. Barreira and J. Schmeling, *Sets of "non-typical" points have full topological entropy and full Hausdorff dimension*, Israel J. Math. **116** (2000), 29–70.
[38] L. Barreira and C. Valls, *Smoothness of invariant manifolds for nonautonomous equations*, Comm. Math. Phys., to appear.
[39] V. Belykh, *Qualitative Methods of the Theory of Nonlinear Oscillations in Point Systems*, Gorki Univ. Press (1980).
[40] M. Benedicks and L. Carleson, *The dynamics of the Hénon map*, Ann. of Math. (2) **133** (1991), 73–169.
[41] M. Benedicks and L.-S. Young, *Sinai–Bowen–Ruelle measure for certain Hénon maps*, Invent. Math. **112** (1993), 541–576.
[42] M. Benedicks and L.-S. Young, *Markov extensions and decay of correlations for certain Hénon maps*, Astérisque (2000).
[43] G. Benettin, L. Galgani, A. Giorgilli and J.-M. Strelcyn, *Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them*, Meccanica **15** (1980), 9–20.
[44] G. Birkhoff, *Extensions of Jentzsch's theorem*, Trans. Amer. Math. Soc. **85** (1957), 219–227.
[45] M. Blank, G. Keller and C. Liverani, *Ruelle–Perron–Frobenius spectrum for Anosov maps*, Nonlinearity **15** (2001), 1905–1973.
[46] J. Bochi, *Genericity of zero Lyapunov exponents*, Ergodic Theory Dynam. Systems **22** (2002), 1667–1696.
[47] J. Bochi and M. Viana, *Uniform* (*projective*) *hyperbolicity or no hyperbolicity*: *A dichotomy for generic conservative maps*, Ann. Inst. H. Poincaré Anal. Non Linéaire **19** (2002), 113–123.
[48] J. Bochi and M. Viana, *Lyapunov exponents*: *How frequently are dynamical systems hyperbolic*? Modern Dynamical Systems and Applications, Cambridge Univ. Press (2004), 271–297.
[49] J. Bochi and M. Viana, *The Lyapunov exponents of generic volume preserving and symplectic systems*, Ann. of Math. (2), to appear.
[50] C. Bonatti, X. Gómez-Mont and M. Viana, *Généricité d'exposants de Lyapunov non-nuls pour des produits déterministes de matrices*, Ann. Inst. H. Poincaré Anal. Non Linéaire **20** (2003), 579–624.
[51] C. Bonatti, C. Matheus, M. Viana and A. Wilkinson, *Abundance of stable ergodicity*, Comment. Math. Helv. **79** (2004), 753–757.
[52] C. Bonatti and M. Viana, *SRB measures for partially hyperbolic systems whose central direction is mostly contracting*, Israel J. Math. **115** (2000), 157–193.
[53] C. Bonatti and M. Viana, *Lyapunov exponents with multiplicity* 1 *for deterministic products of matrices*, Ergodic Theory Dynam. Systems **24** (2004), 1295–1330.
[54] F. Bonetto, G. Galavotti and P. Garrido, *Chaotic principle*: *An experimental test*, Phys. D **105** (1997), 226–252.
[55] J. Bourgain, *Positivity and continuity of the Lyapunov exponent for shifts on* T^d *with arbitrary frequency and real analytic potential*, Preprint (2002).
[56] J. Bourgain and S. Jitomirskaya, *Continuity of the Lyapunov exponent for quasi-periodic operators with analytic potential*, J. Stat. Phys. **108** (2002), 1203–1218.
[57] R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, Lecture Notes in Math., Vol. 470, Springer (1975).
[58] M. Brin, *Bernoulli diffeomorphisms with nonzero exponents*, Ergodic Theory Dynam. Systems **1** (1981), 1–7.
[59] M. Brin, *Hölder continuity of invariant distributions*, Smooth Ergodic Theory and Its Applications, A. Katok, R. de la Llave, Ya. Pesin and H. Weiss, eds, Proc. Sympos. Pure Math., Amer. Math. Soc. (2001), 91–93.
[60] A. Broise-Alamichel and Y. Guivarc'h, *Exposants caractéristiques de l'algorithme de Jacobi–Perron et de la transformation associée*, Ann. Inst. Fourier (Grenoble) **51** (2001), 565–686.

[61] H. Bruin, S. Luzzatto and S. van Strien, *Rates of decay of correlation for one-dimensional dynamics*, Ann. Sci. École Norm. Sup. (4) **36** (2003), 621–646.
[62] L. Bunimovich, *Systems of hyperbolic type with singularities*, Modern Problems of Mathematics, Vol. 2, Ya. Sinai, ed., Springer (1989).
[63] L. Bunimovich, *Billiards and other hyperbolic systems*, Dynamical Systems, Ergodic Theory and Applications Series, Ya. Sinai, ed., Encyclopaedia of Mathematical Sciences, Vol. 100, Springer (2000), 192–233.
[64] K. Burns, D. Dolgopyat and Ya. Pesin, *Partial hyperbolicity, Lyapunov exponents and stable ergodicity*, J. Stat. Phys. **108** (2002), 927–942.
[65] K. Burns and M. Gerber, *Continuous invariant cone families and ergodicity of flows in dimension three*, Ergodic Theory Dynam. Systems **9** (1989), 19–25.
[66] K. Burns and R. Spatzier, *Manifolds of nonpositive curvature and their buildings*, Inst. Hautes Études Sci. Publ. Math. **65** (1987), 35–59.
[67] J. Buzzi and G. Keller, *Zeta functions and transfer operators for multidimensional piecewise affine and expanding maps*, Ergodic Theory Dynam. Systems **21** (2001), 689–716.
[68] J. Buzzi and V. Maume-Deschamps, *Decay of correlations for piecewise invertible maps in higher dimensions*, Israel J. Math. **131** (2003), 203–220.
[69] D. Bylov, R. Vinograd, D. Grobman and V. Nemyckii, *Theory of Lyapunov Exponents and Its Application to Problems of Stability*, Izdat. Nauka, Moscow (1966), in Russian.
[70] A. Carverhill, *A nonrandom Lyapunov spectrum for nonlinear stochastic dynamical systems*, Stochastics **17** (1986), 253–287.
[71] C.-Q. Cheng and Y.-S. Sun, *Existence of invariant tori in three dimensional measure-preserving mappings*, Celestial Mech. Dyn. Astronom. **47** (3) (1989/90), 275–292.
[72] N. Chernov, *Local ergodicity of hyperbolic systems with singularities*, Funct. Anal. Appl. **27** (1) (1993), 51–54.
[73] N. Chernov, *Markov approximations and decay of correlations for Anosov flows*, Ann. of Math. (2) **147** (1998), 269–324.
[74] N. Chernov, *Decay of correlations and dispersing billiards*, J. Stat. Phys. **94** (1999), 513–556.
[75] N. Chernov and Ya. Sinai, *Ergodic properties of some systems of* 2*-dimensional discs and* 3*-dimensional spheres*, Russian Math. Surveys **42** (1987), 181–207.
[76] N. Chernov and L.-S. Young, *Decay of correlations for Lorenz gases and hard balls*, Hard Ball Systems and the Lorentz Gas, Encyclopaedia of Mathematical Sciences, Vol. 101, Springer (2000), 89–120.
[77] I. Cornfeld, S. Fomin and Ya. Sinai, *Ergodic Theory*, Springer (1982).
[78] C. Dellago, H. Posch and W. Hoover, *Lyapunov instability in a system of hard disks in equilibrium and nonequilibrium steady states*, Phys. Rev. E **53** (1996), 1485–1501.
[79] C. Dettmann and G. Morriss, *Proof of Lyapunov exponent pairing for systems at constant kinetic energy*, Phys. Rev. E **53** (1996), 5541.
[80] C. Dettmann and G. Morriss, *Hamiltonian formulation of the Gaussian isokinetic thermostat*, Phys. Rev. E **54** (1996), 2495.
[81] W. Doeblin and R. Fortet, *Sur des chaînes à liaisons complétes*, Bull. Soc. Math. France **65** (1937), 132–148.
[82] D. Dolgopyat, *On decay of correlations in Anosov flows*, Ann. of Math. (2) **147** (1998), 357–390.
[83] D. Dolgopyat, *Prevalence of rapid mixing in hyperbolic flows*, Ergodic Theory Dynam. Systems **18** (1998), 1097–1114.
[84] D. Dolgopyat, *On dynamics of mostly contracting diffeomorphisms*, Comm. Math. Phys. **213** (2000), 181–201.
[85] D. Dolgopyat, *Prevalence of rapid mixing—II*: *Topological prevalence*, Ergodic Theory Dynam. Systems **20** (2000), 1045–1059.
[86] D. Dolgopyat, *On mixing properties of compact group extensions of hyperbolic systems*, Israel J. Math. **130** (2002), 157–205.
[87] D. Dolgopyat, H. Hu and Ya. Pesin, *An example of a smooth hyperbolic measure with countably many ergodic components*, Smooth Ergodic Theory and Its Applications, A. Katok, R. de la Llave, Ya. Pesin and H. Weiss, eds, Proc. Sympos. Pure Math., Amer. Math. Soc. (2001), 95–106.
[88] D. Dolgopyat and Ya. Pesin, *Every compact manifold carries a completely hyperbolic diffeomorphism*, Ergodic Theory Dynam. Systems **22** (2002), 409–435.

[89] P. Eberlein, *Geodesic flows on negatively curved manifolds. I*, Ann. of Math. (2) **95** (1972), 492–510.
[90] P. Eberlein, *When is a geodesic flow of Anosov type*? *I*, J. Differential Geometry **8** (1973), 437–463.
[91] P. Eberlein, *Geometry of Nonpositively Curved Manifolds*, Chicago Lectures in Mathematics, Chicago Univ. Press (1996).
[92] P. Eberlein, *Geodesic flows in manifolds of nonpositive curvature*, Smooth Ergodic Theory and Its Applications, A. Katok, R. de la Llave, Ya. Pesin and H. Weiss, eds, Proc. Sympos. Pure Math., Amer. Math. Soc. (2001), 525–571.
[93] L. Eliasson, *Reducibility and point spectrum for linear quasi-periodic skew-products*, Proceedings of the International Congress of Mathematicians II, Berlin, 1998, Doc. Math., Extra Vol. II (1998), 779–787.
[94] A. Fathi, M. Herman and J. Yoccoz, *A proof of Pesin's stable manifold theorem*, Geometric Dynamics, Rio de Janeiro, 1981, J. Palis, ed., Lecture Notes in Math., Vol. 1007, Springer (1983), 177–215.
[95] A. Fathi, F. Laudenbach and V. Poénaru, *Travaux de Thurston sur les surfaces*, Séminaire Orsay, Astérisque 66–67 (1979), 284 pp.
[96] B. Fayad, *Polynomial decay of correlations for a class of smooth flows on the two torus*, Bull. Soc. Math. France **129** (2001), 487–503.
[97] P. Ferrero and B. Schmitt, *Produits aléatories d'opérateurs matrices de transfert*, Probab. Theory Related Fields **79** (1988), 227–248.
[98] M. Field, I. Melbourne and A. Török, *Decay of correlations, central limit theorems, and approximation by Brownian motion for compact Lie group extensions*, Ergodic Theory Dynam. Systems **23** (2003), 87–110.
[99] G. Forni, *Deviation of ergodic averages for area-preserving flows on surfaces of higher genus*, Ann. of Math. (2) **155** (2002), 1–103.
[100] A. Freire and R. Mañé, *On the entropy of the geodesic flow in manifolds without conjugate points*, Invent. Math. **69** (1982), 375–392.
[101] A. Furman, *On the multiplicative ergodic theorem for uniquely ergodic systems*, Ann. Inst. H. Poincaré **33** (1997), 797–815.
[102] H. Furstenberg and Yu. Kifer, *Random matrix products and measures on projective spaces*, Israel J. Math. **46** (1983), 12–32.
[103] P. Garrido and G. Galavotti, *Billiards correlation functions*, J. Stat. Phys. **76** (1994), 549–585.
[104] M. Gerber, *Conditional stability and real analytic pseudo-Anosov maps*, Mem. Amer. Math. Soc. 321, 1985.
[105] M. Gerber and A. Katok, *Smooth models of Thurston's pseudo-Anosov maps*, Ann. Sci. École Norm. Sup. (4) **15** (1982), 173–204.
[106] I. Goldsheid and G. Margulis, *Lyapunov indices of a product of random matrices*, Russian Math. Surveys **44** (1989), 11–71.
[107] S. Gouëzel, *Sharp polynomial estimates for the decay of correlations*, Israel J. Math. **139** (2004), 29–65.
[108] S. Gouëzel and C. Liverani, *Banach spaces adapted to Anosov systems*, Preprint.
[109] R. Gunesch, *Precise volume estimates in nonpositive curvature*, Preprint (2003).
[110] R. Gunesch, *Precise asymptotics for periodic orbits of the geodesic flow in nonpositive curvature*, Preprint (2003).
[111] J. Hadamard, *Sur l'itération et les solutions asymptotiques des équations différentielles*, Bull. Soc. Math. France **29** (1901), 224–228.
[112] H. Hennion, *Sur un théorème spectral et son application aux noyaux lipschitziens*, Proc. Amer. Math. Soc. **118** (1993), 627–634.
[113] M. Hénon, *A two dimensional mapping with a strange attractor*, Comm. Math. Phys. **50** (1976), 69–77.
[114] M. Herman, *Construction d'un difféomorphisme minimal d'entropie topologique non nulle*, Ergodic Theory Dynam. Systems **1** (1981), 65–76.
[115] M. Herman, *Une méthode pour minorer les exposants de Lyapunov et quelques examples montrant le caractére local d'un théorèm d'Arnold et de Moser sur le tore de dimension* 2, Comment. Math. Helv. **58** (1983), 453–502.
[116] M. Herman, *Stabilité topologique des systémes dynamiques conservatifs*, Preprint (1990).
[117] M. Hirayama and Ya. Pesin, *Non-absolutely continuous invariant foliations*, Moscow Math. J., to appear.
[118] M. Hirsch, C. Pugh and M. Shub, *Invariant Manifolds*, Lecture Notes in Math., Vol. 583, Springer (1977).
[119] F. Hofbauer and G. Keller, *Ergodic properties of invariant measures for piecewise monotonic transformations*, Math. Z. **180** (1982), 119–140.

[120] M. Holland, *Slowly mixing systems and intermittency maps*, Ergodic Theory Dynam. Systems, to appear.
[121] E. Hopf, *Statistik der geodätischen linien in mannigfaltigkeiten negativer krümmung*, Ber. Verh. Sächs. Akad. Wiss. Leipzig **91** (1939), 261–304.
[122] L. Hörmander, *Complex Analysis in Several Variables*, Van Nostrand (1966).
[123] H. Hu, *Conditions for the existence of SRB measures for "almost Anosov" diffeomorphisms*, Trans. Amer. Math. Soc. **352** (2000), 2331–2367.
[124] H. Hu, *Decay of correlations for piecewise smooth maps with indifferent fixed points*, Ergodic Theory Dynam. Systems **24** (2004), 495–524.
[125] H. Hu and L.-S. Young, *Nonexistence of SRB measures for some diffeomorphisms that are "almost Anosov"*, Ergodic Theory Dynam. Systems **15** (1995), 67–76.
[126] H. Hu, Ya. Pesin and A. Talitskaya, *Every compact manifold carries a hyperbolic ergodic flow*, Modern Dynamical Systems and Applications, Cambridge Univ. Press (2004).
[127] C. Ionescu-Tulcea and G. Marinescu, *Théorie ergodique pour des classes d'opérations non complétement continues*, Ann. of Math. (2) **52** (1950), 140–147.
[128] M. Jakobson, *Absolutely continuous invariant measures for one-parameter families of one-dimensional maps*, Comm. Math. Phys. **81** (1981), 39–88.
[129] R. Johnson, *On a Floquet theory for almost periodic, two-dimensional linear systems*, J. Differential Equations **37** (1980), 184–205.
[130] R. Johnson, K. Palmer and G. Sell, *Ergodic properties of linear dynamical systems*, SIAM J. Math. Anal. **18** (1987), 1–33.
[131] B. Kalinin and V. Sadovskaya, *On pointwise dimension of non-hyperbolic measures*, Ergodic Theory Dynam. Systems **22** (2002), 1783–1801.
[132] A. Karlsson and G. Margulis, *A multiplicative ergodic theorem and nonpositively curved spaces*, Comm. Math. Phys. **201** (1999), 107–123.
[133] A. Katok, *A conjecture about entropy*, Smooth Dynamical Systems, D. Anosov, ed., Mir, Moscow (1977), 181–203.
[134] A. Katok, *Bernoulli diffeomorphisms on surfaces*, Ann. of Math. (2) **110** (1979), 529–547.
[135] A. Katok, *Lyapunov exponents, entropy and periodic orbits for diffeomorphisms*, Inst. Hautes Études Sci. Publ. Math. **51** (1980), 137–173.
[136] A. Katok in collaboration with K. Burns, *Infinitesimal Lyapunov functions, invariant cone families and stochastic properties of smooth dynamical systems*, Ergodic Theory Dynam. Systems **14** (1994), 757–785.
[137] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge Univ. Press (1995).
[138] A. Katok and J. Lewis, *Global rigidity results for lattice actions on tori and new examples of volume-preserving action*, Israel J. Math. **93** (1996), 253–280.
[139] A. Katok and L. Mendoza, *Dynamical systems with nonuniformly hyperbolic behavior*, *Introduction to the Modern Theory of Dynamical Systems*, A. Katok and B. Hasselblatt, eds, Cambridge Univ. Press (1995).
[140] A. Katok and R. Spatzier, *First cohomology of Anosov actions of higher rank Abelian groups and applications to rigidity*, Inst. Hautes Études Sci. Publ. Math. **79** (1994), 131–156.
[141] A. Katok and R. Spatzier, *Subelliptic estimates of polynomial differential operators and applications to cocycle rigidity*, Math. Res. Lett. **1** (1994), 193–202.
[142] A. Katok and J.-M. Strelcyn, *Invariant Manifolds, Entropy and Billiards; Smooth Maps with Singularities*, with the collaboration of F. Ledrappier and F. Przytycki, Lecture Notes in Math., Vol. 1222, Springer (1986).
[143] G. Keller, *Interval maps with strictly contracting Perron–Frobenius operators*, Internat. J. Bifur. Chaos Appl. Sci. Engrg. **9** (1999), 1777–1783.
[144] K. Khanin and Ya. Sinai, *Mixing for some classes of special flows over rotations of the circle*, Funktsional. Anal. Prilozhen. **26** (1992), 155–169.
[145] Yu. Kifer, *A multiplicative ergodic theorem for random transformations*, J. Anal. Math. **45** (1985), 207–233.
[146] Yu. Kifer, *Ergodic Theory of Random Transformations*, Progress in Probability and Statistics, Vol. 10, Birkhäuser (1986).
[147] G. Knieper, *Volume growth, entropy and the geodesic stretch*, Math. Res. Lett. **2** (1995), 39–58.

[148] G. Knieper, *On the asymptotic geometry of nonpositively curved manifolds*, Geom. Funct. Anal. **7** (1997), 755–782.
[149] G. Knieper, *The uniqueness of the measure of maximal entropy for the geodesic flows on rank* 1 *manifolds*, Ann. of Math. (2) **148** (1998), 291–314.
[150] A. Kocergin, *Mixing in special flows over a rearrangement of segments and in smooth flows on surfaces*, Mat. USSR Sb. **25** (1975), 471–502.
[151] Z. Kowalski, *Ergodic properties of skew products with Lasota–Yorke type maps in the base*, Studia Math. **106** (1993), 45–57.
[152] A. Krámli, N. Simányi and D. Szász, *A "transversal" fundamental theorem for semi-dispersing billiards*, Comm. Math. Phys. **129** (1990), 535–560, erratum in Comm. Math. Phys. **138** (1991), 207–208.
[153] R. Krikorian, *Réductibilité des systèmes produits-croisés à valeurs dans des groupes compacts*, Astérisque 259 (1999).
[154] J. Lagarias, *The quality of the Diophantine approximations found by the Jacobi–Perron algorithm and related algorithms*, Monatsh. Math. **115** (1993), 299–328.
[155] A. Lasota and J. Yorke, *On the existence of invariant measures for piecewise monotonic transformations*, Trans. Amer. Math. Soc. **186** (1973), 481–488.
[156] F. Ledrappier, *Propriétes ergodiques des mesures de Sinaï*, Inst. Hautes Études Sci. Publ. Math. **59** (1984), 163–188.
[157] F. Ledrappier and M. Misiurewicz, *Dimension of invariant measures for maps with exponent zero*, Ergodic Theory Dynam. Systems **5** (1985), 595–610.
[158] F. Ledrappier and J.-M. Strelcyn, *A proof of the estimate from below in Pesin's entropy formula*, Ergodic Theory Dynam. Systems **2** (1982), 203–219.
[159] F. Ledrappier and L.-S. Young, *The metric entropy of diffeomorphisms. I. Characterization of measures satisfying Pesin's entropy formula*, Ann. of Math. (2) **122** (1985), 509–539.
[160] F. Ledrappier and L.-S. Young, *The metric entropy of diffeomorphisms. II. Relations between entropy, exponents and dimension*, Ann. of Math. (2) **122** (1985), 540–574.
[161] Y. Levy, *Ergodic properties of the Lozi map*, Lecture Notes in Math., Vol. 1109, Springer (1985), 103–116.
[162] J. Lewowicz, *Lyapunov functions and topological stability*, J. Differential Equations **38** (1980), 192–209.
[163] J. Lewowicz, *Lyapunov functions and stability of geodesic flows*, Geometric Dynamics, Rio de Janeiro, 1981, J. Palis, ed., Lecture Notes in Math., Vol. 1007, Springer (1983), 463–479.
[164] J. Lewowicz and E. Lima de Sá, *Analytic models of pseudo-Anosov maps*, Ergodic Theory Dynam. Systems **6** (1986), 385–392.
[165] M. Lin, *Mixing for Markov operators*, Z. Wahrscheinlichkeitstheorie Verw. Gebiete **19** (1971), 231–242.
[166] P.-D. Liu and M. Qian, *Smooth Ergodic Theory of Random Dynamical Systems*, Lecture Notes in Math., Vol. 1606, Springer (1995).
[167] C. Liverani, *Decay of correlations*, Ann. of Math. (2) **142** (1995), 239–301.
[168] C. Liverani, *On contact Anosov flows*, Ann. of Math. (2) **159** (2004), 1275–1312.
[169] C. Liverani and M. Wojtkowski, *Ergodicity in Hamiltonian systems*, Dynam. Report. Expositions Dynam. Systems (N.S.), Vol. 4, Springer (1995), 130–202.
[170] S. Luzzatto and M. Viana, *Parameter exclusions in Hénon-like systems*, Russian Math. Surveys **58** (2003), 1053–1092.
[171] A. Lyapunov, *The General Problem of the Stability of Motion*, Taylor & Francis (1992).
[172] R. Mañé, *A proof of Pesin's formula*, Ergodic Theory Dynam. Systems **1** (1981), 95–102; errata in Ergodic Theory Dynam. Systems **3** (1983), 159–160.
[173] R. Mañé, *Lyapunov exponents and stable manifolds for compact transformations*, Geometric Dynamics, Rio de Janeiro, 1981, J. Palis, ed., Lecture Notes in Math., Vol. 1007, Springer (1983), 522–577.
[174] R. Mañé, *Oseledec's theorem from the generic viewpoint*, Proceedings of the International Congress of Mathematicians 1–2, Warsaw 1983 (1984), 1269–1276.
[175] R. Mañé, *Ergodic Theory and Differentiable Dynamics*, Ergebnisse der Mathematik und ihrer Grenzgebiete 3, Vol. 8, Springer (1987).
[176] R. Mañé, *The Lyapunov exponents of generic area preserving diffeomorphisms*, International Conference on Dynamical Systems, Montevideo, 1995, F. Ledrappier, J. Lewowicz and S. Newhouse, eds, Pitman Res. Notes Math. Ser., Vol. 362, Longman (1996), 110–119.

[177] A. Manning, *Topological entropy and the first homology group*, Dynamical Systems, Warwick, 1974, A. Manning, ed., Lecture Notes in Math., Vol. 468, Springer (1975), 185–190.
[178] R. Markarian, *The fundamental theorem of Sinai–Chernov for dynamical systems with singularities*, Dynamical Systems, Santiago, 1990, Pitman Res. Notes Math. Ser., Vol. 285, Longman (1993), 131–158.
[179] R. Markarian, *Non-uniformly hyperbolic billiards*, Ann. Fac. Sci. Toulouse, VI. Sér., Math. **3** (5–6) (1994), 223–257.
[180] R. Markarian, *Billiards with polynomial decay of correlations*, Ergodic Theory Dynam. Systems **24** (2004), 177–197.
[181] V. Maume-Deschamps, *Projective metrics and mixing properties on towers*, Trans. Amer. Math. Soc. **353** (2001), 3371–3389.
[182] H. McCluskey and A. Manning, *Hausdorff dimension for horseshoes*, Ergodic Theory Dynam. Systems **3** (1983), 251–260; errata in Ergodic Theory Dynam. Systems **3** (1983), 319.
[183] V. Millionshchikov, *Les systèmes linéaires d'équations différentielles ordinaires*, Actes du Congrès International des Mathématiciens, Nice, 1970, Vol. 2, Gauthier–Villars (1971), 915–919.
[184] J. Milnor, *On the concept of attractor*, Comm. Math. Phys. **102** (1985), 517–519.
[185] J. Milnor, *Fubini foiled: Katok's paradoxical example in measure theory*, Math. Intelligencer **19** (1997), 30–32.
[186] M. Misiurewicz and F. Przytycki, *Topological entropy and degree of smooth mappings*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys. **25** (6) (1977), 573–574.
[187] M. Misiurewicz and F. Przytycki, *Entropy conjecture for tori*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys. **25** (6) (1977), 575–578.
[188] L. Mora and M. Viana, *Abundance of strange attractors*, Acta Math. **171** (1993), 1–71.
[189] S. Newhouse, *Entropy and volume*, Ergodic Theory Dynam. Systems **8*** (1988), 283–299.
[190] S. Newhouse, *Continuity properties of entropy*, Ann. of Math. (2) **129** (1989), 215–235.
[191] D. Ornstein and B. Weiss, *Geodesic flows are Bernoullian*, Israel J. Math. **14** (1973), 184–198.
[192] V. Oseledets, *A multiplicative ergodic theorem. Liapunov characteristic numbers for dynamical systems*, Trans. Moscow Math. Soc. **19** (1968), 197–221.
[193] W. Parry and M. Pollicott, *The Chebotarov theorem for Galois coverings of Axiom A flows*, Ergodic Theory Dynam. Systems **6** (1986), 133–148.
[194] O. Perron, *Über stabilität und asymptotishes Verhalten der Lösungen eines Systemes endlicher Differenzengleichungen*, J. Reine Angew. Math. **161** (1929), 41–64.
[195] O. Perron, *Die stabilitätsfrage bei Differenzengleichungen*, Math. Z. **32** (1930), 703–728.
[196] Ya. Pesin, *An example of a nonergodic flow with nonzero characteristic exponents*, Funct. Anal. Appl. **8** (1974), 263–264.
[197] Ya. Pesin, *Families of invariant manifolds corresponding to nonzero characteristic exponents*, Math. USSR Izv. **40** (1976), 1261–1305.
[198] Ya. Pesin, *Characteristic Ljapunov exponents, and smooth ergodic theory*, Russian Math. Surveys **32** (1977), 55–114.
[199] Ya. Pesin, *A description of the π-partition of a diffeomorphism with an invariant measure*, Math. Notes **22** (1977), 506–515.
[200] Ya. Pesin, *Geodesic flows on closed Riemannian manifolds without focal points*, Russian Math. Surveys **11** (1977), 1195–1228.
[201] Ya. Pesin, *Formulas for the entropy of the geodesic flow on a compact Riemannian manifold without conjugate points*, Math. Notes **24** (4) (1978), 796–805.
[202] Ya. Pesin, *Dynamical systems with generalized hyperbolic attractors: hyperbolic, ergodic and topological properties*, Ergodic Theory Dynam. Systems **12** (1992), 123–151.
[203] Ya. Pesin, *Dimension Theory in Dynamical Systems: Contemporary Views and Applications*, Chicago Lectures in Mathematics, Chicago Univ. Press (1998).
[204] Ya. Pesin and Ya. Sinai, *Gibbs measures for partially hyperbolic attractors*, Ergodic Theory Dynam. Systems **2** (1982), 417–438.
[205] V. Pliss, *On a conjecture due to Smale*, Differ. Uravn. **8** (1972), 262–282, in Russian.
[206] M. Pollicott, *On the rate of mixing of Axiom A flows*, Invent. Math. **81** (1985), 413–426.
[207] M. Pollicott, *On the mixing of Axiom A attracting flows and a conjecture of Ruelle*, Ergodic Theory Dynam. Systems **19** (1999), 535–548.

[208] C. Pugh, *The $C^{1+\alpha}$ hypothesis in Pesin theory*, Inst. Hautes Études Sci. Publ. Math. **59** (1984), 143–161.
[209] C. Pugh and M. Shub, *Ergodic attractors*, Trans. Amer. Math. Soc. **312** (1989), 1–54.
[210] M. Qian and S. Zhu, *SRB-measures and Pesin's entropy formula for endomorphisms*, Trans. Amer. Math. Soc. **354** (2002), 1453–1471.
[211] M. Raghunathan, *A proof of Oseledec's multiplicative ergodic theorem*, Israel J. Math. **32** (1979), 356–362.
[212] D. Ruelle, *An inequality for the entropy of differentiable maps*, Bol. Soc. Brasil. Mat. **9** (1978), 83–87.
[213] D. Ruelle, *Thermodynamic Formalism. The Mathematical Structures of Classical Equilibrium Statistical Mechanics*, Encyclopedia of Mathematics and its Applications, Vol. 5, Addison–Wesley (1978).
[214] D. Ruelle, *Analyticity properties of the characteristic exponents of random matrix products*, Adv. Math. **32** (1979), 68–80.
[215] D. Ruelle, *Ergodic theory of differentiable dynamical systems*, Inst. Hautes Études Sci. Publ. Math. **50** (1979), 27–58.
[216] D. Ruelle, *Characteristic exponents and invariant manifolds in Hilbert space*, Ann. of Math. (2) **115** (1982), 243–290.
[217] D. Ruelle, *Flots qui ne mélangement pas exponentiellement*, C. R. Acad. Sci. Paris **296** (1983), 191–193.
[218] D. Ruelle, *Smooth dynamics and new theoretical ideas in nonequilibrium statistical mechanics*, J. Stat. Phys. **94** (1999), 393–468.
[219] D. Ruelle and A. Wilkinson, *Absolutely singular dynamical foliations*, Comm. Math. Phys. **219** (2001), 481–487.
[220] M. Rychlik, *Bounded variation and invariant measures*, Studia Math. **76** (1983), 69–80.
[221] R. Sacker and G. Sell, *Lifting properties in skew-product flows with applications to differential equations*, Mem. Amer. Math. Soc., Vol. 190 (1977).
[222] R. Sacker and G. Sell, *A spectral theory for linear differential systems*, J. Differential Equations **27** (1978), 320–358.
[223] O. Sarig, *Subexponential decay of correlations*, Invent. Math. **150** (2002), 629–653.
[224] B. Saussol, *Absolutely continuous invariant measures for multidimensional expanding maps*, Israel J. Math. **116** (2000), 223–248.
[225] F. Schweiger, *The metrical theory of Jacobi–Perron algorithm*, Lecture Notes in Math., Vol. 334, Springer (1973).
[226] G. Sell, *The structure of a flow in the vicinity of an almost periodic motion*, J. Differential Equations **27** (1978), 359–393.
[227] M. Shub, *Dynamical systems, filtrations, and entropy*, Bull. Amer. Math. Soc. (N.S.) **80** (1974), 27–41.
[228] M. Shub and D. Sullivan, *A remark on the Lefschetz fixed point formula for differentiable maps*, Topology **13** (1974), 189–191.
[229] M. Shub and A. Wilkinson, *Pathological foliations and removable zero exponents*, Invent. Math. (2000), 495–508.
[230] M. Shub and R. Williams, *Entropy and stability*, Topology **14** (1975), 329–338.
[231] B. Simon and M. Reed, *Methods of Modern Mathematical Physics. I: Functional Analysis*, Academic Press (1980).
[232] Ya. Sinai, *Dynamical systems with countably-multiple Lebesgue spectrum II*, Amer. Math. Soc. Trans. (2) **68** (1966), 34–88.
[233] Ya. Sinai, *Dynamical systems with elastic reflections*, Russian Math. Surveys **68** (1970), 137–189.
[234] Ya. Sinai and N. Chernov, *Entropy of hard spheres gas with respect to the group of space–time translations*, Tr. Semin. Im. I.G. Petrovskogo **8** (1982), 218–238, in Russian.
[235] L. Stoyanov, *Spectrum of the Ruelle operator and exponential decay of correlations for open billiard flows*, Amer. J. Math. **123** (2001), 715–759.
[236] D. Szász, ed., *Hard Ball Systems and the Lorentz Gas*, Encyclopaedia of Mathematical Sciences, Vol. 101, Springer (2000).
[237] A. Tahzibi, *C^1-generic Pesin's entropy formula*, C. R. Acad. Sci. Paris **335** (2002), 1057–1062.
[238] P. Thieullen, *Fibrés dynamiques asymptotiquement compacts. Exposants de Lyapunov. Entropie. Dimension*, Ann. Inst. H. Poincaré. Anal. Non Linéaire **4** (1987), 49–97.
[239] P. Thieullen, *Ergodic reduction of random products of two-by-two matrices*, J. Anal. Math. **73** (1997), 19–64.
[240] W. Thurston, *On the geometry and dynamics of diffeomorphisms of surfaces*, Preprint.

[241] M. Tsujii, *Regular points for ergodic Sinai measures*, Trans. Amer. Math. Soc. **328** (1991), 747–777.
[242] I. Ugarcovici, *Hyperbolic measures and periodic orbits*, Preprint (2003).
[243] S. Vaienti, *Ergodic properties of the discontinuous sawtooth map*, J. Stat. Phys. **67** (1992), 251–269.
[244] M. Viana, *Strange attractors in higher dimensions*, Bol. Soc. Brasil. Mat. (N.S.) **24** (1993), 13–62.
[245] P. Walters, *Unique ergodicity and matrix products*, Lyapunov Exponents, Proc. Workshop, Bremen, Germany, 1984, Lecture Notes in Math., Vol. 1186, Springer (1986), 37–55.
[246] Q. Wang and L.-S.Young, *Strange attractors with one direction of instability*, Comm. Math. Phys. **218** (2001), 1–97.
[247] Q. Wang and L.-S.Young, *From invariant curves to strange attractors*, Comm. Math. Phys. **225** (2002), 275–304.
[248] M. Wojtkowski, *Invariant families of cones and Lyapunov exponents*, Ergodic Theory Dynam. Systems **5** (1985), 145–161.
[249] M. Wojtkowski, *Monotonicity, J-algebra of Potapov and Lyapunov exponents*, Smooth Ergodic Theory and Its Applications, A. Katok, R. de la Llave, Ya. Pesin and H. Weiss, eds, Proc. Sympos. Pure Math., Amer. Math. Soc. (2001), 499–521.
[250] M. Wojtkowski and C. Liverani, *Conformally symplectic dynamics and symmetry of the Lyapunov spectrum*, Comm. Math. Phys. **194** (1998), 47–60.
[251] Z. Xia, *Existence of invariant tori in volume-preserving diffeomorphisms*, Ergodic Theory Dynam. Systems **12** (1992), 621–631.
[252] J.-C. Yoccoz, *Travaux de Herman sur les tores invariants*, Séminaire Bourbaki, Astérisque 206, exposé 754 (1991–1992), 311–344.
[253] Y. Yomdin, *Volume growth and entropy*, Israel J. Math. **57** (1987), 285–300.
[254] Y. Yomdin, *C^k-resolution of semialgebraic mappings. Addendum to*: "*Volume growth and entropy*", Israel J. Math. **57** (1987), 301–317.
[255] L.-S. Young, *Dimension, entropy and Lyapunov exponents*, Ergodic Theory Dynam. Systems **2** (1982), 109–124.
[256] L.-S. Young, *Bowen–Ruelle–Sinai measures for certain piecewise hyperbolic maps*, Trans. Amer. Math. Soc. **287** (1985), 41–48.
[257] L.-S. Young, *Some open sets of nonuniformly hyperbolic cocycles*, Ergodic Theory Dynam. Systems **13** (1993), 409–415.
[258] L.-S. Young, *Lyapunov exponents for some quasi-periodic cocycles*, Ergodic Theory Dynam. Systems **17** (1997), 483–504.
[259] L.-S. Young, *Statistical properties of dynamical systems with some hyperbolicity*, Ann. of Math. (2) **147** (1998), 585–650.
[260] L.-S. Young, *Recurrence times and rates of mixing*, Israel J. Math. **110** (1999), 153–188.
[261] R. Zimmer, *Ergodic Theory and Semisimple Groups*, Monographs in Mathematics, Vol. 81, Birkhäuser (1984).

This page intentionally left blank

CHAPTER 3

Stochastic-Like Behaviour in Nonuniformly Expanding Maps

Stefano Luzzatto

Mathematics Department, Imperial College, London, UK
E-mail: *stefano.luzzatto@imperial.ac.uk*
url: *htttp://www.ma.ic.ac.uk/~luzzatto*

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

1. Introduction

1.1. *Hyperbolic dynamics*

A feature of many real-life phenomena in areas as diverse as physics, biology, finance, economics, and many others, is the *random-like* behaviour of processes which nevertheless are clearly *deterministic*. On the level of applications this dual aspect has proved very problematic. Specific mathematical models tend to be developed either on the basis that the process is deterministic, in which case sophisticated numerical techniques can be used to attempt to understand and predict the evolution, or that it is random, in which case probability theory is used to model the process. Both approaches lose sight of what is probably the most important and significant characteristic of the system which is precisely that it is deterministic *and* has random-like behaviour. The theory of Dynamical Systems has contributed a phenomenal amount of work showing that it is perfectly natural for completely deterministic systems to behave in a very random-like way and achieving a quite remarkable understanding of the mechanisms by which this occurs.

We shall assume that the state space can be represented by a compact Riemannian manifold M and that the evolution of the process is given by a map $f : M \to M$ which is piecewise differentiable. Following an approach which goes back at least to the first half of the 20th century, we shall discuss how certain statistical properties can be deduced from geometrical assumptions on f formulated explicitly in terms of "*hyperbolicity*" assumptions on the *derivative map* Df of f. This is often referred to as *Hyperbolic Dynamics* or *Smooth Ergodic Theory*. The basic strategy is to construct certain geometrical structures (invariant manifolds, partitions) which imply some statistical/probabilistic properties of the dynamics. A striking and pioneering example of this is the work of Hopf on the ergodicity of geodesic flows on manifolds of negative curvature [80]. The subject has grown enormously since then to become of one of the key areas in the modern theory of Dynamical Systems. This is reflected in the present handbook in which several surveys, see [1–3,5, 7–9] address different facets of the theory in the case of *diffeomorphisms*.

The main focus of these notes will be on the analogous theory for *endomorphisms*. In this case the hyperbolicity conditions reduce to *expansivity* conditions. We shall concentrate here on three particular types of results about expanding maps: the existence of Markov structures, the existence of absolutely continuous invariant probability measures, and estimates on the rates of decay of correlations. See also [13] for a more detailed treatment of the theory and other results such as stochastic stability.

1.2. *Expanding dynamics*

We start with the basic definition of an expanding map.

DEFINITION 1. We say that $f : M \to M$ is (nonuniformly) *expanding* if there exists $\lambda > 0$ such that

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \log \left\| Df^{-1}_{f^i(x)} \right\|^{-1} > \lambda \qquad (*)$$

for almost every $x \in M$. Equivalently, for almost every $x \in M$ there exists a constant $C_x > 0$ such that

$$\prod_{i=0}^{n-1} \left\| Df^{-1}_{f^i(x)} \right\|^{-1} \geqslant C_x e^{\lambda n}$$

for every $n \geqslant 1$.

Notice that the condition $\log \|Df_x^{-1}\|^{-1} > 0$, which is equivalent to $\|Df_x^{-1}\|^{-1} > 1$ and which in turn is equivalent to $\|Df_x^{-1}\| < 1$, implies that *all* vectors in all directions are contracted by the inverse of Df_x and thus that *all* vectors in all directions are *expanded* by Df_x; the intuitively more obvious condition $\log \|Df_x\| > 0$, which is equivalent to $\|Df_x\| > 1$, implies only that there is *at least* one direction in which vectors are expanded by Df_x. Thus a map is nonuniformly expanding if every vector is asymptotically expanded at a uniform exponential rate. The constant C_x can in principle be arbitrarily small and indicates that an arbitrarily large number of iterates may be needed before this exponential growth becomes apparent.

The definition and the corresponding results can be generalized to the case in which the expansivity condition holds only on an invariant set of positive measure instead of on the entire manifold M. In the special case in which condition $(*)$ holds at *every* point x and the constant C can be chosen uniformly positive independent of x we say that f is *uniformly expanding*. Thus *uniformly expanding* is a special case of *nonuniformly expanding*. The terminology is slightly awkward for historical reasons: uniformly expanding maps have traditionally been referred to simply as expanding maps whereas this term should more appropriately refer to the more general (i.e. possibly nonuniformly) expanding case. We shall generally say that f is strictly nonuniformly expanding if f satisfies condition $(*)$ but is strictly *not* uniformly expanding. A basic theme of these notes is to discuss the difference between uniformly and nonuniformly expanding maps: how the nonuniformity affects the results and the ideas and techniques used in the proofs and how different *degrees of nonuniformity* can be quantified.

Nonuniform expansivity is a special case of *nonuniform hyperbolicity*. This concept was first formulated and studied by Pesin [128,129] and has since become one of the main areas of research in dynamical systems, see [46,168,27] and [1] by Barreira and Pesin in this volume, for extensive and in-depth surveys. The formal definition is in terms of *nonzero Lyapunov exponents* which means that the tangent bundle can be decomposed into subbundles in which vectors either contract or expand at an asymptotically exponential rate. Nonuniform *expansivity* corresponds to the case in which all the Lyapunov exponents are positive and therefore all vectors expand asymptotically at an exponential rate. The natural setting for this situation is that of (noninvertible) local diffeomorphisms whereas the theory of nonuniform hyperbolicity has been developed mainly for diffeomorphisms (however see also [142] and [1, Section 5.8]). For greater generality, and also because this has great importance for applications, we shall also allow various kinds of critical and/or singular points for f or its derivative.

1.3. *General overview of the notes*

We first review the basic notions of invariant measure, ergodicity, mixing, and decay of correlations in order to fix the notation and to motivate the results and techniques. In Section 3 we discuss the key idea of a *Markov Structure* and sketch some of the arguments used to study systems which admit such a structure. In Sections 4–6, we give a historical and technical survey of many classes of systems for which results are known, giving references to the original proofs whenever possible, and sketching in varying amounts of details the construction of Markov structures in such systems. In Section 7 we present some recent abstract results which go towards a general theory of nonuniformly expanding maps. In Section 8 we discuss the important problem of verifying the geometric nonuniform expansivity assumptions in specific classes of maps. Finally, in Section 9 we make some concluding remarks and present some open questions and conjectures.

The focus on Markov structures is partly a matter of personal preference; in some cases the results can be proved and/or were first proved using completely different arguments and techniques. Of particular importance is the so-called *Functional–Analytic* approach in which the problems are reformulated and reduced to questions about the spectrum of a certain linear operator on some functional space. There are several excellent survey texts focussing on this approach, see [23,101,163]. In any case, however, it is hard to see how the study of systems in which the hyperbolicity or expansivity is *nonuniform* can be carried out without constructing or defining some kind of subdivision into subsets on which relevant estimates satisfy uniform bounds. The Markov structures to be described below provide one very useful way in which this can be done and give some concrete geometrical structure. It seems very likely that these structures will prove useful in studying many other features of nonuniformly hyperbolic or expanding systems such as their stability, persistence, and even existence in particular settings. Another quite different way to partition a set satisfying nonuniform hyperbolicity conditions is with so-called *Pesin* or *regular* sets, see [1, Section 4.5]. These sets play a very useful role in the general theory of nonuniform hyperbolicity for diffeomorphisms, for example, in the construction of the stable and unstable foliations.

2. Basic definitions

In this section we review, for convenience, the basic definitions of invariant measure, ergodicity, mixing and decay of correlations. We shall also present these definitions in such a way as to motivate the results given later. A more extensive review of these ideas can be found in the general survey [4] in Vol. 1A of this handbook.

We shall always assume that M is a smooth, compact, Riemannian manifold of dimension $d \geqslant 1$. For simplicity we shall call the Riemannian volume *Lebesgue measure*, denote it by m or $|\cdot|$ and assume that it is normalized so that $m(M) = |M| = 1$. We let $f : M \to M$ denote a Lebesgue-measurable map. In practice we shall always assume significantly more regularity on f, e.g., that f is C^2 or at least piecewise C^2, but the main definitions apply in the more general case of f measurable. All measures on M will be assumed to be defined on the Borel σ-algebra of M.

2.1. *Invariant measures*

For a set $A \in M$ and a map $f : M \to M$ we define $f^{-1}(A) = \{x\colon\ f(x) \in A\}$.

DEFINITION 2. We say that a probability measure μ on M is *invariant* under f if

$$\mu\big(f^{-1}(A)\big) = \mu(A)$$

for every μ-measurable set $A \subset M$.

A given measure can be invariant for many different maps. For example Lebesgue measure on the circle $M = \mathbb{S}^1$ is invariant for the identity map $f(\theta) = \theta$, the rotation $f(\theta) = \theta + \alpha$ for any $\alpha \in \mathbb{R}$, and the covering map $f(\theta) = \kappa\theta$ for any $\kappa \in \mathbb{N}$. Similarly, for a given point $p \in M$, the Dirac-δ measure δ_p defined by

$$\delta_p(A) = \begin{cases} 1, & p \in A, \\ 0, & p \notin A, \end{cases}$$

is invariant for any map f for which $f(p) = p$. On the other hand, a given map f can admit many invariant measures. For example *any* probability measure is invariant for the identity map $f(x) = x$ and, more generally, any map which admits multiple fixed or periodic points admits as invariant measures the Dirac-δ measures supported on such fixed points or their natural generalizations distributed along the orbit of the periodic points. There exist also maps that do not admit any invariant probability measures. However some mild conditions, e.g., continuity of f, do guarantee that there exists at least one.

A first step in the application of the theory and methods of ergodic theory is to introduce some ways of distinguishing between the various invariant measures. We do this by introducing various properties which such measures may or may not satisfy. Unless we specify otherwise we shall use μ to denote a generic invariant probability measure for a given unspecified map $f : M \to M$.

2.2. *Ergodicity*

DEFINITION 3. We say that μ is *ergodic* if there does not exist a measurable set A with

$$f^{-1}(A) = A \quad \text{and} \quad \mu(A) \in (0, 1).$$

In other words, any fully invariant set A, i.e. a set for which $f^{-1}(A) = A$, has either zero or full measure. This is a kind of *indecomposability* property of the measure. If such a set existed, its complement $B = A^c$ would also be fully invariant and , in particular, both A and B would be also forward invariant: $f(A) = A$ and $f(B) = B$. Thus no point originating in A could ever intersect B and vice-versa and we essentially have two independent dynamical systems.

Simple examples such as the Dirac-δ_p measure on a fixed point p are easily shown to be ergodic, but in general this is a highly nontrivial property to prove. A lot of the techniques and methods to be described below are fundamentally motivated by the basic question of whether some relevant invariant measures are ergodic. It is known that Lebesgue measure is ergodic for circle rotations $f(\theta) = \theta + \alpha$ when α is irrational and for covering maps $f(\theta) = \kappa\theta$ when $\kappa \in \mathbb{N}$ is $\geqslant 2$ (the proof of ergodicity for the latter case will be sketched below). Irrational circle rotations are very special because they do not admit any other invariant measures besides Lebesgue measure. On the other hand covering maps have infinitely many periodic points and thus admit infinitely many invariant measures. It is sometimes easier to show that certain examples are not ergodic. This is clearly true for example for Lebesgue measure and the identity map since any subset is fully invariant. A less trivial example is the map $f : [0, 1] \to [0, 1]$ given by

$$f(x) = \begin{cases} 2x & \text{if } 0 \leqslant x \leqslant 1/4, \\ -2x + 1 & \text{if } 1/4 \leqslant x < 1/2, \\ 2x - 1/2 & \text{if } 1/2 \leqslant x \leqslant 3/4, \\ -2x + 5/2 & \text{if } 3/4 \leqslant x \leqslant 1. \end{cases}$$

Lebesgue measure is invariant, but the intervals $[0, 1/2)$ and $[1/2, 1]$ are both backward (and forward) invariant. This example can easily be generalized by defining two different Lebesgue measure preserving transformations mapping each of the two subintervals $[0, 1/2)$ and $[1/2, 1]$ into themselves.

The fundamental role played by the notion of ergodicity is due to the well-known and classical *Birkhoff Ergodic Theorem*. We give here only a special case of this result.

THEOREM [30,31]. *Let $f : M \to M$ be a measurable map and let μ be an ergodic invariant probability measure for f. Then, for any function $\varphi : M \to \mathbb{R}$ in $\mathcal{L}^1(\mu)$, i.e. such that $\int \varphi \, d\mu < \infty$, and for μ almost every x we have*

$$\frac{1}{n} \sum_{i=1}^{n} \varphi\big(f^i(x)\big) \to \int \varphi \, d\mu.$$

In particular, for any measurable set $A \subset M$, letting $\varphi = \mathbb{1}_A$ be the characteristic function of A, we have for μ almost every $x \in M$,

$$\frac{\#\{1 \leqslant j \leqslant n\colon\ f^j(x) \in A\}}{n} \to \mu(A). \tag{1}$$

Here $\#\{1 \leqslant j \leqslant n\colon\ f^j(x) \in A\}$ denotes the cardinality of the set of indices j for which $f^j(x) \in A$. Thus the *average proportion of time* which the orbit of a typical point spends in A converges *precisely* to the μ-measure of A. Notice that the convergence of this proportion as $n \to \infty$ is in itself an extremely remarkable and nonintuitive result. The fact that the limit is given a priori by $\mu(A)$ means in particular that this limit is *independent of the specific initial condition x*. Thus μ-almost every initial condition has the same statistical distribution in space and this distribution depends only on μ and not even on the map f, except implicitly for the fact that μ is ergodic and invariant for f.

2.3. *Absolute continuity*

Some care needs to be taken when applying Birkhoff's ergodic theorem to maps which admit several ergodic invariant measures. Consider, for example, the circle map $f(\theta) = 10\theta$. This maps preserves Lebesgue measure and also has several fixed points, e.g., $p = 0.2222\ldots$, on which we can consider the Dirac-δ_p measure. Both these measures are ergodic. Thus an application of Birkhoff's theorem says that "almost every" point spends an average proportion of time converging to $m(A)$ in the set A but also that "almost every" point spends an average proportion of time converging to $\delta_p(A)$ in the set A. If $m(A) \neq \delta_p(A)$ this may appear to generate a contradiction.

The crucial observation here is that the notion of *almost every* point is always understood *with respect to a particular measure*. Thus Birkhoff's ergodic theorem asserts that for a given measure μ there exists a set $\tilde{M} \subset M$ with $\mu(\tilde{M}) = 1$ such that the convergence property holds for every $x \in \tilde{M}$ and in general it may not be possible to identify $\tilde{M}$ explicitly. Conversely, if $X \in M$ satisfies $\mu(X) = 0$ then no conclusion can be drawn about whether (1) holds for any point of X. Returning to the example given above we have the following situation: the convergence (1) of the time averages to $\delta_p(A)$ can be guaranteed only for points belonging to a minimal set of full measure. But in this case this set reduces to the single point p for which (1) clearly holds. On the other hand the single point p clearly has zero Lebesgue measure and thus the convergence (1) to $m(A)$ is not guaranteed by Birkhoff's Theorem. Thus there is no contradiction.

An important point therefore is that the information provided by Birkhoff's ergodic theorem depends on the measure μ under consideration. Based on the premise that Lebesgue is the given "physical" measure and that we consider a satisfactory description of the dynamics one which accounts for a sufficiently large set of points from the point of view of Lebesgue measure, it is clear that if μ is a Dirac-δ measure on a fixed point it gives essentially no useful information. On the other hand, if μ is Lebesgue measure itself then we do get a convergence result that holds for Lebesgue almost every starting condition. The invariance of Lebesgue measure is a very special property but much more generally we can ask about the existence of ergodic invariant measures μ which are *absolutely continuous* with respect to m.

DEFINITION 4. μ is absolutely continuous with respect to m if

$$m(A) = 0 \quad \text{implies} \quad \mu(A) = 0$$

for every measurable set $A \subset M$.

In this case, Birkhoff's theorem implies that (1) holds for all points belonging to a set $\tilde{M} \subset M$ with $\mu(\tilde{M}) = 1$ and the absolute continuity of μ with respect to m therefore implies that $m(\tilde{M}) > 0$. Thus the existence of an ergodic *absolutely continuous invariant probability* (*acip*) μ implies some control over the asymptotic distribution of at least a set of positive Lebesgue measure. It also implies that such points tend to have a dynamics which is nontrivial in the sense that it is distributed over some relatively large subset of the space as opposed to converging, for example, to some attracting fixed point. Thus it

indicates that there is a minimum amount of inherent *complexity* as well as *structure*. This motivates the basic question:

1. Under what conditions does f admit an ergodic acip?

This question is already addressed explicitly by Hopf [79] for invertible transformations. Interestingly he formulates some conditions in terms of the existence of what are essentially some induced transformations, similar in some respects to the Markov structures to be defined below.[1] In these notes we shall discuss what is effectively a generalization of this basic approach.

2.4. *Mixing*

Birkhoff's ergodic theorem is very powerful but it is easy to see that the asymptotic space distribution given by (1) does not necessarily tell the whole story about the dynamics of a given map f. Indeed these conclusions depend not on f but simply on the fact that Lebesgue measure is invariant and ergodic. Thus from this point of view the dynamics of an irrational circle rotation $f(\theta) = \theta + \alpha$ and of the map $f(\theta) = 2\theta$ are indistinguishable. However it is clear that they give rise to very different kinds of dynamics. In one case, for example, nearby points remain nearby for all time, whereas in the other they tend to move apart at an exponential speed. This creates a kind of *unpredictability* in one case which is not present in the other.

DEFINITION 5. We say that an invariant probability measure μ is *mixing* if

$$\left|\mu\big(A \cap f^{-n}(B)\big) - \mu(A)\mu(B)\right| \to 0$$

as $n \to \infty$, for all measurable sets $A, B \subseteq M$.

Notice that mixing implies ergodicity and is therefore a stronger property. Thus a natural follow up to question 1 is the following. Suppose that f admits an ergodic *acip* μ.

2. Under what conditions is μ mixing?

Early work in ergodic theory in the 1940's considered the question of the genericity of the mixing property is in various spaces of systems [75,137,138,78,89] but, as with ergodicity, in specific classes of systems it is generally easier to show that a system is not mixing rather than that it is mixing. For example it is immediate that irrational circle rotations are not mixing. On the other hand it is nontrivial that maps of the form $f(\theta) = \kappa\theta$ for integers $\kappa \geqslant 2$ are mixing.

To develop an intuition for the concept of mixing, notice that mixing is equivalent to the condition

$$\left|\frac{\mu(A \cap f^{-n}(B))}{\mu(B)} - \mu(A)\right| \to 0$$

[1] Hopf's result is the following: suppose that for every measurable partition $\mathcal{P}$ of the manifold M and every stopping time function p such that the images $f^{p(\omega)}(\omega)$ for $\omega \in \mathcal{P}$ are all disjoint, the union of all images has full measure. Then f admits an absolutely continuous invariant probability measure.

as $n \to \infty$, for all measurable sets $A, B \subseteq M$, with $\mu(B) \neq 0$. In this form there are two natural interpretations of mixing, one geometrical and one probabilistic. From a geometrical point of view, recall that $\mu(f^{-n}(B)) = \mu(B)$ by the invariance of the measure. Then one can think of $f^{-n}(B)$ as a "redistribution of mass" and the mixing condition says that for large n the proportion of $f^{-n}(B)$ which intersects A is just proportional to the measure of A. In other words $f^{-n}(B)$ is spreading itself uniformly with respect to the measure μ. A more probabilistic point of view is to think of $\mu(A \cap f^{-n}(B))/\mu(B)$ as the conditional probability of having $x \in A$ given that $f^n(x) \in B$, i.e. the probability that the occurrence of the event B today is a consequence of the occurrence of the event A n steps in the past. The mixing condition then says that this probability converges to the probability of A, i.e. asymptotically, there is no causal relation between the two events. This is why we say that a mixing system exhibits *stochastic-like* or *random-like* behaviour.

2.5. *Decay of correlations*

It turns out that mixing is indeed a quite generic property at least under certain assumptions which will generally hold in the examples we shall be interested in. Thus apparently very different systems admit mixing *acip*'s and become, in some sense, statistically indistinguishable at this level of description. Thus it is natural to want to dig deeper in an attempt relate finer statistical properties with specific geometric characteristics of systems under considerations. One way to do this is to try to distinguish systems which mix at different *speeds*. To formalize this idea we need to generalize the definition of mixing. Notice first of all that the original definition can be written in integral form as

$$\left| \int \mathbb{1}_{A \cap f^{-n}(B)} \, d\mu - \int \mathbb{1}_A \, d\mu \int \mathbb{1}_B \, d\mu \right| \to 0,$$

where $\mathbb{1}_X$ denotes the characteristic function of the set X. This can be written in the equivalent form

$$\left| \int \mathbb{1}_A (\mathbb{1}_B \circ f^n) \, d\mu - \int \mathbb{1}_A \, d\mu \int \mathbb{1}_B \, d\mu \right| \to 0$$

and this last formulation now admits a natural generalization by replacing the characteristic functions with arbitrary measurable functions.

DEFINITION 6. For real valued measurable functions $\varphi, \psi : M \to \mathbb{R}$ we define the *correlation function*[2]

$$\mathcal{C}_n(\varphi, \psi) = \left| \int \psi (\varphi \circ f^n) \, d\mu - \int \psi \, d\mu \int \varphi \, d\mu \right|.$$

[2]The derivation of the correlation function from the definition of mixing as given here does not perhaps correspond to the historical development. I believe that the notion of decay of correlation arose in the context of statistical mechanics and was not directly linked to abstract dynamical systems framework until the work of Bowen, Lebowitz, Ruelle and Sinai in the 1960's and 1970's [152,32,153,126,33].

In this context, the functions φ and ψ are often called *observables*. If μ is mixing, the correlation function decays to zero whenever the observables ϕ, ψ are characteristic functions. It is possible to show that indeed it decays also for many other classes of functions. We then have the following very natural question. Suppose that the measure μ is mixing, fix two observables φ, ψ, and let $\mathcal{C}_n = \mathcal{C}_n(\varphi, \psi)$.

3. Does $\mathcal{C}_n$ decay at a specific *rate* depending only on f?

The idea behind this question is that a system may have an intrinsic *rate of mixing* which reflects some characteristic geometrical structures. It turns out that an intrinsic rate does sometimes exist and is in some cases possible to determine, but only by restricting to a suitable class of observables. Indeed, a classical result says that even in the "best" cases it is possible to choose subsets A, B such that the correlation function $\mathcal{C}_n(\mathbb{1}_A, \mathbb{1}_B)$ of the corresponding characteristic functions decays at an arbitrarily slow rate. Instead positive results exist in many cases by restricting to, for example, the space of observables of bounded variation, or Hölder continuous, or even continuous with non-Hölder modulus of continuity. Once the space $\mathcal{H}$ of observables has been fixed, the goal is to show that there exists a sequence $\gamma_n \to 0$ (e.g., $\gamma_n = e^{-\alpha n}$ or $\gamma_n = n^{-\alpha}$ for some $\alpha > 0$) depending only on f and $\mathcal{H}$, such that for any two $\varphi, \psi \in \mathcal{H}$ there exists a constant $C = C(\varphi, \psi)$ (generally depending on the observables φ, ψ) such that

$$\mathcal{C}_n \leqslant C\gamma_n$$

for all $n \geqslant 1$. Ideally we would like to show that $\mathcal{C}_n$ actually decays like γ_n, i.e. to have both lower and upper bounds, but this is known only in some very particular cases. Most known results at present are upper bounds and thus when we say that the correlation functions decays at a certain rate we will usually mean that it decays *at least* at that rate. Also, most known results deal with Hölder continuous observables and thus, to simplify the presentation, we shall assume that we are dealing with this class unless we mention otherwise.

We shall discuss below several examples of systems whose correlations decay at different rates, for example, *exponential*, *polynomial* or even *logarithmic*, and a basic theme of these notes will be gain some understanding about *how* and *why* such differences occur and what this tells us about the system.

3. Markov structures

In this section we define the notion of a "Markov structure" and sketch the proof of the fact that the existence of a Markov structure implies the existence of an absolutely continuous invariant probability measure.

DEFINITION 7. $f : M \to M$ admits (or, is) a *Markov map* if there exists a finite or countable partition $\mathcal{P}$ (mod 0) of M into open sets with smooth boundaries such that $f(\omega) = M$ for every partition element $\omega \in \mathcal{P}$ and $f|_\omega$ is a continuous nonsingular bijection.

We recall that a partition mod 0 of M means that Lebesgue almost every point belongs to the interior of some partition elements. Also, $f|_\omega$ is nonsingular if $|A| > 0$ implies

$|f(A)| > 0$ for every (measurable) $A \subset \omega$. These two conditions together immediately imply that the full forward orbit of almost every point always lies in the interior of some partition element. The condition $f(\omega) = M$ is a particularly strong version of what is generally referred to as the Markov property where it is only required that the image of each ω be a continuous nonsingular bijection onto some union of partition elements and not necessarily all of M; see, for example, [4] for general definitions. The stronger requirement we use here is sometimes called a Bernoulli or Gibbs–Markov property. A significant generalization of this definition allows the partition element to be just measurable sets and not necessarily open; the general results to be given below apply in this case also. However we shall not need this for any of the applications which we shall discuss.

A natural but extremely far-reaching generalization of the notion of a Markov map is the following

DEFINITION 8. $f : M \to M$ admits an *induced Markov map* if there exists an open set $\Delta \subset M$, a partition $\mathcal{P}$ (mod 0) of Δ and a return time function $R : \Delta \to \mathbb{N}$, piecewise constant on each element of $\mathcal{P}$, such that the induced map $F : \Delta \to \Delta$ defined by $F(x) = f^{R(x)}$ is a Markov map.

Again, the condition that Δ is open is not strictly necessary. Clearly if f is Markov to begin with, it trivially admits an induced Markov map with $\Delta = M$ and $R \equiv 1$. However, as we shall see below, this is a much more general definition and many systems turn out to admit an induced Markov map. The notion of induced Markov map is a fairly classical notion in ergodic theory. The first use of this concept, in the specific context of nonuniformly expanding dynamics, is probably due to Jakobson in the late 1970's and early 1980's, see Section 3.2 in [6]. Since then it has played an increasingly important role in the theory and, in some sense, the main theme of this survey is precisely to describe the development of the theory of nonuniformly expanding maps from this point of view. Recent results suggest that the existence of an induced Markov map (with some additional conditions to be stated below) is essentially equivalent to nonuniform expansivity.

For the rest of this section we shall suppose that $F : \Delta \to \Delta$ is an induced Markov map associated to some map $f : M \to M$. We call $\mathcal{P}$ the Markov partition associated to the Markov map F. Since $\mathcal{P}$ is assumed to be countable, we can define an indexing set $\omega = \{0, 1, 2, \ldots\}$ of the Markov partition $\mathcal{P}$. Then, for any finite sequence $a_0 a_1 a_2 \ldots a_n$ with $a_i \in \mathcal{I}$, we can define the *cylinder set of order n* by

$$\omega^{(n)}_{a_0 a_1 \ldots a_n} \left\{ x\colon\ F^i(x) \in \omega_{a_i} \text{ for } 0 \leqslant i \leqslant n \right\}.$$

Inductively, given $\omega_{a_0 a_1 \ldots a_{n-1}}$, then $\omega_{a_0 a_1 \ldots a_n}$ is the part of $\omega_{a_0 a_1 \ldots a_{n-1}}$ mapped to ω_{a_n} by F^n. The cylinder sets define refinements of the partition $\mathcal{P}$. We let $\omega^{(0)}$ denote generic elements of $\mathcal{P}^{(0)} = \mathcal{P}$ and $\omega^{(n)}$ denote generic elements of $\mathcal{P}^{(n)}$. Notice that by the nonsingularity of the map F on each partition element and the fact that $\mathcal{P}$ is a partition mod 0, it follows that each $\mathcal{P}^{(n)}$ is also a partition mod 0 and that Lebesgue almost every point in Δ falls in the interior of some partition element of $\mathcal{P}$ for all future iterates. In particular almost every $x \in \Delta$ has an associated infinite symbolic sequence $\underline{a}(x)$ determined by the future iterates of x in relation to the partition $\mathcal{P}$. To get the much more sophisticated results on the statistical properties of f we need first of all the following two additional conditions.

DEFINITION 9. $F : \Delta \to \Delta$ has *integrable* or *summable* return times if

$$\int_\Delta R(x)\,dx = \sum_{\omega \in P} |\omega| R(\omega) < \infty.$$

DEFINITION 10. $F : \Delta \to \Delta$ has the (volume) *bounded distortion* property if there exists a constant $\mathcal{D} > 0$ such that for all $n \geqslant 1$ and any measurable subset $\tilde{\omega}^{(n)} \subset \omega^{(n)} \in \mathcal{P}^{(n)}$ we have

$$\frac{1}{\mathcal{D}} \frac{|\tilde{\omega}^{(n)}|}{|\omega^{(n)}|} \leqslant \frac{|f^n(\tilde{\omega}^{(n)})|}{|f^n(\omega^{(n)})|} \leqslant \mathcal{D} \frac{|\tilde{\omega}^{(n)}|}{|\omega^{(n)}|}. \tag{2}$$

This means that the relative measure of subsets of a cylinder set of any level n are preserved up to some factor $\mathcal{D}$ under iteration by f^n. A crucial observation here is that the constant $\mathcal{D}$ is independent of n. Thus in some sense the geometrical structure of any subset of Δ reoccurs at every scale inside each partition element of $\mathcal{P}^{(n)}$ up to some bounded distortion factor. This is in principle a very strong condition but we shall see below that it is possible to verify it in many situations. We shall discuss in the next section some techniques for verifying this condition in practice. First of all we state the first result of this section.

THEOREM 1. *Suppose that $f : M \to M$ admits an induced Markov map satisfying the bounded distortion property and having summable return times. Then it admits an ergodic absolutely continuous invariant probability measure μ.*

This result goes back to the 1950's and is often referred to as the *Folklore Theorem* of dynamics. We will sketch below the main ideas of the proof, see also [6, Section 3.1] for a particularly compact proof.

First we state a much more recent result which applies in the same setting but takes the conclusions much further in the direction of mixing and rates of decay of correlations. First of all we shall assume without loss of generality that the greatest common divisor of all values taken by the return time function R is 1. If this were not the case all return times would be multiples of some integer $k \geqslant 2$ and the measure μ given by the theorem stated above would clearly not be mixing. If this is the case however, we could just consider the map $\tilde{f} = f^k$ and the results to be stated below will apply to $\tilde{f}$ instead of f. We define the *tail of the return times* as the measure of the set

$$R_n = \{x \in \Delta\colon\ R(x) > n\}$$

of points whose return times is strictly larger than n. The integrability condition implies that $R(x) < \infty$ for almost every point and thus

$$|R_n| \to 0$$

as $n \to \infty$. However there is a range of possible *rates* of decay of $|R_n|$ all of which are compatible with the integrability condition. L.-S. Young observed and proved that a bound on the decay of correlations for Hölder continuous observables can be obtained from bounds on the rate of decay of the tail of the return times.

THEOREM 2 [169,170]. *Suppose that $f : M \to M$ admits an induced Markov map satisfying the bounded distortion property and having summable return times. Then it admits an ergodic* (*and mixing*) *absolutely continuous invariant probability measure. Moreover, the correlation function for Hölder continuous observables satisfies the following bounds*:

Exponential tail: *If $\exists \alpha > 0$ such that $|R_n| = \mathcal{O}(e^{-\alpha n})$, then $\exists \tilde{\alpha} > 0$ such that $\mathcal{C}_n = \mathcal{O}(e^{-\tilde{\alpha} n})$.*

Polynomial tail: *If $\exists \alpha > 1$ such that $|R_n| = \mathcal{O}(n^{-\alpha})$, then $\mathcal{C}_n = \mathcal{O}(n^{-\alpha+1})$.*

Other papers have also addressed the question of the decay of correlations for similar setups mainly using spectral operator methods [169,36,115,37,116]. We remark that the results about the rates of decay of correlations generally require an a priori slightly stronger form of bounded distortion than that given in (2). The proof in [170] uses a very geometrical/probabilistic *coupling argument* which appears to be quite versatile and flexible. Variations of the argument have been applied to prove the following generalizations which apply in the same setting as above (in both cases we state only a particular case of the theorems proved in the cited papers).

The first one extends Young's result to arbitrarily slow rates of decay. We say that $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ is *slowly varying* (see [10]) if for all $y > 0$ we have $\lim_{x\to\infty} \rho(xy)/\rho(x) = 0$. A simple example of a slowly varying function is the function $\rho(x) = e^{(\log x)/(\log\log x)}$. Let $\hat{R}_n = \sum_{\hat{n} \geqslant n} R_{\hat{n}}$.

THEOREM 3 [77]. *The correlation function for Hölder continuous observables satisfies the following bound*:

Slowly varying tail: *If $\hat{R}_n = \mathcal{O}(\rho(n))$ where ρ is a monotonically decreasing to zero, slowly varying, C^∞ function, then $\mathcal{C}_n = \mathcal{O}(\rho(n))$.*

The second extends Young's result to observables with very weak, non-Hölder, modulus of continuity. We say that $\psi : I \to \mathbb{R}$ has a *logarithmic modulus of continuity γ* if there exists $C > 0$ such that for all $x, y \in I$ we have

$$\big|\psi(x) - \psi(y)\big| \leqslant C\big|\log|x - y|\big|^{-\gamma}.$$

For both the exponential and polynomial tail situations we have the following

THEOREM 4 [110]. *There exists $\alpha > 0$ such that for all γ sufficiently large and observables with logarithmic modulus of continuity γ, we have $\mathcal{C}_n = \mathcal{O}(n^{-\alpha})$.*

These general results indicate that the rate of decay of correlations is linked to what is in effect the *geometrical structure* of f as reflected in the tail of the return times for the

induced map F. From a technical point of view they shift the problem of the statistical properties of f to the problem of the geometrical structure of f and thus to the (still highly nontrivial) problem of showing that f admits an induced Markov map and of estimating the tail of the return times of this map. The construction of an induced map in certain examples is relatively straightforward and essentially canonical but the most interesting constructions require statistical arguments to even show that such a map exists and to estimate the tail of the return times. In these cases the construction is not canonical and it is usually not completely clear to what extent the estimates might depend on the construction.

We now give a sketch of the proof of Theorem 1. The proofs of Theorems 2–4 are in a similar spirit and we refer the interested reader to the original papers. We assume throughout the next few sections that $F : \Delta \to \Delta$ is the Markov induced map associated to $f : I \to I$ and $\mathcal{P}^{(n)}$ are the family of cylinder sets generated by the Markov partition $\mathcal{P} = \mathcal{P}^{(0)}$ of Δ. We first define a measure ν on Δ and show in that it is F-invariant, ergodic, and absolutely continuous with respect to Lebesgue. Then we define the measure μ on I in terms of ν and show that it is f-invariant, ergodic, and absolutely continuous.

3.1. *The invariant measure for F*

We start with a preliminary result which is a consequence of the bounded distortion property.

3.1.1. *The measure of cylinder sets* A straightforward but remarkable consequence of the bounded distortion property is that the measure of cylinder sets tends to zero uniformly.

LEMMA 3.1.

$$\max\{|\omega^{(n)}|;\ \omega^{(n)} \in \mathcal{P}^{(n)}\} \to 0 \quad \text{as } n \to 0.$$

Notice that in the one-dimensional case, the measure of an interval coincides with its diameter and so this implies in particular that the diameter of cylinder sets tends to zero, implying the essential uniqueness of the symbolic representation of itineraries.

PROOF. It is sufficient to show that there exists a constant $\tau \in (0, 1)$ such that for every $n \geqslant 0$ and every $\omega^{(n)} \subset \omega^{(n-1)}$ we have

$$|\omega^{(n)}| \big/ |\omega^{(n-1)}| \leqslant \tau. \tag{3}$$

Applying this inequality recursively then implies $|\omega^{(n)}| \leqslant \tau |\omega^{(n-1)}| \leqslant \tau^2 |\omega^{(n-2)}| \leqslant \cdots \leqslant \tau^n |\omega^0| \leqslant \tau^n |\Delta|$. To verify (3) we shall show that

$$1 - \frac{|\omega^{(n)}|}{|\omega^{(n-1)}|} = \frac{|\omega^{(n-1)}| - |\omega^{(n)}|}{|\omega^{(n)}|} = \frac{|\omega^{(n-1)} \setminus \omega^{(n)}|}{|\omega^{(n)}|} \geqslant 1 - \tau. \tag{4}$$

To prove (4) let first of all $\delta = \max_{\omega \in \mathcal{P}} |\omega| < |\Delta|$. Then, from the definition of cylinder sets we have that $F^n(\omega^{(n-1)}) = \Delta$ and that $F^n(\omega^{(n)}) \in \mathcal{P} = \mathcal{P}^{(0)}$, and therefore $|F^n(\omega^{(n)})| \leqslant \delta$

or, equivalently, $|F^n(\omega^{(n-1)} \setminus \omega^{(n)})| \geqslant |\Delta| - \delta > 0$. Thus, using the bounded distortion property we have

$$\frac{|\omega^{(n-1)} \setminus \omega^{(n)}|}{|\omega^{(n)}|} \geqslant \frac{1}{\mathcal{D}} \frac{|F^n(\omega^{(n-1)} \setminus \omega^{(n)})|}{|F^n(\omega^{(n)})|} \geqslant \frac{|\Delta| - \delta}{|\Delta|\mathcal{D}}$$

and (4) follows choosing $\tau = 1 - ((|\Delta| - \delta)/|\Delta|\mathcal{D})$. □

The next property actually follows only from the conclusions of Lemma 3.1 rather than from the bounded distortion property itself. It essentially says that it is possible to "zoom in" to any given set of positive measure.

LEMMA 3.2. *For any $\varepsilon > 0$ and any Borel set A with $|A| > 0$ there exists $n \geqslant 1$ and $\omega^{(n)} \in \mathcal{P}^{(n)}$ such that*

$$\left|A \cap \omega^{(n)}\right| \geqslant (1 - \varepsilon)\left|\omega^{(n)}\right|.$$

PROOF. Fix some $\varepsilon > 0$. Suppose first of all that A is compact. Then, using the properties of Lebesgue measure it is possible to show that for any $\eta > 0$ there exists an integer $n \geqslant 1$ and a collection $\mathcal{I}_\eta = \{\omega^n\} \subset \mathcal{P}^{(n)}$ such that $A \subset \bigcup_{\omega_\eta} \omega^{(n)}$ and $|\omega_\eta| \leqslant |A| + \eta$. Now suppose by contradiction that $|\omega^{(n)} \cap A| \leqslant (1 - \varepsilon)|\omega^{(n)}|$ for every $\omega^{(n)} \in \omega_\eta$ for any given $\eta > 0$. Using that fact that the $\omega^{(n)} \in \omega_\eta$ are disjoint and thus $\sum |\omega^{(n)}| = |\omega_\eta|$, this implies that

$$|A| = \sum_{\omega^{(n)} \in \omega_\eta} \left|\omega^{(n)} \cap A\right| \leqslant (1 - \varepsilon) \sum_{\omega^{(n)} \in \omega_\eta} \left|\omega^{(n)}\right| \leqslant (1 - \varepsilon)\big(|A| + \eta\big).$$

Since η can be chosen arbitrarily small after fixing ε this gives a contradiction. If A is not compact we can approximate if from below in measure by compact sets and repeat essentially the same argument. □

3.1.2. *Absolute continuity* The following estimate also follows immediately from the bounded distortion property. It says that the absolute continuity property of F on partition elements is preserved up to arbitrary scale with uniform bounds.

LEMMA 3.3. *Let $A \subset \Delta$ and $n \geqslant 1$. Then*

$$\left|F^{-n}(A)\right| \leqslant \mathcal{D}|A|.$$

PROOF. The Markov property implies that $F^{-n}(A)$ is a union of disjoint sets each contained in the interior of some element $\omega^{(n)} \in \mathcal{P}^{(n)}$. Moreover, each $\omega^{(n)}$ is mapped by F^n to Δ with uniformly bounded distortion, thus we have $|F^{-n}(A) \cap \omega^{(n)}|/\omega^{(n)} \leqslant \mathcal{D}|A|/|\Delta|$ or, equivalently, $|F^{-n}(A) \cap \omega^{(n)}| \leqslant \mathcal{D}|A||\omega^{(n)}|/|\Delta|$. Therefore

$$\left|F^{-n}(A)\right| = \sum_{\omega^{(n)} \in \mathcal{P}^{(n)}} \left|F^{-n}(A) \cap \omega^{(n)}\right| \leqslant \frac{\mathcal{D}|A|}{|\Delta|} \sum_{\omega^{(n)} \in \mathcal{P}^{(n)}} \left|\omega^{(n)}\right| = \mathcal{D}|A|. \qquad \square$$

3.1.3. *The pull-back of a measure* For any $n \geqslant 1$ and Borel $A \subseteq \Delta$, let

$$\nu_n(A) = \frac{1}{n} \sum_{i=0}^{n} \big|F^{-i}(A)\big|.$$

It is easy to see that ν_n is a probability measure on Δ and absolutely continuous with respect to Lebesgue. Moreover, Lemma 3.3 implies that the absolute continuity property is uniform in n and A in the sense that $\nu_n(A) \leqslant \mathcal{D}|A|$ for any A and for any $n \geqslant 1$. By some standard results of functional analysis, this implies the following

LEMMA 3.4. *There exists a probability measure ν and a subsequence $\{\nu_{n_k}\}$ such that, for every measurable set A,*

$$\nu_{n_k}(A) \to \nu(A) \leqslant \mathcal{D}|A|. \tag{5}$$

In particular A is absolutely continuous with respect to Lebesgue.

3.1.4. *Invariance* To show that ν is F-invariant, let $A \subset \Delta$ be a measurable set. Then, by (5) we have

$$\begin{aligned}\nu\big(F^{-1}(A)\big) &= \lim_{k\to\infty} \frac{1}{n_k} \sum_{i=0}^{n_k-1} \big|F^{-(i+1)}(A)\big| \\ &= \lim_{k\to\infty} \left[\frac{1}{n_k} \sum_{i=0}^{n_k-1} \big|F^{-1}(A)\big| - \frac{|A|}{n_k} + \frac{|F^{-n_k}(A)|}{n_k}\right].\end{aligned}$$

Since $|A|$ and $|f^{-n}(A)|$ are both uniformly bounded by 1, we have $|A|/n_k \to 0$ and $|F^{-n_k}(A)|/n_k \to 0$ as $k \to 0$. Therefore

$$\begin{aligned}\nu\big(F^{-1}(A)\big) &= \lim_{k\to\infty} \left[\frac{1}{n_k} \sum_{i=0}^{n_k-1} \big|F^{-1}(A)\big| - \frac{|A|}{n_k} + \frac{|F^{-n_k}(A)|}{n_k}\right] \\ &= \lim_{k\to\infty} \frac{1}{n_k} \sum_{i=0}^{n_k-1} \big|F^{-1}(A)\big| = \nu(A).\end{aligned}$$

Therefore ν is F-invariant.

3.1.5. *Ergodicity and uniqueness* Let $A \subset I$ be a measurable set with $F^{-1}(A) = A$ and $\mu(A) > 0$. We shall show that $\nu(A) = |A| = 1$. This implies both ergodicity and uniqueness of ν. Indeed, if $\tilde{\nu}$ were another such measure invariant absolutely continuous measure, there would be have to be a set B with $F^{-1}(B) = B$ and $\tilde{\nu}(B) = 1$. But in this case we

would have also $|B| = 1$ and thus $A = B \bmod 0$. This is impossible since two absolutely continuous invariant measures must have disjoint support.

To prove that $|A| = 1$, let $A^c = \Delta \setminus A$ denote the complement of A. Notice that $x \in A^c$ if and only if $F(x) \in A^c$ and therefore $F(A^c) = A^c$. By Lemma 3.2, for any $\varepsilon > 0$ there exists some $n \geqslant 1$ and $\omega^{(n)} \in \mathcal{P}^{(n)}$ such that $|A \cap \omega^{(n)}| \geqslant (1-\varepsilon)|\omega^{(n)}|$ and therefore

$$\left|A^c \cap \omega^{(n)}\right| \leqslant \varepsilon \left|\omega^{(n)}\right|.$$

Using that fact that $F^n(\omega^{(n)}) = I$ and the invariance of A^c have $F^n(\omega^{(n)} \cap A^c) = A^c$. The bounded distortion property then gives

$$|A^c| = \frac{|F^n(\omega^{(n)} \cap A^c)|}{|F^n(\omega^{(n)})|} \leqslant \mathcal{D}\frac{|\omega^{(n)} \cap A^c|}{|\omega^{(n)}|} \leqslant \mathcal{D}\varepsilon.$$

Since ε is arbitrary this implies $|A^c| = 0$ and thus $|A| = 1$.

3.2. *The invariant measure for f*

We now show how to define a probability measure μ which is invariant for the original map f and satisfies all the required properties.

3.2.1. *The probability measure μ* We let ν_ω denote the restriction of ν to the partition element $\omega \in \mathcal{P}$, i.e. for any measurable set $A \subset \Delta$ we have $\nu_\omega(A) = \nu(A \cap \omega)$. Then $\nu(A) = \sum_{\omega \in \mathcal{P}} \nu_\omega(A)$. Then, for any measurable set $A \subseteq M$ (we no longer restrict our attention to Δ) we define

$$\hat{\mu}(A) = \sum_{\omega \in \mathcal{P}} \sum_{j=0}^{R(\omega)-1} \nu_\omega\big(f^{-j}(A)\big).$$

Notice that this is a sum of nonnegative terms and is uniformly bounded since

$$\begin{aligned}\hat{\mu}(A) \leqslant \hat{\mu}(M) &= \sum_{\omega \in \mathcal{P}} \sum_{j=0}^{R(\omega)-1} \nu_\omega\big(f^{-j}(M)\big) = \sum_{\omega \in \mathcal{P}} \sum_{j=0}^{R(\omega)-1} \nu_\omega(M) \\ &= \sum_{\omega \in \mathcal{P}} R(\omega)\nu(\omega) < \infty\end{aligned}$$

by the assumption on the summability of the return times. Thus it defines a finite measure on M and from this we define a probability measure by normalizing to get

$$\mu(A) = \hat{\mu}(A)/\hat{\mu}(M).$$

3.2.2. *Absolute continuity* The absolute continuity of μ is an almost immediate consequence of the definition and the absolute continuity of μ. Indeed, $|A| = 0$ implies $\nu(A) = 0$ which implies $\nu_\omega(A) = 0$ for all $I\omega \in \mathcal{P}$, which therefore implies that we have

$$\sum_{j=0}^{R(\omega)-1} \nu_\omega\big(f^{-j}(A)\big) = 0$$

and therefore $\mu(A) = 0$.

3.2.3. *Invariance* Recall first of all that by definition $f^{R(\omega)}(\omega) = \Delta$ for any $\omega \in \mathcal{P}$. Therefore, for any $A \subset M$ we have

$$f^{-R(\omega)}(A) \cap \omega = F|_\omega^{-1}(A) \cap \omega,$$

where $F|_\omega^{-1}$ denotes the inverse of the restriction $F|_\omega$ of F to ω (notice that $f^{-R(\omega)}(A) \cap \omega = \emptyset$ if $A \cap \Delta = \emptyset$). In particular, using the invariance of ν under F, this gives

$$\sum_{\omega\in\mathcal{P}} \nu\big(f^{-R(\omega)}(A) \cap \omega\big) = \sum_{\omega\in\mathcal{P}} \nu\big(F^{-1}|_\omega(A) \cap \omega\big) = \nu\big(F^{-1}(A)\big) = \nu(A).$$

Using this equality we get, for any measurable set $A \subseteq I$,

$$\begin{aligned}
\mu\big(f^{-1}(A)\big) &= \sum_{\omega\in\mathcal{P}} \sum_{j=0}^{R(\omega)-1} \nu_\omega\big(f^{-(j+1)}(A)\big) \\
&= \sum_{\omega\in\mathcal{P}} \sum_{j=0}^{R(\omega)-1} \nu\big(f^{-(j+1)}(A) \cap \omega\big) \\
&= \sum_{\omega\in\mathcal{P}} \nu\big[\big(f^{-1}(A) \cap \omega\big) + \cdots + \big(f^{-R(\omega)}(A) \cap \omega\big)\big] \\
&= \sum_{\omega\in\mathcal{P}} \sum_{j=1}^{R(\omega)-1} \nu\big(f^{-j}(A) \cap \omega\big) + \sum_{\omega\in\mathcal{P}} \big(f^{-R(\omega)}(A) \cap \omega\big) \\
&= \sum_{\omega\in\mathcal{P}} \sum_{j=1}^{R(\omega)-1} \nu\big(f^{-j}(A) \cap \omega\big) + \nu(A) \\
&= \sum_{\omega\in\mathcal{P}} \sum_{j=0}^{R(\omega)-1} \nu\big(f^{-j}(A) \cap \omega\big) = \mu(A).
\end{aligned}$$

3.2.4. *Ergodicity and uniqueness* Ergodicity of μ follows immediately from the ergodicity of ν since every fully invariant set for of positive measure must intersect the image

of some partition element ω and therefore must have positive (and therefore full) measure for ν and therefore must have full measure for μ. Notice however that we can only claim a limited form of uniqueness for the measure μ. Indeed, the support of μ is given by

$$\text{supp}(\mu) = \bigcup_{\omega \in \mathcal{P}} \bigcup_{j=0}^{R_k - 1} f^j(I_k)$$

which is the union of all the images of all partition elements. Then μ is indeed the unique ergodic absolutely continuous invariant measure on this set. However in a completely abstract setting there is no way of saying that $\text{supp}(\mu) = M$ nor that there may not be other relevant measures in $M \setminus \text{supp}(\nu)$.

3.3. *Expansion and distortion estimates*

The application of the abstract results discussed above to specific examples involves three main steps:

- Combinatorial construction of the induced map;
- Verification of the bounded distortion property;
- Estimation of the tail of the return times function and verification of the integrability of the return times.

We shall discuss some of these step in some detail in relation to some of the specific case as we go through them below. Here we just make a few remarks concerning the bounded distortion property and in particular the crucial role played by *regularity* and *derivative* conditions in these calculations.

We begin with a quite general observation which relates the bounded distortion condition to a property involving the derivative of F. Let $F : \Delta \to \Delta$ be a Markov map which is continuously differentiable on each element of the partition $\mathcal{P}$. We let $\det DF^n$ denote the determinant of the derivative of the map F^n.

DEFINITION 11. We say that F has *uniformly bounded derivative distortion* if there exists a constant $\mathcal{D} > 0$ such that for all $n \geqslant 1$ and $\omega \in P^{(n)}$ we have

$$\text{Dist}(f^n, \omega) := \max_{x,y \in I^{(n)}} \log \frac{\det DF^n(x)}{\det DF^n(y)} \leqslant \mathcal{D}. \tag{6}$$

Notice that this is just the infinitesimal version of the bounded distortion property and indeed it is possible to show that this condition implies the bounded distortion property. In the one-dimensional setting and assuming $J \subset \omega$ to be an open set, this implication follows immediately from the Mean Value Theorem. Indeed, in one dimension the determinant of the derivative is just the derivative itself. Thus, the Mean Value Theorem implies that there exists $x \in I\omega$ such that $|Df^n(x)| = |DF^n(\omega)|/|\omega|$ and $y \in J$ such that $|DF(y)| = |DF^n(J)|/|J|$. Therefore

$$\frac{|\omega|}{|J|} \frac{|F^n(J)|}{|F^n(\omega)|} = \frac{|F^n(J)|/|J|}{|F^n(\omega)|/|\omega|} = \frac{|DF^n(y)|}{|DF^n(x)|} \leqslant \mathcal{D}. \tag{7}$$

To verify (6) we use the chain rule to write

$$\log \frac{|\det DF^n(x)|}{|\det DF^n(y)|} = \log \prod_{i=0}^{n-1} \frac{|\det DF(F^i(x))|}{|\det DF(F^i(x))|} = \sum_{i=0}^{n-1} \log \frac{|\det DF(F^i(x))|}{|\det DF(F^i(y))|}.$$

Now adding and subtracting $|\det DF(F^i(y))|/|\det DF(F^i(y))|$ and using that fact that $\log(1+x) < x$ for $x > 0$ gives

$$\begin{aligned} \log \frac{|\det DF(F^i(x))|}{|\det DF(F^i(y))|} &\leqslant \log\left(\frac{|\det DF(F^i(x)) - \det DF(F^i(y))|}{|\det DF(F^i(y))|} + 1 \right) \\ &\leqslant \frac{|\det DF(F^i(x)) - \det DF(F^i(y))|}{|\det DF(F^i(y))|}. \end{aligned}$$

Therefore we have

$$\log \frac{|\det DF^n(x)|}{|\det DF^n(y)|} \leqslant \sum_{i=0}^{n-1} \frac{|\det DF(F^i(x)) - \det DF(F^i(y))|}{|\det DF(F^i(y))|}. \tag{8}$$

The inequality (8) gives us the basic tool for verifying the required distortion properties in particular examples.

4. Uniformly expanding maps

In this section we discuss maps which are *uniformly* expanding.

4.1. *The smooth/Markov case*

We say that f is uniformly expanding if there exist constants $C, \lambda > 0$ such that for all $x \in M$, all $v \in T_x M$, and all $n \geqslant 0$, we have

$$\left\| Df_x^n(v) \right\| \geqslant C e^{\lambda n} \|v\|.$$

We remark once again that this is a special case of the nonuniform expansivity condition.

THEOREM 5. *Let $f : M \to M$ be C^2 uniformly expanding. Then there exists a unique acip μ* [136,68,125,139,22,96,164,97]. *The measure μ is mixing and the correlation function decays exponentially fast* [153,127,33,140].

The references given here use a variety of arguments some of which use the remarkable observation that uniformly expanding maps are intrinsically *Markov* in the strong sense given above, with $\Delta = M$, a finite number of partition elements and return time $R \equiv 1$

(this is particularly easy to see in the case of one-dimensional circle maps $f:\mathbb{S}^1 \to \mathbb{S}^1$). Thus the main issue here is the verification of the distortion condition.

One way to show this is to show that there is a uniform upper bound independent of n for the sum in (8) above. Indeed, notice first of all that the expansivity condition implies in particular that $|\det DF(F^i(y))| \geqslant Ce^{\lambda i} \geqslant C > 0$ for every y, and the C^2 regularity condition implies that $\det Df$ is Lipschitz: there exists $L > 0$ such that $|\det DF(F^i(x)) - \det DF(F^i(y))| \leqslant L|F^i(x) - F^i(y)|$ for all $x, y \in M$. Substituting these inequalities into (8) we get

$$\sum_{i=0}^{n-1} \frac{|\det DF(F^i(x)) - \det DF(F^i(y))|}{|\det DF(F^i(y))|} \leqslant \frac{L}{C} \sum_{i=0}^{n-1} \big|F^i(x) - F^i(y)\big|. \tag{9}$$

The next, and final, step uses the expansivity condition as well as, implicitly, the Markov property in a crucial way. Indeed, let $\operatorname{diam} M$ denote the diameter of M, i.e. the maximum distance between any two points in M. The definition of $\mathcal{P}^{(n)}$ implies that ω is mapped diffeomorphically to M by F^n and thus

$$|\operatorname{diam} M| \geqslant \big|F^n(x) - F^n(y)\big| \geqslant Ce^{\lambda(n-i)}\big|F^i(x) - F^i(y)\big|$$

for every $i = 0, \ldots, n-1$. Therefore

$$\sum_{i=0}^{n-1} \big|F^i(x) - F^i(y)\big| \leqslant \frac{\operatorname{diam} M}{C} \sum_{i=0}^{n-1} e^{-\lambda(n-i)} \leqslant \frac{\operatorname{diam} M}{C} \sum_{i=0}^{\infty} e^{-\lambda i}. \tag{10}$$

Substituting back into (9) and (8) gives a bound for the distortion which is independent of n.

The regularity condition on $\det DF$ can be weakened somewhat but not completely. There exist examples of one-dimensional circle maps $f:\mathbb{S}^1 \to \mathbb{S}^1$ which are C^1 uniformly expanding (and thus Markov as above) but for which the uniqueness of the absolutely continuous invariant measure fails [134,56], essentially due to the failure of the bounded distortion calculation. On the other hand, the distortion calculation above goes through with minor modifications as long as $\det DF$ is just Hölder continuous. In some situations, such as the one-dimensional *Gauss map* $f(x) = x^{-1} \bmod 1$ which is Markov but for which the derivative Df is not even Hölder continuous, one can compensate by taking advantage of the large derivative. Then it is possible to show directly that the right-hand side of (8) is uniformly bounded, even though (9) does not hold.

4.2. *The non-Markov case*

The general (non-Markov) piecewise expanding case is significantly more complicated and even the existence of an absolutely continuous invariant measure is no longer guaranteed [98,70,135,159,49]. One possible problem is that the images of the discontinuity set can be very badly distributed and cause havoc with any kind of *structure*. In the Markov case

this does not happen because the set of discontinuities gets mapped to itself by definition. Also the possibility of components being *translated* in different directions can destroy on a global level the local expansiveness given by the derivative. Moreover, where results exist for rates of decay of correlations, they do not always apply to the case of Hölder continuous observables, as technical reasons sometimes require that different functions spaces be considered which are more compatible with the discontinuous nature of the maps. We shall not explicitly comment on the particular classes of observables considered in each case.

In the one-dimensional case these problems are somewhat more controllable and relatively simple conditions guaranteeing the existence of an ergodic invariant probability measure can be formulated even in the case of a countable number of domains of smoothness of the map. These essentially require that the size of the image of all domains on which the map is C^2 be strictly positive and that certain conditions on the second derivative are satisfied [98,11,34,35]. In the higher-dimensional case, the situation is considerably more complicated and there are a variety of possible conditions which can be assumed on the discontinuities. The conditions of [98] were generalized to the two-dimensional context in [90] and then to arbitrary dimensions in [69,48,160]. There are also several other papers which prove similar results under various conditions, we mention [12,51,53,50,148,54]. In [47,62] it is shown that conditions sufficient for the existence of a measure are *generic* in a certain sense within the class of piecewise expanding maps.

Estimates for the decay of correlations have been proved for non-Markov piecewise smooth maps, although again the techniques have had to be considerably generalized. In terms of setting up the basic arguments and techniques, a similar role to that played by [98] for the existence of absolutely continuous invariant measures can be attributed to [91, 76,143] for the problem of decay of correlations in the one-dimensional context. More recently, alternative approaches have been proposed and implemented in [99,100,169]. The approach of [169] has proved particularly suitable for handling some higher-dimensional cases such as [52] in which assumptions on the discontinuity set are formulated in terms of *topological pressure* and [18,73] in which they are formulated as *geometrical nondegeneracy* assumptions and *dynamical* assumptions on the *rate of recurrence* of typical points to the discontinuities. The construction of an induced Markov map is combined in [64] with Theorem 4 to obtain estimates for the decay of correlations of non-Hölder observables for Lorenz-like expanding maps. We remark also that the results of [18,73] apply to more general piecewise nonuniformly expanding maps, see Section 7. It would be interesting to understand the relation between the assumptions of [18,73] and those of [52].

5. Almost uniformly expanding maps

Perhaps the simplest way to relax the uniform expansivity condition is to allow some fixed (or periodic) point p to have a neutral eigenvalue, e.g., in the one-dimensional setting $|Df(p)| = 1$, while still requiring all other vectors in all directions over the tangent spaces of all points to be strictly expanded by the action of the derivative (though of course not uniformly since the expansion must degenerate near the point p). Remarkably this can have extremely dramatic consequences on the dynamics.

There are some recent results for higher-dimensional systems [131,81,71] but a more complete picture is available the one-dimensional setting and thus we concentrate on this case. An initial motivation for these kinds of examples arose from the concept of *intermittency* in fluid dynamics. A class of one-dimensional maps expanding everywhere except at a fixed point was introduced by Manneville and Pomeau in [114] as a model of intermittency since numerical studies showed that orbits tend to spend a long time *trapped* in a neighbourhood of the fixed point with relatively short *bursts of chaotic activity* outside this neighbourhood. Recent work shows that indeed, these long periods of inactivity near the fixed point are a key to slowing down the mixing process and obtaining examples of systems with subexponential decay of correlations.

We shall consider interval maps f which are piecewise C^2 with a C^1 extension to the boundaries of the C^2 domains and for which the derivative is strictly greater than 1 everywhere except at a fixed point p (which for simplicity we can assume lies at the origin) where $Df(p) = 1$. For definiteness, let us suppose that on a small neighbourhood of 0 the map takes the form

$$f(x) \approx x + x^2\phi(x),$$

where $\approx$ means that the terms on the two sides of the expression as well as their first and second order derivatives converge as $x \to 0$. We assume, moreover, that ϕ is C^∞ for $x \neq 0$; the precise form of ϕ determines the precise degree of *neutrality* of the fixed point, and in particular affects the second derivative $D^2 f$. It turn out that it plays a crucial role in determining the mixing properties and even the very existence of an absolutely continuous invariant measure. For the moment we assume also a strong Markov property: each domain of regularity of f is mapped bijectively to the whole interval. The following result shows that the situation can be drastically different from the uniformly expanding case.

THEOREM 6 [130]. *If f is C^2 at p (e.g., $\phi(x) \equiv 1$) then f does not admit any acip.*

Note that f has the same topological behaviour as a uniformly expanding map, typical orbits continue to wander densely on the whole interval, but the proportion of time which they spend in various regions tends to concentrate on the fixed point, so that, asymptotically, typical orbits spend all their time near 0. It turns out that in this situation there exists an infinite (σ-finite) absolutely continuous invariant measure which gives finite mass to any set not containing the fixed point and infinite mass to any neighbourhood of p [154].

The situation changes if we relax the condition that f be C^2 at p and allow the second derivative $D^2 f(x)$ to diverge to infinity as $x \to p$. This means that the derivative increases quickly as one moves away from p and thus nearby points are repelled at a faster rate. This is a very subtle change but it makes all the difference.

THEOREM 7. *If ϕ is of the form $\phi(x) = x^{-\alpha}$ for some $\alpha \in (0, 1)$, then* [130,83] *f admits an ergodic acip μ and* [84,102,170,131,147,82,72] *μ is mixing with decay of correlations*

$$\mathcal{C}_n = \mathcal{O}\big(n^{1-1/\alpha}\big).$$

If $\phi(1/x) = \log x \log^{(2)} x \dots \log^{(r-1)} x (\log^{(r)} x)^{1+\alpha}$ *for some* $r \geqslant 1$, $\alpha \in (-1, \infty)$, *where* $\log^{(r)} = \log\log\dots\log$ *repeated* r *times, then* [77] f *admits a mixing acip* μ *with decay of correlations*

$$\mathcal{C}_n = \mathcal{O}\big(\log^{(r)} n\big)^{-\alpha}.$$

Thus, the existence of an absolutely continuous invariant measure as in the uniformly expanding case has been recovered, but the exponential rate of decay of correlation has not. We can think of the indifferent fixed point as having the effect of *slowing down* this process by trapping nearby points for disproportionately long time. The estimates in [147, 82,72] include lower bounds as well as upper bounds. The approach in [170] using Markov induced maps applies also to non-Markov cases and [77] can also be generalized to these cases.

The proofs of Theorem 7 do not use directly the fact that f is nonuniformly expanding. Indeed the fact that f is nonuniformly expanding does not follow automatically from the fact that the map is expanding away from the fixed point p. However we can use the existence of the *acip* to show that this condition is satisfied. Indeed by Birkhoff's Ergodic Theorem, typical points spend a large proportion of time near p but also a positive proportion of time in the remaining part of the space. More formally, by a simple application of Birkhoff's Ergodic Theorem to the function $\log|Df(x)|$, we have that, for μ-almost every x,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \log\big|Df\big(f^i(x)\big)\big| \to \int \log|Df|\,d\mu > 0.$$

The fact that $\int \log|Df|\,d\mu > 0$ follows from the simple observation that μ is absolutely continuous, finite, and that $\log|Df| > 0$ except at the neutral fixed point.

6. One-dimensional maps with critical points

The general theory of one-dimensional maps is extremely advanced and sophisticated, see the survey [6] in this volume. We shall concentrate here on the particular case of nonuniformly expanding one-dimensional maps. This is another class of systems which can exhibit various rates of decay of correlations, but where the mechanism for producing these different rates is significantly more subtle. The most general set-up is that of a piecewise smooth one-dimensional map $f : I \to I$ with some finite set $\mathcal{C}$ of *critical/singular points* at which $Df = 0$ or $Df = \pm\infty$ and/or at which f may be discontinuous. There are at least two ways to quantify the "uniformity": of the expansivity of f in ways that get reflected in different rates of decay of correlations:

- To consider the rate of growth of the derivatives along the orbits of the critical points;
- To consider the *average* rate of growth of the derivative along *typical* orbits.

In this section we will concentrate on the first, somewhat more concrete, approach and describe the main results which have been obtained over the last 20–25 years. We shall

focus specifically on the smooth case since this is where most results have been obtained. Some partial generalization to the piecewise smooth case can be found in [104]. The second approach is somewhat more abstract but also more general since it extends naturally to the higher-dimensional context where critical points are not so well defined and/or cannot play such a fundamental role. The main results in this direction will be described in Section 7 below in the framework of a general theory of nonuniformly expanding maps.

6.1. *Unimodal maps*

We consider the class of C^3 interval maps $f : I \to I$ with some finite set $\mathcal{C}$ of nonflat critical points. We recall that c is a critical point if $Df(c) = 0$; the critical point is nonflat if there exists an $0 < \ell < \infty$ called the *order* of the critical point, such that $|Df(x)| \approx |x - c|^{\ell-1}$ for x near c; f is unimodal if it has only one critical point, and multimodal if it has more than one. Several results to be mentioned below have been proved under a standard technical *negative Schwarzian derivative* condition which is a kind of convexity assumption on the derivative of f, see [117] for details. Recent results [93] indicate that this condition is often superfluous and thus we will not mention it explicitly.

The first result on the statistical properties of such maps goes back to Ulam and von Neumann [161] who showed that the *top* unimodal quadratic map, $f(x) = x^2 - 2$ has an *acip*. Notice that this map is actually a Markov map but does not satisfy the bounded distortion condition due to the presence of the critical point. It is possible to construct an induced Markov map for f which does satisfy this condition and gives the result, but Ulam and von Neumann used a more direct approach, taking advantage of the fact that f is a Chebyshev polynomial and thus in particular using that fact that f is C^1 conjugate to a piecewise-linear uniformly expanding Markov map, for which Lebesgue measure is invariant and ergodic. This implies that the pull-back of the Lebesgue measure by the conjugacy is an *acip* for f. A similar strategy was exploited also in [45,141]. However, the existence of a smooth conjugacy is extremely rare and such an approach is not particularly effective in general.

More general and more powerful approaches have allowed the existence of an *acip* to be proved under increasingly general assumptions on the behaviour of the critical point. Let

$$D_n(c) = \left|Df^n\big(f(c)\big)\right|.$$

Notice that the derivative along the critical orbit needs to be calculated starting from the critical value and not from the critical point itself for otherwise it would be identically 0.

THEOREM 8. *Let $f : I \to I$ be a unimodal map with negative Schwarzian derivative. Then f admits an ergodic acip if the following conditions hold* (*each condition is implied by the preceding ones*):

- *The critical point is pre-periodic* [85];
- *The critical point is nonrecurrent* [85,119];
- $D_n \to \infty$ *exponentially fast* [60,123];
- $D_n \to \infty$ *sufficiently fast so that* $\sum_n D_n^{1/\ell} < \infty$ [124];
- $D_n \to \infty$ [43].

If $D_n \to \infty$ exponentially fast then some power of f is mixing and exhibits exponential decay of correlations [92] (*and* [167] *with additional bounded recurrence assumptions on the critical point*).

Notice that the condition of [43] is extremely weak. In fact they show that it is sufficient for D_n to be eventually bounded below by some constant depending only on the order of the critical point. However even this condition is not optimal as there are examples of maps for which $\liminf D_n = 0$ but which still admit an ergodic *acip*. It would be interesting to know whether an optimal condition is even theoretically possible: it is conceivable that a complete characterization of maps admitting acip's in terms of the behaviour of the critical point is not possible because other subtleties come into play.

6.2. *Multimodal maps*

Many arguments and techniques used in the context of unimodal maps have turned out to be almost too sophisticated for their own good and difficult to generalize to the multimodal setting. Until recently there have been almost no results available, not even on the existence of *acip*'s, for multimodal maps. A significant breakthrough was achieved by implementing the strategy of constructing induced Markov maps and estimating the rate of decay of the tail. This strategy yields also estimates for various rates of decay in the unimodal case and extends very naturally to the multimodal case.

THEOREM 9 [42]. *Let f by a multimodal map with a finite set of critical points of order ℓ and suppose that*

$$\sum_n D_n^{-1/(2\ell-1)} < \infty$$

for each critical point c. Then there exists an ergodic acip μ for f. Moreover, some power of f is mixing and the correlation function decays at the following rates:

Polynomial case: *If there exists $C > 0,\ \tau > 2\ell - 1$ such that*

$$D_n(c) \geqslant Cn^\tau$$

for all $c \in \mathcal{C}$ and $n \geqslant 1$, then, for any $\tilde{\tau} < \frac{\tau-1}{\ell-1} - 1$, we have

$$\mathcal{C}_n = \mathcal{O}\big(n^{-\tilde{\tau}}\big).$$

Exponential case: *If there exist $C, \beta > 0$ such that*

$$D_n(c) \geqslant Ce^{\beta n}$$

for all $c \in \mathcal{C}$ and $n \geqslant 1$, then there exist $\tilde{\beta} > 0$ such that

$$\mathcal{C}_n = \mathcal{O}\big(e^{-\tilde{\beta}n}\big).$$

These results gives previously unknown estimates for the decay of correlations even for unimodal maps in the quadratic family. For example, they imply that the so-called *Fibonacci maps* [112] exhibit decay of correlation at rates which are faster than any polynomial. It seems likely that these estimates are essentially optimal although the argument only provides upper bounds. The general framework of [110] applies to these cases to provide estimates for the decay of correlation for observable which satisfy weaker than Hölder conditions on the modulus of continuity. The condition for the existence of an *acip* have recently been weakened to the summability condition $\sum D_n^{-1/\ell} < \infty$ and to allow the possibility of critical points of different orders [44]. Some improved technical expansion estimates have been also obtained in [59] which allow the results on the decay of correlations to apply to maps with critical points of different orders.

Based on these results, the conceptual picture of the causes of slow rates of decay of correlations appears much more similar to the case of maps with indifferent fixed points than would appear at first sight: we can think of the case in which the rate of growth of D_n is subexponential as a situation in which the critical orbit is *neutral* or *indifferent* and points which land close to the critical point tend to remain close to ("trapped" by) its orbit for a particularly long time. During this time orbits are behaving "nongenerically" and are not distributing themselves over the whole space as uniformly as they should. Thus the mixing process is delayed and the rate of decay of correlations is correspondingly slower. When D_n grows exponentially, the critical orbit can be thought of (and indeed *is*) a nonperiodic *hyperbolic repelling* orbit and nearby points are pushed away exponentially fast. Thus there is no significant loss in the rate of mixing, and the decay of correlations is not significantly slowed down notwithstanding the presence of a critical point.

6.3. *Benedicks–Carleson maps*

We give here a sketch of the construction of the induced Markov map for a class of unimodal maps. We shall try to give a conceptually clear description of the main steps and ingredients required in the construction. The details of the argument are unfortunately particularly technical and a lot of notation and calculations are carried out only to formally verify statements which are intuitively obvious. It is very difficult therefore to be at one and the same time conceptually clear and technically honest. We shall therefore concentrate here on the former approach and make some remarks about the technical details which we omit or present in a simplified form.

Let

$$f_a(x) = x^2 - a$$

for $x \in I = [-2, 2]$ and

$$a \in \Omega_\varepsilon = [2 - \varepsilon, 2]$$

for some $\varepsilon > 0$ sufficiently small. The assumptions and the details of the proof require the introduction of several additional constants, some intrinsic to the maps under consideration

and some auxiliary for the purposes of the argument. In particular we suppose that there is a $\lambda \in (0, \log 2)$ and constants

$$\lambda \gg \alpha \gg \hat{\delta} \gg \delta > 0,$$

where $x \gg y$ means that y must be sufficiently small relative to x. Finally, to simplify the notation we also let $\beta = \alpha/\lambda$. We restrict ourselves to parameter values $a \in \Omega_\varepsilon$ which satisfy the *Benedicks–Carleson* conditions:

Hyperbolicity: There exist $C > 0$ such that

$$D_n \geqslant Ce^{\lambda n}, \quad \forall n \geqslant 1;$$

Slow recurrence:

$$|c_n| \geqslant e^{-\alpha n}, \quad \forall n \geqslant 1.$$

In Section 8 on p. 304 we sketch a proof of the fact that these conditions are satisfied for a positive measure set of parameters in Ω_ε (for *any* $\lambda \in (0, \log 2)$ and *any* $\alpha > 0$). They are therefore reasonably generic conditions. Assuming them here will allow us to present in a compact form an almost complete proof. During the discussion we shall make some comments about how the argument can be modified to deal with slower rates of growth of D_n and arbitrary recurrence patterns of the critical orbit.

We remark that the overall strategy as well as several details of the construction in the two arguments (one proving that the hyperbolicity and slow recurrence conditions occur with positive probability and the other proving that they imply the existence of an *acip*) are remarkably similar. This suggests a deeper, yet to be fully understood and exploited, relationship between the structure of dynamical space and that of parameter space.

We let

$$\Delta = (\delta, \delta) \subset (-\hat{\delta}, \hat{\delta}) = \hat{\Delta}$$

denote δ and $\hat{\delta}$ neighbourhood of this critical point c. The aim is to construct a Markov induced map

$$F : \Delta \to \Delta.$$

We shall do this in three steps. We first define an induced map $f^p : \Delta \to I$ which is essentially based on the time during which points in Δ shadow the critical orbit. The shadowing time p is piecewise constant on a countable partition of Δ but the images of partition elements can be arbitrarily small. Then we define an induced map $f^E : \Delta \to I$ which is still not Markov but has the property that the images of partition elements are uniformly large. Finally we define the Markov induced map $F^R : \Delta \to \Delta$ as required.

6.4. *Expansion outside Δ*

Before starting the construction of the induced maps, we state a lemma which gives some derivative expansion estimates outside the critical neighbourhood Δ.

LEMMA 6.1. *There exists a constant $C > 0$ independent of δ such that for $\varepsilon > 0$ sufficiently small, all $a \in \Omega_\varepsilon$, $f = f_a$, $x \in I$ and $n \geqslant 1$ such that $x, f(x), \ldots, f^{n-1}(x) \notin \Delta$ we have*

$$\left|Df^n(x)\right| \geqslant \delta e^{\lambda n}$$

and if, moreover, $f^n(x) \in \hat{\Delta}$ and/or $x \in f(\hat{\Delta})$ then

$$\left|Df^n(x)\right| \geqslant C e^{\lambda n}.$$

Notice that the constant C and the exponent λ do not depend on δ or $\hat{\delta}$. This allows us to choose $\hat{\delta}$ and δ small in the following argument without worrying about this affecting the expansivity estimates given here. In general of course both the constants C and λ depend on the size of this neighbourhood and it is an extremely useful feature of this particular range of parameter values that they do not. In the context of the quadratic family these estimates can be proved directly using the smooth conjugacy of the top map f_2 with the piecewise affine tent map, see [161,103]. However there are general theorems in one-dimensional dynamics to the effect that one has uniform expansivity outside an arbitrary neighbourhood of the critical point under extremely mild conditions [113] and this is sufficient to treat the general case in [42].

6.5. *Shadowing the critical orbit*

We start by defining a partition of the critical neighbourhoods Δ and $\hat{\Delta}$. For any integer $r \geqslant 1$ let $I_r = [e^{-r}, e^{-r+1})$ and $I_{-r} = (-e^{-r+1}, -e^{-r}]$ and, for each $r \geqslant r_{\hat{\delta}} + 1$, let $\hat{I}_r = I_{r-1} \cup I_r \cup I_{r+1}$. We can suppose without loss of generality that $r_\delta = \log \delta^{-1}$ and $r_{\hat{\delta}} = \log \hat{\delta}^{-1}$ are integers. Then

$$\Delta = \{0\} \cup \bigcup_{|r| \geqslant r_\delta + 1} I_r 0 \quad \text{and} \quad \hat{\Delta} = \{0\} \cup \bigcup_{|r| \geqslant r_{\hat{\delta}} + 1} I_r.$$

This is one of the minor technical points of which we do not give a completely accurate description. Strictly speaking, the distortion estimates to be given below require a further subdivision of each I_r into r^2 subintervals of equal length. This does not affect significantly any of the other estimates. A similar partition is defined in Section 8.1.2 on p. 307 in somewhat more detail. We remark also that the need for the two neighbourhoods Δ and $\hat{\Delta}$ will not become apparent in the following sketch of the argument. We mention it however because it is a crucial technical detail: the region $\hat{\Delta} \setminus \Delta$ acts as a *buffer zone* in which we

can choose to apply the derivative estimates of Lemma 6.1 or the shadowing argument of Lemma 6.2 according to which one is more convenient in a particular situation.

Now let

$$p(r) = \max\{k : |f^{j+1}(x) - f^{j+1}(c)| \leqslant e^{-2\alpha j},\ \forall x \in \hat{I}_r,\ \forall j < k\}.$$

This definition was essentially first formulated in [28] and [29]. The key characteristic is that it guarantees a bounded distortion property which in turn allows us to make several estimates based on information about the derivative growth along the critical orbit. Notice that the definition in terms of α is based crucially on the fact that the critical orbit satisfies the slow recurrence condition. We mention below how this definition can be generalized.

LEMMA 6.2. *For all points $x \in \hat{I}_r$ and $p = p(r)$ we have*

$$|Df^{p+1}(x)| = |Df(x)| \cdot |Df^p(x_0)| \geqslant e^{(1-7\beta)r}.$$

Recall that $\beta = \alpha/\lambda$ can be chosen arbitrarily small.

PROOF. First of all, using the bounded recurrence condition, the definition of the binding period, and arguing as in the distortion estimates for the uniformly expanding maps above, it is not difficult to show that there is a constant $\mathcal{D}_1$, depending on α but independent of r and δ, such that for all $x_0, y_0 \in f(\hat{I}_r)$ and $1 \leqslant k \leqslant p$,

$$\left|\frac{Df^k(x_0)}{Df^k(y_0)}\right| \leqslant \mathcal{D}_1. \tag{11}$$

Using the definition of p this implies

$$e^{-2\alpha(p-1)} \geqslant |x_{p-1} - c_{p-1}| \geqslant \mathcal{D}_1^{-1}|Df^{p-1}(c_0)||x_0 - c_0| \geqslant \mathcal{D}_1^{-1}e^{\lambda(p-1)}e^{-2r}$$

and thus $\mathcal{D}_1 e^{-2\alpha p}e^{2\alpha} \geqslant e^{\lambda p}e^{-\lambda}e^{-2r}$. Rearranging gives

$$p + 1 \leqslant \frac{\log \mathcal{D}_1 + 2\alpha + 2\lambda + 2r}{\lambda + 2\alpha} \leqslant \frac{3r}{\lambda} \tag{12}$$

as long as we choose δ so that r_δ is sufficiently large in comparison to the other constants, none of which depend on δ. Moreover,

$$\mathcal{D}e^{-2r}|Df^p(x_0)| \geqslant \mathcal{D}|x_0 - c_0||Df^p(x_0)| \geqslant |x_p - c_p| \geqslant e^{-2\alpha p}$$

and therefore, using (12), we have $|Df^p(x_0)| \geqslant \mathcal{D}^{-1}e^{2r}e^{-2\alpha p} \geqslant \mathcal{D}^{-1}e^{(2-6\alpha/\lambda)r}$. Since $x \in \hat{I}_r$ we have $|Df(x)| = 2|x - c| \geqslant 2e^{-(r+2)}$ and therefore

$$|Df^{p+1}(x)| = |Df^p(x_0)||Df(x)| \geqslant e^{-2}\mathcal{D}^{-1}e^{(2-6\beta)r}e^{-r} \geqslant e^{-2}\mathcal{D}^{-1}e^{(1-6\beta)r}.$$

This implies the result as long as we choose r_δ large enough. □

Thus we have a first induced map $F_p : \Delta \to I$ given by $F_p(x) = f^{p(x)}(x)$ where $p(x) = p(r)$ for $x \in I_{\pm r}$ which is uniformly expanding. Indeed notice that $Df^{p(x)}(x) \to \infty$ as $x \to c$. However there is no reason for which this map should satisfy the Markov property and indeed, an easy calculation shows that the images of the partition elements are $\sim e^{-7\beta r} \to 0$ and thus not even of uniform size.

The notion of shadowing can be generalized without any assumptions on the recurrence of the critical orbit in the following way, see [42]: let $\{\gamma_n\}$ be a monotonically decreasing sequence with $1 > \gamma_n > 0$ and $\sum \gamma_n < \infty$. Then for $x \in \Delta$, let

$$p(x) := \max\bigl\{p\colon \bigl|f^k(x) - f^k(c)\bigr| \leqslant \gamma_k \bigl|f^k(c) - c\bigr|,\ \forall k \leqslant p-1\bigr\}.$$

A simple variation of the distortion calculation used above shows that the summability of γ_n implies that (11) holds with this definition also. Analogous bounds on D_n will reflect the rate growth of the derivative along the critical orbit. If the growth of D_n is subexponential, the binding period will last much longer because the interval $|f^k(x) - f^k(c)|$ is growing at a slower rate. The generality of the definition means that it is more natural to define a partition I_p as the "level sets" of the function $p(x)$. The drawback is that we have much less control over the precise size of these intervals and their distance from the critical point. Some estimates of the tail $\{x > p\}$ can be obtained and it turns out that these are closely related to the rate of growth of D_n and to those of the return time function for the final induced Markov map. This is because the additional two steps, the escape time and the return time occur exponentially fast. Thus *the only bottleneck is the delay caused by the long shadowing of the critical orbit.*

6.6. *The escape partition*

Now let $J \subset I$ be an arbitrary interval (which could also be Δ itself). We want to construct a partition $\mathcal{P}$ of J and a stopping time $E : J \to \mathbb{N}$ constant on elements of $\mathcal{P}$ with the property that for each $\omega \in \mathcal{P}$, $f^{E(\omega)}(\omega) \approx \delta$. We think of δ as being our definition of *large scale*; we call $E(\omega)$ the *escape* time of ω, we call the interval $f^{E(\omega)}(\omega)$ and *escape interval*, and call $\mathcal{P}$ the *escape time* partition of J.

The construction is carried out inductively in the following way. Let $k \geqslant 1$ and suppose that the intervals with $E < k$ have already been defined. Let ω be a connected component of the complement of the set $\{E < k\} \subset J$. We consider the various cases depending on the position of $f^k(\omega)$. If $f^k(\omega)$ contains $\hat{\Delta} \cup I_{r_{\hat{\delta}}} \cup I_{-r_{\hat{\delta}}}$ then we subdivide ω into three subintervals satisfying the required properties, and let $E = k$ on each of them. If $f^k(\omega) \cap \Delta = \emptyset$ we do nothing. If $f^k(\omega) \cap \Delta \neq \emptyset$ but $f^k(\omega)$ does not intersect more than two adjacent I_r's then we say that k is an *inessential return* and define the corresponding *return depth* by $r = \max\{|r|\colon\ f^k(\omega) \cap I_r \neq \emptyset\}$. If $f^k(\omega) \cap \Delta \neq \emptyset$ and $f^k(\omega)$ intersects at least three elements of $\mathcal{I}$, then we simply subdivide ω into subintervals ω_r in such a way that each ω_r satisfies $I_r \subset f^k(\omega_r) \subset \hat{I}_r$. For $r > r_\delta$ we say that ω_r has an essential return at time k with an associated return depth r. For all other r, the ω_r are escape intervals, and for these intervals we set $E = k$. Finally we consider one more important case. If $f^{k(\omega)}$ contains Δ as well as (at least) the two adjacent partition elements then we subdivide as

described above except for the fact that we keep together that portion of ω which maps exactly to Δ. We let this belong to the escape partition $\mathcal{P}$ but, as we shall see, we also let it belong to the final partition associated to the full induced Markov map.

This purely combinatorial algorithm is designed to achieve two things, neither one of which follows immediately from the construction:

(1) Guarantee uniformly bounded distortion on each partition element up to the escape time;
(2) Guarantee that almost every point eventually belongs to the interior of a partition element $\omega \in \mathcal{P}$.

We shall not enter into the details of the distortion estimates here but discuss the strategy for showing that $\mathcal{P}$ is a partition mod 0 of J. Indeed this follows from a much stronger estimate concerning the tail of the escape time function: there exists a constant $\gamma > 0$ such that for any interval J with $|J| \geqslant \delta$ we have

$$\left|\left\{x \in J\colon\ E(x) \geqslant n\right\}\right| = \sum_{\substack{\omega' \in \mathcal{P} \\ E(\omega') \geqslant n}} |\omega'| \lesssim e^{-\gamma n}|J|. \tag{13}$$

The argument for proving (13) revolves fundamentally around the combinatorial information defined in the construction. More specifically, for $\omega \in \mathcal{P}$ let $r_1, r_2, \ldots, r_s$ denote the sequence of return depths associated to essential return times occurring before $E(\omega')$, and let $\mathcal{E}(\omega) = \sum_{i=0}^{s} r_s$. Notice that this sequence may be empty if ω escapes without intersecting Δ, in this case we set $\mathcal{E}(\omega) = 0$. We now split the proof into three steps:

Step 1. *Relation between escape time and return depths.* The first observation is that the escape time is bound by a constant multiple of the sum of the return depths: there exists a κ depending only on λ such that

$$E(\omega) \lesssim \kappa \mathcal{E}(\omega). \tag{14}$$

Notice that a constant T_0 should also be added to take care of the case in which $\mathcal{E}(\omega) = 0$, corresponding to the situation in which ω has an escape the first time that iterates of ω intersect Δ. Since ω is an escape, it has a minimum size and the exponential growth outside Δ gives a uniform bound for the maximum number of iterates within which such a return must occur. Since this constant is uniform it does not play a significant role and we do not add it explicitly to simplify the notation. For the situation in which $\mathcal{E}(\omega) > 0$ it is sufficient to show that each essential return with return depth r has the next essential return or escape within at most κr iterations. Again this follow from the observation that the derivative is growing exponentially on average during all these iterations: we have exponential growth outside Δ and also exponential growth on average during each complete inessential binding period. This implies (14). From (14) we then have

$$\sum_{\substack{\omega \in \mathcal{P}(\omega) \\ E(\omega) \geqslant n}} |\omega| \lesssim \sum_{\substack{\omega \in \mathcal{P}(\omega) \\ \mathcal{E}(\omega) \geqslant n/\kappa}} |\omega|. \tag{15}$$

Thus it is enough to estimate the right-hand side of (15) which is saying that there is an exponentially small probability of having a large total accumulated return depth before escaping, i.e. most intervals escape after relatively few and shallow return depths. The strategy is perfectly naive and consists of showing that the size of interval with a certain return depth $\mathcal{E}$ is exponentially small in $\mathcal{E}$ and that there cannot be too many so that their total sum is still exponentially small.

Step 2. *Relation between size of ω and return depth.* The size of each partition element can be estimated in terms of the essential return depths in a very coarse, i.e. nonsharp, way which is nevertheless sufficient for your purposes. The argument relies on the following observation. Every return depth corresponds to a return which is followed by a binding period. During this binding period there is a certain overall growth of the derivative. During the remaining iterates there is also derivative growth, either from being outside Δ or from the binding period associated to some inessential return. Therefore a simple application of the Mean Value Theorem gives

$$|\omega| \lesssim e^{-\frac{1}{2}\lambda\mathcal{E}(\omega)}. \tag{16}$$

Step 3. *The cardinality of the ω with a certain return depth.* It therefore remains only to estimate the cardinality of the set of elements ω which can have the same value of $\mathcal{E}$. To do this, notice first of all that we have a bounded multiplicity of elements of $\mathcal{P}(\omega)$ which can share exactly the same sequence of return depths. More precisely this corresponds to the number of escaping intervals which can arise at any given time from the subdivision procedure described above, and is therefore less than $r_{\hat{\delta}}$. Moreover, every return depth is bigger than r_δ and therefore for a given sequence $r_1, \ldots, r_s$ we must have $s \leqslant \mathcal{E}/r_\delta$. Therefore letting $\eta = r_\delta^{-1}$, choosing δ sufficiently small the result follows from the following fact: Let $N_{k,s}$ denote the number of sequences $(t_1, \ldots, t_s)$, $t_i \geqslant 1$ for all i, $1 \leqslant i \leqslant s$, such that $\sum_{i=1}^{s} t_i = k$. Then, for all $\hat{\eta} > 0$, there exists $\eta > 0$ such that for any integers s, k with $s < \eta k$ we have

$$N_{k,s} \leqslant e^{\hat{\eta}k}. \tag{17}$$

Indeed, applying (17) we get that the total number of possible sequences is $N_k = \sum_{s=1}^{\eta\mathcal{E}} N_{\mathcal{E},s} \leqslant \eta\mathcal{E}e^{\hat{\eta}\mathcal{E}} \leqslant e^{2\hat{\eta}\mathcal{E}}$. Taking into account the multiplicity of the number of elements sharing the same sequence we get the bound on this quantity as $\leqslant r_{\hat{\delta}} e^{2\hat{\eta}\mathcal{E}} \leqslant e^{3\hat{\eta}\mathcal{E}}$. Multiplying this by (16) and substituting into (15) gives the result.

To prove (17), notice first of all that $N_{k,s}$ can be bounded above by the number of ways to choose s balls from a row of $k + s$ balls, thus partitioning the remaining k balls into at most $s + 1$ disjoint subsets. Notice also that this expression is monotonically increasing in s, and therefore

$$N_{k,s} \leqslant \binom{k+s}{s} = \binom{k+s}{k} \leqslant \binom{(1+\eta)k}{k} = \frac{[(1+\eta)k]!}{(\eta k)!k!}.$$

Using Stirling's approximation formula $k! \in [1, 1 + \frac{1}{4k}]\sqrt{2\pi k}k^k e^{-k}$, we have $N_{k,s} \leqslant \frac{[(1+\eta)k]^{(1+\eta)k}}{(\eta k)^{\eta k}k^k} \leqslant (1+\eta)^{(1+\eta)k}\eta^{-\eta k} \leqslant \exp\{(1+\eta)k\log(1+\eta) - \eta k\log\eta\} \leqslant \exp\{((1+\eta)\eta - \eta\log\eta)k\}$. Clearly $(1+\eta)\eta - \eta\log\eta \to 0$ as $\eta \to 0$. This completes the proof of (13).

6.7. *The return partition*

Finally we need to construct the full induced Markov map. To do this we simply start with Δ and construct the escape partition $\mathcal{P}$ of Δ. Notice that this is a refinement of the binding partition into intervals I_r. Notice also that the definition of this escape partition allows as a special case the possibility that $f^{E(\omega)}(\omega) = \Delta$. In this case of course ω satisfies exactly the required properties and we let it belong by definition to the partition $\mathcal{Q}$ and define the return time of ω as $R(\omega) = E(\omega)$. Otherwise we consider each escape interval $J = f^{E(\omega)}(\omega)$ and use it as a starting interval for constructing and escape partition and escape time function. Again some of the partition elements constructed in this way will actually have returns to Δ. These we define to belong to $\mathcal{Q}$ and let their return time be the sum of the two escape times, i.e. the total number of iterations since they left Δ, so that $f^{R(\omega)}(\omega) = \Delta$. For those that do not return to Δ we repeat the procedure. We claim that almost every point of Δ eventually belongs to an element which returns to Δ in a good (Markov) way at some point and that the tail estimates for the return time function are not significantly affected, i.e. they are still exponential.

The final calculation to support this claim is based on the following fairly intuitive observation. Once an interval ω has an escape, it has reached large scale and therefore it will certainly cover Δ after some uniformly bounded number of iterations. In particular it contains some subinterval $\tilde{\omega} \subset \omega$ which has a return to δ with at most this uniformly bounded number of iterates after the escape. Moreover, and crucially, the proportion of $\tilde{\omega}$ in ω is uniformly bounded below, i.e. there exists a constant $\xi > 0$ independent of ω such that

$$|\tilde{\omega}| \geqslant \xi\,|\omega|. \tag{18}$$

Using this fact we are now ready to estimate the tail of return times, $|\{\omega \in \mathcal{Q} \mid R(\omega) > n\}|$. The argument is again based on taking into account some combinatorial information related to the itinerary of elements of the final partition $\mathcal{Q}$. In particular we shall keep track of the *number of escape times* which occur before time n for all elements whose return is greater than n. First of all we let

$$\mathcal{Q}^{(n)} = \big\{\omega \in \mathcal{Q} \mid R(\omega) \geqslant n\big\}. \tag{19}$$

Then, for each $1 \leqslant i \leqslant n$ we let

$$\mathcal{Q}_i^{(n)} = \big\{\omega \in \mathcal{Q}^{(n)} \mid E_{i-1}(\omega) \leqslant n < E_i(\omega)\big\} \tag{20}$$

be the set of partition elements in $\mathcal{Q}^{(n)}$ who have exactly i escapes. Amongst those we distinguish those with a specific escape combinatorics. More precisely, for $(t_1, \dots, t_i)$ such that $t_j \geqslant 1$ and $\sum t_j = n$, let

$$\mathcal{Q}_i^{(n)}(t_1, \dots, t_i) = \left\{ \omega \in \mathcal{Q}_i^{(n)} \mid \sum_{j=1}^{k} t_j = E_k(\omega),\ 1 \leqslant k \leqslant i-1 \right\}. \tag{21}$$

We then fix some small $\eta > 0$ to be determined below and write

$$\left|\left\{\omega \in \mathcal{Q} \mid R(\omega) \geqslant n\right\}\right| = \sum_{i \leqslant n} \left|\mathcal{Q}_i^{(n)}\right| = \sum_{i \leqslant \eta n} \left|\mathcal{Q}_i^{(n)}\right| + \sum_{\eta n < i \leqslant n} \left|\mathcal{Q}_i^{(n)}\right|. \tag{22}$$

By (18) we have $|\mathcal{Q}_i^{(n)}| \lesssim (1-\xi)^i$, which gives

$$\sum_{\eta n < i \leqslant n} \left|\mathcal{Q}_i^{(n)}\right| \lesssim \sum_{\eta n < i \leqslant n} (1-\xi)^i \lesssim (1-\xi)^{\eta n} \approx e^{-\gamma_\xi n} \tag{23}$$

for some $\gamma_\xi > 0$. Now let $\omega \subset \tilde{\omega} \in \mathcal{P}_{E_i}$ be one of the nonreturning parts of an interval $\tilde{\omega}$ that had its ith escape at time E_i. Note that $f^{E_i}(\mathcal{P}_{E_{i+1}}|_\omega) = \mathcal{P}(f^{E_i}(\omega))$.

Therefore

$$\sum_{\substack{\omega' \subset \omega \\ E_{i+1}(\omega') \geqslant E_i + n}} |\omega'| \lesssim e^{-\gamma n} |\omega|, \tag{24}$$

where γ is as in (13). Let $\mathcal{Q}_i^{(n)}$ denote the set of intervals $\omega \in \mathcal{Q}$ that have precisely i escapes before time n then

$$\sum_{\omega \in \mathcal{Q}^{(n)}(E_1, \dots, E_i)} |\omega| \lesssim e^{-\gamma n} |\Delta|. \tag{25}$$

Therefore using again the combinatorial counting argument and the inequality (17) we get

$$\begin{aligned} \sum_{i \leqslant \eta n} = \left|\mathcal{Q}_i^{(n)}\right| &= \sum_{i \leqslant \eta n} \sum_{(t_1, \dots, t_i)} \left|\mathcal{Q}_i^{(n)}(t_1, \dots, t_i)\right| \lesssim \sum_{i \leqslant \eta n} N_{n,i} e^{-\gamma n} |\Delta| \\ &\lesssim e^{\hat{\eta} n} e^{-\gamma n}. \end{aligned} \tag{26}$$

Recall that by (17) $\hat{\eta}$ can be chosen arbitrarily small by choosing η small. Thus, combining (23) and (26) and substituting into (22) we get

$$\left|\left\{\omega \in \mathcal{Q} \mid R(\omega) \geqslant n\right\}\right| \lesssim e^{-\gamma_\xi n} + e^{(\hat{\eta} - \gamma) n} \lesssim e^{-cn}.$$

7. General theory of nonuniformly expanding maps

The intuitive picture which emerges from the examples discussed above is that of a *default* exponential mixing rate for uniformly expanding systems and Hölder continuous observables. However it is clear that general nonuniformly expanding systems can exhibit a variety of rates of decay. Sometimes these rates can be linked to properties of specific *neutral* orbits which can *slow down* the mixing process. However it is natural to ask whether there is some intrinsic information related to the very definition of nonuniform expansivity which determines the rate of decay of correlation. We recall that f is nonuniformly expanding if there exists $\lambda > 0$ such that for almost every $x \in M$,

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \log \left\| Df^{-1}_{f^i(x)} \right\|^{-1} > \lambda. \tag{$*$}$$

Although the constant $\lambda > 0$ is uniform for Lebesgue almost every point, the convergence to the lim inf is not generally uniform.

A *measure of nonuniformity* has been proposed in [16] based precisely on the idea of quantifying the rate of convergence. The measure has been shown to be directly linked to the rate of decay of correlations in [17] in the one-dimensional setting and in [18] in arbitrary dimensions, in the case of polynomial rates of decay. Recently the theory has been extended to cover the exponential case as well [73]. We give here the precise statements.

7.1. *Measuring the degree of nonuniformity*

7.1.1. *The critical set* Let $f : M \to M$ be a (piecewise) C^2 map. For $x \in M$ we let Df_x denote the derivative of f at x and define $\|Df_x\| = \max\{\|Df_x(v)\|\colon\ v \in T_xM,\ \|v\| = 1\}$. We suppose that f fails to be a local diffeomorphism on some zero measure *critical* set $\mathcal{C}$ at which f may be discontinuous and/or Df may be discontinuous and/or singular and/or blow up to infinity. Remarkably, all these cases can be treated in a unified way as *problematic* points as will be seen below. In particular we can define a natural generalization of the nondegeneracy (nonflatness) condition for critical points of one-dimensional maps.

DEFINITION 12. The *critical set* $\mathcal{C} \subset M$ is *nondegenerate* if $m(\mathcal{C}) = 0$ and there is a constant $\beta > 0$ such that for every $x \in M \setminus \mathcal{C}$ we have $\operatorname{dist}(x, \mathcal{C})^\beta \lesssim \|Df_x v\|/\|v\| \lesssim \operatorname{dist}(x, \mathcal{C})^{-\beta}$ for all $v \in T_xM$, and the functions $\log \det Df$ and $\log \|Df^{-1}\|$ are *locally Lipschitz* with Lipschitz constant $\lesssim \operatorname{dist}(x, \mathcal{C})^{-\beta}$.

From now on we shall always assume these nondegeneracy conditions. We remark that the results to be stated below are nontrivial even when the critical set $\mathcal{C}$ is empty and f is a local diffeomorphism everywhere. For simplicity we suppose also that f is topologically transitive, i.e. there exists a point x whose orbit is dense in M. Without the topological transitivity condition we would just get that the measure μ admits a finite number of ergodic components and the results to be given below would then apply to each of its components.

7.1.2. *Expansion and recurrence time functions* Since we have no geometrical information about f we want to show that the statistical properties such as the rate of decay of correlations somehow depends on abstract information related to the nonuniform expansivity condition only. Thus we make the following

DEFINITION 13. For $x \in M$, we define the *expansion time function*

$$\mathcal{E}(x) = \min\left\{ N\colon \frac{1}{n}\sum_{i=0}^{n-1} \log \left\| Df^{-1}_{f^i(x)} \right\|^{-1} \geqslant \lambda/2,\ \forall n \geqslant N \right\}.$$

By condition $(*)$ this function is defined and finite almost everywhere. It measures the amount of time one has to wait before the uniform exponential growth of the derivative kicks in. If $\mathcal{E}(x)$ was uniformly bounded, we would essentially be in the uniformly expanding case. In general it will take on arbitrarily large values and not be defined everywhere. If $\mathcal{E}(x)$ is *large* only on a *small* set of points, then it makes sense to think of the map as being not very nonuniform, whereas, if it is large on a large set of points it is in some sense, very nonuniform. We remark that the choice of $\lambda/2$ in the definition of the expansion time function $\mathcal{E}(x)$ is fairly arbitrary and does not affect the asymptotic rate estimates. Any positive number smaller than λ would yield the same results.

We also need to assume some dynamical conditions concerning the rate of recurrence of typical points near the critical set. We let $d_\delta(x, \mathcal{C})$ denote the *δ-truncated* distance from x to $\mathcal{C}$ defined as $d_\delta(x, \mathcal{C}) = d(x, \mathcal{C})$ if $d(x, \mathcal{C}) \leqslant \delta$ and $d_\delta(x, \mathcal{C}) = 1$ otherwise.

DEFINITION 14. We say that f satisfies the property of *subexponential recurrence* to the critical set if for any $\varepsilon > 0$ there exists $\delta > 0$ such that for Lebesgue almost every $x \in M$,

$$\limsup_{n\to+\infty} \frac{1}{n}\sum_{j=0}^{n-1} -\log \operatorname{dist}_\delta\big(f^j(x), \mathcal{C}\big) \leqslant \varepsilon. \qquad (**)$$

We remark that although condition $(**)$ might appear to be a very technical condition, it is actually quite natural and in fact *almost* necessary. Indeed, suppose that an absolutely continuous invariant measure μ did exist for f. Then, a simple application of Birkhoff's Ergodic theorem implies that condition $(**)$ is equivalent to the integrability condition

$$\int_M \big|\log \operatorname{dist}_\delta(x, \mathcal{S})\big|\, d\mu < \infty$$

which is simply saying that the invariant measure does not give too much weight to a neighbourhood of the discontinuity set.

Again, we want to differentiate between different degrees of recurrence in a similar way to the way we differentiated between different degrees of nonuniformity of the expansion.

DEFINITION 15. For $x \in M$, we define the *recurrence time function*

$$\mathcal{R}(x) = \min\left\{ N \geqslant 1\colon \frac{1}{n}\sum_{j=0}^{n-1} -\log \operatorname{dist}_\delta\bigl(f^j(x), \mathcal{C}\bigr) \leqslant 2\varepsilon,\ \forall n \geqslant N \right\}.$$

Then, for a map satisfying both conditions $(*)$ and $(**)$ we let

$$\Gamma_n = \bigl\{x\colon \mathcal{E}(x) > n \text{ or } \mathcal{R}(x) > n\bigr\},$$

Notice that $\mathcal{E}(x)$ and $\mathcal{R}(x)$ are finite almost everywhere and thus $\Gamma_n \to 0$. It turns out that the rate of decay of $|\Gamma_n|$ is closely related to the rate of decay of correlations. In the statement of the theorem we let $\mathcal{C}_n$ denote the correlation function for Hölder continuous observables.

THEOREM 10. *Let $f : M \to M$ be a transitive C^2 local diffeomorphism outside a nondegenerate critical set $\mathcal{C}$, satisfying conditions $(*)$ and $(**)$. Then*

(1) [15] *f admits an acip μ. Some power of f is mixing.*

(2) [16,18,17] *Suppose that there exists $\gamma > 0$ such that*

$$|\Gamma_n| = \mathcal{O}(n^{-\gamma}).$$

Then

$$\mathcal{C}_n = \mathcal{O}\bigl(n^{-\gamma+1}\bigr).$$

(3) [73] *Suppose that there exists $\gamma > 0$ such that*

$$|\Gamma_n| = \mathcal{O}(e^{-\gamma n}).$$

Then there exists $\gamma' > 0$ such that

$$\mathcal{C}_n = \mathcal{O}\bigl(e^{-\gamma' n}\bigr).$$

7.2. *Viana maps*

A main application of the general results described above are a class of maps known as *Viana* or *Alves–Viana* maps. Viana maps were introduced in [162] as an example of a class of higher-dimensional systems which are strictly *not* uniformly expanding but for which the nonuniform expansivity condition is satisfied and, most remarkably, is *persistent* under small C^3 perturbations, which is not the case for any of the examples discussed above. These maps are defined as skew-products on a two-dimensional cylinder of the form $f : \mathbb{S}^1 \times \mathbb{R} \to \mathbb{S}^1 \times \mathbb{R}$,

$$f(\theta, x) = \bigl(\kappa\theta, x^2 + a + \varepsilon \sin 2\pi\theta\bigr),$$

where ε is assumed sufficiently small and a is chosen so that the one-dimensional quadratic map $x \mapsto x^2 + a$ for which the critical point lands after a finite number of iterates onto a hyperbolic repelling periodic orbit (and thus is a *good* parameter value and satisfies the nonuniform expansivity conditions as mentioned above). The map $\kappa\theta$ is taken modulo 2π, and the constant κ is a positive integer which was required to be $\geqslant 16$ in [162] although it was later shown in [55] that any integer $\geqslant 2$ will work. The sin function in the skew product can also be replaced by more general Morse functions.

THEOREM 11. *Viana maps*

- [162] *satisfy* $(*)$ *and* $(**)$. *In particular they are nonuniformly expanding*;
- [12,19] *are topologically mixing and have a unique ergodic acip* (*with respect to two-dimensional Lebesgue measure*);
- [18] *have super-polynomial decay of correlations*: *for any* $\gamma > 0$ *we have*

$$\mathcal{C}_n = \mathcal{O}(n^{-\gamma});$$

- [24,73] *have stretched exponential decay of correlations*: *there exists* $\gamma > 0$ *such that*

$$\mathcal{C}_n = \mathcal{O}\big(e^{-\gamma\sqrt{n}}\big).$$

8. Existence of nonuniformly expanding maps

An important point which we have not yet discussed is the fact that the verification of the nonuniform expansivity assumptions is a highly nontrivial problem. For example, the verification that Viana maps are nonuniformly expanding is one of the main results of [162]. Only in some special cases can the required assumptions be verified directly and easily. The definition of nonuniform expansivity is in terms of asymptotic properties of the map which are therefore intrinsically not checkable in any given finite number of steps. The same is true also for the derivative growth assumptions on the critical orbits of one-dimensional maps as in Theorems 8 on p. 290 and 9 on p. 291. A perfectly legitimate question is therefore whether these conditions actually do occur for any map at all. Moreover, recent results suggest that this situation is at best extremely rare in the sense that the set of one-dimensional maps which have attracting periodic orbits, and in particular do not have an *acip*, is *open and dense* in the space of all one-dimensional maps [74,111,94,150, 95]. However, this topological point of view is only one way of defining "genericity" and it turns out that for general one-parameter families of one-dimensional maps, the set of parameters for which an *acip* does exist can have *positive Lebesgue measure* (even though it may be topologically *nowhere dense*).

We give here a fairly complete sketch (!) of the argument in a special case, giving the complete description of the combinatorial construction and just brief overview of how the analytic estimates are obtained. For definiteness and simplicity we focus on the family

$$f_a(x) = x^2 - a$$

for $x \in I = [-2, 2]$ and

$$a \in \Omega_\varepsilon = [2 - \varepsilon, 2]$$

for some $\varepsilon > 0$.

THEOREM 12 [86]. *For every $\eta > 0$ there exists an $\varepsilon > 0$ and a set $\Omega^* \subset \Omega_\varepsilon$ such that for all $a \in \Omega^*$, f_a admits an ergodic absolutely continuous invariant probability measure, and such that*

$$|\Omega^*| > (1 - \eta)|\Omega_\varepsilon| > 0.$$

There exist several generalizations of this result for families of smooth maps [86,28,144, 155,118,157,158,87,165,149,88] and even to families with completely degenerate (flat) critical points [156] and to piecewise smooth maps with critical points [108,107]. The arguments in the proofs are all fundamentally of a probabilistic nature and the conclusions depend on the fact that if f is nonuniformly expanding for a large number of iterates n then it has a "high probability" of being nonuniformly expanding for $n + 1$ iterates. Thus, by successively deleting those parameters which fail to be nonuniformly expanding up to some finite number of iterates, one has to delete smaller and smaller proportions. Therefore a positive proportion survives all exclusions.

In Section 8.1 we give the formal inductive construction of the set Ω^*. In Sections 8.2 and 8.3 we prove the two main technical lemmas which give expansion estimates for orbit starting respectively outside and inside some critical neighbourhood. In Section 8.4 we prove the inductive step in the definition of Ω^* and in Section 8.5 we obtain the lower bound on the size of $|\Omega^*|$.

The proof involves several constants, some intrinsic to the family under consideration and some auxiliary for the purposes of the construction. The relationships between these constants and the order in which they are chosen is quite subtle and also crucial to the argument. However this subtlety cannot easily be made explicit in such a sketch as we shall give here. We just mention therefore that there are essentially only two intrinsic constants: λ which is the expansivity exponent outside some (in fact any) critical neighbourhood, and ε which is the size of the parameter interval Ω_ε. λ can be chosen *first* and is essentially arbitrary as long as $\lambda \in (0, \log 2)$; ε needs to be chosen *last* to guarantee that the auxiliary constants can be chosen sufficiently small. The main auxiliary constants are λ_0 which can be chosen arbitrarily in $(0, \lambda)$ and which gives the *target* Lyapunov exponent of the critical orbit for good parameters, and

$$\lambda \gg \alpha \gg \hat{\delta} = \delta^\iota \gg \delta > 0$$

which are chosen in the order given and sufficiently small with respect to the previous ones. During the proof we will introduce also some "second order" auxiliary constants which depend on these. Finally we shall use the constant $C > 0$ to denote a generic constant whose specific value can in different formulae.

8.1. *The definition of Ω^**

We let $c_0 = c_0(a) = f_a(0)$ denote the *critical value* of f_a and for $i \geqslant 0$, $c_i = c_i(a) = f^i(c_0)$. For $n \geqslant 0$ and $\omega \subseteq \Omega$ let $\omega_n = \{c_n(a);\ a \in \omega\} \subseteq I$. Notice that for $a = 2$ the critical value maps to a fixed point. Therefore iterates of the critical point for parameter values sufficiently close to 2 remain in an arbitrarily small neighbourhood of this fixed point for an arbitrarily long time. In particular it is easy to see that all the inductive assumptions to be formulated below hold for all $k \leqslant N$ where N can be taken arbitrarily large if ε is small enough. This observation will play an important role in the very last step of the proof.

8.1.1. *Inductive assumptions* Let $\Omega^{(0)} = \Omega$ and $\mathcal{P}^{(0)} = \{\Omega^{(0)}\}$ denote the trivial partition of Ω. Given $n \geqslant 1$ suppose that for each $k \leqslant n-1$ there exists a set $\Omega^{(k)} \subseteq \Omega$ satisfying the following properties.

Combinatorics: For the moment we describe the combinatorial structure as abstract data, the geometrical meaning of this data will become clear in the next section. There exists a partition $\mathcal{P}^{(k)}$ of $\Omega^{(k)}$ into intervals such that each $\omega \in \mathcal{P}^{(k)}$ has an associated *itinerary* constituted by the following information. To each $\omega \in \mathcal{P}^{(k)}$ is associated a sequence $0 = \theta_0 < \theta_1 < \cdots < \theta_r \leqslant k$, $r = r(\omega) \geqslant 0$ of *escape times*. Escape times are divided into three categories, i.e. *substantial*, *essential*, and *inessential*. Inessential escapes possess no combinatorial feature and are only relevant to the analytic bounded distortion argument to be developed later. Substantial and essential escapes play a role in splitting itineraries into segments in the following sense. Let $0 = \eta_0 < \eta_1 < \cdots < \eta_s \leqslant k$, $s = s(\omega) \geqslant 0$ be the maximal sequence of substantial and essential escape times. Between any of the two η_{i-1} and η_i (and between η_s and k) there is a sequence $\eta_{i-1} < \nu_1 < \cdots < \nu_t < \eta_i$, $t = t(\omega, i) \geqslant 0$ of *essential return times* (or *essential returns*) and between any two essential returns ν_{j-1} and ν_j (and between ν_t and η_i) there is a sequence $\nu_{j-1} < \mu_1 < \cdots < \mu_u < \nu_j$, $u = u(\omega, i, j) \geqslant 0$ of *inessential return times* (or *inessential returns*). Following essential and inessential return (respectively escape) there is a time interval $[\nu_j + 1, \nu_j + p_j]$ (respectively $[\mu_j + 1, \mu_j + p_j]$) with $p_j > 0$ called the *binding period*. A binding period cannot contain any return and escape times. Finally, associated to each essential and inessential return time (respectively escape) is a positive integer r called the *return depth* (respectively *escape depth*).

Bounded Recurrence: We define the function $\mathcal{E}^{(k)} : \Omega^{(k)} \to \mathbb{N}$ which associates to each $a \in \Omega^{(k)}$ the total sum of all essential return depths of the element $\omega \in \mathcal{P}^{(k)}$ containing a in its itinerary up to and including time k. Notice that $\mathcal{E}^{(k)}$ is constant on elements of $\mathcal{P}^{(k)}$ by construction. Then, for all $a \in \Omega^{(k)}$,

$$\mathcal{E}^{(k)}(a) \leqslant \alpha k. \tag{BR}_k$$

Slow Recurrence: For all $a \in \Omega^{(k)}$ and all $i \leqslant k$ we have

$$\left|c_i(a)\right| \geqslant e^{-\alpha i}. \tag{SR}_k$$

Notice that α can be chosen arbitrarily small as long as ε is small in order for this to hold for all $i \leqslant N$.

Hyperbolicity: For all $a \in \Omega^{(k)}$,

$$\left|\left(f_a^{k+1}\right)'(c_0)\right| \geqslant Ce^{\lambda_0(k+1)}. \qquad (EG)_k$$

Bounded Distortion: Critical orbits with the same combinatorics satisfy uniformly comparable derivative estimates: For every $\omega \in \mathcal{P}^{(k)}$, every pair of parameter values $a, b \in \omega$ and every $j \leqslant \nu + p + 1$ where ν is the last return or escape before or equal to time k and p is the associated binding period, we have

$$\frac{|(f_a^j)'(c_0)|}{|(f_b^j)'(c_0)|} \leqslant \mathcal{D} \quad \text{and} \quad \frac{|c_j'(a)|}{|c_j'(b)|} \leqslant \mathcal{D}. \qquad (BD)_k$$

Moreover, if k is a substantial escape a similar distortion estimate holds for all $j \leqslant l$ (l is the next chopping time) replacing $\mathcal{D}$ by $\tilde{\mathcal{D}}$ and ω by any subinterval $\omega' \subseteq \omega$ which satisfies $\omega_l' \subseteq \Delta^+$. In particular for $j \leqslant k$, the map $c_j : \omega \to \omega_j = \{c_j(a)\colon\ a \in \omega\}$ is a bijection.

8.1.2. *Definition of $\Omega^{(n)}$ and $\mathcal{P}^{(n)}$* For $r \in \mathbb{N}$, let $I_r = [e^{-r}, e^{-r+1})$, $I_{-r} = -I_r$ and define

$$\Delta^+ = \{0\} \cup \bigcup_{|r| \geqslant r_{\delta^+}+1} I_r \quad \text{and} \quad \Delta = \{0\} \cup \bigcup_{|r| \geqslant r_\delta + 1} I_r,$$

where $r_\delta = \log \delta^{-1}$, $r_{\delta^+} = \iota \log \delta^{-1}$. We can suppose without loss of generality that $r_\delta, r_{\delta^+} \in \mathbb{N}$. For technical reasons related to the distortion calculation we also need to subdivide each I_r into r^2 subintervals of equal length. This defines partitions $\mathcal{I}, \mathcal{I}^+$ of Δ^+ with $\mathcal{I} = \mathcal{I}^+|_\Delta$. An interval belonging to either one of these partitions is of the form $I_{r,m}$ with $m \in [1, r^2]$. Let $I_{r,m}^\ell$ and $I_{r,m}^\rho$ denote the elements of $\mathcal{I}^+$ adjacent to $I_{r,m}$ and let $\hat{I}_{r,m} = I_{r,m}^\ell \cup I_{r,m} \cup I_{r,m}^\rho$. If $I_{r,m}$ happens to be one of the extreme subintervals of $\mathcal{I}^+$ then let $I_{r,m}^\ell$ or $I_{r,m}^\rho$, depending on whether $I_{r,m}$ is a left or right extreme, denote the intervals $(-\delta^\iota - \frac{\delta^\iota}{(\log \delta^{-\iota})^2} - \delta^\iota]$ or $[\delta^\iota, \delta^\iota + \frac{\delta^\iota}{(\log \delta^{-\iota})^2})$, respectively. We now use this partition to define a refinement $\hat{\mathcal{P}}^{(n)}$ of $\mathcal{P}^{(n-1)}$. Let $\omega \in \mathcal{P}^{(n-1)}$. We distinguish two different cases.

Nonchopping times: We say that n is a nonchopping time for $\omega \in \mathcal{P}^{(n-1)}$ if one (or more) of the following situations occur:

(1) $\omega_n \cap \Delta^+ = \emptyset$;

(2) n belongs to the binding period associated to some return or escape time $\nu < n$ of ω;

(3) $\omega_n \cap \Delta^+ \neq \emptyset$ but ω_n does not intersect more than two elements of the partition $\mathcal{I}^+$.

In all three cases we let $\omega \in \hat{\mathcal{P}}^{(n)}$. In cases (1) and (2) no additional combinatorial information is added to the itinerary of ω. In case (3), if $\omega_n \cap (\Delta \cup I_{\pm r_\delta}) \neq \emptyset$ (respectively, $\omega_n \subset \Delta^+ \setminus (\Delta \cup I_{\pm r_\delta})$, we say that n is an inessential return time (respectively inessential escape time) for $\omega \in \hat{\mathcal{P}}^{(n)}$. We define the corresponding depth by $r = \max\{|r|\colon\ \omega_n \cap I_r \neq \emptyset\}$.

Chopping times: In all remaining cases, i.e. if $\omega_n \cap \Delta^+ \neq \emptyset$ and ω_n intersects at least three elements of $\mathcal{I}^+$, we say that n is a chopping time for $\omega \in \mathcal{P}^{(n-1)}$. We define a natural subdivision

$$\omega = \omega^\ell \cup \bigcup_{(r,m)} \omega^{(r,m)} \cup \omega^\rho,$$

so that each $\omega_n^{(r,m)}$ fully contains a unique element of $\mathcal{I}_+$ (though possibly extending to intersect adjacent elements) and ω_n^ℓ and ω_n^ρ are components of $\omega_n \setminus (\Delta^+ \cap \omega_n)$ with $|\omega_n^\ell| \geqslant \delta^\iota/(\log \delta^{-\iota})^2$ and $|\omega_n^\rho| \geqslant \delta^\iota/(\log \delta^{-\iota})^2$. If the connected components of $\omega_n \setminus (\Delta^+ \cup \omega_n)$ fail to satisfy the above condition on their length we just glue them to the adjacent interval of the form $w_n^{(r,m)}$. By definition we let each of the resulting subintervals of ω be elements of $\hat{\mathcal{P}}^{(n)}$. The intervals ω^ℓ, ω^ρ and $\omega^{(r,m)}$ with $|r| < r_\delta$ are called *escape components* and are said to have an *substantial escape* and *essential escape*, respectively, at time n. The corresponding values of $|r| < r_\delta$ are the associated *essential escape depths*. All other intervals are said to have an *essential return* at time n and the corresponding values of $|r|$ are the associated *essential return depths.* We remark that partition elements $I_{\pm r_\delta}$ do not belong to Δ but we still say that the associated intervals $\omega^{(\pm r_\delta, m)}$ have a return rather than an escape.

This completes the definition of the partition $\hat{\mathcal{P}}^{(n)}$ of $\Omega^{(n-1)}$ and of the function $\mathcal{E}^{(n)}$ on $\Omega^{(n-1)}$. We define

$$\Omega^{(n)} = \left\{a \in \Omega^{(n-1)} \colon \mathcal{E}^{(n)}(a) \leqslant \alpha n\right\}. \tag{27}$$

Notice that $\mathcal{E}^{(n)}$ is constant on elements of $\hat{\mathcal{P}}^{(n)}$. Thus $\Omega^{(n)}$ is the union of elements of $\hat{\mathcal{P}}^{(n)}$ and we can define

$$\mathcal{P}^{(n)} = \hat{\mathcal{P}}^{(n)}|_{\Omega^{(n)}}.$$

Notice that the combinatorics and the recurrence condition $(BR)_n$ are satisfied for every $a \in \Omega^{(n)}$ by construction. In Section 8.4 we shall prove that conditions $(EG)_n$, $(SR)_n$, $(BD)_n$ all hold for $\Omega^{(n)}$. Then we define

$$\Omega^* = \bigcap_{n \geqslant 0} \Omega^{(n)}.$$

In particular, for every $a \in \Omega^*$, the map f_a has an exponentially growing derivative along the critical orbit and thus, in particular, by Lemma 8 admits an ergodic *acip*. In Section 8.5 we prove that $|\Omega^*| > 0$.

We recall that a sketch of the proof of the existence of an induced Markov map under precisely the hyperbolicity and slow recurrence assumptions given here is carried out in Section 6.3. As mentioned there, the strategy for construction of the induced Markov map is remarkably similar to the strategy for the construction carried out here for estimating the probability that such conditions hold. The deeper meaning of this similarity is not clear.

8.2. *Expansion outside the critical neighbourhood*

On some deep level, the statement in Theorem 12 depends essentially on the following result which we have already used in Section 6.3.

LEMMA 8.1. *There exists a constant $C > 0$ independent of δ such that for $\varepsilon > 0$ sufficiently small, all $a \in \Omega_\varepsilon$, $f = f_a$, $x \in I$ and $n \geqslant 1$ such that $x, f(x), \ldots, f^{n-1}(x) \notin \Delta$ we have*

$$\left|Df^n(x)\right| \geqslant \delta e^{\lambda n}$$

and if, moreover, $f^n(x) \in \Delta^+$ and/or $x \in f(\Delta^+)$ then

$$\left|Df^n(x)\right| \geqslant Ce^{\lambda n}.$$

In the proof of the theorem we will use some other features of the quadratic family and of the specific parameter interval Ω_ε but it is arguable that they are inessential and that the statement of Lemma 8.1 are to a certain extent sufficient conditions for the argument. It would be very interesting to try to prove the main theorem using only the properties stated in Lemma 8.1. On a general "philosophical" level, this is based on the general principle, which goes back at least to the pioneering paper of Jakobson [86], that uniform hyperbolicity for *all parameters* in *most of the state space* implies nonuniform hyperbolicity in *all the state space* for *most parameters*.

8.3. *The binding period*

Next we make precise the definition of the *binding period* which is part of the combinatorial information given above. Let $k \leqslant n-1$, $\omega \in \mathcal{P}^{(k)}$ and suppose that k is an essential or inessential return or escape time for ω with return depth r. Then we define the *binding period* of ω_k as

$$p(\omega_k) = \min_{a\in\omega}\left\{p\left(c_k(a)\right)\right\},$$

where

$$p\left(c_k(a)\right) = \min\left\{i\colon \left|c_{k+1+i}(a) - c_i(a)\right| \geqslant e^{-2\alpha i}\right\}. \tag{28}$$

This is the time for which the future orbit of $c_k(a)$ can be thought of as *shadowing* or being *bound to* the orbit of the critical point (that is, in some sense, the number of iterations for which the orbit of $c(a)$ repeats its early history after the kth iterate). We will obtain some estimates concerning the length of this binding period and the overall derivative growth during this time.

LEMMA 8.2. *There exist constants $\tau_0 > 0$ and $\gamma_1 \in (0,1)$ such that the following holds. Let $k \leqslant n-1$, $\omega \in \mathcal{P}^{(k)}$ and suppose that k is an essential or inessential return or escape time for ω with return depth r. Let $p = p(\omega_k)$. Then for every $a \in \omega$ we have*

$$p \leqslant \tau_0 \log \left|c_k(a)\right|^{-1} < k \tag{29}$$

and

$$\left|Df^{p+1}\big(c_k(a)\big)\right| \geqslant Ce^{r(1-\gamma_1)} \geqslant Ce^{\frac{1-\gamma_1}{\tau_0}(p+1)} \tag{30}$$

and, if k is an essential return or an essential escape, then

$$|\omega_{k+p+1}| \geqslant Ce^{-\gamma_1 r}. \tag{31}$$

To simplify the notation we write $x = c_k(a)$ and $x_0 = c_{k+1}(a)$ and omit the dependence on the parameter a where there is no risk of confusion. The first step in the proof is to obtain a *bounded distortion* estimate during binding periods: there exists a constant $\mathcal{D}_1(\alpha_0, \alpha_1)$ independent of x, such that for all $a \in \omega$, all $y_0, z_0 \in [x_0, c_0]$ and all $0 \leqslant j \leqslant \min\{p-1, k\}$ we have

$$\left|\frac{(f^j)'(z_0)}{(f^j)'(y_0)}\right| < \mathcal{D}_1.$$

This follows from the standard distortion calculations as in (8) on p. 285, using the upper bound $e^{-2\alpha i}$ from the definition of binding in the numerator and the lower bound $e^{-\alpha i}$ from the bounded recurrence condition $(SR)_k$ in the denominator. Notice that for this reason the distortion bound is formally calculated for iterates $j \leqslant \min\{p-1, k\}$ (the bounded recurrence condition cannot be guaranteed for iterates larger than k). The next step however gives an estimate for the duration of the binding period and implies that $p < k$ and therefore the distortion estimates do indeed hold throughout the duration of the binding period. The basic idea for the upper bound on p is simple. The length of the interval $[x_0, c_0]$ is determined by the length of the interval $[x, c]$ which is $c_k(a)$. The exponential growth of the derivative along the critical orbit and the bounded distortion imply that this interval is growing exponentially fast. The condition which determines the end of the binding period is shrinking exponentially fast. Some standard mean value theorem estimates using these two facts give the result. Finally the average derivative growth during the binding is given by the combined effect of the small derivative of order $c_k(a)$ at the return to the critical neighbourhood and the exponential growth during the binding period. The result then intuitively boils down to showing that the binding period is long enough to (over) compensate the small derivative at the return.

The final statement in the lemma requires some control over the way that the derivatives with respect to the parameter are related to the standard derivatives with respect to a point. This is a fairly important point which will be used again and therefore we give a more formal statement.

LEMMA 8.3. *There exists a constant $\mathcal{D}_2 > 0$ such that for any $1 \leqslant k \leqslant n-1$, $\omega \in \mathcal{P}^{(k-1)}$ and $a \in \omega$ we have*

$$\mathcal{D}_2 \geqslant \frac{|c_k'(a)|}{|Df_a^k(c_0)|} \geqslant \mathcal{D}_2$$

and, for all $1 \leqslant i < j \leqslant k+1$, there exists $\tilde{a} \in \omega$ such that

$$\frac{1}{\mathcal{D}_2}\big|Df_{\tilde{a}}^{j-i}\big(c_i(\tilde{a})\big)\big| \leqslant \frac{|\omega_j|}{|\omega_i|} \leqslant \mathcal{D}_2\big|Df_{\tilde{a}}^{j-i}\big(c_i(\tilde{a})\big)\big|. \tag{32}$$

PROOF. The second statement is a sort of *parameter mean value theorem* and follows immediately from the first one and the standard mean value theorem. To prove the first one let $F : \Omega \times I \to I$ be the function of two variables defined inductively by $F(a,x) = f_a(x)$ and $F^k(a,x) = F(a, f_a^{k-1}x)$. Then, for $x = c_0$, we have

$$c_k'(a) = \partial_a F^k(a, c_0) = \partial_a F\big(a, f_a^{k-1}c_0\big) = -1 + f_a'(c_{k-1})c_{k-1}'(a).$$

Iterating this expression gives

$$\begin{aligned} -c_k'(a) = 1 &+ f_a'(c_{k-1}) + f_a'(c_{k-1})f_a'(c_{k-2}) + \cdots \\ &+ f_a'(c_{k-1})f_a'(c_{k-2})\ldots f_a'(c_1)f_a'(c_0) \end{aligned}$$

and dividing both sides by $(f^k)'(c_0) = f_a'(c_{k-1})f_a'(c_{k-2})\ldots f_a'(c_1)f_a'(c_0)$ gives

$$\frac{c_k'(a)}{(f_a^k)'(c_0)} = 1 + \sum_{i=1}^{k} \frac{1}{(f^i)'(c_0)}. \tag{33}$$

The result then depends on making sure that the sum on the right-hand side is bounded away from -1. Since the critical point spends an arbitrarily large number N of iterates in an arbitrarily small neighbourhood of a fixed point at which the derivative is -4 we can bound an arbitrarily long initial part of this sum by $-1/2$. By the exponential growth condition the tail of the sum is still geometric and by taking N large enough we can make sure that this tail is less than $1/2$ in absolute value. □

Returning to the proof of Lemma 8.2 we can use the parameter/space derivative bound to extend the derivative expansion result to the entire interval ω_k and therefore to estimate the growth of this interval during the binding period.

8.4. *Positive exponents in dynamical space*

Using a combination of the expansivity estimates outside Δ and the binding period estimates for returns to Δ it is possible to prove the inductive step stated above.

The slow recurrence condition is essentially an immediate consequence of the parameter exclusion condition.

The exponential growth condition relies on the following crucial and nontrivial observation: *the overall proportion of bound iterates is small*. This follows from the parameter exclusion condition which bounds the total sum of return depths (an estimate is required to show that inessential return do not contribute significantly to the total) and the binding period estimates which show that the length of the binding period is bounded by a fraction of the return depth. This implies that the overall derivative growth is essentially built up from the free iterates outside Δ and this gives an overall derivative growth at an exponential rate independent of n.

The bounded distortion estimates again starts with the basic estimate as in (8) on p. 285. By Lemma 8.3 it is sufficient to prove the estimate for the space derivatives Df^k; intuitively this is saying that critical orbits with the same combinatorics satisfy the same derivative estimates. The difficulty here is that although the images of parameter intervals ω are growing exponentially, they do not satisfy a uniform *backward exponential* bound as required to carry out the step leading to (9) on p. 286; also images of ω can come arbitrarily close to the critical point and thus the denominator does not admit any uniform bounds. The calculation therefore is technically quite involved and we refer the reader to published proofs such as [103] for the details. Here we just mention that the argument involves decomposing the sum into "pieces" corresponding to free and bound iterates and estimating each one independently, and taking advantage of the subdivision of the critical neighbourhood into interval I_r each of which is crucially further subdivided into further r^2 subintervals of equal length. This implies that the contribution to the distortion of each return is at most of the order of $1/r^2$ instead of order 1 and allows us to obtain the desired conclusion using the fact that $1/r_2$ is summable in r.

8.5. *Positive measure in parameter space*

Recall that $\hat{\mathcal{P}}^{(n)}$ is the partition of $\Omega^{(n-1)}$ which takes into account the dynamics at time n and which restricts to the partition $\mathcal{P}^{(n)}$ of $\Omega^{(n)}$ after the exclusion of a certain elements of $\hat{\mathcal{P}}^{(n)}$. Our aim here is to develop some combinatorial and metric estimates which will allow us to estimate the measure of parameters to be excluded at time n.

The first step is to take a fresh look at the combinatorial structure and "reformulate it" in a way which is more appropriate. To each $\omega \in \hat{\mathcal{P}}^{(n)}$ is associated a sequence $0 = \eta_0 < \eta_1 < \cdots < \eta_s \leqslant n$, $s = s(\omega) \geqslant 0$ of escape times and a corresponding sequence of escaping components $\omega \subseteq \omega^{(\eta_s)} \subseteq \cdots \subseteq \omega^{(\eta_0)}$ with $\omega^{(\eta_i)} \subseteq \Omega^{(\eta_i)}$ and $\omega^{(\eta_i)} \in \mathcal{P}^{(\eta_i)}$. To simplify the formalism we also define some "fake" escapes by letting $\omega^{(\eta_i)} = \omega$ for all $s+1 \leqslant i \leqslant n$. In this way we have a well-defined parameter interval $\omega^{(\eta_i)}$ associated to $\omega \in \hat{\mathcal{P}}^{(n)}$ for each $0 \leqslant i \leqslant n$. Notice that for two intervals $\omega, \tilde{\omega} \in \hat{\mathcal{P}}^{(n)}$ and any $0 \leqslant i \leqslant n$, the corresponding intervals $\omega^{(\eta_i)}$ and $\tilde{\omega}^{(\eta_i)}$ are either disjoint or coincide. Then we define

$$Q^{(i)} = \bigcup_{\omega \in \hat{\mathcal{P}}^{(n)}} \omega^{(\eta_i)}$$

and let $\mathcal{Q}i = \{\omega^{(\eta_i)}\}$ denote the natural partition of $Q^{(i)}$ into intervals of the form $\omega^{(\eta_i)}$. Notice that $\Omega^{(n-1)} = Q^{(n)} \subseteq \cdots \subseteq Q^{(0)} = \Omega^{(0)}$ and $\mathcal{Q}n = \hat{\mathcal{P}}^{(n)}$ since the number s of escape times is always strictly less than n and therefore in particular $\omega^{(\eta_n)} = \omega$ for all $\omega \in \hat{\mathcal{P}}^{(n)}$. For a given $\omega = \omega^{(\eta_i)} \in \mathcal{Q}i$, $0 \leqslant i \leqslant n-1$, we let

$$Q^{(i+1)}(\omega) = \left\{\omega' = \omega^{(\eta_{i+1})} \in \mathcal{Q}i + 1\colon\ \omega' \subseteq \omega\right\}$$

denote all the elements of $\mathcal{Q}i + 1$ which are contained in ω and let $\mathcal{Q}i + 1(\omega)$ denote the corresponding partition. Then we define a function $\Delta\mathcal{E}i : Q^{(i+1)}(\omega) \to \mathbb{N}$ by

$$\Delta\mathcal{E}i(a) = \mathcal{E}\eta_{i+1}(a) - \mathcal{E}\eta_i(a).$$

This gives the total sum of all essential return depths associated to the itinerary of the element $\omega' \in \mathcal{Q}i + 1(\omega)$ containing a, between the escape at time η_i and the escape at time η_{i+1}. Clearly $\Delta\mathcal{E}i(a)$ is constant on elements of $\mathcal{Q}i + 1(\omega)$. Finally we let

$$\mathcal{Q}i + 1(\omega, R) = \left\{\omega' \in \mathcal{Q}^{(i+1)}\colon\ \omega' \subseteq \omega,\ \Delta\mathcal{E}i(\omega') = R\right\}.$$

Notice that the entire construction given here depends on n. The main motivation for this construction and is the following

LEMMA 8.4. *There exists a constant $\gamma_0 \in (0, 1-\gamma_1)$ such that the following holds. For all $i \leqslant n-1$, $\omega \in \mathcal{Q}i$ and $R \geqslant 0$ we have*

$$\sum_{\tilde{\omega} \in \mathcal{Q}i+1(\omega,R)} |\tilde{\omega}| \leqslant e^{(\gamma_1+\gamma_0-1)R}|\omega|. \tag{34}$$

This says essentially that the probability of accumulating a large total return depth between one escape and the next is exponentially small. The strategy for proving this result is straightforward. We show first of all that for $0 \leqslant i \leqslant n-1$, $\omega \in \mathcal{Q}i$, $R \geqslant 0$ and $\tilde{\omega} \in \mathcal{Q}i + 1(\omega, R)$ we have

$$|\tilde{\omega}| \leqslant e^{(\gamma_1-1)R}|\omega|. \tag{35}$$

The proof is not completely straightforward but depends on the intuitively obvious fact that an interval which has a deep return must necessarily be very small (since it is only allowed to contain at most three adjacent partition elements at the return). Notice, moreover, that this statement on its own is not sufficient to imply (34) as there could be many small intervals which together add up to a lot of intervals having large return. However we can control to some extent the multiplicity of these intervals and show that we can choose an arbitrarily small γ_0 (by choosing the critical neighbourhood Δ sufficiently small) so that for all $0 \leqslant i \leqslant n-1$, $\omega \in \mathcal{Q}^{(i)}$ and $R \geqslant r_\delta$, we have

$$\#\mathcal{Q}_n^{(i+1)}(\omega, R) \leqslant e^{\gamma_0 R}. \tag{36}$$

This depends on the observation that each ω has an essentially unique (uniformly bounded multiplicity) sequence of return depths. Thus the estimate can be approached via purely combinatorial arguments very similar to those used in relation to Equation (17). Choosing δ small means the sequences of return depths have terms bounded below by r_δ which can be chosen large, and this allows the exponential rate of increase of the combinatorially distinct sequences with R to be taken small. Combining (35) and (36) immediately gives (34).

Now choose some $\gamma_2 \in (0, 1-\gamma_0-\gamma_1)$ and let

$$\gamma = \gamma_0 + \gamma_1 + \gamma_2 > 0.$$

For $0 \leqslant i \leqslant n-1$, and $\omega \in \mathcal{Q}i$, write

$$\sum_{\omega' \in \mathcal{Q}i+1(\omega)} e^{\gamma_2 \Delta\mathcal{E}i(\omega')}|\omega'| = \sum_{\omega' \in \mathcal{Q}i+1(\omega,0)} |\omega'| + \sum_{R \geqslant r_\delta} e^{\gamma_2 R} \sum_{\omega' \in \mathcal{Q}i+1(\omega,R)} |\omega'|.$$

By (34) we then have

$$\sum_{\omega' \in \mathcal{Q}i+1(\omega,0)} |\omega'| + \sum_{R \geqslant r_\delta} e^{\gamma_2 R} \sum_{\omega' \in \mathcal{Q}i+1(\omega,R)} |\omega'| \leqslant \left(1 + \sum_{R \geqslant r_\delta} e^{(\gamma_0+\gamma_1+\gamma_2-1)R}\right)|\omega|. \tag{37}$$

Since $\mathcal{E}n = \Delta\mathcal{E}0 + \cdots + \Delta\mathcal{E}n-1$ and $\Delta\mathcal{E}i$ is constant on elements of $\mathcal{Q}i$ we can write

$$\sum_{\omega \in \mathcal{Q}n} e^{\gamma_2 \mathcal{E}n(\omega)}|\omega| = \sum_{\omega^{(1)} \in \mathcal{Q}1(\omega^{(0)})} e^{\gamma_2 \Delta\mathcal{E}0(\omega^{(1)})} \sum_{\omega^{(2)} \in \mathcal{Q}2(\omega^{(1)})} e^{\gamma_2 \Delta\mathcal{E}1(\omega^{(2)})} \cdots$$
$$\sum_{\omega^{(n-1)} \in \mathcal{Q}n-1(\omega^{(n-2)})} e^{\gamma_2 \Delta\mathcal{E}n-1(\omega^{(n-1)})} \sum_{\omega=\omega^{(n)} \in \mathcal{Q}n} e^{\gamma_2 \Delta\mathcal{E}n-1(\omega^{(n)})}|\omega|.$$

Notice the *nested* nature of the expression. Applying (37) repeatedly gives

$$\int_{\Omega^{(n-1)}} e^{\gamma_2 \mathcal{E}n} = \sum_{\omega \in \mathcal{Q}n} e^{\gamma_2 \mathcal{E}n(\omega)}|\omega| \leqslant \left(1 + \sum_{R \geqslant r_\delta} e^{(\gamma-1)R}\right)^n |\Omega|. \tag{38}$$

The definition of $\Omega^{(n)}$ gives

$$\left|\Omega^{(n-1)}\right| - \left|\Omega^{(n)}\right| = \left|\Omega^{(n-1)} \setminus \Omega^{(n)}\right| = \left|\left\{\omega \in \hat{\mathcal{P}}^{(n)} = \mathcal{Q}n\colon\ e^{\gamma_2 \mathcal{E}n} \geqslant e^{\gamma_2 \alpha n}\right\}\right|$$

and therefore using Chebyshev's inequality and (38) we have

$$\left|\Omega^{(n-1)}\right| - \left|\Omega^{(n)}\right| \leqslant e^{-\gamma_2 \alpha n} \int_{\Omega^{(n-1)}} e^{\gamma_2 \mathcal{E}n} \leqslant \left[e^{-\gamma_2 \alpha}\left(1 + \sum_{R \geqslant r_\delta} e^{(\gamma-1)R}\right)\right]^n |\Omega|$$

which implies

$$\left|\Omega^{(n)}\right| \geqslant \left|\Omega^{(n-1)}\right| - \left[e^{-\gamma_2 \alpha}\left(1 + \sum_{R \geqslant r_\delta} e^{(\gamma-1)R}\right)\right]^n |\Omega|$$

and thus

$$|\Omega^*| \geqslant \left(1 - \sum_{j=N}^{\infty} \left[e^{-\gamma_2 \alpha} \left(1 + \sum_{R \geqslant r_\delta} e^{(\gamma-1)R}\right)\right]^j\right) |\Omega|.$$

Choosing N sufficiently large, by taking ε sufficiently small, guarantees that the right-hand side is positive.

9. Conclusion

In this final section we make some concluding remarks and present some questions and open problems.

9.1. *What causes slow decay of correlations?*

The general theory described in Section 7 is based on a certain way of quantifying the intrinsic nonuniformity of f which does not rely on identifying particular critical and/or neutral orbits. However, the conceptual picture according to which slow rates of decay are caused by a slowing down process due to the presence of neutral orbits can also be generalized. Indeed, the abstract formulation of the concept of a neutral orbit is naturally that of an orbit with a zero Lyapunov exponent. The definition of nonuniform expansivity implies that almost all orbits have uniformly positive Lyapunov exponents but this does not exclude the possibility of some other point having a zero Lyapunov exponent. It seems reasonable to imagine that a point with a zero Lyapunov exponent could *slow down* the overall mixing process in a way which is completely analogous to the specific examples mentioned above. Therefore we present here, in a heuristic form, a natural conjecture.

CONJECTURE 1. *Suppose f is nonuniformly expanding. Then f has exponential decay of correlations if and only it has no invariant measures with all zero Lyapunov exponents.*

An attempt to state this conjecture in a precise way reveal several subtle points which need to be considered. We discuss some of these briefly. Let $\mathcal{M}$ denote the space of all probability f-invariant measures μ on M which satisfy the integrability condition $\int \log \|Df_x\| \, d\mu < \infty$. Then by standard theory, see also [1, Section 5.8], we can apply a version of Oseledets' Theorem for noninvertible maps which says that there exist constants $\lambda_1, \ldots, \lambda_k$ with $k \leqslant d$, and a measurable decomposition $T_x M = E_x^1 \oplus \cdots \oplus E_x^k$ of the tangent bundle over M such that the decomposition is invariant by the derivative and such that for all $j = 1, \ldots, k$ and for all non zero vectors $v^{(j)} \in E_x^j$ we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \log \left\| Df_x^n \left(v^{(j)}\right)\right\| = \lambda_j.$$

The constants $\lambda_1, \ldots, \lambda_k$ are called the *Lyapunov exponents* associated to the measure μ. The definition of nonuniform expansivity implies that all Lyapunov exponents associated to the *acip* μ are $\geqslant \lambda$ and thus uniformly positive, but it certainly does not exclude the possibility that there exist some other (singular with respect to Lebesgue) invariant probability measure with some zero Lyapunov exponent. This is the case, for example, for the maps of Section 5 for which the Dirac measure on the indifferent fixed point has a zero Lyapunov exponent.

Thus one way to state precisely the above conjecture is to claim that f has exponential decay of correlations if and only if all Lyapunov exponents associated to all invariant probability measure in $\mathcal{M}$ are uniformly positive. Of course, a priori, there may also be some exceptional points, not typical for any measure in $\mathcal{M}$, along whose orbit the derivative expands subexponentially and which therefore might similarly have a slowing down effect. Also it may be that one zero Lyapunov exponent along one specific direction may not have a significant effect whereas a measure for which all Lyapunov exponents were zero would. Positive results in the direction of this conjecture include the remarkable observation that local diffeomorphisms for which all Lyapunov exponents for all measures are positive, must actually be uniformly expanding [14,57,58] and thus in particular have exponential decay of correlations. Moreover, in the context of one-dimensional smooth maps with critical points it is known that in the unimodal case exponential growth of the derivative along the critical orbit (the Collet–Eckmann condition) implies uniform hyperbolicity on periodic orbits [120] which in turn implies that all Lyapunov exponents of all measures as positive [41] and the converse is also true [122]. Thus Conjecture 1 is true in the one-dimensional unimodal setting.

We remark that the assumption of nonuniform expansion is crucial here. There are several examples of systems which have exponential decay of correlations but clearly have invariant measures with zero Lyapunov exponents, e.g., partially hyperbolic maps or maps obtained as time-1 maps of certain flows [65–67]. These examples however are not nonuniformly expanding, and are generally *partially hyperbolic* which means that there are two continuous subbundles such that the derivative restricted to one subbundle has very good expanding properties or contracting properties and the other subbundle has the zero Lyapunov exponents. For reasons which are not at all clear, this might be *better* from the point of view of decay of correlations than a situation in which all the Lyapunov exponents of the absolutely continuous measure are positive but there is some *embedded* singular measure with zero Lyapunov exponent slowing down the mixing process. Certainly there is still a lot to be understood on this topic.

9.2. *Stability*

The results on the existence of nonuniformly expanding maps for open sets or positive measure sets of parameters are partly *stability* results. They say that certain properties of a system, e.g., being nonuniformly expanding, are stable in a certain sense. We mention here two other forms of stability which can be investigated.

9.2.1. *Topological rigidity* The notion of (nonuniform) expansivity is, a priori, completely metrical: it depends on the differentiable structure of f and most constructions and estimates related to nonuniform expansivity require delicate metric distortion bounds. However the statistical properties we deduce (the existence of an *acip*, the rate of decay of correlation) are objects and quantities which make sense in a much more general setting. A natural question therefore is whether the metric properties are really necessary or just very useful conditions and to what extent the statistical properties might depend only on the underlying topological structure of f. We recall that two maps $f : M \to M$ and $g : N \to N$ are *topologically conjugate*, $f \sim g$, if there exists a homeomorphism $h : M \to N$ such that $h \circ f = g \circ h$. We say that a property of f is *topological* or depends only on the *topological structure* of f if it holds for all maps in the topological conjugacy class of f.

The existence of an absolutely continuous invariant measure is clearly not a topological invariant in general: if μ_f is an *acip* for f then we can define $\mu_g = h^*\mu_f$ by $\mu_g(A) = \mu_f(h^{-1}(A))$ which gives an invariant probability measure but not absolutely continuous unless the conjugating homeomorphism h is itself absolutely continuous. For example, the map of Theorem 6 has no *acip* even though it is topologically conjugate to a *uniformly* expanding Markov map. However it turns out, quite remarkably, that there are many situations in the setting of one-dimensional maps with critical points in which the existence of an *acip* is indeed a topological property (although there are also examples in which it is not [39]). Topological conditions which imply the existence of an *acip* for unimodal maps were given in [38,146,40]. In [121] (bringing together results of [122,133]) it was shown that the exponential growth condition along the critical orbit for unimodal maps (which in particular implies the existence of an *acip*, see Theorem 8) is a topological property. A counterexample to this result in the multimodal case was obtained in [132]. However it was shown in [109] that in the general multimodal case, if all critical points are *generic* with respect to the *acip*, then the existence of an *acip* still holds for all maps in the same conjugacy class (although not necessarily the genericity of the critical points).

We emphasize that all these results do not rely on showing that all conjugacies in question are absolutely continuous. Rather they depend on the existence of some topological property which forces the existence of an *acip* in each map in the conjugacy class. These *acip*'s are generally *not* mapped to each other by the conjugacy.

9.2.2. *Stochastic stability* Stochastic stability is one way to formalize the idea that the statistical properties of a dynamical systems are *stable* under small random perturbations. There are several positive results on stochastic stability for uniformly expanding [166,26, 61] and nonuniformly expanding maps in dimension 1 [25,21] and higher [20,19]. See [13] for a comprehensive treatment of the results.

9.3. *Nonuniform hyperbolicity and induced Markov maps*

The definition of nonuniform hyperbolicity in terms of conditions $(*)$ and $(**)$ given above are quite natural as they are assumptions which do not a priori require the existence of an invariant measure. However they do imply the existence of an *acip* μ which has all positive Lyapunov exponents. Thus the system (f, μ) is also nonuniformly expanding in the more

abstract sense of *Pesin theory*, see [1]. The systematic construction of induced Markov maps in many examples and under quite general assumptions, as described above, naturally leads to the question of whether such a construction is always possible in this abstract setting. Since the existence of an induced Markov map implies nonuniform expansivity this would essentially give an equivalent characterization of nonuniform expansivity. A general result in this direction has been given for smooth one-dimensional maps in [145]. It would be interesting to extend this to arbitrary dimension. A generalization to nonuniformly hyperbolic surface diffeomorphisms is work in progress [105].

It seems reasonable to believe that the scope of application of induced Markov towers may go well beyond the statistical properties of a map f. The construction of the induced Markov map in [145], for example, is primarily motivated by the study if the Hausdorff dimension of certain sets. A particularly interesting application would be a generalization of the existence (parameter exclusion) argument sketched in Section 8. Even in a very general setting, with no information about the map f except perhaps the existence of an induced Markov map, it is natural to ask about the possible existence of induced Markov maps for small perturbations of f. If, moreover, the existence of an induced Markov map were essentially equivalent to nonuniform expansivity then this would be a question about the persistence of nonuniform expansivity under small perturbations.

CONJECTURE 2. *Suppose that f is nonuniformly expanding. Then sufficiently small perturbations of f have positive probability of also being nonuniformly expanding.*

Using the Markov induced maps one could define, even in a very abstract setting, a *critical region* Δ formed by those points that have very large return time. Then outside Δ one would have essentially uniform expansivity and these, as well as the Markov structure, would essentially persist under small perturbations. One could then perturb f and, up to parameter exclusions, try to show that the Markov structure can be extended once again to the whole of Δ for some nearby map g.

9.4. *Verifying nonuniform expansivity*

The verification of the conditions of nonuniform hyperbolicity are a big problem on both a theoretical and a practical level. As mentioned in Section 8, for the important class of one-dimensional maps with critical point, nonuniform expansivity occurs with positive probability but for sets of parameters which are topologically negligible and thus essentially impossible to pinpoint exactly. The best we can hope for is to show they occur with "very high" probability in some given small range of parameter values.

However even this is generally impossible with the available techniques. Indeed, all existing argument rely on choosing a *sufficiently small* parameter interval *centred* on some *sufficiently good* parameter value. The closeness to this parameter value is then used to obtain the various conditions which are required to start the induction. However the problem then reduces to showing that such a good parameter value exists in the particular parameter interval of interest, and this is again both practically and theoretically impossible in general. Moreover, even if such a parameter value was determined (as in the special case of

the "top" quadratic map) existing estimates do not control the size of the neighbourhood in which the good parameters are obtained nor the relative proportion of good parameters. For example, there are no explicit bounds for the actual measure of the set of parameters in the quadratic family which have an *acip*. A standard coffee-break joke directed towards authors of the papers on the existence of such maps is that so much work has gone into proving the existence of a set of parameters which as far as we know might be infinitesimal. Moreover, there just does not seem to be any even heuristic argument for believing that such a set is or is not very small. Thus, for no particular reason other than a reaction to these coffee-break jokers (!), we formulate the following

CONJECTURE 3. *The set of parameters in the quadratic family which admit an acip is* "*large*".

For definiteness let us say that "large" means at least 50% but it seems perfectly reasonable to expect even 80% or 90%, and of course we mean here those parameters between the Feigenbaum period doubling limit and the top map. An obvious strategy for proving (or disproving) this conjecture would be to develop a technique for estimating the proportion of maps having an *acip* in any given small one-parameter family of maps. The large parameter interval of the quadratic family could then be subdivided into small intervals each and the contribution of each of these small intervals could then be added up.

A general technique of this kind would also be interesting in a much broader context of applications. As mentioned in the introduction, many real-life systems appear to have a combination of deterministic and random-like behaviour which suggests that some form of expansivity and/or hyperbolicity might underly the basic driving mechanisms. In modelling such a system it seems likely that one may obtain a parametrized family and be interested in a possibly narrow range of parameter values. It would be desirable therefore to be able to obtain a rigorous prove of the existence of stochastic like behaviour such as mixing with exponential decay of correlations in this family and to be able to estimate the probability of such behaviour occurring. An extremely promising strategy has been proposed recently by K. Mischaikow. The idea is to combine nontrivial numerical estimates with the geometric and probabilistic parameter exclusion argument discussed above. Indeed the parameter exclusion argument, see Section 8, relies fundamentally on an induction which shows that the probability of being excluded at time n are exponentially small in n. The implementation of this argument however also requires several delicate relations between different system constants to be satisfied and in particular no exclusions to be required before some sufficiently large N so that the exponentially small exclusions occurring for $n > N$ cannot cumulatively add up to the full measure of the parameter interval under consideration. The assumption of the existence of a particularly good parameter value a^* and the assumption that the parameter interval is a sufficiently small neighbourhood of a^* are used in all existing proofs to make sure that certain constants can be chosen arbitrarily small or arbitrarily large thus guaranteeing that the necessary relations are satisfied. Mischaikow's suggestion is to reformulate the induction argument in such a way that the inductive assumptions can, at least in principle, be explicitly verified computationally. This requires the dependence of all the constants in the argument to be made completely explicit in such a way that the

inductive assumptions boil down to a finite set of open conditions on the family of maps which can be verified with finite precision in finite time.

Besides the interest of the argument in this particular setting this could perhaps develop into an extremely fruitful interaction between the "numerical" and the "geometric/probabilistic" approach to Dynamical Systems, and contribute significantly to the applicability of the powerful methods of Dynamical Systems to the solution and understanding of real-life phenomena.

REMARK. Since this survey was written, significant progress has been made on the general strategy proposed by Mischaikow, mentioned above. A first explicit rigorous lower bound for the set of good parameters in the quadratic family has been obtained in [106]. The argument is completely analytic and restricted to a small parameter interval near the "top" of the family. The paper, however, contains a general formula for the relative measure of good parameters in terms of a small number of constants, all explicitly computable within a finite number of steps. The most difficult constants to calculate are those related to the expansivity outside a critical neighbourhood, as in Lemma 8.1 above. These require some nontrivial numerical analysis arguments which have recently been developed in [63]. It can be expected therefore that some reasonable estimates for the overall proportion of good parameters in the quadratic family will be obtained by combining these two arguments. The author has recently become aware of numerical computations of Simó and Tatjer [151] which provide reasonably reliable lower bounds for the set of regular parameters in the quadratic family, of the order of about 10% of the interval of parameters for which interesting dynamics occurs.

Acknowledgements

Thanks to Gerhard Keller, Giulio Pianigiani, Mark Pollicott, Omri Sarig, and Benoit Saussol for useful comments on a preliminary version of this paper. Simon Cedervall made some useful contributions to some parts of Section 6.3. Thanks also to the editors, Boris Hasselblatt and Anatole Katok, for their support and encouragement. Thanks also the referee who suggested several references for Sections 6 and 8.

References

Surveys in volume 1A and this volume

[1] L. Barreira and Ya. Pesin, *Smooth ergodic theory and nonuniformly hyperbolic dynamics*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 57–263.

[2] N. Chernov, *Invariant measures for hyperbolic dynamical systems*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 321–407.

[3] B. Hasselblatt, *Hyperbolic dynamical systems*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 239–319.

[4] B. Hasselblatt and A. Katok, *Principal structures*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 1–203.

[5] B. Hasselblatt and Ya. Pesin, *Partially hyperbolic dynamical systems*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 1–55.
[6] M. Jakobson and G. Świątek, *One-dimensional maps*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 599–666.
[7] G. Knieper, *Hyperbolic dynamics and Riemannian geometry*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 453–545.
[8] M. Pollicott, *Periodic orbits and zeta functions*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 409–452.
[9] E. Pujals and M. Sambarino, *Homoclinic bifurcations, dominated splitting and robust transitivity*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 327–378.

Other sources

[10] J. Aaronson, *An introduction to infinite ergodic theory*, Mathematical Surveys and Monographs, Vol. 50, Amer. Math. Soc. (1997).
[11] R.L. Adler, *F-expansions revisited*, Recent Advances in Topological Dynamics, Proc. Conf., Yale Univ., New Haven, Conn., 1972, in honor of Gustav Arnold Nedlund, Lecture Notes in Math., Vol. 318, Springer, Berlin (1973), 1–5.
[12] J.F. Alves, *SRB measures for non-hyperbolic systems with multidimensional expansion*, Ann. Sci. École Norm. Sup. (4) **33** (2000), 1–32.
[13] J.F. Alves, *Statistical analysis of nonuniformly expanding dynamical systems* (2003), Preprint.
[14] J.F. Alves, V. Araujo and B. Saussol, *On the uniform hyperbolicity of some nonuniformly hyperbolic systems*, Proc. Amer. Math. Soc. **131** (4) (2003), 1303–1309.
[15] J.F. Alves, C. Bonatti and M. Viana, *SRB measures for partially hyperbolic systems whose central direction is mostly expanding*, Invent. Math. **140** (2000), 351–398.
[16] J.F. Alves, S. Luzzatto and V. Pinheiro, *Markov structures and decay of correlations for non-uniformly expanding maps on compact manifolds of arbitrary dimension*, Electron. Res. Announc. Amer. Math. Soc. **9** (2003), 26–31.
[17] J.F. Alves, S. Luzzatto and V. Pinheiro, *Lyapunov exponents and rates of mixing for one-dimensional maps*, Ergodic Theory Dynam. Systems **24** (2004), 637–657, DOI 10.1017/S0143385703000579.
[18] J.F. Alves, S. Luzzatto and V. Pinheiro, *Markov structures and decay of correlations for non-uniformly expanding maps*, Ann. Inst. H. Poincaré Anal. Non Linéaire (2004), to appear.
[19] J.F. Alves and M. Viana, *Statistical stability for robust classes of maps with non-uniform expansion*, Ergodic Theory Dynam. Systems **22** (2002), 1–32.
[20] V. Araújo, *Infinitely many stochastically stable attractors*, Nonlinearity **14** (2001), 583–596.
[21] V. Araújo, M. Viana and S. Luzzatto, *Weak expansion implies stochastic stability*, in progress.
[22] A. Avez, *Propriétés ergodiques des endomorphisms dilatants des variétés compactes*, C. R. Acad. Sci. Paris Sér. A–B **266** (1968), A610–A612 (in French).
[23] V. Baladi, *Decay of correlations*, Smooth Ergodic Theory and Its Applications, Seattle, WA, 1999, Proc. Sympos. Pure Math., Vol. 69, Amer. Math. Soc., Providence, RI (2001), 297–325.
[24] V. Baladi and S. Gouëzel, *A note on stretched exponential decay of correlations for the Viana–Alves map*, http://front.math.ucdavis.edu/math.DS/0311189, Preprint.
[25] V. Baladi and M. Viana, *Strong stochastic stability and rate of mixing for unimodal maps*, Ann. Sci. École Norm. Sup. (4) **29** (1996), 483–517.
[26] V. Baladi and L.-S. Young, *On the spectra of randomly perturbed expanding maps*, Comm. Math. Phys. **156** (1993), 355–385.
[27] L. Barreira and Y.B. Pesin, *Lyapunov exponents and smooth ergodic theory*, University Lecture Series, Vol. 23, Amer. Math. Soc., Providence, RI (2002), ISBN 0-8218-2921-1.
[28] M. Benedicks and L. Carleson, *On iterations of* $1 - ax^2$ *on* $(-1, 1)$, Ann. of Math. **122** (1985), 1–25.
[29] M. Benedicks and L. Carleson, *The dynamics of the Hénon map*, Ann. of Math. **133** (1991), 73–169.
[30] G.D. Birkhoff, *Proof of the ergodic theorem*, Proc. Nat. Acad. Sci. USA **17** (1931), 656–660.

[31] G.D. Birkhoff, *What is the ergodic theorem?* Amer. Math. Monthly **49** (1942), jstor, 222–226.
[32] R. Bowen, *Markov partitions for Axiom A diffeomorphisms*, Amer. J. Math. **92** (1970), 725–747.
[33] R. Bowen, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Lecture Notes in Math., Vol. 470, Springer-Verlag, Berlin (1975).
[34] R. Bowen, *Bernoulli maps of the interval*, Israel J. Math. **28** (1977), 161–168.
[35] R. Bowen, *Invariant measures for Markov maps of the interval*, Comm. Math. Phys. **69** (1979), 1–17, with an afterword by Roy L. Adler and additional comments by Caroline Series.
[36] X. Bressaud, *Subshifts on an infinite alphabet*, Ergodic Theory Dynam. Systems **19** (1999), 1175–1200.
[37] X. Bressaud, R. Fernández and A. Galves, *Decay of correlations for non-Hölderian dynamics. A coupling approach*, Electron. J. Probab. **4** (3) (1999), 19 pp. (electronic).
[38] H. Bruin, *Topological conditions for the existence of invariant measures for unimodal maps*, Ergodic Theory Dynam. Systems **14** (1994), 433–451.
[39] H. Bruin, *The existence of absolutely continuous invariant measures is not a topological invariant for unimodal maps*, Ergodic Theory Dynam. Systems **18** (1998), 555–565.
[40] H. Bruin, *Topological conditions for the existence of absorbing Cantor sets*, Trans. Amer. Math. Soc. **350** (1998), 2229–2263.
[41] H. Bruin and G. Keller, *Equilibrium states for S-unimodal maps*, Ergodic Theory Dynam. Systems **18** (1998), 765–789.
[42] H. Bruin, S. Luzzatto and S. van Strien, *Decay of correlations in one-dimensional dynamics*, Ann. Sci. École Norm. Sup. **36** (2003), 621–646.
[43] H. Bruin, W. Shen and S. van Strien, *Invariant measures exist without a growth condition*, Comm. Math. Phys. **241** (2003), 287–306.
[44] H. Bruin and S. van Strien, *Existence of absolutely continuous invariant probability measures for multimodal maps*, Global Analysis of Dynamical Systems, Inst. Phys., Bristol (2001), 433–447.
[45] L.A. Bunimovič, *On a transformation of the circle*, Mat. Zametki **8** (1970), 205–216 (in Russian).
[46] L.A. Bunimovich, I.P. Cornfeld, R.L. Dobrushin, M.V. Jakobson, N.B. Maslova, Ya.B. Pesin, Ya.G. Sinaĭ, Yu.M. Sukhov and A.M. Vershik, *Dynamical systems. II*, Encyclopaedia of Mathematical Sciences, Vol. 2, Springer-Verlag, Berlin (1989), Ergodic theory with applications to dynamical systems and statistical mechanics; edited and with a preface by Sinaĭ, translated from the Russian.
[47] J. Buzzi, *Absolutely continuous invariant measures for generic multi-dimensional piecewise affine expanding maps*, Internat. J. Bifur. Chaos Appl. Sci. Engrg. **9** (1999), 1743–1750, Discrete dynamical systems.
[48] J. Buzzi, *Absolutely continuous invariant probability measures for arbitrary expanding piecewise* **R***-analytic mappings of the plane*, Ergodic Theory Dynam. Systems **20** (2000), 697–708.
[49] J. Buzzi, *No or infinitely many a.c.i.p. for piecewise expanding C^r maps in higher dimensions*, Comm. Math. Phys. **222** (2001), 495–501.
[50] J. Buzzi, *Thermodynamical formalism for piecewise invertible maps: absolutely continuous invariant measures as equilibrium states*, Smooth ergodic theory and its applications, Seattle, WA, 1999, Proc. Sympos. Pure Math., Vol. 69, Amer. Math. Soc., Providence, RI (2001), 749–783 (in English, with English and French summaries).
[51] J. Buzzi and G. Keller, *Zeta functions and transfer operators for multidimensional piecewise affine and expanding maps*, Ergodic Theory Dynam. Systems **21** (2001), 689–716.
[52] J. Buzzi and V. Maume-Deschamps, *Decay of correlations for piecewise invertible maps in higher dimensions*, Israel J. Math. **131** (2002), 203–220.
[53] J. Buzzi, F. Paccaut and B. Schmitt, *Conformal measures for multidimensional piecewise invertible maps*, Ergodic Theory Dynam. Systems **21** (2001), 1035–1049.
[54] J. Buzzi and O. Sarig, *Uniqueness of equilibrium measures for countable Markov shifts and multidimensional piecewise expanding maps*, Ergodic Theory Dynam. Systems **23** (2003), 1383–1400.
[55] J. Buzzi, O. Sester and M. Tsujii, *Weakly expanding skew-products of quadratic maps*, Ergodic Theory Dynam. Systems **23** (2003), 1401–1414.
[56] J.T. Campbell and A.N. Quas, *A generic C^1 expanding map has a singular S-R-B measure*, Comm. Math. Phys. **221** (2001), 335–349, DOI 10.1007/s002200100491.
[57] Y. Cao, *Non-zero Lyapunov exponents and uniform hyperbolicity*, Nonlinearity **16** (2003), 1473–1479.
[58] Y. Cao, I. Rios and S. Luzzatto, *A minimum principle for Lyapunov exponents and a higher-dimensional version of a theorem of Mané*, Qual. Theory Dynam. Systems (2005), to appear.

[59] S. Cedervall, *Expansion estimates and decay of correlations for multimodal maps* (2004), in progress.
[60] P. Collet and J.-P. Eckmann, *Positive Lyapunov exponents and absolute continuity for maps of the interval*, Ergodic Theory Dynam. Systems **3** (1983), 13–46.
[61] W.J. Cowieson, *Stochastic stability for piecewise expanding maps in* $\mathbf{R}^d$, Nonlinearity **13** (2000), 1745–1760.
[62] W.J. Cowieson, *Absolutely continuous invariant measures for most piecewise smooth expanding maps*, Ergodic Theory Dynam. Systems **22** (2002), 1061–1078.
[63] S. Day, H. Kokubu, K. Mischaikow and H. Oka, *An algorithm for computing expansivity properties of one-dimensional maps* (2005), in progress.
[64] K. Diaz-Ordaz, *Decay of correlations for non-Hölder observables for one-dimensional expanding Lorenz-like maps*, Discrete Contin. Dynam. Systems (2005), to appear.
[65] D. Dolgopyat, *On decay of correlations in Anosov flows*, Ann. of Math. (2) **147** (1998), 357–390.
[66] D. Dolgopyat, *Prevalence of rapid mixing in hyperbolic flows*, Ergodic Theory Dynam. Systems **18** (1998), 1097–1114.
[67] D. Dolgopyat, *On dynamics of mostly contracting diffeomorphisms*, Comm. Math. Phys. **213** (2000), 181–201.
[68] A.O. Gel'fond, *A common property of number systems*, Izv. Akad. Nauk SSSR Ser. Mat. **23** (1959), 809–814 (in Russian).
[69] P. Góra and A. Boyarsky, *Absolutely continuous invariant measures for piecewise expanding* C^2 *transformation in* $\mathbf{R}^N$, Israel J. Math. **67** (1989), 272–286.
[70] P. Góra and B. Schmitt, *Un exemple de transformation dilatante et* C^1 *par morceaux de l'intervalle, sans probabilité absolument continue invariante*, Ergodic Theory Dynam. Systems **9** (1989), 101–113 (in French, with English summary).
[71] S. Gouëzel, *Statistical properties of a skew product with a curve of neutral points* (2003), Preprint.
[72] S. Gouëzel, *Sharp polynomial estimates for the decay of correlations*, Israel J. Math. **139** (2004), 29–65.
[73] S. Gouëzel, *Decay of correlations for nonuniformly expanding systems* (2004), arXiv:math.DS/0401184.
[74] J. Graczyk and G. Świątek, *Generic hyperbolicity in the logistic family*, Ann. of Math. (2) **146** (1997), 1–52.
[75] P.R. Halmos, *In general a measure preserving transformation is mixing*, Ann. of Math. (2) **45** (1944), 786–792.
[76] F. Hofbauer and G. Keller, *Ergodic properties of invariant measures for piecewise monotonic transformations*, Math. Z. **180** (1982), 119–140.
[77] M. Holland, *Slowly mixing systems and intermittency maps*, Ergodic Theory Dynam. Systems **25** (1) (2005), 133–159.
[78] J.C. Holladay, *On the existence of a mixing measure*, Proc. Amer. Math. Soc. **8** (1957), 887–893.
[79] E. Hopf, *Theory of measure and invariant integrals*, Trans. Amer. Math. Soc. **34** (1932), 373–393.
[80] E. Hopf, *Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung*, Ber. Verh. Sächs. Akad. Wiss. Leipzig **91** (1939), 261–304 (in German).
[81] H. Hu, *Statistical properties of some almost hyperbolic systems*, Smooth Ergodic Theory and Its Applications, Seattle, WA, 1999, Proc. Sympos. Pure Math., Vol. 69, Amer. Math. Soc., Providence, RI (2001), 367–384.
[82] H. Hu, *Decay of correlations for piecewise smooth maps with indifferent fixed points*, Ergodic Theory Dynam. Systems **24** (2004), 495–524.
[83] H.Y. Hu and L.-S. Young, *Nonexistence of SBR measures for some diffeomorphisms that are "almost Anosov"*, Ergodic Theory Dynam. Systems **15** (1995), 67–76.
[84] S. Isola, *Renewal sequences and intermittency*, J. Statist. Phys. **97** (1999), 263–280.
[85] M.V. Jakobson, *Topological and metric properties of one-dimensional endomorphisms*, Sov. Math. Dokl. **19** (1978), 1452–1456.
[86] M.V. Jakobson, *Absolutely continuous invariant measures for one-parameter families of one-dimensional maps*, Comm. Math. Phys. **81** (1981), 39–88.
[87] M.V. Jakobson, *Piecewise smooth maps with absolutely continuous invariant measures and uniformly scaled Markov partitions*, Proc. Sympos. Pure Math. **69** (2001), 825–881.
[88] M.V. Jakobson, *Parameter choice for families of maps with many critical points*, Modern Dynam. Systems Appl. (2004).

[89] M. Kac and H. Kesten, *On rapidly mixing transformations and an application to continued fractions*, Bull. Amer. Math. Soc. **64** (1958), 283–283; correction **65** (1958), 67.
[90] G. Keller, *Ergodicité et mesures invariantes pour les transformations dilatantes par morceaux d'une région bornée du plan*, C. R. Acad. Sci. Paris Sér. A–B **289** (1979), A625–A627 (in French, with English summary).
[91] G. Keller, *Un théorème de la limite centrale pour une classe de transformations monotones par morceaux*, C. R. Acad. Sci. Paris Sér. A–B **291** (1980), A155–A158 (in French, with English summary).
[92] G. Keller and T. Nowicki, *Spectral theory, zeta functions and the distribution of periodic points for Collet–Eckmann maps*, Comm. Math. Phys. **149** (1992), 31–69.
[93] O.S. Kozlovski, *Getting rid of the negative Schwarzian derivative condition*, Ann. of Math. (2) **152** (2000), 743–762.
[94] O.S. Kozlovski, *Axiom A maps are dense in the space of unimodal maps in the C^k topology*, Ann. of Math. (2) **157** (2003), 1–43.
[95] O. Kozlovski, W. Shen and S. van Strien, *Rigidity for real polynomials* (June 2003), http://www.maths.warwick.ac.uk/~ strien/Publications/rigid4june.ps, Preprint.
[96] K. Krzyżewski and W. Szlenk, *On invariant measures for expanding differentiable mappings*, Studia Math. **33** (1969), 83–92.
[97] A. Lasota, *Invariant measures and functional equations*, Aequationes Math. **9** (1973), 193–200.
[98] A. Lasota and J.A. Yorke, *On the existence of invariant measures for piecewise monotonic transformations*, Trans. Amer. Math. Soc. **186** (1973), 481–488 (1974).
[99] C. Liverani, *Decay of correlations for piecewise expanding maps*, J. Statist. Phys. **78** (1995), 1111–1129.
[100] C. Liverani, *Decay of correlations*, Ann. of Math. (2) **142** (1995), 239–301.
[101] C. Liverani, *Invariant measures and their properties. A functional analytic point of view* (2004), http://www.mat.uniroma2.it/liverani/rev.html, Preprint.
[102] C. Liverani, B. Saussol and S. Vaienti, *Conformal measure and decay of correlation for covering weighted systems*, Ergodic Theory Dynam. Systems **18** (1998), 1399–1420.
[103] S. Luzzatto, *Bounded recurrence of critical points and Jakobson's theorem*, The Mandelbrot Set, Theme and Variations, London Math. Soc. Lecture Note Ser., Vol. 274, Cambridge Univ. Press, Cambridge (2000), 173–210.
[104] S. Luzzatto, M. Holland and K. Diaz-Ordaz, *Mixing properties of one-dimensional Lorenz-like maps with critical points and discontinuities with infinite derivative*, Stochastics and Dynamics (2005), to appear.
[105] S. Luzzatto and F. Sanchez-Salas, *Markov structures for nonuniformly hyperbolic surface diffeomorphisms*, in progress.
[106] S. Luzzatto and H. Takahashi, *Computable conditions for the occurrence of non-uniform hyperbolicity in families of one-dimensional maps* (2005), Preprint.
[107] S. Luzzatto and W. Tucker, *Non-uniformly expanding dynamics in maps with singularities and criticalities*, Inst. Hautes Études Sci. Publ. Math. (1999), 179–226.
[108] S. Luzzatto and M. Viana, *Positive Lyapunov exponents for Lorenz-like families with criticalities*, Astérisque (2000), xiii, 201–237, Géométrie complexe et systèmes dynamiques (Orsay, 1995) (in English, with English and French summaries).
[109] S. Luzzatto and L. Wang, *Topological invariance of generic non-uniformly expanding multimodal maps*, http://front.math.ucdavis.edu/math.DS/0307030, Preprint.
[110] V. Lynch, *Decay of correlations for non-Hölder continuous observables*, Discrete Contin. Dynam. Systems (2005), to appear.
[111] M. Lyubich, *Dynamics of quadratic polynomials. I, II*, Acta Math. **178** (1997), 185–247, 247–297.
[112] M. Lyubich and J. Milnor, *The Fibonacci unimodal map*, J. Amer. Math. Soc. **6** (1993), 425–457.
[113] R. Mañé, *Hyperbolicity, sinks and measure in one-dimensional dynamics*, Comm. Math. Phys. **100** (1985), 495–524.
[114] P. Manneville and Y. Pomeau, *Intermittent transition to turbulence in dissipative dynamical systems*, Comm. Math. Phys. **74** (1980), 189–197.
[115] V. Maume-Deschamps, *Correlation decay for Markov maps on a countable state space*, Ergodic Theory Dynam. Systems **21** (2001), 165–196.
[116] V. Maume-Deschamps, *Projective metrics and mixing properties on towers*, Trans. Amer. Math. Soc. **353** (2001), 3371–3389 (electronic).

[117] W. de Melo and S. van Strien, *One-dimensional dynamics: The Schwarzian derivative and beyond*, Bull. Amer. Math. Soc. (N.S.) **18** (1988), 159–162.
[118] W. de Melo and S. van Strien, *One-dimensional dynamics*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], Vol. 25, Springer-Verlag, Berlin (1993).
[119] M. Misiurewicz, *Absolutely continuous measures for certain maps of an interval*, Inst. Hautes Études Sci. Publ. Math. **53** (1981), 17–51.
[120] T. Nowicki, *A positive Liapunov exponent for the critical value of an S-unimodal mapping implies uniform hyperbolicity*, Ergodic Theory Dynam. Systems **8** (1988), 425–435.
[121] T. Nowicki and F. Przytycki, *Topological invariance of the Collet–Eckmann property for S-unimodal maps*, Fund. Math. **155** (1998), 33–43.
[122] T. Nowicki and D. Sands, *Non-uniform hyperbolicity and universal bounds for S-unimodal maps*, Invent. Math. **132** (1998), 633–680.
[123] T. Nowicki and S. van Strien, *Absolutely continuous invariant measures for C^2 unimodal maps satisfying the Collet–Eckmann conditions*, Invent. Math. **93** (1988), 619–635.
[124] T. Nowicki and S. van Strien, *Invariant measures exist under a summability condition for unimodal maps*, Invent. Math. **105** (1991), 123–136.
[125] W. Parry, *On the β-expansions of real numbers*, Acta Math. Acad. Sci. Hungar. **11** (1960), 401–416 (in English, with Russian summary).
[126] O. Penrose and J.L. Lebowitz, *On the exponential decay of correlation functions*, Comm. Math. Phys. **39** (1974), 165–184.
[127] V.T. Perekrest, *Exponential mixing in C-systems*, Uspekhi Mat. Nauk **29** (1974), 181–182 (in Russian).
[128] Ya. Pesin, *Families of invariant manifolds corresponding to non-zero characteristic exponents*, Math. USSR Izv. **10** (1976), 1261–1302.
[129] Ya.B. Pesin, *Characteristic Lyapunov exponents and smooth ergodic theory*, Russian Math. Surveys **324** (1977), 55–114.
[130] G. Pianigiani, *First return map and invariant measures*, Israel J. Math. **35** (1980), 32–48.
[131] M. Pollicott and M. Yuri, *Statistical properties of maps with indifferent periodic points*, Comm. Math. Phys. **217** (2001), 503–520.
[132] F. Przytycki, J. Rivera-Letelier and S. Smirnov, *Equivalence and topological invariance of conditions for non-uniform hyperbolicity in the iteration of rational maps*, Invent. Math. **151** (2003), 29–63.
[133] F. Przytycki and S. Rohde, *Porosity of Collet–Eckmann Julia sets*, Fund. Math. **155** (1998), 189–199.
[134] A.N. Quas, *Non-ergodicity for C^1 expanding maps and g-measures*, Ergodic Theory Dynam. Systems **16** (1996), 531–543.
[135] A.N. Quas, *Most expanding maps have no absolutely continuous invariant measure*, Studia Math. **134** (1999), 69–78.
[136] A. Rényi, *Representations for real numbers and their ergodic properties*, Acta Math. Acad. Sci. Hungar. **8** (1957), 477–493.
[137] V. Rohlin, *A "general" measure-preserving transformation is not mixing*, Doklady Akad. Nauk SSSR (N.S.) **60** (1948), 349–351 (in Russian).
[138] M. Rosenblatt, *A central limit theorem and a strong mixing condition*, Proc. Nat. Acad. Sci. USA **42** (1956), 43–47.
[139] D. Ruelle, *Statistical mechanics of a one-dimensional lattice gas*, Comm. Math. Phys. **9** (1968), 267–278.
[140] D. Ruelle, *A measure associated with axiom-A attractors*, Amer. J. Math. **98** (1976), 619–654.
[141] D. Ruelle, *Applications conservant une mesure absolument continue par rapport à dx sur* $[0, 1]$, Comm. Math. Phys. **55** (1977), 47–51 (in French, with English summary).
[142] D. Ruelle, *Characteristic exponents and invariant manifolds in Hilbert space*, Ann. of Math. (2) **115** (1982), 243–290.
[143] M. Rychlik, *Bounded variation and invariant measures*, Studia Math. **76** (1983), 69–80.
[144] M.R. Rychlik, *Another proof of Jakobson's theorem and related results*, Ergodic Theory Dynam. Systems **8** (1988), 93–109.
[145] F.J. Sánchez-Salas, *Dimension of Markov towers for non uniformly expanding one-dimensional systems*, Discrete Contin. Dynam. Systems **9** (2003), 1447–1464.
[146] D. Sands, *Topological conditions for positive Lyapunov exponents* (1995), http://topo.math.u-psud.fr/~sands/Papers/thesis.ps, Preprint.

[147] O. Sarig, *Subexponential decay of correlations*, Invent. Math. **150** (2002), 629–653.
[148] B. Saussol, *Absolutely continuous invariant measures for multidimensional expanding maps*, Israel J. Math. **116** (2000), 223–248.
[149] S. Senti, *Dimension of weakly expanding points for quadratic maps*, Bull. Soc. Math. France **131** (2003), 399–420.
[150] W. Shen, *On the metric properties of multimodal interval maps and C^2 density of Axiom A*, Preprint.
[151] C. Simó and J.C. Tatjer, *Windows of attraction of the logistic map*, European Conference on Iteration Theory (Batschuns, 1989), World Scientific, River Edge, NJ (1991), 335–342.
[152] Ja.G. Sinai, *Markov partitions and U-diffeomorphisms*, Funkcional. Anal. i Priložen. **2** (1968), 64–89 (in Russian).
[153] Ja.G. Sinaĭ, *Gibbs measures in ergodic theory*, Uspekhi Mat. Nauk **27** (1972), 21–64 (in Russian).
[154] M. Thaler, *Transformations on* $[0, 1]$ *with infinite invariant measures*, Israel J. Math. **46** (1983), 67–96.
[155] Ph. Thieullen, C. Tresser and L.-S. Young, *Positive Lyapunov exponent for generic one-parameter families of unimodal maps*, J. Anal. Math. **64** (1994), 121–172.
[156] H. Thunberg, *Positive exponent in families with flat critical point*, Ergodic Theory Dynam. Systems **19** (1999), 767–807.
[157] M. Tsujii, *A proof of Benedicks–Carleson–Jacobson theorem*, Tokyo J. Math. **16** (1993), 295–310.
[158] M. Tsujii, *Positive Lyapunov exponents in families of one-dimensional dynamical systems*, Invent. Math. **111** (1993), 113–137.
[159] M. Tsujii, *Piecewise expanding maps on the plane with singular ergodic properties*, Ergodic Theory Dynam. Systems **20** (2000), 1851–1857.
[160] M. Tsujii, *Absolutely continuous invariant measures for expanding piecewise linear maps*, Invent. Math. **143** (2001), 349–373.
[161] S. Ulam and J. von Neumann, *On combination of stochastic and deterministic processes*, Bull. Amer. Math. Soc. **53** (1947), 1120.
[162] M. Viana, *Multidimensional nonhyperbolic attractors*, Inst. Hautes Études Sci. Publ. Math. **85** (1997), 63–96.
[163] M. Viana, *Stochastic dynamics of deterministic systems*, Lecture Notes XXI Braz. Math. Colloq., IMPA, Rio de Janeiro (1997).
[164] M.S. Waterman, *Some ergodic properties of multi-dimensional f-expansions*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **16** (1970), 77–103.
[165] J.-C. Yoccoz, *Jakobson's Theorem* (2001), Manuscript.
[166] L.-S. Young, *Stochastic stability of hyperbolic attractors*, Ergodic Theory Dynam. Systems **6** (1986), 311–319.
[167] L.-S. Young, *Decay of correlations for certain quadratic maps*, Comm. Math. Phys. **146** (1992), 123–138.
[168] L.-S. Young, *Ergodic theory of differentiable dynamical systems*, Real and Complex Dynamical Systems, Hillerød, 1993, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., Vol. 464, Kluwer Academic, Dordrecht (1995), 293–336.
[169] L.-S. Young, *Statistical properties of dynamical systems with some hyperbolicity*, Ann. of Math. (2) **147** (1998), 585–650.
[170] L.-S. Young, *Recurrence times and rates of mixing*, Israel J. Math. **110** (1999), 153–188.

CHAPTER 4

Homoclinic Bifurcations, Dominated Splitting, and Robust Transitivity

Enrique R. Pujals
IMPA, Rio de Janeiro, RJ, Brazil
E-mail: *enrique@impa.br*

Martin Sambarino
CMAT-Facultad de Ciencias, Montevideo, Uruguay
E-mail: *samba@cmat.edu.uy*

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

This page intentionally left blank

1. Introduction

For a long time (mainly since Poincaré) it has been a goal in the theory of dynamical systems to describe the dynamics from the generic viewpoint, that is, describing the dynamics of "big sets" (residual, dense, etc.) of the space of all dynamical systems.

It was briefly thought in the sixties that this could be realized by the so-called hyperbolic systems with the assumption that the tangent bundle over the limit set $L(f)$ (see definition in Section 4.3) splits into two complementary subbundles $T_{L(f)}M = E^s \oplus E^u$ such that vectors in E^s (respectively E^u) are uniformly forward (respectively backward) contracted by the tangent map Df (see Chapter 1, Principal structures (Hasselblatt and Katok), in Volume 1A of this handbook). Under this assumption, the limit set decomposes into a finite number of transitive sets such that the asymptotic behavior of any orbit is described by the dynamics in the trajectories in those finitely many transitive sets. Moreover, this topological description allows to get the statistical behavior of the system. In other words, hyperbolic dynamics on the tangent bundle characterizes the dynamics over the manifold from a geometrical–topological and statistical point of view.

Nevertheless, uniform hyperbolicity was soon realized to be a property less universal than it was initially thought: it was shown that there are open sets in the space of dynamics which are nonhyperbolic. The initial mechanism to show this nonhyperbolic robustness (see [1,90]) was the existence of open sets of diffeomorphisms exhibiting hyperbolic periodic points of different indices (i.e. different dimension of their stable manifolds) inside a transitive set. Indeed, Shub's construction leads to an open set of transitive diffeomorphisms on T^4 exhibiting hyperbolic periodic points of different indices. Related to this is the notion of hetero-dimensional cycle where two periodic points of different indices are linked through the intersection of their stable and unstable manifolds (notice that at least one of the intersections is nontransversal).

In all the above examples the underlying manifolds must be of dimension at least three and so the case of surfaces was still unknown at the time. It was shown through the seminal works of Newhouse (see [59,61,63]) that hyperbolicity was not dense in the space of C^r-diffeomorphisms of compact surfaces (however, let us point out that in the C^1-topology it is still open!). The underlying mechanism here was the presence of a homoclinic tangency leading to the nowadays so-called "Newhouse phenomenon", i.e. residual subsets of diffeomorphisms displaying infinitely many periodic attractors.

These results naturally pushed some aspects of the theory of dynamical systems in different directions:

1. The study of the dynamical phenomena obtained from homoclinic bifurcations (i.e. the unfolding of homoclinic tangencies or hetero-dimensional cycles);
2. The characterization of universal mechanisms that could lead to robustly (meaning any perturbation of the initial system) nonhyperbolic behavior;
3. The study and characterization of isolated transitive sets that remain transitive for all nearby systems (*robust transitivity*);
4. The dynamical consequences of weaker forms of hyperbolicity.

As we will show, these problems are all related to each other. In many cases, such relations provide a conceptual framework, as the hyperbolic theory did for the case of transverse homoclinic orbits.

Besides, we mentioned that two basic mechanisms were found to the obstruction of hyperbolicity, namely *hetero-dimensional cycles* and *homoclinic tangencies*. In the early 80's Palis conjectured that these are very common in the complement of the hyperbolic systems:

PALIS CONJECTURE.

1. *Every C^r-diffeomorphism of a compact manifold M can be C^r-approximated by one which is hyperbolic or by one exhibiting a hetero-dimensional cycle or by one exhibiting a homoclinic tangency*;
2. *When M is a two-dimensional compact manifold every C^r-diffeomorphism of M can be C^r-approximated by one which is hyperbolic or by one exhibiting a homoclinic tangency.*

This conjecture may be considered as a starting point to obtain a generic description of C^r-diffeomorphisms. If it turns out to be true, we may focus on the two mechanisms mentioned above in order to understand the dynamics. Nevertheless, the unfolding of these homoclinic bifurcations is still mainly a local study.

The above conjecture was proved [79] (see Section 4.2) for the case of surfaces and the C^1-topology. When the manifold has dimension greater than two, it is still open despite some progress, but still in the C^1-topology. The case of higher topologies (C^r, $r \geqslant 2$) is, at this point, far from being solved.

We would like to emphasize that this chapter attempts to give some insight into the problems mentioned above but it is not meant to be a complete survey of the subject. For this, we apologize both to the reader and to those whose works we have not cited.

We also want to point out that in the case of conservative systems, a successful theory has been developed where the KAM theorem is the major highlight. In the present chapter we focus on nonconservative systems; however, in Section 5.4 we mention some results for the conservative case.

In Section 2.2, we introduce the notion of dominated splitting that has been a cornerstone for addressing the conjecture of stability and the above-mentioned conjecture. Also this notation plays an important role in the characterization of robustly transitive systems.

In Section 4.1 we focus the attention on diffeomorphisms of surfaces and we address the Palis conjecture for surface diffeomorphisms.

In Section 5 the examples of nonhyperbolic robustly transitive systems and their relation with the notion of dominated splitting are explained.

In the last section, some of the previous issues are revisited for the case of flow.

Many of the definitions used here can be found in Chapter 1, Principal structures (B. Hasselblatt and A. Katok), in Volume 1A of this handbook.

2. A weaker form of hyperbolicity: Dominated splitting

2.1. *Introduction*

In the theory of differentiable dynamics, i.e. the study of the asymptotic behavior of orbits $\{f^n(x)\}$ when $f : M \to M$ is a diffeomorphism of a compact Riemannian manifold M,

one may say that a fundamental problem is to understand how the dynamics of the tangent map Df controls or determines the underlying dynamics of f.

So far, this program has been solved for hyperbolic dynamics by the so-called Spectral Decomposition Theorem (see Theorem 4.3.1 in Section 4.3).

There were, basically, two ways to relax hyperbolicity. One, called partial hyperbolicity, allows the tangent bundle to split into Df-invariant subbundles $TM = E^s \oplus E^c \oplus E^u$, and the behavior of vectors in E^s, E^u is similar to the hyperbolic case, but vectors in E^c may be neutral for the action of the tangent map. In the other, nonuniform hyperbolicity (or Pesin theory), where the tangent bundle splits for points a.e. with respect to some invariant measure, and vectors are asymptotically contracted or expanded at a rate that may depend on the base point.

Since the latter case considers a measure-theoretic framework, one cannot expect to obtain a description from the point of view of topological dynamics (see Barreira–Pesin in the present volume for an account of nonuniform hyperbolicity). In the former, there is no general theory regarding its consequences for the topological dynamics (although there are many important results, mostly from the ergodic point of view; see, for instance, Chapter 1, Partially hyperbolic dynamical systems (Hasselblatt and Pesin), in this handbook, and [24, 76,3,29]).

There is also another category which includes the partially hyperbolic systems: *dominated splitting*. Although partially hyperbolic systems arose in a natural way (time-one maps of Anosov flows, frame flows, group extensions), the concept of dominated splitting was introduced independently by Mañé, Liao and Pliss, as a first step in the attempt to prove that structurally stable systems satisfy a hyperbolic condition on the tangent map. In fact, under the assumption of C^1-structural stability, the closure of the periodic points exhibits dominated splitting. Nevertheless, in the last decades there has been a large amount of research on this concept showing that it is not just a technical concept but appears naturally when one tries to understand robust phenomena.

It is believed that a robust dynamic phenomenon must be reflected in the tangent map. This turns out to be true, for instance, in the case of structural stability, robust transitivity and lack of tangencies. These last two subjects will be discussed in the next sections.

After that, a natural question arises: what are the properties of a system having dominated splitting? In other words, is it possible to describe the dynamics of a system having dominated splitting?

In few words, we could say that a general strategy to deal with many of the problems mentioned before is the following: first, it is shown that for some kind of robust phenomenon (stability, transitivity, lack of tangencies) the limit set exhibits dominated splitting; later, one tries to extract information from this kind of dynamics in the tangent bundle. However, this main issue is far from being understood except for C^2-surface maps as we will see in Sections 4.1 and 4.3.

2.2. *Definition and general remarks*

DEFINITION 2.2.1. An f-invariant set Λ is said to have *dominated splitting* of index i (or just dominated splitting) if there is a decomposition of its tangent bundle into two invariant subbundles $T_\Lambda M = E \oplus F$, such that:

1. $\dim E(x) = i$;
2. There exist $C > 0$ and $0 < \lambda < 1$ such that

$$\left\| Df^n_{/E(x)} \right\| \left\| Df^{-n}_{/F(f^n(x))} \right\| \leqslant C\lambda^n, \quad \text{for all } x \in \Lambda,\ n \geqslant 0.$$

The constant λ is called a constant of domination.

NOTE. It is assumed that none of the subbundles are trivial (otherwise, the other one has a uniform hyperbolic behavior: contracting or expanding).

Let us explain briefly the meaning of the above definition. It says that, for n large, the "greatest expansion" of Df^n on E is less than the "greatest contraction" of Df^n on F and by a factor that becomes exponentially small with n. In other words, every direction not belonging to E must converge exponentially fast under iteration of Df to the direction F.

As in the hyperbolic case, the subbundles E and F depend continuously on the base point. The following lemma is elementary but nevertheless important.

LEMMA 2.2.1. *Let Λ be an f-invariant set exhibiting dominated splitting of index i. Then $\bar{\Lambda}$ (the closure of Λ) exhibits a dominated splitting of index i (with the same constant of domination).*

Notice that on an invariant set different dominated splittings may coexist. As a trivial example consider a periodic point p of period n having a simple Lyapunov spectrum, i.e. all the eigenvalues of $Df^n_{/T_pM}$ have (algebraic) multiplicity one. In this case the periodic point admits a dominated splitting of index i for any $i = 1, \ldots, \dim M - 1$.

It is straightforward to check that a hyperbolic set has a dominated splitting. On the other hand, it is trivial to find examples of sets having dominated splitting without being hyperbolic. A simple one is a nonhyperbolic periodic point where the spectrum is not contained in S^1. Another example is an attracting (or repelling) invariant closed curve supporting an irrational rotation. More sophisticated examples are partially hyperbolic systems. See Section 5 for this kind of examples.

Another simple yet useful property is that dominated splitting cannot be destroyed by small perturbations:

LEMMA 2.2.2. *Let $f : M \to M$ be a diffeomorphism and let Λ be a compact set having dominated splitting $T_\Lambda M = E \oplus F$. Then, there exist an open set U containing Λ and a C^1-neighborhood $\mathcal{U}(f)$ such that for any $g \in \mathcal{U}(f)$ there is a dominated splitting over any compact g-invariant set $\Lambda_g \subset U$. Moreover, the constant of domination can be chosen uniformly. The set U is called an admissible neighborhood of Λ.*

2.3. *Sufficient conditions for dominated splitting*

In this section we will explain a technique mainly developed by Mañé to show the existence of a dominated splitting.

First we recall a fundamental result in C^1-dynamics: the so-called generic C^1-closing lemma proved by C. Pugh (see [74] and [75]).

THEOREM 2.3.1. *$\Omega(f) = \overline{\mathrm{Per}(f)}$ for a C^1-generic (i.e. residual) set of diffeomorphisms.*

This result opens the possibility of carrying information about the periodic points to the whole nonwandering set. Recall that also generically it follows that the periodic points are hyperbolic, so for each periodic point p there is a natural hyperbolic decomposition of the tangent bundle over the periodic orbit in two complementary directions $E^s \oplus E^u$. Therefore, a natural question arises: *can the tangent bundle decomposition over the periodic points be extended to the nonwandering set*? Here dominated splitting plays the key role. The main idea is that if the angles of the eigenspaces of hyperbolic periodic points are bounded away from zero in a robust way then there must be a dominated splitting. Let us explain this in a more precise way, and we refer to [47] for details. We first introduce a simple yet powerful perturbation technique (in the C^1-topology) due to Franks.

LEMMA 2.3.1 [35, Lemma 1.1]. *Let M be a closed n-manifold and $f : M \to M$ be a C^1-diffeomorphism, and let a neighborhood $\mathcal{U}(f)$ of f be given. Then, there exist $\mathcal{U}_0(f) \subset \mathcal{U}(f)$ and $\delta > 0$ such that if $g \in \mathcal{U}_0(f)$, $S \subset M$ is a finite set, $S = \{p_1, p_2, \ldots, p_m\}$ and L_i, $i = 1, \ldots, m$, are linear maps $L_i : TM_{p_i} \to TM_{f(p_i)}$ satisfying $\|L_i - D_{p_i} g\| \leqslant \delta$, $i = 1, \ldots, m$, then there exists $\tilde{g} \in \mathcal{U}(f)$ satisfying $\tilde{g}(p_i) = g(p_i)$ and $D_{p_i}\tilde{g} = L_i$, $i = 1, \ldots, m$. Moreover, if U is any neighborhood of S then we may choose $\tilde{g}$ so that $\tilde{g}(x) = g(x)$ for all $x \in \{p_1, p_2, \ldots, p_m\} \cup (M \backslash U)$.*

These results say, for instance, that any small perturbation of the linear maps along a periodic orbit can be realized through a diffeomorphism nearby. Now let us introduce the definition of angle between subspaces.

DEFINITION 2.3.1. Let E and F be two subspaces of $\mathbb{R}^d$ such that $E \oplus F = \mathbb{R}^d$. Hence $\dim(F) = \dim(E^\perp)$ and F is the graph of the linear map $L : E^\perp \to E$ defined as follows: given $w \in F$ there exists a unique pair of vectors $v \in E^\perp$ and $u \in E$, such that $v + u = w$. Define $L(v) = u$ obtaining that $\mathrm{graph}(L) = F$. We define the angle $\angle(E, F)$ between E and F as $\|L\|^{-1}$. In particular, $\angle(E, E^\perp) = +\infty$.

DEFINITION 2.3.2. Let $f : M \to M$ be a diffeomorphism. The index of a hyperbolic periodic point p of f is the dimension of the stable manifold of p. We will denote by $\mathrm{Per}_i(f)$ the set of hyperbolic periodic points of index i.

The next theorem is essentially due to Mañé [48]. The techniques are rather involved so we will give an outline of the main ideas to help the reader to go through the details.

THEOREM 2.3.2. *Assume that there is a neighborhood $\mathcal{U}(f)$ such that*

$$\angle\big(E^s(p, g), E^u(p, g)\big) > \gamma > 0$$

for every $g \in \mathcal{U}(f)$ *and* $p \in \mathrm{Per}_i(g)$. *Then there is dominated splitting of index* i *over* $\mathrm{Per}_i(f)$.

OUTLINE OF PROOF. The best way to understand the argument is by contraposition: absence of dominated splitting forces the presence of arbitrarily small angles.

The first step is to prove the domination at the end of the period of each periodic point, meaning the following: there exits $0 < \lambda < 1$ such that if p is a hyperbolic periodic point of index i of f then:

$$\left\| Df^{n_p}_{/E^s(p)} \right\| \left\| Df^{-n_p}_{/E^u(p)} \right\| \leqslant \lambda^{n_p},$$

where n_p is the period of p. Otherwise we do not have uniform contraction at the end of the period on $Df^{n_p}_{/E^s(p)}$ nor on $Df^{-n_p}_{/E^u(p)}$. Thus, we can perturb to obtain a hyperbolic periodic point p_g of index i of a diffeomorphism g close to f having one eigenvalue less but close to one and another eigenvalue greater but also close to one. From this it is possible to construct arbitrarily small perturbations for which the angle between the subspaces E^s and E^u is arbitrarily small.

For the second and last step, let us assume that we do not have a dominated splitting over $\mathrm{Per}_i(f)$. Then there is a periodic point p of f whose period is arbitrarily large (say k), and for some m large we have

$$\left\| Df^j | E^s(p) \right\| \cdot \left\| Df^{-j} | E^u\big(f^j(p)(x)\big) \right\| \geqslant 1/2, \quad 1 \leqslant j \leqslant m.$$

In some sense, this means that the action of the action of Df along the piece of orbit $p, f(p), \ldots, f^m(p)$ is neutral, although at the end of the period we have domination, that we may assume is due to contraction at the end of the period on E^s.

Next, define $T_i : T_{f^i(p)}M \to T_{f^i(p)}M$ such that $T_i | E^s(f^i(p)) = (1 + \epsilon)\mathrm{Id}$ and $T_i | E^u(f^i(p)) = \mathrm{Id}$ and $L_i = T_i \circ Df_{f^{i-1}(p)}$. Here, the number ϵ is small enough in order to still have contraction on $E^s(p)$ at the end of the period, i.e. under $L_{n-1} \circ \cdots \circ L_0$. Let us see the effect of this perturbation: take a space S close to $E^u(p)$. Then, due to the expansion added in E^s along $p, f(p), \ldots, f^m(p)$, we have that $L_m \circ \cdots \circ L_0(S)$ will "fall down" to the direction of E^s, meaning that the angle between $L_m \circ \cdots \circ L_0(S)$ and $E^s(f^m(p))$ is small. On the other hand, due to the domination, at the end of the period we will have that $L_n \circ \cdots \circ L_0(S)$ is "up" again, meaning that $L_n \circ \cdots \circ L_0(S)$ and $E^u(p)$ are close again. Thus, adding another perturbation on L_n so that $L_n \circ \cdots \circ L_0(S) = S$ and using Franks lemma we find g near f so that p is hyperbolic periodic point of g and $E^s(p, g) = E^s(p, f)$ and $E^u(p, g) = S$, and we have "destroyed" the angle at $g^m(p) = f^m(p)$ (i.e. we have shown that it is less than γ). □

3. Homoclinic tangencies

3.1. *The dynamics near a homoclinic tangency*

In this section we will mention some of the dynamical phenomena related to the presence of homoclinic tangencies.

For that, first we recall that the stable and unstable sets

$$W^s(p) = \{y \in M\colon \operatorname{dist}(f^n(y), f^n(p)) \to 0 \text{ as } n \to \infty\},$$
$$W^u(p) = \{y \in M\colon \operatorname{dist}(f^n(y), f^n(p)) \to 0 \text{ as } n \to -\infty\}$$

are C^r-injectively immersed submanifolds when p is a hyperbolic periodic point of f.

DEFINITION 3.1.1. Let $f : M \to M$ be a diffeomorphism. We say that f exhibits a homoclinic tangency if there is a hyperbolic periodic point p of f such that the stable and unstable manifolds of p have a nontransverse intersection.

It is important to say that a homoclinic tangency is (locally) easily destroyed by a small perturbation of the invariant manifolds. To get open sets of diffeomorphisms where each system exhibits an homoclinic tangency, Newhouse studied systems where the homoclinic tangency is associated to an invariant hyperbolic set with the property that it has large fractal dimension. In fact, he studied the intersection of the local stable and unstable manifolds of a hyperbolic set (for instance, a classical horseshoes), where this kind of hyperbolic sets, roughly speaking, can be visualized as a product of two Cantor sets with the property that the fractal dimension of these Cantor sets (more specifically, the thickness) is large. Newhouse's construction depends on how the fractal dimension varies with perturbation of the dynamics and actually this is the main reason that his construction works in the C^2-topology. In fact, Newhouse's construction is based on the continuous dependence of the fractal dimension (thickness) on C^2-perturbations. A similar construction in the C^1-topology leading to same phenomena is unknown (results in the opposite direction can be found in [95]).

After the seminal works of Newhouse, no other results were obtained in the direction to understand the dynamics induced by unfolding homoclinic tangencies, specially in the case of one-parameter families. In the study of bifurcations of a generic one-parameter family of surface diffeomorphisms having a generic homoclinic tangency at a parameter value, the arithmetic difference of two Cantor sets appears in a natural way (see [69]). The Cantor sets that appear in this context are regular on the line, that is, they are defined by expansive maps and have a sort of self-similarity property, which means, roughly speaking, that a small part of them is diffeomorphic to a large part with uniformly bounded distortion. In this context, in [68] some results were obtained on homoclinic bifurcations associated to a basic set which ensures full density of hyperbolicity on the parameter line, provided that the Hausdorff dimension of the basic set is less than one. Palis, in [66], conjectured that for generic pairs of regular Cantor sets of the real line, either their arithmetic difference has Lebesgue measure equal to zero or otherwise it contains an interval. The latter statement should correspond in the context of homoclinic bifurcations to open sets of tangencies. Regarding the second part of Palis conjecture, a partial result was obtained in [71] and later in [58] it was proved that if the sum of the Hausdorff dimensions of two Cantor sets is greater than one, then in almost all cases, there exist translations of these Cantor sets whose intersection has Hausdorff dimension greater than one.

Other fundamental dynamic prototypes were found in the context of this bifurcation, namely the so-called cascade of bifurcations, the Hénon-like strange attractor [11,50] (infinitely many coexisting ones [27]), and superexponential growth of periodic points [40].

Despite the rich dynamics that appear after the unfolding of a homoclinic tangency, Palis conjectured (see [67]) that for a generic one-parameter family of surface maps unfolding a homoclinic tangency, the set of parameter values corresponding to diffeomorphisms with infinitely many sinks or infinitely many Hénon-like attractors has (Lebesgue) measure zero.

In this direction it is announced in [72] that given a surface diffeomorphisms f_0 such that the maximal invariant set in an open set V is the union of a horseshoe and a quadratic tangency between the stable and unstable foliations of this horseshoe such that the dimension of the horseshoe is larger than but close to one, then for most diffeomorphisms f close to f_0, the maximal f-invariant set in V is a nonuniformly hyperbolic horseshoe, with dynamics of the same type as met in Hénon attractors. In particular, most diffeomorphisms (in a measure-theoretical point of view for parameters in one-parameter families) do not exhibit attracting periodic points.

In higher dimension, many of the previous results were generalized (see [70,84,57]).

As it was mentioned at the beginning when we referred to the Newhouse's techniques, all the previous results hold when at least C^2-diffeomorphisms are considered. However, the coexistence of infinitely many sinks or sources for C^1-diffeomorphisms on three-dimensional manifolds was obtained in [18].

3.2. *Dominated splitting versus tangencies*

In this section we will deal with diffeomorphism that cannot be approximated by ones exhibiting a homoclinic tangency.

DEFINITION 3.2.1. Let $f : M \to M$ be a C^r-diffeomorphism. We say that f is C^r far from tangencies if there exists a neighborhood $\mathcal{U}(f)$ in the C^r-topology such that no $g \in \mathcal{U}(f)$ exhibits a homoclinic tangency (see Definition 3.1.1).

The next theorem says that the lack (in a robust way) of the presence of homoclinic tangency guarantees the existence of a dominated splitting. It was originally proved in [79] for surface diffeomorphisms and extended to higher dimensions by L. Wen in [97].

THEOREM 3.2.1. *Let $f : M \to M$ be a diffeomorphism which is C^1 far from tangencies. Then* $\mathrm{Per}_i(f)$ (*see Definition* 2.3.2) *has dominated splitting of index i, where $i = 1, \ldots, \dim M - 1$.*

IDEA OF PROOF. We will give some insight of the proof of the above theorem in the case of surface diffeomorphisms and at the end we will comment on higher dimensions. Based on Theorem 2.3.2, it is enough to show that the angles of the eigenspaces of hyperbolic periodic points of index i (index 1 for surfaces) are uniformly bounded away from zero in a C^1-neighborhood of f. Thus, arguing by contradiction, we must show that if these angles can be arbitrarily small then we can find a diffeomorphism nearby exhibiting a homoclinic

tangency. So let p be a hyperbolic periodic point such that the stable and unstable spaces have an arbitrarily small angle. For the sake of simplicity, assume that the stable and unstable manifolds of p coincides in a neighborhood of p with the stable and unstable spaces in a local chart. The idea is to push a fundamental domain in the local stable manifold to have a tangency with the unstable manifold. Nevertheless, some estimates must be done and we refer to [79] to see the details. The idea in higher dimension is the same. However, there is the main difference that makes the proof much more technically involved: in a hyperbolic periodic point of index 1 of a surface diffeomorphism the size (or shape) of a fundamental domain on the local stable manifold can be estimated from $Df_{/E^s(p)}$ while in higher dimension, the shape or "eccentricity" could be out of control. To overcome this difficulty, it is required that no homoclinic tangency can be created associated to *any* hyperbolic periodic point and not just those of index i. In other words, a bad shape or eccentricity on a fundamental domain of a hyperbolic periodic point of index i is related to a bad angle (after some perturbation) between the stable and unstable subspaces of another periodic point of different index. □

4. Surface diffeomorphisms

4.1. *Dynamical consequences of the dominated splitting. Sufficient conditions for hyperbolicity*

The presence of homoclinic tangencies has many analogies with the presence of critical points for one-dimensional endomorphisms. On one hand, homoclinic tangencies correspond in the one-dimensional setting to preperiodic critical points and it is known that their bifurcation leads to complex dynamics. On the other hand, Mañé (see [49]) showed that for regular and generic one-dimensional endomorphisms, the absence of critical points is enough to guarantee hyperbolicity. This result raises the question about the dynamical properties of surface maps exhibiting no homoclinic tangencies. As dominated splitting prevents the presence of tangencies, we could say that domination plays for surface diffeomorphisms the role that the noncritical behavior does for one-dimensional endomorphisms.

One may ask whether a set having dominated splitting is hyperbolic. Two necessary conditions follow trivially: all the periodic points in the set must be hyperbolic and no attracting (repelling) closed invariant (periodic) curve supporting an irrational rotation is in the set.

The next result says that these two conditions are also sufficient as long as the diffeomorphism is smooth enough. It is the analog of a one-dimensional theorem by Mañé (see [49]).

THEOREM 4.1.1 [79]. *Let $f \in \mathrm{Diff}^2(M^2)$ and assume that $\Lambda \subset \Omega(f)$ is a compact invariant set exhibiting a dominated splitting and such that each periodic point in Λ is hyperbolic. Then, $\Lambda = \Lambda_1 \cup \Lambda_2$ where Λ_1 is a hyperbolic set and Λ_2 consists of a finite union of normally hyperbolic periodic simple closed curves $\mathcal{C}_1, \ldots, \mathcal{C}_n$ such that $f^{m_i} : \mathcal{C}_i \to \mathcal{C}_i$ is conjugate to an irrational rotation (m_i denotes the period of $\mathcal{C}_i$).*

The proof of the above theorem exceeds the scope of this chapter. Nevertheless, let us mention the main four steps:

Step I. It is sufficient to prove that Λ is hyperbolic under the above conditions and to assume that Λ has neither a normally hyperbolic invariant curve supporting an irrational rotation nor attracting or repelling periodic points.

Step II. Due to the dominated splitting, the fact that f is C^2, and that all the periodic points are of saddle type, two locally invariant manifolds $W^{cs}_\epsilon(x)$, $W^{cu}_\epsilon(x)$, which are of class C^2, pass through each point $x \in \Lambda$. One of the main problems is that, although these manifolds are locally invariant, they do not have (a priori) any dynamic meaning at all.

Step III. Considering the fact that these central manifolds are one dimensional and of class C^2, we are able to prove that they have a dynamic meaning. Indeed, $W^{cs}_\epsilon(x)$ is a stable manifold and $W^{cu}_\epsilon(x)$ is an unstable manifold.

Step IV. For every $x \in \Lambda$ we have

$$\sum_{n\geqslant 0} \left| f^n\left(W^{cs}_\epsilon(x)\right)\right| < \infty \quad \text{and} \quad \sum_{n\geqslant 0} \left| f^{-n}\left(W^{cu}_\epsilon(x)\right)\right| < \infty,$$

where $|\cdot|$ means length. This is enough to prove the hyperbolicity on Λ.

In the next subsection we will give an explanation of Step I, the main technique to obtain Step III, and finally we will give a rather complete proof of Theorem 4.1.1 in a particular case (see Theorem 4.1.4).

4.1.1. *Getting rid of attracting and repelling closed curves and periodic attractors–repellers* First it is proved that the number of closed curves that could appear in a set having dominated decomposition is finite. This is established by showing that the basin of attraction of each attracting curve has a uniform size. More precisely one proves that the diameter of the curves do not go to zero (otherwise the splitting $E \oplus F$ must have a "singularity"). Afterwards, since $f^{m_n} : \mathcal{C}_n \to \mathcal{C}_n$ is conjugate to an irrational rotation, they support only one invariant measure with zero Lyapunov exponent along the F direction and this implies (using the domination) that the fiber E is contractive, which implies that the basin of attraction of $\mathcal{C}_n$ has uniform size. So, if there are infinitely many such curves, we will get an intersection of two different basins, which is a contradiction. Therefore, these normally hyperbolic closed curves are isolated in our set. Notice that the periodic attractors or repellers are isolated in Λ as well. Thus, if we remove all the periodic attractors or repellers and all the invariant closed curves supporting an irrational rotation, we obtain a compact invariant set $\tilde{\Lambda}$ as in the next theorem:

THEOREM 4.1.2. *Let $f \in \mathrm{Diff}^2(M^2)$ and let $\tilde{\Lambda}$ be a compact invariant set having a dominated splitting $T_{/\Lambda}M = E \oplus F$ such that the periodic points of f in $\tilde{\Lambda}$ are hyperbolic of saddle type* (*i.e. of index* 1) *and $\tilde{\Lambda}$ does not contain any normally hyperbolic periodic simple closed curve $\mathcal{C}$ such the restriction to it is conjugate to an irrational rotation. Then $\tilde{\Lambda}$ is a hyperbolic set.*

Further explanation must be done to conclude Theorem 4.1.1 from Theorem 4.1.2: we must show that Λ has finitely many periodic attractors or repellers. This is done by showing

that, if Λ had infinitely many of them, then they would accumulate (in the Hausdorff sense) in $\tilde{\Lambda}$, a contradiction.

4.1.2. *Nonexistence of wandering intervals* In this section we provide a useful tool in order to understand dynamic properties of one-dimensional central manifolds under the assumption of dominated splitting. The results of this section hold not only for surface diffeomorphism but for diffeomorphisms on manifolds of any dimension, provided the dominated splitting is codimension one, that is, one of the subbundles of the splitting has dimension one, say F. Although there is not much difference, we still assume that we are working in dimension two.

DEFINITION 4.1.1. Let $f : M \to M$ be a C^2-diffeomorphism, Λ a compact invariant set having dominated splitting and V an admissible neighborhood (see Lemma 2.2.2). Let U be an open set containing Λ such that $\bar{U} \subset V$. We say that a C^2-arc I in M (i.e. a C^2-embedding of the interval $(-1, 1)$) is a δ-*E-arc* or δ-*E-interval* provided that the following two conditions hold:

1. $f^n(I) \subset U$, $n \geqslant 0$, and $|f^n(I)| \leqslant \delta$ for all $n \geqslant 0$;
2. $f^n(I)$ is always transverse to the E-direction.

In other words, a δ-E-arc is an arc that does not grow in length in the future and always remains transversal to the E subbundle.

The next result is a fundamental tool in the understanding the behavior of the action of the differential map when expansion or contraction along a subspace is known on a large step. It is an important tool for studying sets with a dominated splitting. Although this tool is purely arithmetic, we will state it in a different way for our future purposes.

LEMMA 4.1.1 (Pliss Lemma [73]). *Given a diffeomorphism f and $0 < \gamma_1 < \gamma_2$, there exist $N = N(\gamma_1, \gamma_2, f)$ and $c = c(\gamma_1, \gamma_2, f) > 0$ with the following property: given $x \in M$, a subspace $S \subset T_x M$ such that for some $n \geqslant N$ we have (denoting $S_i = Df^i(S)$)*

$$\prod_{i=0}^{n} \|Df_{/S_i}\| \leqslant \gamma_1^n,$$

there exist $0 \leqslant n_1 < n_2 < \cdots < n_l \leqslant n$ such that

$$\prod_{i=n_r}^{j} \|Df_{/S_i}\| \leqslant \gamma_2^{j-n_r}, \quad r = 1, \ldots, l,\ n_r \leqslant j \leqslant n.$$

Moreover, $l \geqslant cn$.

The integers n_i from the lemma above are known in the recent literature as *hyperbolic times*.

The next theorem characterizes the dynamic of a δ-E-arc. It says that such an arc cannot be "wandering".

THEOREM 4.1.3. *There exists δ_0 such that if I is a δ-E-interval with $\delta \leqslant \delta_0$, then one of the following properties holds*:

1. *$\omega(I) = \bigcup_{x\in I} \omega(x)$ is a normally hyperbolic periodic simple closed curve $\mathcal{C}$ and $f^m_{/\mathcal{C}} : \mathcal{C} \to \mathcal{C}$ (where m is the period of $\mathcal{C}$) is conjugate to an irrational rotation;*
2. *$\omega(I) \subset \mathrm{Per}(f_{/V})$ where $\mathrm{Per}(f_{/V})$ is the set of the periodic points of f in V. Moreover, one of the periodic points in $\omega(I)$ is not hyperbolic of saddle type.*

SKETCH OF PROOF. In the sequel, λ will be the constant of domination.

Step 1. First, for each n we take I_n the maximal δ-E-arc that contains $f^n(I)$. It is proved that there are infinitely many n_i's such that for I_{n_i} we get that E is a uniform contracting direction (which could be not true for every I_n). This implies that for any $x \in I_{n_i}$ there is a stable manifold $W^s_\epsilon(x)$ of uniform size. So we can consider the box $W^s_\epsilon(I_{n_i}) = \bigcup_{x\in I_{n_i}} W^s_\epsilon(x)$. Furthermore, for these n_i's it is possible to compare (uniformly for any box) the one-dimensional length of I_{n_i} with the two-dimensional volume of the box $W^s_\epsilon(I_{n_i})$, i.e. there is a constant K such that $K \operatorname{vol}(W^s_\epsilon(I_{n_i})) \geqslant |I_{n_i}|$, where K is independent of n_i.

Step 2. There exist $n_i < n_j$ such that

$$W^s_\epsilon(I_{n_i}) \cap W^s_\epsilon\big(f^{n_j-n_i}(I_{n_i})\big) \neq \emptyset. \tag{1}$$

If this is not the case, a contradiction with the maximality of the I_{n_i} is shown.

Step 3. The fact that $W^s_\epsilon(I_{n_i}) \cap W^s_\epsilon(f^{n_j-n_i}(I_{n_i})) \neq \emptyset$ for some $n_i < n_j$ will imply the conclusion of the theorem.

Now we will explain these steps in more detail. Coming back to the intervals I_n, consider the sequence of positive integers n_j such that $|I_{n_j}| \geqslant |I_k|$ for any $k > n_j$.

Observe that this implies that for any $k > 0$ there is some $x_k \in I_{n_j}$ such that $\|Df^k_{|F_{x_k}}\| \leqslant 1$ and since the iterates of I_{n_j} remain small (less than δ) it follows that there is β small such that $\|Df^k_{|F_x}\| < (1+\beta)^k$ for any $x \in I_{n_j}$. Using the domination property and if β is small enough (which is obtained taking δ_0 small), we get that

$$\big\|Df^k_{|E_x}\big\| < \big[\lambda(1+\beta)\big]^k = \lambda_1^k \quad \text{for any } x \in I_{n_j} \text{ and } k \geqslant 0, \tag{2}$$

where $\lambda_1 < 1$. Now consider all the positive integers n_i such that (2) holds. In particular, every point in I_{n_i} has a stable manifold of uniform size.

Let λ_2 be such that $\lambda < \lambda_2 < \lambda_1 < 1$. Consider $N = N(\lambda_2, \lambda_1)$ from Pliss Lemma 4.1.1. It follows (assuming for simplicity that $n_{i+1} - n_i \geqslant N = 1$) that:

$$\big\|Df^{n_{i+1}-j}_{|E_x}\big\| > \lambda_2^j \quad \text{for any } x \in f^j(I_{n_i}) \text{ and } 0 \leqslant j < n_{i+1} - n_i. \tag{3}$$

This implies that the F direction behaves as an expanding direction for iterates between n_i and n_{i+1}. In fact, (3) implies that given $0 \leqslant j < n_{i+1} - n_i$, we have

$$\big\|Df^{-(n_{i+1}-j)}_{|F_x}\big\| < \left(\frac{\lambda}{\lambda_2}\right)^j$$

for any $x \in I_{n_{i+1}}$. In particular,

$$|I_{-(n_{i+1}-j)}| < \left(\frac{\lambda}{\lambda_2}\right)^j |I_{n_{i+1}}|. \tag{4}$$

We are about to finish Step 2. Let us assume that $W^s_\epsilon(I_{n_r}) \cap W^s_\epsilon(I_{n_j}) = \emptyset$ for every r, j. Using that we can compare volume with length.

$$\sum_{i>0}^{\infty} |I_{n_i}| < \infty,$$

which together with (4) implies

$$\sum_{n>0}^{\infty} |I_n| < \infty,$$

and arguing as in Schwarz's proof of the Denjoy Theorem for some n large, we may find an arc J_n containing properly each I_n such that J_n is a δ-interval, which is a contradiction with the maximality of I_n for every n.

Now let us explain Step 3. Let $m = n_j - n_i$. If $|f^{km}(I_{n_i})| \to 0$ as $k \to \infty$, then $\omega(I_{n_i})$ consists of a periodic orbit. Indeed, if $|f^{km}(I_{n_i})| \to 0$, then $|f^k(I_{n_i})| \to 0$ as $k \to \infty$. Let p be an accumulation point of $f^k(I_{n_i})$, that is, $f^{k_j}(I_{n_i}) \to p$ for some $k_j \to \infty$, and so, $f^{k_j+m}(I_{n_i}) \to f^m(p)$. But by the property we are assuming, i.e. $W^s_\epsilon(I_{n_i}) \cap W^s_\epsilon(f^{n_j-n_i}(I_{n_i})) \neq \emptyset$, we have $f^{k_j+m}(I_{n_i}) \to p$, implying that p is a periodic point. Thus, for any $x \in I_{n_i}$ we have that $\omega(x)$ consists only of periodic orbits, and so $\omega(x)$ is single periodic orbit p. Since $|f^k(I_{n_i})| \to 0$ we conclude that $\omega(I_{n_i})$ is the orbit of the periodic point p. By the way we choose I_{n_i}, we have $f^{n_i}(I) \subset I_{n_i}$ and so $\omega(I)$ consists of a periodic orbit, as the thesis of the theorem requires.

On the other hand, if $|f^{km}(I_{n_i})|$ does not go to zero, we take a sequence k_j such that $f^{k_j m}(I_{n_i}) \to L$ for some arc L (which is at least C^1, and has F as its tangent direction). Now $f^{(k_j+1)m}(I_{n_i}) \to L'$ and $f^m(L) = L'$. Moreover, by (1), $L \cup L'$ is an interval (with F as its tangent direction). Let

$$J = \bigcup_{n \geqslant 0} f^{nm}(L).$$

We claim that there are only two possibilities: either J is an arc or a simple closed curve. To prove this, notice that $f^{nm}(L)$ is a δ-E-interval for any $n \geqslant 0$. In particular, for any $x \in J$ there exists $\epsilon(x)$ such that $W^{cs}_{\epsilon(x)}(x)$ is stable manifold for x, and so

$$W(J) = \bigcup_{x \in J} W^{cs}_{\epsilon(x)}(x)$$

is a neighborhood of J.

We only have to show that, given $x \in J$, there exists a neighborhood $U(x)$ such that $U(x) \cap J$ is an arc. This implies that J is a simple closed curve or an interval. Thus, take $x \in J$, in particular $x \in f^{n_1 m}(L)$. Take U an open interval, $x \in U \subset f^{n_1 m}(L)$ and let $U(x)$ be a neighborhood of x such that $U(x) \subset W(J)$ and such that $U(x) \cap L_1 \subset U$ where L_1 is any interval containing $f^{n_1 m}(L)$, transverse to the E-direction and $|L_1| \leqslant 2\delta_0$ (this is always possible if δ_0 is small). Now let $y \in J \cap U(x)$. We have to prove that $y \in U$. There is n_2 such that $y \in f^{n_2 m}(L)$. Since

$$f^{n_1 m}(L) = \lim_j f^{k_j m + n_1 m}(I_{n_i}),$$

$$f^{n_2 m}(L) = \lim_j f^{k_j m + n_2 m}(I_{n_i})$$

and both have nonempty intersection with $U(x)$, we conclude for some j that $f^{k_j m + n_1 m}(I_{n_i})$ and $f^{k_j m + n_2 m}(I_{n_i})$ are linked by a local stable manifold. Hence $f^{n_1 m}(L) \cup f^{n_2 m}(L)$ is an arc L_1 transverse to the E-direction with $|L_1| \leqslant 2\delta_0$. Therefore $y \in U(x) \cap L_1 \subset U$ as desired, completing the proof that J is an arc or a simple closed curve.

In case J is an arc, since $f^m(J) \subset J$, it follows that for any $x \in I$, $\omega(x)$ is an ω-limit point of a point in J, hence a periodic orbit, completing the proof in this case. On the other hand, if J is a simple closed curve, which is of class C^2 because it is normally hyperbolic (attractive), then we have two possibilities. If $f^m_{/J} : J \to J$ has rational rotation number, then we can see that $\omega(I_{n_i})$ consists of a union of periodic points, and the same happens to I. If $f^m_{/J} : J \to J$ has irrational rotation number, then it is conjugate to an irrational rotation, and denoting $\mathcal{C} = J$, we have that $\omega(I)$ is as in the first property of the thesis of the theorem. □

DEFINITION 4.1.2. We say that the point x is Lyapunov stable (in the future) if given $\epsilon > 0$ there exists $\delta > 0$ such that $f^n(B_\delta(x)) \subset B_\epsilon(f^n(x))$ for any positive integer n.

As a consequence of Theorem 4.1.3 we get:

COROLLARY 4.1.1. *Let $f : M \to M$ be a C^2-diffeomorphism of a finite-dimensional compact Riemannian manifold M and let Λ be a set having a codimension-one dominated splitting. Then there exists a neighborhood V of Λ such that if $f^n(x) \in V$ for any positive integer n and x is Lyapunov stable, one of the following holds*:

1. *$\omega(x)$ is a periodic orbit*;
2. *$\omega(x)$ is a normally attractive periodic curve supporting an irrational rotation.*

4.1.3. *Outline of proof of Theorem 4.1.1 in a particular case* The proof of Theorem 4.1.1 and in particular the last "two steps" are extremely involved, so we will explain them in a case where some of the main ideas are present. Indeed, let us give a rather complete proof of the following theorem (compare with Theorem 4.1.2):

THEOREM 4.1.4. *Let $f \in \mathrm{Diff}^2(M^2)$ where M^2 is the two torus. Assume that $\Omega(f) = M$ has dominated splitting $TM = E^s \oplus F$, E^s is a contractive subbundle and all the periodic points are hyperbolic. Then f is Anosov.*

The proof (or outline of proof) will be done through several lemmas. In what follows we always assume that we are under the conditions of the above theorem.

LEMMA 4.1.2. *Through any $x \in M$ passes an arc $W^{cu}_{\epsilon}(x)$ of class C^2 and always tangent to F.*

SKETCH OF PROOF. Since F is integrable, we do have an arc $W^{cu}_{\epsilon}(x)$ through x and tangent to F. In order to prove that it is of class C^2, by [38], we need to prove the so-called 2-domination, that is, there exist $\lambda < 1$ and a positive constant C such that for any positive integer n it follows that $\frac{|Df^n_{|E}|}{|Df^n_{|F}|^2} < C\lambda^n$. □

LEMMA 4.1.3. *Under the above conditions the distribution F is uniquely integrable and determines an unstable foliation, i.e. if x, y belong to the same leaf we have that* $\operatorname{dist}(f^{-n}(x), f^{-n}(y)) \to 0$ *as* $n \to \infty$.

SKETCH OF PROOF. Let $x \in M$ be any point and denote by $W^{cu}_{\epsilon}(x)$ any arc of length ϵ containing x and tangent to F at every point. Let us show that, given δ, there exists ϵ_1 such that $|f^{-n}(W^{cu}_{\epsilon_1}(x))| < \delta$ for any $n \geqslant 0$ (where $|\cdot|$ means length). Otherwise, there exist $\epsilon_n \to 0$ and $m_n \to \infty$ such that

$$\left|f^{-j}\left(W^{cu}_{\epsilon_n}(x)\right)\right| \leqslant \delta, \quad 0 \leqslant j \leqslant n, \quad \text{and} \quad \left|f^{-m_n}\left(W^{cu}_{\epsilon_n}(x)\right)\right| = \delta.$$

Taking an accumulation arc I of $f^{-m_n}(W^{cu}_{\epsilon_n}(x))$ we conclude that $T_x I = F(x)$ for any $x \in I$ and $|f^n(I)| \leqslant \delta$ for any $n \geqslant 0$. That is, I is a δ-E-arc and from Theorem 4.1.3 we get a contradiction with our assumptions. In the same way it can be shown that $|f^{-n}(W^{cu}_{\epsilon}(x))| \to 0$ as $n \to \infty$. The conclusion of the lemma follows from [38]. □

From the lemma above, we conclude that f is conjugate to an Anosov diffeomorphism. Indeed, a simple way to prove it is by showing that f is expansive and by [45] our claim follows. In particular we have a Markov partition of M, say $\mathcal{R} = \{R_1, \ldots, R_m\}$. From this partition we can construct another Markov partition by taking preimages and intersecting with the original one, i.e. for any $n > 0$ we define $\mathcal{R}_n = \{f^{-n}(R_i) \cap R_j,\ 1 \leqslant i, j \leqslant m, R_i \in \mathcal{R}\}$ as a new Markov partition (whose "unstable" size decreases with n).

Let R_i be any rectangle of the Markov partition $\mathcal{R}$. For $x \in \operatorname{int}(R_i)$ we denote by $J_i(x)$ the connected component of $W^u(x) \cap R_i$ that contains x. This $J_i(x)$ is an arc with endpoints in the "stable boundary" of R_i.

LEMMA 4.1.4. *There exists $K > 0$ such that if R_i is any rectangle of the Markov partition and $x \in \operatorname{int}(R_i)$ we have that*

$$\sum_{j=0}^{j=n} \left|f^{-j}\left(J_i(x)\right)\right| < K$$

as long as $f^{-j}(x) \notin R_i$, $1 \leqslant j \leqslant n$.

SKETCH OF PROOF. Take $z_k \in R_k$ for every rectangle of the Markov partition and let $I_k = J_k(z_k)$. Denote by $\Pi_k : R_k \to I_k$ the projection along the stable foliation. Since the stable foliation is C^1 (because its codimension is one) we conclude that there exists $C > 0$ such that if l_u is any arc in the unstable foliation contained in R_k then

$$C^{-1}|l_u| \leqslant \left|\Pi(l_u)\right| \leqslant C|l_u|.$$

Now, let $x \in R_i$ and let n be as in the lemma. We claim that if for some $1 \leqslant j_1 < j_2 \leqslant n$ we have that $f^{-j_1}(x)$ and $f^{-j_2}(x)$ belong to the same rectangle R_k of the partition then $\Pi_k(f^{-j_1}(J_i(x)))$ and $\Pi_k(f^{-j_2}(J_i(x)))$ are disjoint arcs in I_k. Otherwise, from the properties of the Markov partitions we conclude that $f^{-(j_2-j_1)}(x) \in R_i$, a contradiction. Therefore

$$\sum_{j=0}^{j=n} \left|f^{-j}\left(J_i(x)\right)\right| \leqslant C \sum_{k=1}^{k=m} |I_k| = K. \qquad \square$$

Let us observe that K from the lemma above is "independent" of the Markov partition, i.e. the same K works if in the lemma we replace the Markov partition $\mathcal{R}$ by the partition $\mathcal{R}_n$.

Recall that we want to prove that f is Anosov, that is, we would like to show that the distribution F is expanding.

LEMMA 4.1.5. *If f is not Anosov, then there exists a set Λ_0 which is not hyperbolic but with the property that any compact invariant proper subset of Λ_0 is hyperbolic. Moreover, Λ_0 is a transitive set.*

SKETCH OF PROOF. The existence of Λ_0 follows from Zorn's lemma. Besides, if for any $x \in \Lambda_0$ we have $\alpha(x) \subsetneq \Lambda_0$ then it follows that $\|Df^{-n}_{/F(x)}\| \to 0$, i.e. Λ_0 is hyperbolic. Thus, there is some x such that $\alpha(x) = \Lambda_0$; in other words, Λ_0 is transitive. $\square$

At this point we come to the heart of the proof of the theorem. We will show that a transitive set Λ_0 such that any compact invariant proper subset is hyperbolic (and which is not an invariant circle supporting an irrational rotation) must be hyperbolic, leading to a contradiction with Lemma 4.1.5. For the sake of simplicity, we will explain the argument assuming a supplementary condition: there is $x \in \Lambda_0$ such that $x \notin \omega(x)$.

In this case we can take a Markov partition so that if $\mathcal{R}(x)$ is the rectangle of the Markov partition $\mathcal{R}$ containing x then we have that $\mathcal{R}(x) \cap \{f^n(x)\colon\ n \geqslant 1\} = \emptyset$. On the other hand, since Λ_0 is transitive, there exist $x_n \in \Lambda_0 \cap \mathcal{R}(x)$ and $m_n \to \infty$ such that $f^{-j}(x_n) \notin \mathcal{R}(x)$, $1 \leqslant j < m_n$, and $f^{-m_n}(x_n) \in \mathcal{R}(x)$. Fix $k = m_n$ large enough so that $|f^{-n}(J_{\mathcal{R}(x)}(y))| < r$ for $n \geqslant k$, where r is small enough. This condition will be explained later on.

Let R be the connected component of $f^{-m_n}(\mathcal{R}(x)) \cap \mathcal{R}(x)$ that contains x_n, that is $R = \mathcal{R}_{m_n}(x)$. Now, let $y \in R \cap \Lambda_0$ and assume that $f^{-j}(y) \notin R$, $1 \leqslant j < m$, and $f^{-m}(y) \in R$. Let us show that $\|Df^{-m}_{/F(y)}\| < 1/2$. First, notice that $m \geqslant k$. Thus,

$$
\begin{aligned}
&\left\| Df^{-m}_{/F(y)} \right\| \\
&\quad \leqslant \left\| Df^{-k}_{/F(f^{-(m-k)}(y))} \right\| \cdot \left\| Df^{-(m-k)}_{/F(y)} \right\| \\
&\quad \leqslant \frac{|f^{-k}(J_{\mathcal{R}(x)}(f^{-(m-k)}(y)))|}{|J_{\mathcal{R}(x)}(f^{-(m-k)}(y))|} \exp(K_0 K) \frac{|f^{-(m-k)}(J_R(y))|}{|J_R(y)|} \exp(K_0 K) \\
&\quad \leqslant \left| f^{-(m-k)}\big(J_R(y)\big) \right| \frac{|f^{-k}(J_{\mathcal{R}(x)})|}{|J_{\mathcal{R}(x)}(f^{-(m-k)}(y))|} \frac{1}{|J_{\mathcal{R}(x)}(f^{-(m-k)}(y))|} \exp(2K_0 K) \\
&\quad \leqslant rC \frac{1}{L} \exp(2K_0 K).
\end{aligned}
$$

Let us explain the constants above: C is the length distortion along the projection of the stable foliation, L is the minimum length of $J_{\mathcal{R}(x)}(z)$, $z \in \mathcal{R}(x)$, and K_0 is a Lipschitz constant of $\log(Df)$ along the unstable foliation. Hence, if r is small enough we obtain our claim. Now we may conclude. Let z be any point in Λ_0. If $\alpha(z) \subsetneq \Lambda_0$ it follows that $\|Df^{-n}_{F(z)}\| \to 0$ as $n \to \infty$. On the other hand, if $\alpha(x) = \Lambda_0$ then for some n_0 we have that $y = f^{-n_0}(z) \in R$ and returns to it infinitely many times. Since at each return time the derivative is less than $1/2$, we are done.

4.2. *Generic results in the C^1-topology and consequences*

The first result we would like to state is a proof of the Palis conjecture for surface diffeomorphisms in the C^1-topology:

THEOREM 4.2.1 [79]. *Let M^2 be a two-dimensional compact manifold and let $f \in \mathrm{Diff}^1(M^2)$. Then, f can be C^1-approximated either by a diffeomorphism exhibiting a homoclinic tangency or by an Axiom* A *diffeomorphism.*

OUTLINE OF PROOF. Let f be a C^1-diffeomorphism far from tangencies (see Definition 3.2.1). We would like to show that f can be approximated by an Axiom A diffeomorphism. We may assume that f satisfies some generic (i.e. residual) conditions: satisfies Lemma 2.3.1, it is Kupka–Smale and it is a continuity point of the maps $g \to \Gamma_i(g) = \overline{\mathrm{Per}_i(g)}$, $i = 0, 1, 2$. From Theorem 3.2.1 we get that $\mathrm{Per}_1(f)$ has dominated splitting. From this, it may be proved that any C^1-diffeomorphism close to f has a dominated splitting over the nonwandering set, except perhaps finitely many periodic attractors or repellers. Thus, take g, a C^2-diffeomorphism, which is C^1-close to f, is Kupka–Smale and has no invariant curves supporting an irrational rotation (this is C^r-generic). For this diffeomorphism g we apply Theorem 4.1.1 and we conclude that $\Omega(g)$ is hyperbolic. Since it is also Kupka–Smale, and g is a surface diffeomorphism we can conclude that g is an Axiom A diffeomorphism. □

An almost immediate consequence is:

COROLLARY 4.2.1. *Let MS be the set of Morse–Smale diffeomorphisms, $\overline{MS}$ its closure, and consider $\mathcal{U} = \mathrm{Diff}^1(M^2) - \overline{MS}$. Then, there exists an open and dense set $\mathcal{R}$ in $\mathcal{U}$ such that every $f \in \mathcal{R}$ has a transverse homoclinic orbit. In particular, the closure of the interior of the set formed by the diffeomorphisms having (topological) zero entropy is equal to $\overline{MS}$.*

The following theorem is a result obtained by Mañé in [48].

THEOREM 4.2.2. *There exists a residual set $\mathcal{R}$ in $\mathrm{Diff}^1(M^2)$ such that for any $f \in \mathcal{R}$ one of the following holds*:

1. *f is Axiom* A;
2. *f has infinitely many periodic attractors*;
3. *f has infinitely many periodic repellers*.

SKETCH OF PROOF. Let $\mathcal{R}_1$ be the residual set in $\mathrm{Diff}^1(M^2)$ such that any $f \in \mathcal{R}_1$ is a continuity point of the maps $g \to \Gamma_i(g) = \overline{P_i(g)}$, $i = 0, 2$. If $f \in \mathcal{R}_1$ does not satisfy either 2 nor 3 above then f and any nearby diffeomorphism has finitely many periodic attractors and repellers, and hence cannot be approximated by one exhibiting a homoclinic tangency. So, it can be approximated by an Axiom A one. □

The above theorem has a different extension to higher dimension (see [19]).

In [80] the following theorem was also obtained.

THEOREM 4.2.3. *Let M^2 be a two-dimensional compact manifold and let $f \in \mathrm{Diff}^\infty(M^2)$ such that its topological entropy is not locally constant. Then f can be C^1-approximated by a diffeomorphism exhibiting a homoclinic tangency.*

At this point it is important to mention that any $C^{1+\alpha}(\alpha > 0)$-diffeomorphism of a two-dimensional manifold with positive topological entropy that has an invariant set has a closed invariant set Γ such that $f|\Gamma$ is topologically conjugate to a topological Markov shift and the topological entropy $h(f|\Gamma) > 0$ (see [42] for details).

4.3. *Spectral decomposition*

In this section we will go further in the understanding the dynamics of dominated splitting over the limit set on surfaces.

DEFINITION 4.3.1. The limit set of $f : M \to M$ is

$$L(f) = \overline{\bigcup_{x \in M} \big(\omega(x) \cup \alpha(x)\big)},$$

where $\omega(x)$ and $\alpha(x)$ are the ω-limit and α-limit sets, respectively.

We will now state a classical theorem in hyperbolic dynamics (see [92] and Chapter 3, Hyperbolic dynamical systems (Hasselblatt), in Volume 1A of this handbook).

THEOREM 4.3.1 (Spectral Decomposition Theorem). *Let $f : M \to M$ be a diffeomorphism such that $L(f)$ is hyperbolic. Then the periodic points are dense in $L(f)$ and we can decompose the limit set into a finitely many compact transitive invariant disjoint sets $L(f) = \Lambda_1 \cup \Lambda_2 \cup \cdots \cup \Lambda_k$. Moreover, each Λ_i, $i = 1, \ldots, k$, can be decomposed into finitely many compact disjoint sets $\Lambda_i = \Lambda_{i1} \cup \cdots \cup \Lambda_{in_i}$ such that $f(\Lambda_{ij}) = \Lambda_{i\tilde{j}}$ where $\tilde{j} = j + 1 (\mathrm{mod}\, n_i)$ and $f^{n_i}_{/\Lambda_{ij}}$ is topologically mixing (indeed, Λ_{ij} is a homoclinic class, i.e. an equivalence class of $x \sim y :\Leftrightarrow W^u(x)$ transversely intersects $W^s(y)$ or vice versa).*

REMARK 4.3.1. The Spectral Decomposition Theorem is usually stated for Axiom A diffeomorphisms. However, we think that stating it when the limit set is hyperbolic is the right setting (notice that when f is Axiom A, then $L(f) = \Omega(f)$).

A similar description can be obtained for surface diffeomorphisms having dominated splitting over the limit set $L(f)$ as long as the system is smooth enough (C^2):

THEOREM 4.3.2 [81]. *Let $f \in \mathrm{Diff}^2(M^2)$ and assume that $L(f)$ has a dominated splitting. Then $L(f)$ can be decomposed into $L(f) = \mathcal{I} \cup \tilde{\mathcal{L}}(f) \cup \mathcal{R}$ such that*

1. *$\mathcal{I}$ is a set of periodic points with bounded periods and contained in a disjoint union of finitely many normally hyperbolic periodic arcs or simple closed curves;*
2. *$\mathcal{R}$ is a finite union of normally hyperbolic periodic simple closed curves supporting an irrational rotation;*
3. *$\tilde{\mathcal{L}}(f)$ can be decomposed into a disjoint union of finitely many compact invariant and transitive sets. The periodic points are dense in $\tilde{\mathcal{L}}(f)$ and at most finitely many of them are nonhyperbolic periodic points. The (basic) sets above are the union of finitely many (nontrivial) homoclinic classes. Furthermore, $f/\tilde{\mathcal{L}}(f)$ is expansive.*

Roughly speaking, the above theorem says that the dynamics of a C^2-diffeomorphism having dominated splitting can be decomposed into two parts: one where the dynamics consists on periodic and almost periodic motions $(\mathcal{I}, \mathcal{R})$ with the diffeomorphism acting equicontinuously; and another, where the dynamics are expansive and similar to the hyperbolic case.

Two immediate consequences follow from the previous theorem. First, any C^2-diffeomorphism with dominated splitting over $L(f)$ with a sequence of periodic points p_i with unbounded periods must exhibit a nontrivial homoclinic class $H(p_{i_0}, f) \neq \emptyset$ and hence:

COROLLARY 4.3.1. *The topological entropy of a C^2-diffeomorphism of a compact surface having dominated splitting over $L(f)$ and having a sequence of periodic points with unbounded periods is positive.*

Second, using Theorem 3.2.1 and the above one, it can be proved that:

COROLLARY 4.3.2. *Let $f \in \mathrm{Diff}^2(M^2)$ having infinitely many sinks or sources with unbounded period. Then, f can be C^1-approximated by a diffeomorphism exhibiting a homoclinic tangency.*

Let us comment briefly the proof of Theorem 4.3.2. The starting point is Theorem 4.1.1: the breakdown of hyperbolicity of a system with dominated splitting is due either to the presence of irrational rotations or to the presence of nonhyperbolic periodic points (and they could be extremely degenerate). The following theorem provides a way to deal with the presence of nonhyperbolic periodic points.

THEOREM 4.3.3. *Let $f : M \to M$ be a C^2-diffeomorphism of a two-dimensional compact Riemannian manifold M and let Λ be a compact invariant set having dominated splitting. Then, there exists an integer $N_1 > 0$ such that any periodic point $p \in \Lambda$ whose period is greater than N_1 is a hyperbolic periodic point of saddle type* (*i.e. index* 1).

In Section 4.3.1 we will give an outline of the proof of previous theorem. Now, denote by Per_1^N the set of hyperbolic periodic point of index 1 with period greater than N.

THEOREM 4.3.4. *Let $f \in \mathrm{Diff}^2(M^2)$ and assume that $\overline{\mathrm{Per}_1(f)}$ has dominated splitting. Then, there exists $N > 0$ such that $\overline{\mathrm{Per}_1^N(f)}$ can be decomposed into the disjoint union of finitely many homoclinic classes. Moreover, $\overline{\mathrm{Per}_1^N(f)}$ contains at most finitely many nonhyperbolic periodic points and $f_{/\overline{\mathrm{Per}_1^N(f)}}$ is expansive.*

Thus, the final step in the proof Theorem 4.3.2 is to show that $\tilde{\mathcal{L}}(f) \subset \overline{\mathrm{Per}_1^N(f)}$. As the reader might guess, these final steps seem to be similar to the hyperbolic case. In fact they are, but let us explain why. In the hyperbolic case, the description of the dynamics follows from a fundamental tool: at each point there are transverse invariant manifolds of uniform size and these manifolds have a dynamic meaning (points in the "stable" one are asymptotic to each other in the future, and points in the "unstable" one are asymptotic to each other in the past). Under the sole assumption of dominated splitting, even if locally invariant manifolds do exist, they do not have any dynamic meaning at all. However, in the two-dimensional case, using the fact that these locally invariant manifolds are one dimensional together with smoothness, we are able to prove that these manifolds already have a dynamic meaning, perhaps not of uniform size, but enough to proceed to a description of the dynamics.

4.3.1. *Sketch of proof of Theorem 4.3.3* Arguing by contradiction, assume that the conclusion of Theorem 4.3.3 is not true. Then, there exists a sequence p_n of periodic points whose periods are unbounded and they are not hyperbolic periodic points of saddle type. Let Λ_0 be the set of limit points of the orbits of the points p_n, i.e.

$$\Lambda_0 = \bigcap_{m \geqslant 0} \overline{\bigcup_{n \geqslant m} \mathcal{O}(p_n)}.$$

This set is compact invariant and, since it is a subset of Λ, has a dominated splitting.

Assume first that either all the periodic points in Λ_0 are hyperbolic or Λ_0 does not contain any periodic points at all. Then, by Theorem 4.1.1, we conclude that Λ_0 is a union of a hyperbolic set and a finite union of periodic simple normally hyperbolic closed curves. Since given a neighborhood of Λ_0 there exists n_0 such that, for any $n \geqslant n_0$, the orbit of p_n is contained in this neighborhood, we get a contradiction. In fact, the orbits of p_n cannot accumulate on the periodic simple closed curves since they are normally hyperbolic (attracting or repelling curves). Thus, Λ_0 is a hyperbolic set and so the maximal invariant set in an admissible compact neighborhood of Λ_0 is hyperbolic as well. In particular, for sufficiently large n, p_n lies on this maximal invariant set and so it must be a hyperbolic periodic point of saddle type, a contradiction and so our assumption is false.

Therefore, Λ_0 must contain a nonhyperbolic periodic point p, and we have that the orbits of a subsequence of $\{p_n\}$ (with unbounded periods) accumulate on p. This contradicts the following result:

THEOREM 4.3.5. *Let f be a C^2-diffeomorphism of a compact surface M and $\Lambda \subset \Omega(f)$ be a compact set having a dominated splitting. Let $p \in \Lambda$ be a nonhyperbolic periodic point and denote by N_p its period. Then, there exists a neighborhood U_p of p such that any periodic point $q \in \Lambda$ with period greater than $2N_p$ and whose orbit intersects U_p is a hyperbolic periodic point of saddle type.*

SKETCH OF PROOF. Let p be a nonhyperbolic periodic point and q a periodic point whose orbit goes through a very small neighborhood of p. We would like to show that q is hyperbolic of saddle type, that is, $\|Df^{n_q}_{|F_q}\| > 1$ and $\|Df^{n_q}_{|E_q}\| < 1$ where n_q is the period of q.

The idea is to split the orbit of q into pieces outside the neighborhood of p and inside it.

On one hand, we show that outside any neighborhood of p, the derivative along the F-direction for any trajectory is uniformly bounded away from zero, i.e. $\|Df^n_{|F_x}\| > c > 0$ for $f^i(x) \notin U_p$, $i = 1, \ldots, n$ (notice that this does not rule out that q might be a periodic attractor).

On the other hand, when a trajectory is going through a tiny neighborhood of p, not only does it not lose expansion (although the derivative of p along the F-direction might be one) but it has a good expansion along the F-direction from the first time that the point goes into U_p until the last time that remains in it (even if the exponential rate is close to one), i.e. if $f(x) \notin U_p$, $x, \ldots, f^{-n}(x) \in U_p$ and $f^{-(n+1)}(x) \notin U_p$ then $\|Df^n_{|F_x}\| > 2/c$. Let us explain this idea better.

First we consider a small central unstable segment J containing x. Observe that since a long trajectory of x is inside U_p, J is close to the central unstable manifold of p. Let us consider a segment I in a fundamental domain of the central unstable manifold of p, obtained as the "projection of J on the central unstable manifold of p along the central stable foliation". We show that the lengths of $f^{-k}(I)$ and $f^{-k}(J)$ are uniformly comparable for

any $1 \leqslant k \leqslant n$ and we conclude that

$$\left\| Df^{-n}_{/F(x)} \right\| \leqslant \frac{|f^{-n}(J)|}{|J|} \exp\left(K_0 \sum_{j=0}^{n-1} \left| f^{-j}(J) \right| \right)$$

$$\approx \frac{|f^{-n}(I)|}{|I|} \exp\left(K_0 \sum_{j=0}^{n-1} \left| f^{-j}(I) \right| \right) \leqslant \frac{|f^{-n}(I)|}{|I|} \exp\left(K_0 \left| W^{cu}_{\epsilon}(p) \right| \right).$$

Since for n large $|f^{-n}(I)|$ is arbitrarily small, we get the desired inequality. □

4.4. *Dynamical determinant in the presence of a dominated splitting*

Few results have been stated about the existence of an ergodic measure for invariant sets exhibiting a dominated splitting for a surface map. In some cases, the scheme proposed in [39] could be applied for systems with dominated splitting. In that paper it is proved that for a system that is Anosov on the whole torus except over a neutral fixed point, an SRB measure with absolutely continuous conditional measure on unstable manifolds does not exist.

A u-Gibbs state is an ergodic invariant probability measure, whose induced measures along the Pesin unstable manifolds are absolutely continuous with respect to Lebesgue measure (see Chapter 1, Partially hyperbolic dynamical systems (Hasselblatt and Pesin), in this handbook). An invariant probability measure μ is called a physical measure if there is a set of positive Lebesgue measure of points x such that $\frac{1}{n} \sum_{k=0}^{n-1} \delta_{f^k(x)}$ weakly converges to μ as $n \to \infty$. We say that a compact invariant set Λ is an attracting set if there is an open neighborhood U such that $\Lambda = \bigcap_{n>0} f^n(U)$. Let Λ_j be a basic set in the decomposition given in the previous section, and assume that Λ_j is an attracting set which does not contain any periodic points with unstable eigenvalue equal to 1. In this case, the results in [23] may be adapted to $f_{|\Lambda_j}$, proving that it possesses a single uμ-Gibbs state which is also a physical measure. Let us call SRB measure a uμ-Gibbs state which is also a physical measure. The results in [26] (see also [33]) indicate that the unique SRB measure μ furnished before has exponential rates of mixing (for Lipschitz observables).

In [7], a real analytic compact surface diffeomorphism f is considered, for which the tangent space over the limit set admits a dominated splitting. Baladi et al. studied its dynamical determinant $d_f(z)$,

$$d_f(z) = \exp - \sum_{n \geqslant 1} \frac{z^n}{n} \sum_{x \in \mathrm{Fix}^*\, f^n} \frac{1}{|\det(Df^n(x) - \mathrm{Id})|},$$

where $\mathrm{Fix}^*\, f^n$ denotes the (finite) set of fixed points of f^n with no zero Lyapunov exponents. The results in the present section indicate that the definition for $d_f(z)$ given above is appropriate under the assumption of dominated decomposition. Moreover, it is proved that $d_f(z)$ is either an entire function or an analytic function in a "slit plane". In particular, once

the contribution of the hyperbolic sinks has been removed, $d_f(z)$ is analytic in a disc of radius larger than one in all cases where it is known that f admits a unique exponentially mixing SRB probability measure.

The analysis of the dynamical determinant uses the dichotomy proved in Theorem 4.3.2 which says that $\Omega = \mathcal{I} \cup \tilde{\mathcal{L}}(f) \cup \mathcal{R}$, where $\mathcal{R}$ is a finite union of normally hyperbolic closed curves on which f is conjugate to an irrational rotation, $\mathcal{I}$ is contained in a finite union of normally hyperbolic arcs on which some iterate of f is the identity, and *either* $f|_\Lambda$ is Axiom A *or* there is a finite spectral decomposition of Λ ($\Lambda = \bigcup_{1 \leqslant j \leqslant n} \Lambda_j$ together with local product structure) and hyperbolicity is violated on Λ (only) through the presence of finitely many periodic orbits with a single zero Lyapunov exponent.

In order to state the result more precisely in an efficient way, first notice that $d_f(z)$ coincides with $d_{f|\Lambda}(z)$ and then observe that $\Lambda = \Lambda' \cup P$ where P is a finite union of isolated (hyperbolic) sinks and sources with multipliers (eigenvalues of Df^p, $p \geqslant 1$, the period) λ_E, λ_F. It is easy to check that the dynamical determinant $d_{f|P}$ is a finite product of

$$d_{f|\text{sink}}(z) = \prod_{j=0}^{\infty} \prod_{k=0}^{\infty} \left(z^P - \lambda_E^{-j} \lambda_F^{-k}\right), \quad |\lambda_{E,F}| < 1,$$

$$d_{f|\text{source}}(z) = \prod_{j=0}^{\infty} \prod_{k=0}^{\infty} \left(z^P - \lambda_E^{-j+1} \lambda_F^{-k+1}\right), \quad |\lambda_{E,F}| > 1.$$

The infinite products above clearly converge, and define entire functions with an obvious zero-set (in particular, $d_{f|\text{sink}}(z)$ is zero-free in the open unit disk and admits a single zero on the closed disk, which is simple at $z = 1$, while $d_{f|\text{source}}(z)$ admits a first zero at $1/(\lambda_E \lambda_F)$ which is inside the open disk).

So, we may therefore concentrate on the dynamical determinant of $f|_\Lambda$. For that, we take the set $\mathcal{N}$ composed of all nonhyperbolic periodic points p of f. For each $p \in \mathcal{N}$ we take $P = P(p)$ being the period of p, λ_E, λ_F, the eigenvalues of $Df^P(p)$ and the following subset $\Sigma(p)$ of $\mathbb{C}$ which is defined as:

1. $\{z \mid z^P \in [-1, 1]\}$ if $\lambda_F = -1$ and $|\lambda_E| < 1$;
2. $\{z \mid z^P \in [\min\{0, \Lambda_E\}, 1]\}$ if $\lambda_F = 1$ and $|\lambda_E| < 1$;
3. $\{z \mid z^P \in \lambda_F^{-1}[-1, 1]\}$ if $\lambda_E = -1$ and $|\lambda_F| > 1$;
4. $\{z \mid z^P \in \lambda_F^{-1}[\min\{0, \Lambda_F^{-1}\}, 1]\}$ if $\lambda_E = 1$ and $|\lambda_F| > 1$.

Now, the result can be summarized as follows:

THEOREM 4.4.1. *Let $f : M \to M$ be a real analytic diffeomorphism of a compact analytic Riemannian surface. Assume that f admits a dominated splitting over the nonwandering set Ω and let Λ be the* (*almost*) *hyperbolic component in the spectral decomposition of Ω. Let Λ_j be an element in the spectral decomposition of Λ which is not an isolated sink or source, and let $\mathcal{N}_j$ be the finite* (*possibly empty*) *set of nonhyperbolic periodic points*

in Λ_j. Then $d_{f|\Lambda_j}(z)$ is holomorphic in the plane, slit plane, or multiply slit plane defined by

$$\left\{ z \in \mathbb{C} \mid 1/z \notin \bigcup_{p \in \mathcal{N}} \Sigma(p) \right\}.$$

In other words:

1. If all periodic points in Λ_j are hyperbolic, then $d_{f|\Lambda_j}(z)$ is an entire function with no zeroes in the open unit disc. The point $z = 1$ is a zero if and only if Λ_j is an attractor. It is then a simple zero, and the only zero of unit modulus if Λ_j is mixing.
2. If there exist periodic points in Λ_j with multipliers $\lambda_E < 1$ and $|\lambda_F| = \pm 1$, then $d_f(z)$ is analytic and nonzero in the disc of radius 1, with a possibly nonpolar singularity at $z = \pm 1$, and it admits an analytic extension to a (multiply) slit plane.
3. If there exist periodic points with multipliers $|\lambda_E| = \pm 1$ and $\lambda_F > 1$ in Λ_j, but no periodic orbits with both Lyapunov exponents nonpositive, letting λ_F be the multiplier of smallest modulus, then $d_f(z)$ is analytic and nonzero in the disc of radius 1, and it may be analytically extended to the disc of radius $|\lambda_F| > 1$, with a possibly nonpolar singularity at $z = \lambda_F$, and a further analytic extension to a (multiply) slit plane.

SKETCH OF PROOF. Recall the spectral decomposition stated for $\Omega(f)$ presented above. In the first case, since $\mathcal{R}$ does not contain any periodic orbits and the periodic points in $\mathcal{I}$ have one Lyapunov exponent zero, the results of Rugh (see [86,87]) on the dynamical determinants of hyperbolic analytic maps immediately imply that $d_f(z)$ is an entire function. The key point in Rugh's analysis, inspired by Ruelle's [85] seminal study (in the case when the dynamical foliations are analytic), was to express $d_f(z)$ as a quotient of the Grothendieck–Fredholm determinants of two nuclear operators, proving also that zeros in the denominator are always canceled by the numerator.

In the second case, other techniques to investigate $d_f(z)$ are used. Note first that if f is real analytic with dominated decomposition over Λ, then the set $\mathcal{I}$ in the spectral decomposition from [79] is a finite union of closed curves, with $\mathcal{I} \cap \Lambda = \emptyset$. (It is easy to construct examples where $\mathcal{I}$ is not empty: just take a real analytic flow on the sphere with both poles as sources and the equator as limit set.) In a nutshell, $d_f(z)$ is morally the determinant of a transfer operator, the building blocks of which are either "good", i.e. of the hyperbolic type studied in [86,87], or approximate direct products of a one-dimensional hyperbolic operator and a one-dimensional parabolic operator describing the local jet at nonhyperbolic periodic points. For the parabolic operator, we adapt the analysis of one-dimensional analytic dynamics with neutral fixed points, also due to Rugh [88], to this setting. The crucial tool to do this is an approximate Fatou coordinate for parabolic points. We prove that $d_f(z)$ is an analytic function in a "slit plane". □

5. Nonhyperbolic robustly transitive systems

As we said in the introduction, there are manifolds of dimension at least 3 supporting diffeomorphisms that are both robustly transitive and nonhyperbolic ($r \geqslant 1$). Many of these examples show a weaker form of hyperbolicity: they are either partially hyperbolic or they exhibit a dominated splitting. The notion of partial hyperbolicity and different examples of it are discussed in Section 5.1. The nonhyperbolic robustly transitive examples are described in Section 5.2. In Section 5.3 it is proved that having a dominated splitting is a necessary condition for C^1-robust transitivity. The results are discussed in Section 5.3. We point out that the mentioned result is only known in the C^1-topology.

5.1. *Partial hyperbolicity*

We start with the definition of partial hyperbolicity.

DEFINITION 5.1.1. Given a closed invariant set Λ we say that it is strongly partially hyperbolic if $T_\Lambda M = E^{ss} \oplus E^c \oplus E^{uu}$ and there are constants $0 < \sigma^{-1} < \gamma^{-1} < 1 < \gamma < \sigma$ such that

$$\|Df_{|E^{ss}_x}\| < \sigma^{-1} < \sigma < \left\|Df^{-1}_{|E^{uu}}\right\|^{-1} \quad \forall x \in \Lambda,$$
$$\|Df_{|E^{ss}_x}\| \left\|Df^{-1}_{|E^c_{f^{-1}(x)}}\right\| < \gamma^{-1}, \quad \|Df_{|E^c_x}\| \left\|Df^{-1}_{|E^{uu}_{f^{-1}(x)}}\right\| < \gamma^{-1} \quad \forall x \in \Lambda.$$

We say that Λ is partially hyperbolic if $T_\Lambda M = E^s \oplus E^{cu}$ and there are constants $0 < \sigma < \gamma < 1$ such that

$$\|Df_{|E^s_x}\| < \sigma \quad \forall x \in \Lambda,$$
$$\|Df_{|E^s_x}\| \left\|Df^{-1}_{|E^c_{f^{-1}(x)}}\right\| < \gamma \quad \forall x \in \Lambda.$$

It is well known that the subbundles E^{ss} and E^{uu} are uniquely integrable and hence we have two foliations in M called the (strong) stable one, denoted by $\mathcal{F}^{ss}(f)$ and the (strong) unstable one, denoted by $\mathcal{F}^{uu}(f)$ which are tangent to E^{ss} and E^{uu}, respectively. We will denote by $\mathcal{F}^{ss}(x, f)$ and by $\mathcal{F}^{uu}(x, f)$ the leaves of these foliations passing through the point x.

On the other hand, every diffeomorphism $g : M \to M$ sufficiently C^1-close to a partially hyperbolic diffeomorphism f is also partially hyperbolic and therefore it has two invariant foliations $\mathcal{F}^{ss}(g)$, $\mathcal{F}^{uu}(g)$.

The subbundles E^c, $E^{cs} = E^{ss} \oplus E^c$ and $E^{cu} = E^c \oplus E^{uu}$ (called center, center-stable and center-unstable subbundle, respectively) are not (in general) integrable. However, we can choose a continuous family of locally invariant manifolds tangent to them.

For references about these results, see Chapter 1, Partially hyperbolic dynamical systems (Hasselblatt and Pesin), in this handbook.

DEFINITION 5.1.2. Let $f : M \to M$ be a C^r-diffeomorphism. We say that f is *robustly transitive* if there is a C^r-neighborhood $\mathcal{U}(f)$ of f such that every $g \in \mathcal{U}(f)$ is transitive (i.e. has a dense orbit).

We will consider different constructions of robustly transitive partially hyperbolic systems. We list some of them:

1. Product and skew products of a hyperbolic system with a nonhyperbolic one.
2. Derived from Anosov: bifurcation of Anosov maps or maps isotopic to an Anosov system and perturbations of these maps.
3. Time-one map of an Anosov flow. In this category we have to consider the following types of Anosov flows:
   (a) Anosov flows which are suspensions of an Anosov map;
   (b) Anosov flows which are not suspensions.

The first type of examples is built over a product manifold where product diffeomorphisms act. More precisely, we take transitive Anosov diffeomorphisms A on M and nonhyperbolic transitive diffeomorphisms g on a manifold N with the property that the largest and weakest expansion of Dg is dominated by the hyperbolic direction of A. In other words, given the splitting $TM = E^s \oplus E^u$ for A, it is assumed that

$$\frac{\|DA_{|E^s}\|}{m\{Dg\}} < \lambda \quad \text{and} \quad \frac{\|Dg\|}{\|DA_{|E^u}\|} < \lambda$$

for some $\lambda < 1$ and where $m\{L\}$ is the minimum norm of L, i.e. $m\{L\} = \|L^{-1}\|^{-1}$. Then, the product map is considered:

$$A \times g : M \times N \to M \times N,$$
$$(A \times g)(x, y) = \big(A(x), g(y)\big).$$

Observe that the map is partially hyperbolic (because g is not hyperbolic) and the central subbundle is given by TN. Moreover, the central foliation has the property that each central leaf is the manifold N. This example is transitive, not necessarily robustly transitive, and we will show in the next section that it can be perturbed to get a robustly transitive system.

A similar construction is the so-called *skew product* which consists of choosing a map

$$G : M \to \mathrm{Diff}^r(N)$$

with the property that for any $x \in M$ it follows that

$$\frac{\|DA_{|E^s}\|}{m\{D[G(x)]\}} < \lambda \quad \text{and} \quad \frac{\|D[G(x)]\|}{\|DA_{|E^u}\|} < \lambda.$$

Then, the following map is considered:

$$A \times G : M \times N \to M \times N,$$
$$(A \times G)(x, y) = \big(A(x), G(x)(y)\big).$$

Observe that for this type of examples, the manifold $M \times N$ is foliated by leaves homeomorphic to N and this foliation is normally hyperbolic. Therefore, it follows that for any perturbation, there is a central foliation whose central leaves are also homeomorphic to N.

In some cases, it is possible to prove that this kind of system (as we will show later) can be either approximated by a robustly transitive system or by an Axiom A system.

The second kind of example cannot be built as a product and the central leaves are not necessarily compact. In fact, in some cases the central foliation is uniquely integrable and each central leaf is dense; in this case, we say that the central foliation is minimal. To obtain these examples, an Anosov map is taken and a deformation of the system isotopic to the initial map is performed, with the property that some direction remains hyperbolic and other fails to be hyperbolic. See Mañé's example explained in the next subsection.

The dynamical properties of the third type of example depend strongly on whether they are or are not suspensions. The suspension case is constructed from the suspension (with constant roof one) of an Anosov map A acting on M. The suspension flow φ_s^A exhibits a splitting $E^s \oplus [\partial_s \varphi_s^A] \oplus E^u$ where E^s and E^u are the invariant directions of A and $[\partial_s \varphi_s^A]$ is the flow direction.

The time-one map φ_1^A is not hyperbolic since the flow trajectory is neither expanded not contracted. In this case, observe that the central leaves are flow trajectories. Some of the trajectories are compact (the ones associated to periodic trajectories and induced by periodic orbits of the map A) and some other trajectories are not compact.

It is also possible to get examples of Anosov flows which are not suspensions. For that, we refer to [22] and the well-known geodesic flow on manifolds of negative curvature.

In all the previous constructions it is necessary to check that the system (or at least a perturbation of it) is robustly transitive.

5.2. *Examples of nonhyperbolic robustly transitive systems*

The first examples of robustly nonhyperbolic systems (examples of robustly transitive systems which are not Anosov) were given by M. Shub (see [90]), who considered skew products on the 4-torus of an Anosov with a Derived from Anosov diffeomorphisms. Then, R. Mañé (see [47]) reduced the dimension of such examples by showing that certain Derived of Anosov diffeomorphisms on the 3-torus are robustly transitive.

For a long time these examples remained the unique known ones. Later, Diaz, studying the unfolding of hetero-dimensional cycles, produced maximal invariant sets that are robustly transitive and nonhyperbolic. These ideas were pushed, in [17], where new classes of nonhyperbolic robustly transitive diffeomorphisms were presented on manifolds other than the n-dimensional torus and satisfying a weaker form of hyperbolicity. Moreover, a general construction of maximal invariant set which is robustly transitive and nonhyperbolic was also shown. In resume, the examples given in [17] are basically:

1. Perturbations of the time-one map of any transitive Anosov flow.
2. Perturbations of $(A, \mathrm{id}_N) : \mathbb{T}^n \times N \to \mathbb{T}^n \times N$, where A is a transitive Anosov diffeomorphism on the n-dimensional torus $\mathbb{T}^n$ and N is any compact manifold. In this example the nonhyperbolic central direction is chosen tangent to the fibers $\{., N\}$ providing in this way examples of robustly transitive diffeomorphism having central direction with arbitrary dimension.

3. Perturbations of $(f, \mathrm{id}_N) : \Lambda \times N \to \Lambda \times N$, where Λ is a hyperbolic basic piece for f and N is any compact manifold.

Later, in [23], examples without hyperbolic subbundles were introduced:

1. Examples of robustly transitive diffeomorphisms on T^3 without any stable bundle, that is, $TM = E^{cs} \oplus E^u$, where E^{cs} is indecomposable and nonhyperbolic.
2. Examples of robustly transitive diffeomorphisms on T^4 without any hyperbolic bundle (stable or unstable), that is, $TM = E^{cs} \oplus E^{cu}$ is a dominated splitting where E^{cs} and E^{cu} are indecomposable and nonhyperbolic.

In the next subsections we give a rough idea how these examples are constructed and we outline the arguments that prove that these systems are robustly transitive. We follow the construction initially done by the authors (performing some modification in some cases) and in Section 5.2.5 a condition is given that allows to obtain the same examples from a different point of view.

5.2.1. *Shub's example* Let $f : \mathbb{T}^2 \to \mathbb{T}^2$ be an Anosov diffeomorphism having two fixed points p and q. Since f is Anosov, $T\mathbb{T}^2 = E^{ss} \oplus E^{uu}$ with $\|Df_{/E^{ss}}\| < \lambda < 1$ and $\|Df^{-1}_{/E^{uu}}\| < \lambda$.

Now, consider a smooth family of torus diffeomorphisms $g_x : \mathbb{T}^2 \to \mathbb{T}^2$ indexed by $x \in \mathbb{T}^2$ such that:

- $T\mathbb{T}^2 = E^s(g_x) \oplus E^c(g_x)$ invariant under $D(g_x)$ with $\|D(g_x)_{/E^s(g_x)}\| < \mu < \mu_1 < 1$ and $\mu < \mu_1 < \|D(g_x)/E^c(g_x)\| \leqslant \mu^{-1}$;
- g_x preserves cone fields C^s and C^{cu} for all $x \in \mathbb{T}^2$;
- g_p is Anosov and g_q is a DA (derived from Anosov) map;
- $g_x(p) = p$ for every x, and p is an attractor for g_q.

It is assumed (taking a power of f if necessary) that $\lambda < \mu$. Next, let us define the map on $\mathbb{T}^4$ which is a skew product and is the candidate to be robustly transitive:

$$F : \mathbb{T}^2 \times \mathbb{T}^2 \to \mathbb{T}^2 \times \mathbb{T}^2, \quad F(x, y) = \big(f(x), g_x(y)\big).$$

It is not difficult to show that F is partially hyperbolic: $T_{(x,y)}\mathbb{T}^4 = E^{ss}(x, y) \oplus E^s(x, y) \oplus E^c(x, y) \oplus E^u(x, y)$. Set $E^s = E^{ss} \oplus E^s$. Let us show that the stable foliation (tangent to $E^{ss} \oplus E^s$) is minimal. First observe that

$$W^s\big(\{p\} \times \mathbb{T}^2\big) = \bigcup_{z \in \mathbb{T}^2} W^{ss}(p, z) = W^{ss}(p, f) \times \mathbb{T}^2$$

and hence is dense in $\mathbb{T}^2 \times \mathbb{T}^2$. Moreover, since g_p is Anosov, we have that

$$W^s(p, z) = \bigcup_{y \in W^s(z, g_p)} W^{ss}(p, y)$$

is dense in $\mathbb{T}^2 \times \mathbb{T}^2$ for all $(p, z) \in \{p\} \times \mathbb{T}^2$.

On the other hand, if $(z, w) \in \mathbb{T}^2 \times \mathbb{T}^2$ then $W^{uu}((z, w)) \cap W^s(\{p\} \times \mathbb{T}^2) \neq \emptyset$. From this it follows that the stable foliation $\mathcal{F}^{ss}$ (whose leaves are tangent to $E^{ss} \oplus E^s$) is minimal.

Then, it can be proved that this property holds for any small perturbation of the initial map. Observe that if the strong stable foliation is minimal then the system is transitive. Since it is proved that the strong foliation of any (small) perturbation of the initial system is minimal, it follows that the initial system is robustly transitive.

5.2.2. *Mañé's example* Mañé showed that certain Derived-from-Anosov map are robustly transitive. To prove that, first he considered a linear Anosov map $A : \mathbb{T}^3 \to \mathbb{T}^3$ having three real eigenvalues, $0 < \lambda < 1 < \lambda_2 < \lambda_3$. Observe that there are three invariant foliations, each associated to one eigenvalue, which correspond to the projections to $\mathbb{T}^3$ of the eigenspaces associated to each eigenvalue. In particular, the foliation associated to the central eigenvalue (called central foliation) is minimal, i.e. each central leaf is dense.

Now, a Derived-from-Anosov map is obtained by performing a perturbation in a small neighborhood D of the fixed point. More precisely, the Derived-from-Anosov map is obtained in such a way that the following properties hold:

1. The central foliation of the initial Anosov system is preserved;
2. The new system keeps two hyperbolic directions;
3. In the complement of D the new system is equal to the initial one;
4. In the small neighborhood D, two points of different indices appear as a consequences of a bifurcation of the fixed point.

The last item implies that the expansion of the central direction is lost. On one hand, observe that the new systems have the property that the central foliation remains minimal. Moreover, it verifies a property called plaque expansiveness (see [38]). As a consequence of this property, it follows that for any perturbation there exists a unique central foliation which is also minimal. On the other hand, using the facts that the region D is small and the unstable leaves grow exponentially, it is shown that any unstable leaf of a fixed length has a point whose orbit remains in the complement of D and whose central direction therefore expands. This allows to show that any central segment containing this point has length growing to infinity under positive iterates. All these properties together imply transitivity. In fact, given an open set U, an iterate of it will contain a point that expands along the central direction, and so the iterates of U will start to grow along the central direction and using the density of any central leaf, the density of the iterates is obtained.

5.2.3. *Bonatti–Diaz's construction* Now we will explain the arguments used in [30] and [17] to get robustly transitive diffeomorphisms or sets. The authors use a geometrical construction called *blender*. It is said that f has a center-stable blender (center-unstable blender) Γ associated to a hyperbolic saddle p of index k if Γ is a hyperbolic invariant set contained in the homoclinic class of p and there are an open set $\mathcal{D}$ of embeddings of the disk D^{k-1} in M and a C^1-neighborhood $\mathcal{U}$ of f such that, for every $g \in \mathcal{U}$, any disc $D \in \mathcal{D}$ intersects the closure of $W^s_\epsilon(\Gamma_g, g)$ (respectively $W^u_\epsilon(\Gamma_g, g)$), where Γ_g is the continuation of the hyperbolic set Γ for g and $W^s_\epsilon(\Gamma_g, g)$ is the local stable manifold of the invariant set Λ_g and $W^u_\epsilon(\Gamma_g, g)$ is the local unstable manifold of the invariant set Λ_g.

One way to construct a center-blender in a three-dimensional manifold is to construct a hyperbolic maximal invariant set Γ associated to a periodic point p such that for the continuation of p for g close to f it follows that $\Pi^{ss}(W^s_\epsilon(p_g) \cap \Gamma)$ contains an interval,

where Π^{ss} is the projection along the strong stable direction to some central direction contained in the local stable manifold of p.

Now, using these blenders, let us show a general strategy to guarantee robust transitivity. Let f be partially hyperbolic in the whole manifold with nontrivial splitting $TM = E^s \oplus E^c \oplus E^u$ such that:

1. f has two periodic points p and q of indices one and two, respectively;
2. f has a center-unstable blender Γ associated to q;
3. There is a constant $L > 0$ such that any strong stable segment of length larger than L intersects the local unstable manifold Γ, and any strong unstable segment of length larger than L intersects the local stable manifold of q.

It follows from this assumption that f is a nonhyperbolic robustly transitive map. The nonhyperbolicity follows from the presence of points of different indices. Properties 2 and 3 imply transitivity: by property 3, any strong leaf of length larger than L is caught by the blender, and the dynamic is mixed inside it. In fact, given any open set we get that some future iterates will intersect the local stable manifold of the point q (the iterates grow along unstable leaves and by property 3 reach the local stable of q); for any other open set it follows that some negative iterate of it will intersect the blender Γ and hence the unstable manifold of q. These two facts imply that both sets under iteration will intersect. On the other hand, we get that the blender is robust and property 3 also holds for a perturbation of f and this implies the robustness of transitivity.

These arguments are used in [17] to prove that the time-one map of a transitive Anosov flow and the product of an Anosov diffeomorphism by the identity map defined on a compact manifold N are in the closure of the set of robustly transitive diffeomorphisms. In both cases, the dynamics is perturbed along a compact central leaf (along a closed orbit of the flow in the case of the time-one map of an Anosov flow and along a periodic central leaf for the product and skew product) to get a pair of periodic orbits of different index and then the construction of a blender with these points is performed.

5.2.4. *Robustly transitive diffeomorphisms without any hyperbolic directions* We will now explain the ideas (given in [23]) of the construction of robustly transitive diffeomorphisms without any hyperbolic directions. This construction follows closely the example by Mañé.

Consider a linear Anosov map A of the torus $\mathbb{T}^4$ having 4 real eigenvalues, $0 < \lambda_1 < \lambda_2 < 1 < \lambda_3 < \lambda_4$. Then choose A-invariant cone fields C^{cu} corresponding to the expanding eigenvalues λ_3 and λ_4 and C^{cs} around the contracting eigenvalues. Now, take two small boxes C_1 and C_2, and consider a diffeomorphism f coinciding with A outside the boxes C_1 and C_2, and verifying the following:

1. f contains a fixed point p in C_1 (of index 2) with a complex eigenvalue with eigenspace inside C^{cs} and a fixed point q (of index 1) with eigenspace also in C^{cs}. This implies that C^{cs} does not contain a hyperbolic stable direction: the existence of the contracting complex eigenvalue implies that any stable subbundle has to be of dimension 2, but the point being of index 1 implies that such bundle has dimension at most one.

2. f contains a fixed point p_2 in C_2 (of index 2) with an expanding complex eigenvalue and having a fixed point q_2 of index 1. Arguing as before, these properties prevent the existence of a hyperbolic unstable subbundle.

To assure that f is robustly transitive the following properties also hold:

1. Df (respectively Df^{-1}) preserves the cone field C^{cu} (respectively C^{cs}) and uniformly expands the area in this cone field;
2. The restriction of f to the complement of C_1 uniformly expands the vectors in C^{cu};
3. The restriction of f^{-1} to the complement of C_2 uniformly expands the vectors in C^{cs}.

Choosing sufficiently thin cone fields, one obtains that there exists a constant $L > 0$ such that every center-unstable disk D^{cu} (tangent to C^{cu}) of radius larger than L intersects any center-stable disk D^{cs} (tangent to C^{cs}) of radius larger than L.

Arguing as in Mañé's example, but using the uniform expansion of the area in C^{cu} instead of the uniform expansion of the vectors, it is shown that every center-unstable disk D contains a point whose forward orbit remains in the complement of C_1 and this allows to show (as in Mañé example) that $f^n(D)$ contains a center-unstable disk of radius larger than L for every large $n > 0$. The same argument shows that the large negative iterates of any center-stable disk D contain a center-stable disk of radius larger than L. This implies transitivity of f.

5.2.5. *Sufficient conditions for robust transitivity* The mentioned examples of robustly transitive systems which are partially hyperbolic share the strongest property: not only are they robustly transitive but also (at least) one of the strong foliations is robustly minimal, that is, every leaf is dense in the manifold.

DEFINITION 5.2.1. Let $f : M \to M$ be a partially hyperbolic diffeomorphism. It is said that $\mathcal{F}^{ss}(f)$ is *robustly minimal* if there exists a C^1-neighborhood $\mathcal{U}(f)$ such that $\mathcal{F}^{ss}(g)$ is minimal for every diffeomorphism $g \in \mathcal{U}(f)$.

Recall that if (for instance) $\mathcal{F}^{ss}(f)$ is robustly minimal then f is robustly transitive, i.e. every diffeomorphism C^1-close to f is transitive.

All the examples of robust transitivity considered are based either on a property of the initial system or on a geometric construction. But, *does there exist a necessary and sufficient condition on partially hyperbolic systems for transitivity to be equivalent to robust transitivity*? *Can this property be characterized in terms of the dynamics of the tangent map*? This is clear for Anosov maps, where transitivity implies robust transitivity, but what about nonhyperbolic maps?

In [83] partially hyperbolic systems having at least one minimal strong foliation are studied and sufficient conditions are given in order to guarantee that this foliation remains minimal under C^r-perturbations. Before we enunciate the sufficient condition, let us introduce a notation: given a linear isomorphism $L : V \to W$ between normed vector spaces, we denote by $m\{L\}$ the minimum norm of L, i.e. $m\{L\} = \|L^{-1}\|^{-1}$.

DEFINITION 5.2.2 (*Property SH*). Let $f \in \mathrm{Diff}^r(M)$ be a partially hyperbolic diffeomorphism. We say that f exhibits the property SH (or has the property SH), if there exist

$\lambda_0 > 1$, $C > 0$ such that for any $x \in M$ there exists $y^u(x) \in \mathcal{F}^{uu}_1(x, f)$ (the ball of radius 1 in $\mathcal{F}^{uu}(x, f)$ centered at x) for which

$$m\left\{Df^n_{/E^c(f^m(y^u(x)))}\right\} > C\lambda_0^n \quad \text{for any } n > 0,\ m > 0.$$

In other words, it is required that in any disk of radius 1 in any leaf of $\mathcal{F}^{uu}(f)$ there is a point y^u where the central bundle E^c has a uniform expanding behavior along the future orbit of y^u. A nice image of the above is the existence of a hyperbolic set (with E^c being part of the unstable bundle) such that the local stable manifold of this hyperbolic set is a global section to the foliation $\mathcal{F}^{uu}(f)$.

Using this property the next theorem follows.

THEOREM 5.2.1. *Let $f \in \mathrm{Diff}^r(M)$ be a partially hyperbolic diffeomorphism satisfying the property SH and such that the* (*strong*) *stable foliation $\mathcal{F}^{ss}(f)$ is minimal. Then, $\mathcal{F}^{ss}(f)$ is robustly minimal.*

A similar result for the foliation $\mathcal{F}^{uu}(f)$ holds provided that f^{-1} satisfies the property SH. As an immediate consequence any f that satisfies the conditions of the previous theorem is robustly transitive.

It is also possible to use the previous theorem to reconstruct the examples by Shub, Mañé and Bonatti–Diaz.

Let us mention here that in [21] it is shown that C^1-generically, for a robustly transitive diffeomorphism on a three-dimensional manifold M^3 one of the (strong) foliations is minimal. On the other hand, the following natural question remains: is property SH necessary for robustly transitive partially hyperbolic systems?

SKETCH OF THE PROOF OF THEOREM 5.2.1. The proof of Theorem 5.2.1 relies on the three following facts:

1. If the property SH holds then any open set has a positive iterate that contains a central unstable disc of a fixed size;
2. The property SH is robust under perturbation;
3. For any small perturbation of the initial map, it follows that any leaf of the strong stable foliation is "almost" dense.

The last item means that given a positive constant δ, any leaf of the strong stable foliation of any small perturbation is δ-dense. On one hand, by item 2 the property SH holds for any perturbation so by item 1 it follows that any open set has a positive iterate that contains a central unstable disc of a fixed size chosen for the initial system. On the other hand, by item 3, any leaf of the strong stable foliation of any small perturbation is δ-dense. Therefore, if δ is small, given a close perturbation of the initial system, it follows that some positive iterate of any open set intersects any strong stable leaf. So, the strong stable foliation of the perturbed system is also minimal.

This key property that guarantees the robustness of stable foliation of a partially hyperbolic diffeomorphism can be formulated in the following way: *Some Hyperbolicity on the central distribution E^c at some points.*

5.3. *Robust transitivity and dominated splitting*

The previous examples lead to two natural questions:

- *Is there a characterization of robustly transitive sets in terms of the tangent map*?
- *Can we describe the dynamics under the assumption of either partial hyperbolicity or dominated splitting*?

Both questions have a partial answer: it is known that C^1-robust transitivity implies dominated splitting, however, in a generic setting the converse of the previous assertion is unknown.

In the sequel, we will say that a compact invariant set Λ for a diffeomorphism f is a C^1-robustly transitive set, if Λ is a transitive maximal invariant set (i.e. there is U such that $\Lambda = \bigcap_{n\in\mathbb{Z}} f^n(U)$ and Λ has a dense orbit) and for any g C^1-close to f it follows that $\Lambda_g = \bigcap_{n\in\mathbb{Z}} g^n(U)$ is also a transitive set.

The next result, proved in dimension two in [48], dimension three in [32] and in greater dimension in [19], shows that these two questions are closely related. In fact, some kind of dynamics on the tangent bundle is implied by robust transitivity:

THEOREM 5.3.1. *Any robustly transitive set of a C^1-diffeomorphism exhibits a dominated splitting such that its extremal bundles are uniformly volume contracted or expanded*.

In dimension two, the set is hyperbolic (see [48]); in dimension three the sets exhibit a partially hyperbolic splitting (see [32]) and just a dominated splitting in higher dimension.

This last theorem can also be formulated in the following way:

THEOREM 5.3.2 [19]. *There is a residual subset of C^1-diffeomorphisms such that for any diffeomorphism in the residual set any homoclinic class of a periodic point* (*the closure of the intersection of the stable and unstable manifold of it*) *either has dominated splitting or it is contained in the closure of infinitely many sources or sinks*.

In the particular case of dimension two, the following result proved in [48] is more general:

THEOREM 5.3.3 [48]. *There is a residual subset of C^1-diffeomorphisms that either satisfy Axiom* A *or exhibit infinitely many sources or sinks*.

First, we give a sketch of the proof of Theorem 5.3.3. Later we explain how the arguments work in higher dimension.

As in the proof of the dominated splitting versus tangencies, a key element in the proof is that in the C^1-topology any small perturbation of the derivative of a diffeomorphism f along a periodic orbit p can be performed dynamically as a C^1-perturbation of f (see Lemma 2.3.1).

5.3.1. *Sketch of the proof of Theorem 5.3.3* The proof of the theorem in dimension two establishes that given a C^1-diffeomorphism f such that the number of sinks and repellers for any diffeomorphism C^1-close to f is constant, then f is an Axiom A diffeomorphism.

To show that, it is necessary to prove that the limit set is hyperbolic. This is done in two steps:

- *Step* I: the limit set exhibits a dominated splitting;
- *Step* II: this dominated splitting is hyperbolic.

To get the first step, let $\mathcal{P}$ be a set of periodic hyperbolic saddles of a diffeomorphism f. Since we are working in the C^1-topology, the C^1-closing lemma allows us to assume that the periodic points are dense in the limit set. Over $\mathcal{P}$, the natural splitting $T_{\mathcal{P}}M = \bigcup_{p\in\mathcal{P}} E^s(p) \oplus E^u(p)$ is considered where $E^s(p)$ and $E^u(p)$ are the eigenspaces of Df^{n_p}, and n_p is the period of the periodic point p. The goal is to prove that this splitting is a dominated splitting and so can be extended to Λ. To this end, it is shown that if the splitting over the periodic points is not uniformly dominated, then by an arbitrarily small perturbation of the derivative of f along the orbit of some $p \in \mathcal{P}$, a matrix is obtained that has complex eigenvalues of modulus different from one. In other words, by a perturbation either a new sink or a new repeller is obtained, contradicting the hypothesis that no sinks or repellers can be created by perturbation. More precisely, first we use Theorem 2.3.2 that states: *if the splitting $E^s \oplus E^u$ over $\mathcal{P}$ is not dominated, then there is an arbitrarily small perturbation of the derivative of f along the orbit of a point $p \in \mathcal{P}$ such that the angle between the eigenspaces of the corresponding matrix is arbitrarily small.*

Finally, we apply

CLAIM 5.3.1. *Suppose $A \in GL(2, R)$ has two different real eigenvalues whose eigenspaces form an angle less than ϵ. Then there is $t \in [-\epsilon, \epsilon]$ such that the matrix $R_t \circ A$ has a pair of conjugate complex eigenvalues* (R_t *is the rotation by* t).

To prove this claim, identifying the projective space with S^1, observe that A acts on S^1 as a Morse–Smale diffeomorphism with one sink and one repeller close to each other. Then, composing with a small rotation, a new map with irrational rotation is obtained which corresponds to the fact that $R_t \circ A$ has a complex eigenvalue.

After it is proved that the limit set has dominated splitting, it is shown that the splitting is in fact hyperbolic using the fact that neither sinks nor repellers can be created by perturbations. To do this, the *Ergodic Closing Lemma* is used. This lemma states that with full probability any recurrent orbit can be shadowed by a periodic orbit of a system C^1-close to the initial one. Before stating this result, we need some definitions.

Define $B_\varepsilon(f, x)$ as the set of points $y \in M$ such that $d(f^n(x), y) \leqslant \varepsilon$ for some integer n and define $\Sigma(f)$ to be those points $x \in M$ such that for every neighborhood $\mathcal{U} \subset \mathrm{Diff}^1(M)$ of f and for every $\varepsilon > 0$ there are $g \in \mathcal{U}$ and $y \in B_\epsilon(f, x)$ such that $y \in \mathrm{Per}(g)$, $g = f$ on $M \setminus B_\varepsilon(f, x)$ and $d(f^j(x), g^j(y)) \leqslant \varepsilon$ for all j such that $0 \leqslant j \leqslant m$, where m is the g-period of y.

THEOREM 5.3.4 (Ergodic Closing Lemma [48]). *If $f \in \mathrm{Diff}^1(M)$ then $\Sigma(f)$ has measure one for every f-invariant probability measure on the Borel sets of M.*

An equivalent statement of Theorem 5.3.4 is the following:

Let $f \in \mathrm{Diff}^1(M)$ and let μ be an invariant measure of f. Then, there exist a sequence $\{g_n\}$ of C^1-diffeomorphisms converging in the C^1-topology to f and a sequence of mea-

sures $\{\mu_n\}$ invariant for g_n such that each μ_n is supported on a periodic orbit and the sequence of measures converges to μ in the weak topology.

Using this version, it is possible to prove that if one of the subbundles of the dominated splitting is not hyperbolic, then it is possible to create a sink or a repeller for a C^1-perturbation.

Indeed, let us assume that F is not expanding. Then, there is an invariant measure μ such that

$$\int |Df_{|F(x)}|\mu \leqslant 0.$$

By the Ergodic Closing Lemma, for any $\delta > 0$ there is g C^1-close to f and an invariant measure ν supported on a periodic point such that ν is δ-close to μ in the weak topology. Using the facts that the dominated splitting depends continuously on the perturbation and the support of ν is close to the support of μ, it follows that over the support of ν there is a splitting $E_g \oplus F_g$ with the property that the direction F_g is close to F. So

$$\int |Dg_{|F_g(x)}|\nu \leqslant \delta.$$

Since the support of ν is a periodic orbit, it follows that the eigenvalue along the F_g direction is close to 0. Then, by a C^1-perturbation this eigenvalue can be made smaller than zero. From the domination property it follows that the other eigenvalue is also smaller than zero. And this implies that a sink has been created, which is a contradiction to the fact that it is assumed that neither sinks nor repellers can be created by perturbations.

So, it has been proved that the limit set is hyperbolic. Then, the limit set can be decomposed into a finite union of maximal invariant transitive sets. It follows that the no-cycle condition holds; if it is not the case, it is possible to create a tangency by perturbation and this leads to the appearance of either sinks or repellers. Then, if the limit set is hyperbolic and the no-cycle condition holds, from the fact that we are dealing with surface maps it follows that the limit set coincides with the nonwandering set.

5.3.2. *Sketch of the proof of Theorem 5.3.2. Dominated splitting versus homotheties* In the two-dimensional argument presented above we considered a linear cocycle defined over periodic points and its perturbations, without any information about the dynamics of the set of periodic points. These ideas for the two-dimensional case, when they are considered in any dimension, allow to get the following lemma that has a similar proof as in the two-dimensional case:

LEMMA 5.3.1. *Let $\mathcal{P}_k$ be the set of periodic hyperbolic saddles of index k and let $E^s_k(p) \oplus E^u_{n-k}(p)$ be the natural splitting induced by the stable and unstable directions of the periodic points $p \in \mathcal{P}_k$. If this splitting is not dominated, then by an arbitrarily small C^1-perturbation it is possible to change the index of some saddle $p \in \mathcal{P}$.*

Observe that if $2 \leqslant k \leqslant n-2$, the bifurcations mentioned in Lemma 5.3.1 do not break the transitivity and so they may occur in robustly transitive dynamics. In fact this holds in the examples considered. Moreover, C^1-robustly transitive nonhyperbolic diffeomorphisms generically exhibit periodic points of different indices. This follows from the fact that *diffeomorphism is hyperbolic if all periodic points of any C^1-perturbation are hyperbolic.*

Thus, in higher dimension, it is necessary to work out the dynamics generated by different periodic points acting in a homoclinic class. Using this argument it is shown that if there is no dominated splitting then it is possible to perturb the derivative of f along a periodic orbit in order to change the corresponding matrix into a homothety, and hence, after another perturbation, we get a sink or a source for a diffeomorphism close to f.

To perform this argument, first, a dense subset K in $H(p, f)$ of periodic points homoclinically related with p is taken, and we reduce the study to this set of periodic orbits. As any dominated splitting extends to the closure, the lack of dominated splitting on $H(p, f)$ implies also the lack of it for the set K. Then, given two (for simplicity fixed) hyperbolic saddles p and q, it can be assumed that they are homoclinically related (their invariant manifolds intersect transversally) and so there is a horseshoe containing p and q and some points of intersection of their invariant (stable and unstable) manifolds. Using this, it follows that there are periodic points passing first arbitrarily close to p and later to q, and so on. Moreover, it is possible to assume that the time spent in the complement of a neighborhood containing p and q is uniformly bounded. To be more precise, given any family $\{n_1, m_1, \ldots, n_k, m_k\}$ of arbitrarily large positive integers, there is a natural number r (independent of n_i and m_i) and a saddle z of $H(p, f)$ of period $t = kr + \Sigma_i(n_i + m_i)$ whose orbit spends alternately n_i consecutive iterates close to p and m_i consecutive iterates close to q. This means that the derivative of f in z is like

$$Df^t(z) = T_2 \circ Df^{m_k}(q) \circ T_1 \circ Df^{n_k}(p) \circ \cdots \circ T_2 \circ Df^{m_1}(q) \circ T_1 \circ Df^{n_1}(p),$$

where T_1 and T_2 are called the transitions from p to q and from q to p, respectively. These two transitions correspond to a bounded number of iterations and can be adapted by small perturbations in order that their contribution to the product vanishes as n_i and m_i go to infinity, thus the noise introduced by these transitions is negligible, provided that n_i and m_i are large enough. In other words, to multiply derivatives corresponding to different homoclinically related periodic points may make sense and the resulting linear maps (almost) correspond to the derivative along the orbit of another periodic saddle of the homoclinic class.

This kind of argument also shows that in any homoclinic class there is a dense subset of periodic points whose derivative at the period can be turned diagonalizable with positive eigenvalues of multiplicity one, by a small perturbation of the derivative along the orbit.

Now, if no dominated splitting exists at all, we will have the following situation: there are a linear map $A \in GL(R, n)$, linear maps $A_1, A_2, \ldots, A_{n-1}$ in $GL(R, n)$ and a basis $\mathcal{B} = \{v_1, \ldots, v_n\}$ of R^n such that A is diagonal in this basis and if L_i is the space generated by v_i, then A_j keeps the subspaces $L_1, \ldots, L_{j-1}, L_{j+2}, \ldots, L_n$ invariant and permutes the spaces L_j and L_{j+1}. In the previous situation, the matrices A and A_i correspond to the derivatives of f at periodic points p, $\{p_i\}$. The information that L_j and L_{j+1} are permuted follows from the fact that we are assuming that there is no dominated splitting: in

fact, if restricted to a two-dimensional subspace a dominated decomposition in two subbundles does not exist, using Claim 5.3.1, after perturbation we can get that the derivative restricted to this direction acts as a two-dimensional map with complex eigenvalue and so considering iterates of the same map we can assume that there are two invariant subspaces that are interchanged. On the other hand, since we can consider the transition, we can also consider all the possible products of the maps $A, A_1, A_2, \ldots, A_{n-1}$. A simple result of linear algebra shows that some of these products are homotheties. The idea is that, as the matrices A_i produce permutations of the eigenspaces of A which generate the group of all permutations of these eigenspaces, it is possible to mix the multipliers, obtaining a matrix having essentially the same rate of expansion in all directions.

The construction above shows that a robustly transitive set always admits a dominated splitting (so the dominated splitting is defined). To get that the extremal directions are volume expansive, the Ergodic Closing Lemma is applied. Observe that, since E_1 does not admit any dominated subsplitting, the arguments above guarantee that (after a perturbation) we can produce a homothety in E_1. Now, if E_1 is not uniformly volume contracting, by using the Ergodic Closing Lemma we get a homothety (in E_1) at a point which is volume expanding or (at least) whose rate of contraction of volume is close to one. So (after a new perturbation if necessary) we get a homothety by a factor greater than 1. Finally, the domination of the splitting implies that any bundle E_j is expanding, thus this saddle point is a source.

5.4. *Some results for conservative systems*

The same kind of question and dichotomies can be formulated for conservative maps. In [19], the following theorem was proved.

THEOREM 5.4.1. *For volume-preserving maps, C^1-generically the homoclinic classes either have dominated splitting or are accumulated by periodic points at which the derivative is the identity.*

This result is a generalization to higher dimension of a result due to Newhouse (see [62]) for conservative surface diffeomorphisms:

THEOREM 5.4.2. *For a compact two-dimensional manifold M there exists a residual subset $\mathcal{B}$ inside the C^1-conservative diffeomorphisms of M such that if $f \in \mathcal{B}$ then either f is an Anosov system or the elliptic periodic points of f are dense in M.*

In [4] a so-called Pasting Lemma was proved and it states that if f is a $C^{1+\alpha}$-volume-preserving map with a periodic point such that $Df^{n_p} = \mathrm{Id}$ (n_p is the period of p and Id is the identity map) then it is possible to find a C^1-volume-preserving map g C^1-close to f such that p is also a periodic point for g and there is a neighborhood U of p such that g^{n_p} is the identity map in U. As a corollary of this and the result in [19], the following theorem is proved in [4].

THEOREM 5.4.3. *Any robustly transitive volume-preserving C^1-diffeomorphism exhibits a dominated splitting such that its extremal bundles are uniformly volume contracted or expanded.*

It is worth to mention that as a consequence of a closing lemma for chain recurrent points proved in [16], the same paper establishes that:

THEOREM 5.4.4. *C^1-generically the volume-preserving diffeomorphisms are transitive.*

A chain recurrent point is a point such that for every $\epsilon > 0$ there exists a sequence of points $x = x_0, x_1, \ldots, x_n = x_0$ with $\operatorname{dist}(f(x_i) - x_{i+1}) < \epsilon$. The closing lemma for chain recurrent points states that C^1-generically the chain recurrent set coincides with the closure of the set of periodic points.

Also about volume-preserving diffeomorphisms similar dichotomies as the one in Theorem 5.3.2 can be stated. Recalling that for an invariant measure we can obtain the Oseledets splitting and it is possible to calculate the Lyapunov exponents, the following theorem characterizes continuity of Lyapunov exponents for volume-preserving C^1-diffeomorphisms. In fact, in [14] it was proved that:

THEOREM 5.4.5. *Let μ be the normalized Lebesgue measure on a compact manifold M. If $f \in \operatorname{Diff}^1(M)$ is a continuity point for the map*

$$\operatorname{Diff}^1_\mu(M) \to R^d, \quad g \to \big(\lambda_1(g,\mu), \ldots, \lambda_d(g;\mu)\big)$$

then, for almost every point $x \in M$,

1. *either all Lyapunov exponents $\lambda_i(f;x) = 0$ for $1 \leqslant i \leqslant d$,*
2. *or the Oseledets splitting of f is dominated on the orbit of x.*

Many of these results can be obtained also if one replaces robust transitivity by stable ergodicity. In [13], it is proved that C^1-stably ergodic maps in the conservative category are Bernoulli maps exhibiting dominated splitting.

More precisely, let M be a compact manifold of dimension $d \geqslant 2$, and let μ be a volume measure in M. Take $\alpha > 0$ and let $\operatorname{Diff}^{1+\alpha}_\mu(M)$ be the set of μ-preserving $C^{1+\alpha}$-diffeomorphisms, endowed with the C^1-topology. Let $\mathcal{SE} \subset \operatorname{Diff}^{1+\alpha}_\mu(M)$ be the set of stably ergodic diffeomorphisms (i.e. the set of diffeomorphisms such that every sufficiently C^1-close $C^{1+\alpha}$-conservative diffeomorphism is ergodic).

THEOREM 5.4.6 [13]. *There is an open and dense set $\mathcal{R} \subset \mathcal{SE}$ such that if $f \in \mathcal{R}$ then f is nonuniformly hyperbolic, that is, all Lyapunov exponents of f are nonzero. Moreover, every $f \in \mathcal{R}$ admits a dominated splitting.*

REMARK 5.4.1. It is not true that every stably ergodic diffeomorphism can be approximated by a partially hyperbolic system (in the weaker sense), by the examples in [23] and proved in [93].

The proof of Theorem 5.4.6 goes through three steps:

Step 1. A stably ergodic (or stably transitive) diffeomorphism f must have a dominated splitting. This is true because if it does not, [19] permits to perturb f and create a periodic point whose derivative is the identity. Then, using the Pasting Lemma proved in [4] (for which $C^{1+\alpha}$-regularity is an essential hypothesis), one breaks transitivity.

Step 2. A result in [15] (which is a refinement of a technique developed in [91]) gives a perturbation of f such that the sum of the Lyapunov exponents "inside" each of the bundles of the (finest) dominated splitting is nonzero.

Step 3. Using a result proved in [14], we find another perturbation such that the Lyapunov exponents in each of the bundles become almost equal. (If we attempted to make the exponents exactly equal, we could not guarantee that the perturbation is $C^{1+\alpha}$.) Since the sum of the exponents in each bundle varies continuously, we conclude that there are no zero exponents.

5.5. *Robust transitivity and hetero-dimensional cycles*

In dimension higher than two, another kind of homoclinic bifurcation breaks the hyperbolicity: the so-called hetero-dimensional cycles (intersection of the stable and unstable manifolds of points of different indices, see [30] and [31]). More precisely:

DEFINITION 5.5.1. Let M be and let $f \in \mathrm{Diff}^r(M)$ ($r \geqslant 1$). If f has a pair of hyperbolic saddles P and Q with different indices, that is, different dimensions of their unstable subspaces and if $W^s(P)$ and $W^u(Q)$ have nonempty intersection, and the same for $W^u(P)$ and $W^s(Q)$ then we say that f has a *hetero-dimensional cycle* associated to P and Q.

It follows immediately from the definition that hetero-dimensional cycles can only exist in dimension at least 3.

In particular, the unfolding of these cycles implies the existence of striking dynamics such as the appearance of nonhyperbolic robustly transitive sets and the explosion of homoclinic classes (see Theorem 5.5.1 below). Moreover, it is possible to prove the following:

Let $f \in \mathrm{Diff}^1(M)$ and $\Lambda_f(U) = \bigcap_{n\in\mathbb{Z}} g^n(U)$ be maximal invariant nonhyperbolic C^1-robustly transitive sets. Then, there is g C^1-close to f such that g exhibits a hetero-dimensional cycle contained in $\Lambda_g(U) = \bigcap_{n\in\mathbb{Z}} g^n(U)$.

In the same sense, these cycles play the role for the partial hyperbolic theory as transverse intersection play for the hyperbolic theory. In the paper [20] the interplay between hetero-dimensional cycles and robustly transitive systems in any dimension is studied.

To show the richness of hetero-dimensional cycles we present one theorem related to them. For simplicity, we suppose that the periodic points P and Q in the cycle are actually fixed points. In addition, we always assume:

(a) (codimension 1) the saddles P and Q have indices p and $q = p + 1$, respectively,
(b) (quasi-transversality) the manifolds $W^s(P)$ and $W^u(Q)$ intersect transversely, and the intersection between $W^u(P)$ and $W^s(Q)$ is quasi-transverse.

The next theorem asserts that the homoclinic classes of P and Q often explode, and become intermingled (nonempty intersection) when the cycle is unfolded.

We say that a parameterized family $\{f_t\}_{t\in[-1,1]}$ of diffeomorphisms unfolds generically a hetero-dimensional cycle of $f = f_0$ if there are open disks $K_t^u \in W^u(P_t)$ and $K_t^s \in W^s(Q_t)$, depending continuously on t such that $K_0^u \cap K_0^s$ contains a point of quasi-transverse intersection, and the distance between K_t^s and K_t^u increases with positive velocity when t increases. Here P_t and Q_t denote the continuations for f_t of the periodic points P and Q.

THEOREM 5.5.1 [30]. *There is a nonempty open set of C^1-parameterized families of diffeomorphisms $\{f_t\}_{t\in[-1,1]}$ unfolding generically a heteroclinical cycle of $f = f_0$, such that for all small positive t,*

1. *The transverse intersection between $W^s(P_t)$ and $W^u(Q_t)$ is contained in the homoclinic class of Q_t;*
2. *The homoclinic class of P_t is contained in the homoclinic class of Q_t.*

Let d be the dimension of the ambient manifold M. The key fact behind the previous theorem is that, for every small positive t, the $(d-p)$-dimensional manifold $W^s(P_t)$ is contained in the closure of $W^s(Q_t)$. This makes the stable manifold of Q_t, which has dimension $d-p-1$, behave like a manifold of dimension one unit greater. The proof of the theorem relies on a reduction of the dynamics to a family of iterated function systems on the interval.

6. Flows and singular splitting

For flows, a striking example is in [46], given by the solutions of the polynomial vector field in R^3:

$$X(x,y,z) = \begin{cases} \dot{x} = -\alpha x + \alpha y, \\ \dot{y} = \beta x - y - xz \\ \dot{z} = -\gamma z + xy, \end{cases} \tag{5}$$

where α, β, γ are real parameters. This equation was derived by Lorenz from the works done by Saltzman [89] concerning thermal fluid convection. Numerical experiments performed by Lorenz (for $\alpha = 10$, $\beta = 28$ and $\gamma = 8/3$) suggested the existence, in a robust way, of a transitive strange attractor toward which tends a full neighborhood of positive trajectories of the above system. That is, the strange attractor could not be destroyed by any perturbation of the parameters. Most important, the attractor contains an equilibrium point $(0,0,0)$, and hence cannot be hyperbolic. The work of Lorenz raised a number of mathematical questions that are among the leitmotifs in the development of the theory of dynamical systems.

Notably, only now, three and a half decades after this remarkable work, it was proved in [94] that the solutions of (5) satisfy such a property for values α, β, γ near the ones considered by Lorenz.

However, already in the mid-seventies, the existence of robust nonhyperbolic attractors was proved for flows introduced independently in [2,36] (see also [37,97]) which we now call geometric models for Lorenz attractors. In particular, they exhibit, in a robust way, an attracting transitive set with an equilibrium (singularity). Moreover, the properties of these geometric models allow one to extract very complete dynamical information.

A natural question arises: *are such features present for any robustly transitive set of a flow*?

In [54] a positive answer for this question is given:

THEOREM 6.0.1. *C^1-robustly transitive sets with singularities on closed* 3*-manifolds have the following properties*:

1. *There are either proper attractors or proper repellers*;
2. *The eigenvalues at the singularities satisfy the same inequalities as the corresponding ones at the singularity in a Lorenz geometrical model*;
3. *There are partially hyperbolic sets with a volume-expanding central direction*.

The presence of a singularity prevents these attractors from being hyperbolic. But they exhibit a weaker form of hyperbolicity named *singular hyperbolic splitting*. This class of vector fields contains the Axiom A systems, the geometric Lorenz attractors and the singular horseshoes in [44], among other systems. Currently, there is a rather satisfactory and complete description of singular hyperbolic vector fields defined on three-dimensional manifolds (but the panorama in higher dimensions remains open). The first consequence of this result is that every orbit in *any* robust attractor has a direction of exponential divergence from nearby orbits (positive Lyapunov exponent). Another consequence is that robust attractors always admit an invariant foliation whose leaves are forward contracted by the flow, showing that any robust attractor with singularities displays similar properties to those of the geometrical Lorenz model. In particular, in view of the result of Tucker [94], the Lorenz attractor generated by the Lorenz equations much resembles a geometrical one.

Moreover, related to the partial hyperbolic structure for three-dimensional flows, it is proved in a sequel of works that these sets are expansive, the periodic orbits are dense (in the case that the set is an attractor), and it has a spectral decomposition (see [65,9,5]).

On the other hand, for the case of flows, a new kind of bifurcation appears that leads to a new dynamics distinct from the ones for diffeomorphism: the so-called *singular cycles* (cycles involving singularities and periodic orbits, see [8,51,55,53] for examples of dynamics in the sequel of the unfolding of it). In particular, these cycles lead to the creation of robust singular hyperbolic sets as it is shown in the papers [55,53]. Systems exhibiting these cycles are dense among open set of systems exhibiting a singular hyperbolic splitting.

Moreover, recently A. Arroyo and F. Rodriguez Hertz (see [6]), studying the dynamical consequences of the dominated splitting for the Linear Poincaré flow, proved the following:

THEOREM 6.0.2. *Any three-dimensional flow can be C^1-approximated by a flow exhibiting a homoclinic tangency either by a singular cycle or by a hyperbolic one.*

6.1. *Sketch of the proof of Theorem 6.0.1*

In this section we discuss Theorem 6.0.1 stated before, in other words, we will explain the dynamical structure of compact transitive sets (there are dense orbits) of flows on 3-manifolds which are *robust* under small C^1-perturbations.

To state our results in a precise way, let us fix some notations and recall some definitions and results proved elsewhere.

Throughout, M is a boundaryless compact manifold and $\mathcal{X}^r(M)$ denotes the space of C^r-vector fields on M endowed with the C^r-topology, $r \geqslant 1$. Let $X \in \mathcal{X}^r(M)$, and let X_t, $t \in R$, denote the flow induced by X. A compact invariant set Λ of X is *isolated* if there exists an open set $U \supset \Lambda$, called *isolating block*, such that $\Lambda = \bigcap_{t\in R} X_t(U)$. If U above can be chosen such that $X_t(U) \subset U$ for $t > 0$, we say that the isolated set Λ is an *attracting set*.

A compact invariant set Λ of X is *transitive* if it coincides with the ω-limit set of an X-orbit. An *attractor* is a transitive attracting set. A *repeller* is an attractor for the reversed vector field $-X$. An attractor (or repeller) which is not the whole manifold is called *proper*. An invariant set of X is *nontrivial* if it is neither a periodic point nor a singularity. With this concept we define:

An isolated set Λ of a C^1-vector field X is robustly transitive if it has an isolating block U such that

$$\Lambda_Y(U) = \bigcap_{t\in R} Y_t(U)$$

is both transitive and nontrivial for any Y C^1-close to X.

Related to robustly transitive sets the following theorem is proved.

THEOREM 6.1.1. *A robustly transitive set containing singularities of a flow on a closed 3-manifold is either a proper attractor or a proper repeller.*

As a matter of fact, the previous result will follow from a general result on n-manifolds, $n \geqslant 3$, settling sufficient conditions for an isolated set to be an attracting set:

(a) all its periodic points and singularities are hyperbolic and

(b) it robustly contains the unstable manifold of either a periodic point or a singularity.

The previous result is false in dimension greater than three; a counterexample can be obtained by multiplying the geometric Lorenz attractor by a hyperbolic system in such a way that the directions supporting the Lorenz flow be normally hyperbolic. It is false as well in the context of boundary-preserving vector fields on 3-manifolds with boundary [44]. The converse to this theorem is also not true: proper attractors (or repellers) with singularities are not necessarily robustly transitive, even if their periodic points and singularities are hyperbolic in a robust way.

To motivate the next results about singularities of robustly transitive sets for flows displaying singularities, recall that the geometric Lorenz attractor L is a proper robustly transitive set with a hyperbolic singularity σ such that if λ_i, $1 \leqslant i \leqslant 3$, are the eigenvalues of L

at σ, then λ_i, $1 \leqslant i \leqslant 3$, are real and satisfy $\lambda_2 < \lambda_3 < 0 < -\lambda_3 < \lambda_1$ [37]. Inspired by this property we say that a singularity σ is Lorenz-like for X if the eigenvalues λ_i, $1 \leqslant i \leqslant 3$, of $DX(\sigma)$ are real and satisfy $\lambda_2 < \lambda_3 < 0 < -\lambda_3 < \lambda_1$.

If σ is a Lorenz-like singularity for X then the strong stable manifold $W_X^{ss}(\sigma)$ exists. Moreover, $\dim(W_X^{ss}(\sigma)) = 1$, and $W_X^{ss}(\sigma)$ is tangent to the eigenvector direction associated to λ_2. Given a vector field $X \in \mathcal{X}^r(M)$, we let $\Sigma(X)$ be the set of singularities of X. If Λ is a compact invariant set of X we let $\Sigma_X(\Lambda)$ be the set of singularities of X in Λ.

The next result shows that the singularities of robustly transitive sets on closed 3-manifolds are Lorenz-like.

THEOREM 6.1.2. *Let Λ be a robustly transitive set of $X \in \mathcal{X}^1(M)$ with a singularity. Then, either for $Y = X$ or $Y = -X$, every $\sigma \in \Sigma_Y(\Lambda)$ is Lorenz-like for Y and satisfies $W_Y^{ss}(\sigma) \cap \Lambda = \{\sigma\}$.*

In light of these results, a natural question arises: can one achieve a general description of the structure of robust attractors for flows exhibiting singularities? In this direction we prove: if Λ is a robust attractor for X containing singularities then it is partially hyperbolic with volume-expanding central direction. To state this result in a precise way, let us introduce some definitions and notations.

Let Λ be a compact invariant transitive set of $X \in \mathcal{X}^r(M)$, $c > 0$, and $0 < \lambda < 1$. We say that Λ has a (c, λ)-dominated splitting if the bundle over Λ can be written as a continuous DX_t-invariant sum of subbundles $T_\Lambda = E^s \oplus E^{cu}$, such that if $T > 0$ and $x \in \Lambda$ then

(a) E^s is one dimensional;

(b) The bundle E^{cu} contains the direction of X, and

$$\left\| DX_T / E_x^s \right\| \cdot \left\| DX_{-T} / E_{X_T(x)}^{cu} \right\| < c\,\lambda^T.$$

E^{cu} is called the central direction of T_Λ.

A compact invariant transitive set Λ of X is *partially hyperbolic* if Λ has a (c, λ)-dominated splitting $T_\Lambda M = E^s \oplus E^{cu}$ such that the bundle E^s is uniformly contracting, that is, for every $T > 0$, and every $x \in \Lambda$, we have $\|DX_T / E_x^s\| < c\,\lambda^T$.

For $x \in \Lambda$ and $t \in \mathbb{R}$ we let $J_t^c(x)$ be the absolute value of the determinant of the linear map $DX_t / E_x^{cu} : E_x^{cu} \to E_{X_t(x)}^{cu}$. We say that the subbundle E_Λ^{cu} of the partially hyperbolic set Λ is *volume-expanding* if $J_t^c(x) \geqslant ce^{\lambda t}$, for every $x \in \Lambda$ and $t \geqslant 0$ (in this case we say that E_Λ^{cu} is (c, λ)-*volume-expanding* to indicate the dependence on c, λ). With all these definitions in mind, we introduce the following

DEFINITION 6.1.1 (*Singular hyperbolicity*). Let Λ be a compact invariant transitive set of $X \in \mathcal{X}^r(M)$ with singularities. We say that Λ is a *singular hyperbolic set* for X if all the singularities of Λ are hyperbolic, and Λ is partially hyperbolic with volume expanding central direction.

This is the property that describes robust attractors of flows with singularities.

THEOREM 6.1.3. *Robust attractors of $X \in \mathcal{X}^1(M)$ containing singularities are singular hyperbolic sets for X.*

Now we will summarize the sketch of the proof of the previous statement.

In view of the previous results one can assume that every singularity of X is Lorenz-like and has (for instance) two negative eigenvalues. Consider the splitting with three directions defined over the set of regular closed orbits of the vector field, denoted by $E^{ss} \oplus [X] \oplus E^{uu}$, where E^{ss} and E^{uu} are the stable and unstable directions and X is the flow direction. For simplicity let us assume that the set Λ is the closure of the periodic orbits of X. The objective is to extend this splitting to the closure of the regular closed orbits to get a dominated splitting. In general, this is not possible due to the presence of the equilibria; in fact, for periodic orbits getting close to the singularities, the angle between the flow direction and the unstable direction is getting small. Thus it is not possible to extend this splitting in a dominated way to the whole Λ. The idea now is to consider the splitting $E^{ss} \oplus E^c$, where $E^c = [X] \oplus E^{uu}$, defined over the periodic orbits and extend it to its closure (that is the whole set Λ). For that it is enough to prove that such a splitting is dominated. To obtain the domination one proves that given any small neighborhood V of the finite set of singularities of Λ and any point p in a regular closed orbit that does not meet this vicinity, the angle between $E^{ss}(p)$ and $E^c(p)$ is uniformly bounded from below. Otherwise, after a perturbation a repeller or a sink can be created, contradicting the robust transitivity. On the other hand, the analysis near the singularity is the following. Since all the singularities have one-dimensional unstable manifolds and are Lorenz-like, there is a splitting $E^{ss} \oplus E^s \oplus E^{uu}$. Then it is proved that $E^{ss}(p) \to E^{ss}(\sigma)$ and $E^c(p) \to E^s(\sigma) \oplus E^{uu}(\sigma)$ as p approaches the singularity σ. If it is not the case, it is proved after two turns nearby the singularities, by a small perturbation either a sink or a repeller is created, contradicting again the robust transitivity. From the two previous facts it follows that the angles between E^{ss} and E^c are persistently bounded away from zero and as in the case of diffeomorphisms, this allows to prove the domination of the splitting $E^{ss} \oplus E^c$.

6.2. *Dynamical consequences of singular hyperbolicity*

We will list some properties that give us a nice description of the dynamics of robustly transitive sets with singularities, and in particular, for robust attractors.

The first two properties do not depend either on the fact that the set is robustly transitive or an attractor.

LEMMA 6.2.1. *Let Λ be a singular hyperbolic set of $X \in \mathcal{X}^1(M)$. Then any invariant compact set $\Gamma \subset \Lambda$ without singularities is a hyperbolic set.*

Recall that, given $x \in M$, and $v \in T_x M$, the Lyapunov exponent of x in the direction of v is

$$\gamma(x, v) = \liminf_{t \to \infty} \frac{1}{t} \log \| DX_t(x)v \|.$$

We say that x has positive Lyapunov exponent if there is $v \in T_x M$ such $\gamma(x, v) > 0$.

The next two results show that important features of hyperbolic attractors and of the geometric Lorenz attractor are present for singular hyperbolic attractors, and so, for robust attractors with singularities:

LEMMA 6.2.2. *A singular hyperbolic attractor Λ of $X \in \mathcal{X}^1(M)$ has uniform positive Lyapunov exponent at every orbit.*

Currently, there is a rather satisfactory and complete description of singular hyperbolic attractors on three-dimensional manifolds (again, the panorama in higher dimensions remains open). In [5] it is proved that for transitive singular hyperbolic attractors the set of periodic orbits is dense; moreover, it is made of a unique homoclinic class. For that, a theorem of existence of unstable manifolds is obtained, not for the whole set but for a subset which is visited infinitely many times by a residual subset of the attractor. Such theorem has an important consequence that there is a characterization of C^1-robustly transitive attractors with singularities.

More precisely, even in the presence of equilibrium points in a singular hyperbolic attractor there are local unstable manifolds of uniform size, not for the whole attractor but for a subset of it which is visited infinitely many times by points in a residual subset of it. As a first consequence, the following theorem about the set of periodic orbits is obtained. Denote by $\mathrm{Per}(\Lambda)$ the set of periodic orbits of Φ_t in Λ.

THEOREM 6.2.1. *Let Λ be a transitive singular hyperbolic attractor for the flow Φ_t. Then* $\mathrm{cl}(\mathrm{Per}(\Lambda)) = \Lambda$.

The same techniques give the following

THEOREM 6.2.2. *Let Λ be a transitive singular hyperbolic attractor for the flow Φ_t. Then there is a periodic orbit p whose homoclinic class* (*see Theorem* 4.3.1) *is dense in Λ.*

However, there are examples of singular hyperbolic sets which do not have any periodic orbits at all (see [52]). Such sets are not attractors and nor are they robust.

Theorem 6.2.1 is useful also for obtaining statistical properties (recall the definition of SRB measure given in Volume 1A of this handbook). In fact, the following theorem was proved in [28].

THEOREM 6.2.3. *If Λ is a singular hyperbolic transitive attractor of a $C^{1+\alpha}$-vector field, $\alpha > 0$, with dense periodic orbits, then it has an SRB measure.*

Therefore, as a consequence of this result and Theorem 6.2.1, it follows that:

COROLLARY 6.2.1. *If Λ is a singular hyperbolic transitive attractor of a $C^{1+\alpha}$-vector field, $\alpha > 0$, it has an SRB measure.*

A FEW WORDS ABOUT THE PROOF OF THEOREM 6.2.1. Observe that for a singular hyperbolic splitting, the integrability of the strong stable bundle allows us to find stable manifolds of uniform size $\varepsilon_s > 0$ on any point of U; that is, for any $x \in U$ there is a C^1-interval, say $W^s_{\varepsilon_s}(x)$, such that for any point y on it we have that $d(\Phi_t(x), \Phi_t(y)) \to 0$ when $t \to \infty$. In the case of a singularity it corresponds to the stable manifold associated to the strongest contracting eigenvalue $W^{ss}_{\mathrm{loc}}(\sigma)$.

For unstable manifolds such a construction may be impossible on regular points. However, associated to any regular point $x \in \Lambda$ there is a family of two-dimensional sections N_{x_t} whose size depends on the point x_t and which is transversal to the flow; here $x_t := \Phi_t(x)$. According to this section we can write the family of holonomy maps between these transversal sections, for any $t \in \mathbb{R}$,

$$G_x^t : \mathrm{Dom}\left(G_x^t\right) \subset N_x \longrightarrow N_{x_t}.$$

Consider some $\varepsilon > 0$. The unstable manifold of size ε of a regular point $x \in \Lambda$ is

$$\tilde{W}_\varepsilon^u(x) = \left\{ y \in M \mid y \in \mathrm{Dom}\left(G_x^{-t}\right) \text{ and } \mathrm{dist}\left(G_x^{-t}(y), x_{-t}\right) \to 0,\ t \to \infty \right\}.$$

Of course, $\tilde{W}_\varepsilon^u(x) \subset N_x$. On the other hand, there is some $\tilde{\varepsilon} > 0$ for which the central unstable manifold exists on any point x of a singular hyperbolic set Λ; according to [38]. Denote these central manifolds by $W_{\tilde{\varepsilon}}^{cu}(x)$. Also denote by $L(Y)$ the limit set of Y.

Now we can state the theorem about suitable unstable manifolds for a singular hyperbolic attractor, which was the main goal of this chapter.

THEOREM 6.2.4. *Let $X \in \mathcal{X}^1(M)$, $\Lambda \subset M$ be a transitive singular hyperbolic attractor and $U \supset \Lambda$ an open neighborhood contained in its basin of attraction. Then there is a neighborhood $\mathcal{U}(X) \subset \mathcal{X}^1(M)$ such that for all $Y \in \mathcal{U}(X)$ there is a subset $K(Y) \subset \Lambda_Y := \bigcap_{t \geqslant 0} Y_t(U)$, $\varepsilon_u > 0$ and $\lambda_u < 0$ that*

1. *For any $y \in K(Y)$, we have that $W_{\varepsilon_u}^{cu}(y) \cap N_y = \tilde{W}_{\varepsilon_u}^u(y)$;*
2. *For any $y \in \hat{W}_{\varepsilon_u}^u(x)$ there is an unbounded sequence $t_i > 0$ such that*

$$\mathrm{dist}\left(G_x^{-t_i}(y), x_{-t_i}\right) < C \exp(t_i \lambda_u);$$

3. *$\bigcup_{t > T_0} \bigcup_{y \in K} \Phi_{-t}(y)$ is an open and dense set in $\Lambda_Y \cap L(Y)$, for any $T_0 > 0$.*

In order to prove this theorem, first a system of transversal sections associated to the passage through a neighborhood of each equilibrium point is found. Among the systems of transversal sections, some of them present an induced map with a kind of Markovian property and uniform expansion. However, to guarantee the existence of such transversal sections, a deep analysis of the dynamics inside a neighborhood of the singularities and also the combinatorics between them are needed. More precisely, it is necessary to study the local holonomy maps between transversal sections in general.

References

[1] R. Abraham and S. Smale, *Nongenericity of Axiom A and Ω-stability*, Global Analysis, Proc. Sympos. in Pure Math., Vol. 14, Amer. Math. Soc., Providence, RI (1970), 5–8.

[2] V. Afraimovich, V. Bykov and L. Shil'nikov, *On the appearance and structure of the Lorenz attractor*, Dokl. Acad. Sci. USSR **234** (1977), 336–339.

[3] J. Alves, C. Bonatti and M. Viana, *SRB measures for partially hyperbolic systems whose central direction is mostly expanding*, Invent. Math. **140** (2000), 351–398.

[4] A. Arbieto and C. Matheus, *A pasting lemma I: The case of vector fields*, Preprint IMPA (2003), http://www.preprint.impa.br/Shadows/SERIE_A/2003/244.html.
[5] A. Arroyo and E. Pujals, *Dynamical properties of singular hyperbolic attractors*, Preprint IMPA (2004), http://www.preprint.impa.br/Shadows/SERIE_A/2004/292.html.
[6] A. Arroyo and F. Rodriguez Hertz, *Homoclinic bifurcations and uniform hyperbolicity for three-dimensional flows*, Ann. Inst. H. Poincaré Anal. Non Linéaire **20** (5) (2003), 805–841.
[7] V. Baladi, E.R. Pujals and M. Sambarino, *Dynamical zeta functions for analytic surface diffeomorphisms with dominated splitting*, J. Inst. Math. Jussieu **4** (2005), 175–218.
[8] R. Bamon, R. Labarca, R. Mañé and M. Pacifico, *The explosion of singular cycles*, Publ. Math. IHES **78** (1993), 207–232.
[9] S. Bautista and C. Morales, *Existence of periodic orbits for singular-hyperbolic attractors*, Preprint IMPA (2004), http://www.preprint.impa.br/Shadows/SERIE_A/2004/288.html.
[10] S. Bautista, C.A. Morales and M.J. Pacífico, *There is no spectral decomposition for singular-hyperbolic flows*, Preprint IMPA (2004), http://www.preprint.impa.br/Shadows/SERIE_A/2004/278.html.
[11] M. Benedicks and L. Carleson, *The dynamics of the Hénon map*, Ann. of Math. **133** (1991), 73–169.
[12] J. Bochi, *Genericity of zero Lyapunov exponents*, Ergodic Theory Dynam. Systems **22** (2002), 1667–1696.
[13] J. Bochi, B. Fayad and E.R. Pujals, *Dichotomy for conservative robust ergodic maps*, http://arxiv.org/PS_cache/math/pdf/0408/0408344.pdf.
[14] J. Bochi and M. Viana, *The Lyapunov exponents of generic volume preserving and symplectic maps*, Ann. of Math., to appear.
[15] C. Bonatti and A.T. Baraviera, *Removing zero Lyapunov exponents*, Ergodic Theory Dynam. Systems **23** (2003), 1655–1670.
[16] C. Bonatti and S. Crovisier, *Recurrence and genericity*, C. R. Math. Acad. Sci. **336** (10) (2003), 839–844.
[17] C. Bonatti and L.J. Diaz, *Persistence of transitive diffeomorphisms*, Ann. of Math. **143** (1995), 367–396.
[18] C. Bonatti and L. Díaz, *Connexions hétéroclines et généricité d'une infinité de puits et de sources*, Ann. Sci. École Norm. Sup. (4) **32** (1) (1999), 135–150.
[19] C. Bonatti, L.J. Diaz and E.R. Pujals, *A C^1-generic dichotomy for diffeomorphisms: weak form of hyperbolicity or infinitely many sinks or sources*, Ann. of Math. **158** (2003), 355–418.
[20] C. Bonatti, L.J. Diaz, E.R. Pujals and J. Rocha, *Heterodimensional cycles and robust transitivity*, Geometric Methods in Dynamics. I, Astérisque **286** (2003), xix, 187–222.
[21] C. Bonatti, L.J. Díaz and R. Ures, *Minimality of strong stable and unstable foliations for partially hyperbolic diffeomorphisms*, J. Inst. Math. Jussieu **1** (4) (2002), 513–541.
[22] C. Bonatti and R. Langevin, *Un exemple de flot d'Anosov transitif transverse à un tore et non conjugué à une suspension*, Ergodic Theory Dynam. Systems **14** (4) (1994), 633–643.
[23] C. Bonatti and M. Viana, *SRB measures for partially hyperbolic systems whose central direction is mostly contracting*, Israel J. Math. **115** (2000), 157–193.
[24] M.I. Brin and Ja.B. Pesin, *Partially hyperbolic dynamical systems*, Izv. Akad. Nauk SSSR Ser. Mat. **38** (1974), 170–212 (in Russian).
[25] K. Burns, D. Dolgopyat and Ya. Pesin, *Partial hyperbolicity, Lyapunov exponents, and stable ergodicity*, J. Statist. Phys. **108** (2002), 927–942.
[26] A. Castro, *Fast mixing for attractors with a mostly contracting central direction*, Ergodic Theory Dynam. Systems **24** (2004), 17–44.
[27] E. Colli, *Infinitely many coexisting strange attractors*, Ann. Inst. H. Poincaré Anal. Non Linéaire **15** (1998), 539–579.
[28] W. Colmenares, Ph.D. thesis, UFRJ (Universidade Federal do Rio de Janeiro) (2002).
[29] W.J. Cowieson and L.-S. Young, *SRB measures as zero noise limits*, http://www.cims.nyu.edu/∼lsy/papers/Zero.pdf.
[30] L.J. Diaz, *Robust nonhyperbolic dynamics at hetero-dimensional cycles*, Ergodic Theory Dynam. Systems **15** (1995), 291–315.
[31] L.J. Diaz, *Persistence of cycles and nonhyperbolic dynamics at heteroclinic bifurcations*, Nonlinearity **8** (1995), 693–715.
[32] L.J. Diaz, E.R. Pujals and R. Ures, *Partial hyperbolicity and robust transitivity*, Acta Math. **183** (1999), 1–43.

[33] C.I. Doering, *Persistently transitive vector fields on three-dimensional manifolds*, Dynamical Systems and Bifurcation Theory, Rio de Janeiro (1985), 59–89.
[34] D. Dolgopyat and Ya. Pesin, *Every compact manifold carries a completely hyperbolic diffeomorphism*, Ergodic Theory Dynam. Systems **22** (2002), 409–435.
[35] J. Frank, *Necessary conditions for the stability of diffeomorphisms*, Trans. Amer. Math. Soc. **158** (1971), 295–329.
[36] J. Guckenheimer, *A strange, strange attractor*, The Hopf Bifurcation Theorem and Its Applications, J.E. Marsden and M. McCracken, eds, Springer-Verlag (1976), 368–381.
[37] J. Guckenheimer and R. Williams, *Structural stability of Lorenz attractors*, Publ. Math. IHES **50** (1979), 50–59.
[38] M. Hirsch, C. Pugh and M. Shub, *Invariant Manifolds*, Springer Lecture Notes in Math., Vol. 583 (1977).
[39] H. Hu and L.-S. Young, *Nonexistence of SBR measures for some diffeomorphisms that are almost Anosov*, Ergodic Theory Dynam. Systems **15** (1995), 67–76.
[40] V. Kaloshin, *An extension of the Artin–Mazur theorem*, Ann. of Math. (2) **150** (1999), 729–741.
[41] V. Kaloshin and E.R. Pujals, *Superexponential growths of periodic orbits in heteroclinic cycles*, in preparation.
[42] A. Katok, *Lyapunov exponents, entropy and periodic orbits for diffeomorphisms*, Inst. Hautes Études Sci. Publ. Math. **51** (1980), 137–173.
[43] M. Komuro, *Expansive properties of Lorenz attractors*, The Theory of Dynamical Systems and Its Applications to Nonlinear Problems (Kyoto, 1984), World Scientific, Singapore (1984), 4–26.
[44] R. Labarca and M.J. Pacifico, *Stability of singular horseshoes*, Topology **25** (1986), 337–352.
[45] J. Lewowicz, *Persistence in expansive systems*, Ergodic Theory Dynam. Systems **3** (4) (1983), 567–578.
[46] E.N. Lorenz, *Deterministic nonperiodic flow*, J. Atmospheric Sci. **20** (1963), 130–141.
[47] R. Mañé, *Contributions to the stability conjecture*, Topology **17** (1978), 386–396.
[48] R. Mañé, *An ergodic closing lemma*, Ann. of Math. **116** (1982), 503–540.
[49] R. Mañé, *Hyperbolicity, sinks and measure in one-dimensional dynamics*, Comm. Math. Phys. **100** (1985), 495–524.
[50] L. Mora and M. Viana, *Abundance of strange attractors*, Acta Math. **171** (1993), 1–71.
[51] C. Morales, *Lorenz attractors through saddle-node bifurcations*, Ann. Inst. H. Poincaré Anal. Non Linéaire **13** (1996), 589–617.
[52] C. Morales, *A note on periodic orbits for singular-hyperbolic flows*, Discrete Contin. Dynam. Systems **11** (2–3) (2004), 615–619.
[53] C. Morales, M.J. Pacifico and E.R. Pujals, *Strange attractors across the boundary of hyperbolic systems*, Comm. Math. Phys. **211** (2000), 527–558.
[54] C. Morales, M.J. Pacifico and E.R. Pujals, *Robust transitive singular sets for* 3*-flows are partially hyperbolic attractors or repellers*, Ann. of Math. **160** (2) (2004), 375–432.
[55] C. Morales and E.R. Pujals, *Singular strange attractors across the boundary of hyperbolic systems*, Comm. Math. Phys. **211** (1997), 527–558.
[56] C.A. Morales and M.J. Pacifico, *Attractors and singularities robustly accumulated by periodic orbits*, International Conference on Differential Equations, Vols. 1, 2 (Berlin, 1999), World Scientific, River Edge, NJ (2000), 64–67.
[57] C.G. Moreira, J. Palis and M. Viana, *Homoclinic tangencies and fractal invariants in arbitrary dimension*, C. R. Acad. Sci. Paris Sér. I Math. **333** (5) (2001), 475–480.
[58] C.G. Moreira and J.-C. Yoccoz, *Stable intersections of regular Cantor sets with large Hausdorff dimensions*, Ann. of Math. (2) **154** (1) (2001), 45–96.
[59] S. Newhouse, *Non-density of Axiom A(a) on* S^2, Proc. Amer. Math. Soc. Sympos. Pure Math. **14** (1970), 191–202.
[60] S. Newhouse, *Hyperbolic limit sets*, Trans. Amer. Math. Soc. **167** (1972), 125–150.
[61] S. Newhouse, *Diffeomorphism with infinitely many sinks*, Topology **13** (1974), 9–18.
[62] S. Newhouse, *Quasi-elliptic periodic points in conservative dynamical systems*, Amer. J. Math. **99** (5) (1977), 1061–1087.
[63] S. Newhouse, *The abundance of wild hyperbolic sets and nonsmooth stable sets for diffeomorphisms*, Publ. Math. IHES **50** (1979), 101–151.

[64] J.C. Oxtoby and S.M. Ulam, *Measure-preserving homeomorphisms and metrical transitivity*, Ann. of Math. **42** (1941), 874–920.
[65] M.J. Pacifico, E.R. Pujals and M. Viana, *Singular hyperbolic systems are sensitive to initial data*, Preprint.
[66] J. Palis, *Homoclinic orbits, hyperbolic dynamics and dimension of Cantor sets*, The Lefschetz Centennial Conference, Contemp. Math. **58** (III) (1984), 203–216.
[67] J. Palis, *A global view of dynamics and a conjecture on the denseness of finitude of attractors*, Géométrie complexe et systèmes dynamiques (Orsay, 1995), Astérisque **261** (2000), 335–347.
[68] J. Palis and F. Takens, *Hyperbolicity and the creation of homoclinic orbits*, Ann. of Math. **125** (2) (1987), 337–374.
[69] J. Palis and F. Takens, *Hyperbolicity and Sensitive-Chaotic Dynamics at Homoclinic Bifurcations*, Cambridge Univ. Press (1993).
[70] J. Palis and M. Viana, *High dimension diffeomorphisms displaying infinitely many periodic attractors*, Ann. of Math. (2) **140** (1) (1994), 207–250.
[71] J. Palis and J.-C. Yoccoz, *On the arithmetic sum of regular Cantor sets*, Ann. Inst. H. Poincaré Anal. Non Linéaire **14** (4) (1997), 439–456.
[72] J. Palis and J.-C. Yoccoz, *Fers à cheval non uniformément hyperboliques engendrés par une bifurcation homocline et densité nulle des attracteurs*, C. R. Acad. Sci. Paris Sér. I Math. **333** (9) (2001), 867–871.
[73] V.A. Pliss, *On a conjecture due to Smale*, Differ. Uravn. **8** (1972), 268–282.
[74] C. Pugh, *The closing lemma*, Amer. J. Math. **89** (1967), 956–1009.
[75] C. Pugh, *An improved closing lemma and a general density theorem*, Amer. J. Math. **89** (1967), 1010–1021.
[76] C. Pugh and M. Shub, *Stable ergodicity and julienne quasi-conformality*, J. European Math. Soc. (JEMS) **2** (1) (2000), 1–52.
[77] E.R. Pujals, *Tangent bundles dynamics and its consequences*, Proceedings of the International Congress of Mathematicians, Vol. III (Beijing, 2002), Higher Ed. Press, Beijing (2002), 327–338.
[78] E.R. Pujals, F. Rodriguez Hertz and M. Sambarino, *Characterization of partial hyperbolic system on the three dimensional torus*, Preprint.
[79] E.R. Pujals and M. Sambarino, *Homoclinic tangencies and hyperbolicity for surface diffeomorphisms*, Ann. of Math. **151** (2000), 961–1023.
[80] E.R. Pujals and M. Sambarino, *On homoclinic tangencies, hyperbolicity, creation of homoclinic orbits and variation of entropy*, Nonlinearity **13** (2000), 921–926.
[81] E.R. Pujals and M. Sambarino, *On the dynamic of dominated splitting*, Preprint, premat.fing.edu.uy/papers/2003/70.ps.
[82] E.R. Pujals and M. Sambarino, *Codimension one dominated splittings*, Preprint.
[83] E.R. Pujals and M. Sambarino, *Sufficient condition for robustly mixing foliations*, Ergodic Theory Dynam. Systems, to appear.
[84] N. Romero, *Persistence of homoclinic tangencies in higher dimensions*, Ergodic Theory Dynam. Systems **15** (4) (1995), 735–757.
[85] D. Ruelle, *Zeta functions for expanding maps and Anosov flows*, Invent. Math. **34** (1976), 231–242.
[86] H.H. Rugh, *The correlation spectrum for hyperbolic analytic maps*, Nonlinearity **5** (1992), 1237–1263.
[87] H.H. Rugh, *Generalized Fredholm determinants and Selberg zeta functions for Axiom A dynamical systems*, Ergodic Theory Dynam. Systems **16** (1996), 805–819.
[88] H.H. Rugh, *Intermittency and regularized Fredholm determinant*, Invent. Math. **135** (1999), 1–25.
[89] B. Saltzman, *Finite amplitude free convection as an initial value problem*, J. Atmospheric Sci. **19** (1962), 329–341.
[90] M. Shub, *Topologically transitive diffeomorphism of T^4*, Symposium on Differential Equations and Dynamical Systems (University of Warwick, 1968/1969), Lecture Notes in Math., Vol. 206, Springer-Verlag, Berlin (1971), 39–40.
[91] M. Shub and A. Wilkinson, *Pathological foliations and removable zero exponents*, Invent. Math. **139** (2000), 495–508.
[92] S. Smale, *Structurally stable systems are not dense*, Amer. J. Math. **88** (1966), 491–496.
[93] A. Tahzibi, *Stably ergodic diffeomorphisms which are not partially hyperbolic*, Ph.D. thesis IMPA (2002).
[94] W. Tucker, *The Lorenz attractor exists*, C. R. Acad. Sci. Paris **328** (1999), 1197–1202.

[95] R. Ures, *Abundance of hyperbolicity in the C^1 topology*, Ann. Sci. École Norm. Sup. 4 **28** (6) (1995), 747–760.
[96] L. Wen, *Homoclinic tangencies and dominated splittings*, Nonlinearity **15** (5) (2002), 1445–1469.
[97] R. Williams, *The structure of the Lorenz attractors*, Publ. Math. IHES **50** (1979), 73–99.

CHAPTER 5

Random Dynamics

Yuri Kifer

Institute of Mathematics, The Hebrew University, Jerusalem 91904, Israel
E-mail: *kifer@math.huji.ac.il*

Pei-Dong Liu

School of Mathematical Sciences, Peking University, Beijing 100871, PR China
E-mail: *lpd@pku.edu.cn*

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

Introduction

Since the time of Newton it became customary to describe the law of motion of a mechanical system by a solution $X(t)$ of an ordinary differential equation with a given initial condition $X(0) = x$. The dynamical systems ideology developed only in the 20th century suggested to look at the evolution of the whole phase space of initial conditions (and not only of a specific x) under the action of an appropriate group (or semigroup) of transformations F^t, called a flow in view of the natural analogy with hydrodynamics, so that a solution $X(t)$ with an initial condition x can be written as $F^t x$. Among early explicit manifestations of this approach was the celebrated Poincaré recurrence theorem whose statement concerns only almost all initial conditions and it has nothing to say about a specific one.

A similar but much later development occurred with stochastic dynamics. Stochastic differential equations (SDEs) were introduced by Itô at the beginning of the 1940s giving an explicit construction of diffusion processes which were studied in the 1930s by Kolmogorov via partial differential equations and measures in their path spaces. For about 40 years it was customary to consider (random) solutions $X(t, \omega)$ at time $t > 0$ of an SDE with a fixed initial condition $X(0, \omega) = x$ and the distribution of corresponding random paths was usually of prime interest. Around 1980 several mathematicians discovered that solutions of SDEs can also be represented in a similar to the deterministic case form $X(t, \omega) = F^t_\omega x$ where the family F^t_ω is called a stochastic flow (see [107]) and for each $t > 0$ and almost all ω it consists of diffeomorphisms.

With the development of dynamical systems in the 20th century it became increasingly clear that discretizing time and considering iterations of a single transformation is quite beneficial both as a tool to study the original flow generated by an ordinary differential equation, for instance, via the Poincaré first-return map, and as a rich source of new models which cannot appear in the continuous time (especially, ordinary differential equations) framework but provide an important insight into the dynamics which is free from continuous time technicalities. The next step is an observation that the evolution of physical systems is time dependent by its nature, and so they could be better described by compositions of different maps rather than by repeated applications of exactly the same transformation. It is natural to study such problems for typical, in some sense, sequences of maps picked at random in some stationary fashion. This leads to random transformations, i.e., to discrete time random dynamical systems (RDS).

Random transformations were discussed already in 1945 by Ulam and von Neumann [159] and few years later by Kakutani [74] in the framework of random ergodic theorems and their study continued in the 1970s in the framework of relative ergodic theory (see [157] and [109]) but all this attracted only a marginal interest. The appearance of stochastic flows as solutions of SDEs gave a substantial push to the subject and towards the end of the 1980s it became clear that powerful dynamical systems tools united with the probabilistic machinery can produce a scope of results which comprises now the theory of RDS. Emergence of additional structures in SDEs motivated probabilists to have a close look at the theory of smooth dynamical systems. This brought to this subject such notions as Lyapunov exponents, invariant manifolds, bifurcations, etc., which had to be adapted to the random diffeomorphisms setup. Moreover, an introduction of invariant measures of random transformations enables us to speak about such notions as the (relative) entropy,

the variational principle, equilibrium states, and the thermodynamic formalism which were developed in the deterministic case in the second half of the 20th century. The probabilistic state of mind requires here to assume as little as possible about the stochasticity which drives random transformations unlike the approach in the classical ergodic theory where all measure (probability) spaces are usually assumed to be separable.

During last 20 years a lot of work was done on various aspects of RDS and some of it is presented in 5 books [82,118,47,7,49] written on this subject. The theory of RDS found its applications in statistical physics (see [152]), economics (see [155]), meteorology (see [56]) and in other fields. In this survey we describe several important parts of ergodic theory of RDS but we do not try to fulfil an impossible task to cover everything that was done in this subject. This survey consists of 5 sections among which 4 sections exhibit the theory of RDS and Section 5 deals with random perturbations of dynamical systems. Section 1 deals with the general ergodic theory and the topological dynamics of random transformations. The general setup of random transformations together with notations we use in this survey are introduced in Section 1.1 which contains basic results about the measure-theoretic (metric) entropy and generators for random transformations. Section 2 deals with constructions of random stable and unstable manifolds for RDS while Section 3 exhibits results about relations between Lyapunov exponents and the (relative) entropy such as Ruelle's inequality and Pesin's formula for RDS. In short, Sections 2 and 3 describe results which comprise what can be called as Pesin's theory for random diffeomorphisms and endomorphisms whose original deterministic version is exhibited in the article by Barreira and Pesin [1] in this volume. Section 4 exhibits the scope of results related to or relying upon the thermodynamic formalism ideology and constructions adapted to random transformations.

Section 5 about random perturbations of dynamical systems stands quite apart from other sections. The reason for its inclusion to this survey is two-fold. First, some popular models of random perturbations, where we apply at random small perturbations of a given map, lead to random transformations. Secondly, the study of both RDS and random perturbations are motivated to some extent by an attempt to understand various stability properties of dynamical systems. The first paper [135] which rises the problem of stability of dynamical systems under random perturbations appeared already in 1933 but until the 1960s this question had not attracted substantial attention. At that time random perturbations only of dynamical systems with simple dynamics were studied (see [80] and [164]) and only in the 1970s the most interesting case of systems with complicated (chaotic) dynamics had been dealt with (see [81]). Various probabilistic results on diffusion perturbations of systems with simple dynamics can be found in [64]. On the other hand, random perturbations of chaotic dynamical systems are described in [83] (see also [27]). Since then new methods and results have appeared and we will describe also some recent results concerning random perturbations of certain types of nonuniformly hyperbolic systems. We will see also how random perturbations can serve as a tool in computations of chaotic dynamical systems on a computer which, in fact, goes back to Ulam [158].

Among main topics related to RDS which are not covered by this survey are: stochastic bifurcations theory which is not yet complete but some parts of it can be found in [7], topological classification of random cocycles which is described in [47], and infinite-dimensional RDS which play an important role in various models described by partial

differential equations with a random noise, e.g., random force (see, for instance, [50,62, 163]).

1. Basic structures of random transformations

1.1. *Entropy and generators*

The setup of this survey consists of a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ together with a $\mathbb{P}$-preserving invertible map ϑ, of a Polish space $(X, \mathcal{B})$ with the Borel σ-algebra $\mathcal{B}$, and of a set $\mathcal{E} \subset \Omega \times X$ measurable with respect to the product σ-algebra $\mathcal{F} \times \mathcal{B}$. A bundle RDS F over $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ is generated by mappings $F_\omega : \mathcal{E}_\omega \to \mathcal{E}_{\vartheta\omega}$ so that the map $(\omega, x) \to F_\omega x$ is measurable. The family $\{F_\omega,\ \omega \in \Omega\}$ is called a random transformation and each F_ω maps the fiber $\mathcal{E}_\omega = \{x \in X\colon\ (\omega, x) \in \mathcal{E}\}$ to $\mathcal{E}_{\vartheta\omega}$. The map $\Theta : \mathcal{E} \to \mathcal{E}$ defined by $\Theta(\omega, x) = (\vartheta\omega, F_\omega x)$ is called the skew product transformation. Observe that $\Theta^n(\omega, x) = (\vartheta^n\omega, F^n_\omega x)$ where $F^n_\omega = F_{\vartheta^{n-1}\omega} \circ \cdots \circ F_{\vartheta\omega} \circ F_\omega$ for $n \geqslant 0$, $F^0_\omega = \mathrm{id}$, and, if F_ω, $\omega \in \Omega$, are invertible transformations, $F^n_\omega = F^{-1}_{\vartheta^{-n}\omega} \circ \cdots \circ F^{-1}_{\vartheta^{-2}\omega} \circ F^{-1}_{\vartheta^{-1}\omega}$ for $n \leqslant -1$ where $F^{-1}_\omega = (F_\omega)^{-1}$. Clearly, Θ is invertible if and only if all F_ω, $\omega \in \Omega$, are invertible.

Denote by $\mathcal{P}_\mathbb{P}(\Omega \times X)$ the space of probability measures on $\Omega \times X$ having the marginal $\mathbb{P}$ on Ω and set $\mathcal{P}_\mathbb{P}(\mathcal{E}) = \{\mu \in \mathcal{P}_\mathbb{P}(\Omega \times X)\colon\ \mu(\mathcal{E}) = 1\}$. Any $\mu \in \mathcal{P}_\mathbb{P}(\mathcal{E})$ on $\mathcal{E}$ disintegrates $d\mu(\omega, x) = d\mu_\omega(x)\, d\mathbb{P}(\omega)$ (see, for instance, [57, Section 10.2]) where μ_ω are regular conditional probabilities with respect to the σ-algebra $\mathcal{F}_\mathcal{E}$ formed by all sets $(A \times X) \cap \mathcal{E}$ with $A \in \mathcal{F}$. This means that μ_ω is a probability measure on $\mathcal{E}_\omega$ for $\mathbb{P}$-a.a. ω and for any measurable set $R \subset \mathcal{E}$, $\mathbb{P}$-a.s. $\mu_\omega(R(\omega)) = \mu(R|\mathcal{F}_\mathcal{E})$, where $R(\omega) = \{x\colon\ (\omega, x) \in R\}$, and so $\mu(R) = \int \mu_\omega(R(\omega))\, d\mathbb{P}(\omega)$. It is easy to see that $\mu \in \mathcal{P}_\mathbb{P}(\mathcal{E})$ is Θ-invariant if and only if $F_\omega\mu_\omega = \mu_{\vartheta\omega}$ $\mathbb{P}$-almost surely (a.s.) and the space of such measures will be denoted by $\mathcal{I}_\mathbb{P}(\mathcal{E})$.

Let $\mathcal{R} = \{R_i\}$ be a finite or countable partition of $\mathcal{E}$ into measurable sets then $\mathcal{R}(\omega) = \{R_i(\omega)\}$, $R_i(\omega) = \{x \in \mathcal{E}_\omega\colon\ (\omega, x) \in R_i\}$ is a partition of $\mathcal{E}_\omega$. For $\mu \in \mathcal{P}_\mathbb{P}(\Omega \times X)$ the conditional entropy of $\mathcal{R}$ given σ-algebra $\mathcal{F}_\mathcal{E}$ is defined by

$$\begin{aligned} H_\mu(\mathcal{R}|\mathcal{F}_\mathcal{E}) &= -\int \sum_i \mu(R_i|\mathcal{F}_\mathcal{E}) \log \mu(R_i|\mathcal{F}_\mathcal{E})\, d\mathbb{P} \\ &= \int H_{\mu_\omega}\big(\mathcal{R}(\omega)\big)\, d\mathbb{P}(\omega), \end{aligned} \tag{1.1.1}$$

where $H_{\mu_\omega}(\mathcal{A})$ denotes the usual entropy of a partition $\mathcal{A}$. The relative entropy $h^{(r)}_\mu(\Theta)$ of Θ which will be called also the fiber entropy $h_\mu(F)$ of the bundle RDS F with respect to $\mu \in \mathcal{I}_\mathbb{P}(\mathcal{E})$ is defined by the formula

$$h^{(r)}_\mu(\Theta) = h_\mu(F) = \sup_\mathcal{Q} h_\mu(F, \mathcal{Q}), \tag{1.1.2}$$

where

$$h_\mu(F,\mathcal{Q}) = h_\mu^{(r)}(\Theta,\mathcal{Q}) = \lim_{n\to\infty}\frac{1}{n}H_\mu\left(\bigvee_{i=0}^{n-1}\left(\Theta^i\right)^{-1}\mathcal{Q}\,\Big|\,\mathcal{F}_\mathcal{E}\right) \tag{1.1.3}$$

is called the fiber entropy of F or the relative entropy of Θ with respect to a partition $\mathcal{Q}$ and $\bigvee$ denotes the join of partitions. The supremum in (1.1.2) is taken over all finite or countable measurable partitions $\mathcal{Q}=\{Q_i\}$ of $\mathcal{E}$ with finite conditional entropy

$$H_\mu(\mathcal{Q}|\mathcal{F}_\mathcal{E}) < \infty, \tag{1.1.4}$$

and the $\mathbb{P}$-a.s. limit in (1.1.3) exists in view of subadditivity of the conditional entropy (see [142, §5], or [82, Section 2.1]).

Observe that if $\mathcal{Q}=\{Q_i\}$ is a partition of $\mathcal{E}$ then $\mathcal{Q}^{(n)} = \bigvee_{i=0}^{n-1}(\Theta^i)^{-1}\mathcal{Q}$ is a partition of $\mathcal{E}$ consisting of sets $\{Q_j^{(n)}\}$ such that the corresponding partition $\mathcal{Q}^{(n)}(\omega) = \{Q_j^{(n)}(\omega)\}$, $Q_j^{(n)}(\omega) = \{x\colon (\omega,x)\in Q_j^{(n)}\}$ of $\mathcal{E}_\omega$ has the form $\mathcal{Q}^{(n)}(\omega) = \bigvee_{i=0}^{n-1}(F_\omega^i)^{-1}\mathcal{Q}(\vartheta^i\omega)$, where $\mathcal{Q}(\omega)=\{Q_j(\omega)\}$, $Q_j(\omega)=\{x\in\mathcal{E}_\omega\colon (\omega,x)\in Q_j\}$ partitions $\mathcal{E}_\omega$. This together with (1.1.1) yield

$$h_\mu(F,\mathcal{Q}) = \lim_{n\to\infty}\frac{1}{n}\int H_{\mu_\omega}\left(\bigvee_{i=0}^{n-1}\left(F_\omega^i\right)^{-1}\mathcal{Q}\left(\vartheta^i\omega\right)\right)d\mathbb{P}(\omega). \tag{1.1.5}$$

This way of obtaining the fiber entropy is more in line with the spirit of random dynamical systems. It is not difficult to see (see [82, Section 2.1] and [31]) that the resulting entropy remains the same taking the supremum in (1.1.2) only over partitions $\mathcal{Q}$ of $\mathcal{E}$ into sets Q_i of the form $Q_i = (\Omega\times P_i)\cap\mathcal{E}$, where $\mathcal{P}=\{P_i\}$ is a partition of X into measurable sets, so that $Q_i(\omega) = P_i\cap\mathcal{E}_\omega$. The Abramov–Rohlin formula $h_\mu(\Theta) = h_\mu^{(r)}(\Theta) + h_\mathbb{P}(\vartheta)$ (see [2]), where $h_\mu(\Theta)$ and $h_\mathbb{P}(\vartheta)$ are usual entropies of the corresponding measure preserving transformations, relates different entropies emerging in this setup though in many interesting cases only $h_\mu^{(r)}(\Theta)$ is finite among these three (see [82, Theorem II.1.2]). It is easy to see that $a_n(\omega) = H_{\mu_\omega}(\bigvee_{i=0}^{n-1}(F_\omega^i)^{-1}\mathcal{Q}(\vartheta^i\omega))$ is a subadditive process and (1.1.5) together with the subadditive ergodic theorem (see, for instance, [82, §A.2]) yield $h_\mu^{(r)}(\Theta,\mathcal{Q}) = h_\mu(F,\mathcal{Q}) = \int\lim_{n\to\infty}\frac{1}{n}a_n\,d\mathbb{P}$. If, in addition, $\mathbb{P}$ is ergodic then this formula remains true $\mathbb{P}$-a.s. without integrating against $\mathbb{P}$.

Set

$$\hat{\mathcal{Q}}_\mathcal{E}^\infty(\omega) = \bigvee_{i=-\infty}^{\infty}\left(F_\omega^i\right)^{-1}\mathcal{Q}_\mathcal{E}\left(\vartheta^i\omega\right)$$

if Θ is invertible and

$$\hat{\mathcal{Q}}_\mathcal{E}^\infty(\omega) = \bigvee_{i=0}^{\infty}\left(F_\omega^i\right)^{-1}\mathcal{Q}_\mathcal{E}\left(\vartheta^i\omega\right)$$

if Θ is noninvertible.

DEFINITION 1.1.1. Given $\mu \in \mathcal{I}_{\mathbb{P}}(\mathcal{E})$, a countable or finite measurable partition $\mathcal{Q}_{\mathcal{E}}$ of $\mathcal{E}$ is called a fiber μ-generator of F or relative μ-generator of Θ if $\mathbb{P}$-a.s. $\hat{\mathcal{Q}}_{\mathcal{E}}^{\infty}(\omega)$ generates the restriction of the σ-algebra $\mathcal{B}$ to $\mathcal{E}_\omega$ up to sets of μ_ω-measure zero. If $\mathcal{Q}_{\mathcal{E}} = \{Q_i\}$ with $\mathcal{Q}_{\mathcal{E}}(\omega) = \{Q_i(\omega)\}$ and $-\infty \leqslant m \leqslant n \leqslant \infty$ then a sequence $(\xi_m, \xi_{m+1}, \dots, \xi_n)$ will be called the $(\omega, m, n, \mathcal{Q})$-name of $x \in \mathcal{E}_\omega$ if $F_\omega^i x \in Q_{\xi_i}(\vartheta^i \omega)$ for all $i = m, m+1, \dots, n$. An $(\omega, -\infty, \infty, \mathcal{Q})$-name will be called just an $(\omega, \mathcal{Q})$-name.

In the light of this definition it is clear that $\mathcal{Q}$ is a relative μ-generator if there exists a measurable set $\tilde{\mathcal{E}} \subset \mathcal{E}$ with $\mu(\tilde{\mathcal{E}}) = 1$ such that any $x, y \in \tilde{\mathcal{E}}_\omega = \{z \in \mathcal{E}_\omega \colon (\omega, z) \in \tilde{\mathcal{E}}\}$, $x \neq y$ have different $(\omega, \mathcal{Q})$-names.

Using general properties of the conditional entropy the following version of the relative Kolmogorov–Sinai theorem has been proved in a slightly less general setup in [31, Theorem 2.4] (see also [82, Lemma II.1.5]) and it appears in the present setup (with the same proof) in [32, Theorem 2.3.3].

THEOREM 1.1.2. *If $\mathcal{Q}$ is a fiber μ-generator of F then $h_\mu(F) = h_\mu(F, \mathcal{Q})$.*

As usual, we say that ϑ is aperiodic (with respect to $\mathbb{P}$) if $\mathbb{P}\{\omega \in \Omega \colon \exists n,\ \vartheta^n \omega = \omega\} = 0$. The following relative version of Krieger's theorem [105] has been proved in [102].

THEOREM 1.1.3. *Suppose that Θ is invertible, $\mu \in \mathcal{I}_{\mathbb{P}}(\mathcal{E})$ is ergodic, ϑ is aperiodic and $h_\mu(F) < \log \ell$ for some integer $\ell > 1$. Then there exists a relative μ-generator $\mathcal{Q}$ consisting of ℓ sets.*

The proof of Theorem 1.1.3 proceeds in the following way. First, we construct ℓ-element partitions of $\mathcal{E}_\omega$ for ω belonging to a set of $\mathbb{P}$-measure close to 1. This is done via coding using most but not all levels of Rohlin's towers (see [78]) constructed in Ω. These partitions mostly do not change on further steps and we do modifications mainly on the complement ω-set. The partition constructed on the first step carries already most of the entropy. After that we employ repeatedly the second step which enables us to encode points more and more precisely, i.e., to increase the resolution of partitions. This is done using few levels of Rohlin's towers whose number can be estimated via conditional entropies of refinements of partitions.

The aperiodicity assumption on ϑ means that our random transformations are, indeed, random, at least, mildly. If $(\Omega, \mathcal{F}, \mathbb{P})$ is, in addition, a Lebesgue space, i.e., the σ-algebra $\mathcal{F}$ is separable $\mathbb{P}$-mod 0 (which is not a very natural assumption from the probabilistic point of view as we indicated this in Introduction) then we do not have to assume aperiodicity of ϑ and in this case Theorem 1.1.3 is a generalization of Krieger's theorem. Indeed, since ϑ is ergodic it is either aperiodic or purely periodic. In the former case we use Theorem 1.1.3 and in the latter case $h_{\mathbb{P}}(\vartheta) = 0$, and so by [2], $h_\mu(\Theta) = h_\mu(F) < \log \ell$. Then Theorem 1.1.3 follows from Krieger's original result, since, of course, any absolute generating partition for F is also its relative generating partition.

Under the Lebesgue space assumption on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ there is another proof of Theorem 1.1.3 suggested by A. Danilenko after he learned about the proof from [102]. His argument is based on an orbit equivalence of ϑ to a $\mathbb{P}$-preserving invertible

ergodic transformation ϑ' with $h_{\mathbb{P}}(\vartheta')=0$ which holds true assuming that $(\Omega,\mathcal{F},\mathbb{P})$ is a Lebesgue space (see [58]). If $\vartheta'(\omega)=\vartheta^{k(\omega)}\omega$ then defining $\Theta'(\omega,x)=(\vartheta^{k(\omega)}\omega, F_\omega^{k(\omega)}x)$ we obtain from [2] together with [144] that $h_\mu(\Theta')=h_\mu^{(r)}(\Theta')=h_\mu^{(r)}(\Theta)$. Thus $h_\mu(\Theta')<\log\ell$ and applying the original Krieger theorem we obtain an absolute generating partition $\mathcal{P}$ for Θ' with ℓ elements. The proof concludes by the crucial observation that while this partition may not be an absolute generator for Θ it is always a relative generator for Θ, i.e., a fiber generator for F. This follows since knowing ω and the $\mathcal{P}$-name of (ω,x) with respect to Θ also gives us the $\mathcal{P}$-name of (ω,x) with respect to Θ' (the orbit is $\mathcal{F}$-measurable) and this in turn determines (ω,x) uniquely. Although brief, this proof relies on several highly nontrivial theorems including Krieger's theorem itself. On the other hand, the original proof of the relative theorem in [102] is simpler than the proof of the absolute result since no care needs to be taken to encode bases of Rohlin towers. Another advantage of this approach is that essentially the same technique works for the construction of topological generators discussed in Section 1.3 where the orbit equivalence considerations do not help. Finally, the above orbit equivalence arguments require that $(\Omega,\mathcal{F},\mathbb{P})$ be a Lebesgue space whereas in the proof from [102] the random mechanism is quite arbitrary and, in particular, this probability space can be nonseparable, which is natural from the probabilistic point of view.

Next, we state a relative version of another important result in the entropy theory of dynamical systems the Shannon–McMillan–Breiman theorem which is used, in particular, in the proof of Theorem 1.1.3. Its proof in a bit less general setup can be found, for instance, in [82, Theorem 2.5] and in [31, Theorem 4.2] and it appears in the present setup (with, essentially, the same proof) in [32, Theorem 2.2.5]. Let $\mathcal{Q}=\{Q_i\}$ be a finite or countable measurable partition of $\mathcal{E}$ with fiber partitions $\mathcal{Q}(\omega)=\{Q_i(\omega)\}$ of $\mathcal{E}_\omega$. For any $\mu\in\mathcal{P}_{\mathbb{P}}(\Omega\times X)$ the conditional information of $\mathcal{Q}$ given a σ-algebra $\mathcal{F}_{\mathcal{E}}$ is defined by (see [132])

$$\begin{aligned} I_\mu(\mathcal{Q}|\mathcal{F}_{\mathcal{E}})(\omega) &= -\sum_i \mathbb{I}_{Q_i(\omega)}\log\mu(Q_i|\mathcal{F}_{\mathcal{E}})(\omega)\\ &= -\sum_i \mathbb{I}_{Q_i(\omega)}\log\mu_\omega\big(Q_i(\omega)\big)=I_{\mu_\omega}\big(\mathcal{Q}(\omega)\big),\quad \mathbb{P}\text{-a.s.}, \end{aligned}\tag{1.1.6}$$

where $\{\mu_\omega\}$ is the factorization of μ and $\mathbb{I}_A$ denotes the indicator of a set A, i.e., $\mathbb{I}_A(x)=1$ if $x\in A$ and $=0$, otherwise. Recall, that the fibers of the partition $\mathcal{Q}^{(n)}=\bigvee_{i=0}^{n-1}(\Theta^i)^{-1}\mathcal{Q}$ have the form $\mathcal{Q}^{(n)}(\omega)=\bigvee_{i=0}^{n-1}(F_\omega^i)^{-1}\mathcal{Q}(\vartheta^i\omega)$. For any $x\in\mathcal{E}_\omega$ denote by $\mathcal{Q}^{(n)}(\omega,x)$ the element of the partition $\mathcal{Q}^{(n)}(\omega)$ which contains x.

THEOREM 1.1.4. *Suppose that $\mu\in\mathcal{I}_{\mathbb{P}}(\mathcal{E})$ factorizes into $\{\mu_\omega\}$ and let $\mathcal{Q}$ be a finite or countable measurable partition of $\mathcal{E}$ with $H_\mu(\mathcal{Q}|\mathcal{F}_{\mathcal{E}})<\infty$. Then*

$$\lim_{n\to\infty}\frac{1}{n}I_{\mu_\omega}\big(\mathcal{Q}^{(n)}(\omega)\big)=\mathbb{E}_\mu(f|J_\Theta)(\omega),\quad \mu\text{-a.s. and in } L^1(\mathcal{E},\mu),\tag{1.1.7}$$

where $f(\omega,x) = -\log\mu_\omega(Q^{(1)}(\omega,x)|\mathcal{Q}^{(\infty)}(\omega))(x)$ μ*-a.s. where* $\mathcal{Q}^{(\infty)}(\omega)$ *is the* σ*-algebra generated by all* $\mathcal{Q}^{(n)}(\omega)$, $n \geqslant 0$, *and* J_Θ *is the* σ*-algebra of* Θ*-invariant sets. If* μ *is ergodic then*

$$\lim_{n\to\infty}\frac{1}{n}\log\mu_\omega\big(Q^{(n)}(\omega,x)\big) = -h_\mu(F,\mathcal{Q}), \quad \mu\text{-a.s.} \tag{1.1.8}$$

For fiber (relative) isomorphism theory, in particular, the characterization of a fiber isomorphism of random Bernoulli shifts (see Section 4.1) by their fiber entropy we refer the reader to [157] and [67].

1.2. *Topological pressure and variational principle*

In this section we consider continuous RDS acting on compact metric fibers, i.e., in the setup of Section 1.1 we assume that X is a compact metric space, all fibers $\mathcal{E}_\omega \subset X$ are compact, $\mathcal{E}_\omega \neq \emptyset$ $\mathbb{P}$-a.s., and $F_\omega : \mathcal{E}_\omega \to \mathcal{E}_{\vartheta\omega}$ is continuous $\mathbb{P}$-a.s. It follows from a relative version of the Krylov–Bogolyubov theorem or from the Markov–Kakutani fixed point theorem that under these conditions the set of Θ-invariant measures $\mathcal{I}_\mathbb{P}(\mathcal{E})$ is not empty (see [7, Theorem 1.5.10] or [49, Corollary 6.13]). The exposition of this section follows [99] and all proofs of results mentioned below can be found there.

For each $n \in \mathbb{N}$ and a positive random variable $\varepsilon = \varepsilon(\omega)$ define a family of metrics $d^\omega_{\varepsilon,n}$ on $\mathcal{E}_\omega$ by the formula

$$d^\omega_{\varepsilon,n}(x,y) = \max_{0\leqslant k<n}\big(d\big(F^k_\omega x, F^k_\omega y\big)\big(\varepsilon\big(\vartheta^k\omega\big)\big)^{-1}\big), \quad x,y \in \mathcal{E}_\omega,$$

where F^0_ω is the identity map. It is not difficult to see that $\mathcal{E}^{(2)} = \{(\omega,x,y)\colon\ x,y \in \mathcal{E}_\omega\}$ belongs to the product σ-algebra $\mathcal{F}\times\mathcal{B}^2$ (as a graph of a measurable multifunction, see [44, Proposition III.13]). Since for any $k \in \mathbb{N}$ and a number a the set $\{(\omega,x,y) \in \mathcal{E}^{(2)}\colon\ d(F^k_\omega x, F^k_\omega y) \leqslant a\varepsilon(\vartheta^k\omega)\}$ is measurable with respect to this product σ-algebra we conclude that $d^\omega_{\varepsilon,n}(x,y)$ depends measurably on $(\omega,x,y) \in \mathcal{E}^{(2)}$. Denote by $B_x(\omega,\varepsilon,n)$ the closed ball in $\mathcal{E}_\omega$ centered at x of radius 1 with respect to the metric $d^\omega_{\varepsilon,n}$. For $d^\omega_{\varepsilon,1}$ and $B_x(\omega,\varepsilon,1)$ we will write simply d^ω_ε and $B_x(\omega,\varepsilon)$, respectively. We say that $x,y \in \mathcal{E}_\omega$ are (ω,ε,n)-close if $d^\omega_{\varepsilon,n}(x,y) \leqslant 1$.

DEFINITION 1.2.1. A set $Q \subset \mathcal{E}_\omega$ is called (ω,ε,n)-separated if $x,y \in Q$, $x \neq y$ implies $d^\omega_{\varepsilon,n}(x,y) > 1$.

Due to the compactness of $\mathcal{E}_\omega$ there exists a smallest natural number $s_n(\omega,\varepsilon)$ such that $\operatorname{card}(Q) \leqslant s_n(\omega,\varepsilon) < \infty$ for every (ω,ε,n)-separated set Q. Moreover, there always exists a maximal (ω,ε,n)-separated set Q in the sense that for every $x \in \mathcal{E}_\omega$ with $x \notin Q$ the set $Q \cup \{x\}$ is not (ω,ε,n)-separated anymore. If Q is maximal (ω,ε,n)-separated, then $\mathcal{E}_\omega = \bigcup_{x\in Q} B_x(\omega,\varepsilon,n)$.

For each measurable in (ω,x) and continuous in $x \in \mathcal{E}_\omega$ function f on $\mathcal{E}$ set $\|f\|_1 = \int \|f(\omega)\|_\infty\, d\mathbb{P}$ where $\|f(\omega)\|_\infty = \sup_{x\in\mathcal{E}_\omega}|f(\omega,x)|$. Denote by $\mathbb{L}^1_\mathcal{E}(\Omega,\mathcal{C}(X))$ the space

of such functions f with $\|f\|_1 < \infty$. If we identify f and g provided $\|f-g\|_1 = 0$ then $\mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$ becomes a Banach space with the norm $\|\cdot\|_1$. For $f \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$, $n \geqslant 1$, a positive random variable ε, and an (ω, ε, n)-separated set $Q \subset \mathcal{E}_\omega$ set

$$Z_n(\omega, f, Q) = \sum_{x \in Q} \exp\bigl(S_n f(\omega, x)\bigr),$$

where

$$(S_n f)(\omega, x) = \sum_{i=0}^{n-1} f\bigl(\vartheta^i \omega, F^i_\omega x\bigr) = \sum_{i=0}^{n-1} f \circ \Theta^i(\omega, x),$$

and

$$\pi_F(f)(\omega, \varepsilon, n) = \sup\bigl\{Z_n(\omega, f, Q)\colon\ Q \text{ is an } (\omega, \varepsilon, n)\text{-separated subset of } \mathcal{E}_\omega\bigr\}.$$

Since all summands in $Z_n(\omega, f, Q)$ are positive and any (ω, ε, n)-separated set can be completed to a maximal one, the supremum above can be taken only over maximal (ω, ε, n)-separated sets. The following auxiliary result which relies on slightly modified arguments from [31] provides basic measurability properties needed in what follows.

LEMMA 1.2.2. *For any $n \in \mathbb{N}$ and a positive random variable $\varepsilon = \varepsilon(\omega)$ the function $\pi_F(f)(\omega, \varepsilon, n)$ is measurable in ω and for each $\delta > 0$ there exists a family of maximal (ω, ε, n)-separated sets $G_\omega \subset \mathcal{E}_\omega$ satisfying*

$$Z_n(\omega, f, G_\omega) \geqslant (1-\delta)\pi_F(f)(\omega, \varepsilon, n) \tag{1.2.1}$$

and depending measurably on ω in the sense that $G = \{(\omega, x)\colon\ x \in G_\omega\} \in \mathcal{F} \times \mathcal{B}$ which means also that the mapping $\omega \mapsto G_\omega$ is measurable with respect to the Borel σ-algebra induced by the Hausdorff topology on the space $\mathcal{K}(X)$ of compact subsets of X. In particular, the supremum in the definition of $\pi_F(f)(\omega, \varepsilon, n)$ can be taken only over measurable in ω families of (ω, ε, n)-separated sets.

In view of the above result, for each $f \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$ and any positive random variable ε we can set

$$\pi_F(f)(\varepsilon) = \limsup_{n\to\infty} \frac{1}{n} \int \log \pi_F(f)(\omega, \varepsilon, n)\, d\mathbb{P}(\omega). \tag{1.2.2}$$

DEFINITION 1.2.3. The fiber topological pressure of F (or the relative topological pressure of Θ) is the map

$$\pi_F : \mathbb{L}^1_{\mathcal{E}}\bigl(\Omega, \mathcal{C}(X)\bigr) \to \mathbb{R} \cup \{\infty\}, \quad \text{where } \pi_F(f) = \lim_{\varepsilon \to 0} \pi_F(f)(\varepsilon), \tag{1.2.3}$$

and the limit taken over *nonrandom* ε exists since $\pi_F(f)(\varepsilon)$ is monotone in ε and, in fact, $\lim_{\varepsilon\to 0}$ above equals $\sup_{\varepsilon>0}$. The quantity $h_{\mathrm{top}}(F) = h^{(r)}_{\mathrm{top}}(\Theta) = \pi_F(0)$ is called the fiber topological entropy of F or the relative topological entropy of Θ.

We will show at the end of this section that as in the deterministic case (see [162, §9.1]) it is possible to obtain $\pi_F(f)$ via covers in place of (ω, ε, n)-separated sets.

REMARK 1.2.4. Since any two metrics on the compact X compatible with its topology are uniformly continuous with respect to each other then any such metric will yield the same relative topological pressure $\pi_F(f)$ as defined above, i.e., the latter depends only on the topology of X.

REMARK 1.2.5. Sometimes (see, for instance, [109] and [31]) the relative topological pressure is introduced via "weakly" (ω, ε, n)-separated sets Q, where $x, y \in Q$, $x \neq y$ implies $d^{\omega}_{\varepsilon,n}(x, y) \geqslant 1$, which are less convenient in applications but lead, obviously, to the same quantity $\pi_F(f)$. If the supremum in the definition of $\pi_F(f)(\omega, \varepsilon, n)$ is taken over "weakly" (ω, ε, n)-separated sets then the corresponding random variables remain measurable and, in fact, the supremum can be taken over measurable families of such sets.

Definition 1.2.3 provokes immediately two natural questions whether the relative topological pressure remains the same if the order of lim sup and the integral in (1.2.2) is reversed and, also, what happens if $\pi_F(f)$ is defined via random ε.

PROPOSITION 1.2.6. *For any* $f \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$,

$$\begin{aligned}\pi_F(f) &= \int \lim_{\varepsilon\to 0} \limsup_{n\to\infty} \frac{1}{n} \log \pi_F(f)(\omega, \varepsilon, n)\, d\mathbb{P}(\omega) \\ &= \int \lim_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log \pi_F(f)(\omega, \varepsilon, n)\, d\mathbb{P}(\omega) \qquad (1.2.4)\end{aligned}$$

and if $\mathbb{P}$ *is ergodic* (*with respect to the action of* ϑ) *then* (1.2.4) *holds true* $\mathbb{P}$*-a.s. without taking integrals in the right-hand side.*

Next, we compare definitions of the relative topological pressure corresponding to random and nonrandom ε. For any compact subset $Y \subset X$ denote by $N_Y(r)$ the minimal number of closed balls of diameter r which cover Y.

LEMMA 1.2.7. *Let* ε *and* δ *be positive random variables and* $A_{\varepsilon,\delta} = \{\omega\colon \delta(\omega) > \frac{1}{2}\varepsilon(\omega)\}$. *Then for any maximal* (ω, δ, n)*-separated set* Q,

$$\pi_F(f)(\omega, n, \varepsilon) \leqslant \exp\left(\sum_{i=0}^{n-1} R_{\varepsilon,\delta}\left(\vartheta^i\omega\right)\right) Z_n(\omega, f, Q), \qquad (1.2.5)$$

where $R_{\varepsilon,\delta}(\omega) = \kappa_\varepsilon^{(f)}(\omega) + \mathbb{I}_{A_{\varepsilon,\delta}}(\omega)(2\|f(\omega)\|_\infty + \log N_{\mathcal{E}_\omega}(\varepsilon(\omega)))$,

$$\kappa_\varepsilon^{(f)}(\omega) = \sup\big\{\big|f(\omega,x) - f(\omega,y)\big|\colon\, x, y \in \mathcal{E}_\omega,\ d(x,y) \leqslant \varepsilon(\omega)\big\},$$

and $\mathbb{I}_A(\omega) = 1$ *if* $\omega \in A$ *and* $= 0$ *if* $\omega \notin A$.

COROLLARY 1.2.8. *For any* $n \in \mathbb{N}$ *and each positive* δ *let* $Q_\omega^{\delta,n}$ *be a measurably depending on* ω *family of maximal* (ω,δ,n)*-separated sets* (*which exist in view of Lemma* 1.2.2). *Then*

$$\begin{aligned}\pi_F(f) &= \int \lim_{\delta\to 0}\limsup_{n\to\infty}\frac{1}{n}\log Z_n\big(\omega, f, Q_\omega^{\delta,n}\big)\,d\mathbb{P}(\omega)\\ &= \lim_{\delta\to 0}\limsup_{n\to\infty}\frac{1}{n}\int \log Z_n\big(\omega, f, Q_\omega^{\delta,n}\big)\,d\mathbb{P}(\omega).\end{aligned} \tag{1.2.6}$$

Moreover, if $\mathbb{P}$ *is ergodic then* (1.2.6) *remains true* $\mathbb{P}$*-a.s. without taking the integral in the right-hand side and then the first equality in* (1.2.6) *holds true without assuming measurable dependence of* $Q_\omega^{\delta,n}$ *on* ω. *Furthermore, the result remains the same if* lim sup *is replaced by* lim inf.

Observe that $N_Y(r)$ is nonincreasing and right continuous in r and it is lower semicontinuous in Y on the space $\mathcal{K}(X)$ of compact subsets of X considered with the Hausdorff topology. It follows that $\{(Y,r) \in \mathcal{K}(X)\times(0,\infty)\colon\ N_Y(r) \leqslant n\}$ is a closed set in the product topology of $\mathcal{K}(X)\times(0,\infty)$ for any $n \in \mathbb{N}_+$ and, in particular, $N_Y(r)$ is measurable in the pair (Y,r) with respect to the product Borel σ-algebra. Since for any positive random variable $\varepsilon = \varepsilon(\omega)$ the map $\psi : \Omega \mapsto \mathcal{K}(X)\times(0,\infty)$ defined by $\psi(\omega) = (\mathcal{E}_\omega, \varepsilon(\omega))$ is measurable then $N \circ \psi(\omega) = N_{\mathcal{E}_\omega}(\varepsilon(\omega))$ is measurable.

DEFINITION 1.2.9. We say that a positive random variable ε belongs to the class $\mathcal{N}$ and write $\varepsilon \in \mathcal{N}$ if

$$\int \log N_{\mathcal{E}_\omega}\big(\varepsilon(\omega)\big)\,d\mathbb{P}(\omega) < \infty. \tag{1.2.7}$$

Since $N_X(r) \geqslant N_{\mathcal{E}_\omega}(r)$ for all ω this holds true if, in particular,

$$\int \log N_X\big(\varepsilon(\omega)\big)\,d\mathbb{P}(\omega) < \infty.$$

The latter is satisfied if, for instance, the upper Minkowski (box) dimension

$$D_X = \limsup_{r\to 0}\frac{\log N_X(r)}{-\log r}$$

of X is finite and $-\int \log \varepsilon(\omega)\, d\mathbb{P}(\omega) < \infty$. Note also that

$$\log N_Y\big(\min(r_1, r_2)\big) = \max\big(\log N_Y(r_1), \log N_Y(r_2)\big) \leqslant \log N_Y(r_1) + \log N_Y(r_2),$$

and so if $\varepsilon, \delta \in \mathcal{N}$ then $\min(\varepsilon, \delta) \in \mathcal{N}$. Thus $\mathcal{N}$ is a directed set and, clearly, it contains also all positive constants.

The following result which makes sense for deterministic dynamical systems, as well, shows that the same relative topological pressure is obtained taking random $\varepsilon \in \mathcal{N}$ and, essentially, $\mathcal{N}$ is the largest class of positive random variables which yields something reasonable.

PROPOSITION 1.2.10. *Let $f \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$; then*

$$\lim_{\varepsilon \xrightarrow{\mathcal{N}} 0} \pi_F(f)(\varepsilon) = \sup_{\varepsilon \in \mathcal{N}} \pi_F(f)(\varepsilon) = \pi_F(f), \tag{1.2.8}$$

where the left-hand side is the limit along the directed set $\mathcal{N}$. On the other hand, if $2\varepsilon \notin \mathcal{N}$ then $\int \log \pi_F(f)(\omega, \varepsilon, n)\, d\mathbb{P}(\omega) = \infty$ for all $n \in \mathbb{N}$.

Set $\beta_n^{(\varepsilon)}(\omega) = \sup\{\gamma > 0\colon\ d^{\omega}_{\varepsilon,n}(x, y) \leqslant 1$ provided $x, y \in \mathcal{E}_\omega$, $d(x, y) \leqslant \gamma\}$, which is a continuity characteristic of the bundle RDS F. It is easy to see from Lemma 1.2.2 that $\beta_n^{(\varepsilon)}$ is a random variable. For a given $n \in \mathbb{N}_+$ we can replace ϑ by ϑ^n and consider the bundle RDS F^n defined by $(F^n)^k_\omega = F^n_{\vartheta^{(k-1)n}\omega} \circ \cdots \circ F^n_{\vartheta^n\omega} \circ F^n_\omega$.

COROLLARY 1.2.11 (cf. [31, Theorem 5.5]). *Let $f \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$ then for any $n \in \mathbb{N}_+$,*

$$\pi_{F^n}(S_n f) \leqslant n \pi_F(f). \tag{1.2.9}$$

If $\beta_n^{(\varepsilon)} \in \mathcal{N}$ for any constant $\varepsilon > 0$ then

$$\pi_{F^n}(S_n f) = n \pi_F(f). \tag{1.2.10}$$

Next, we turn to the relative variational principle for bundle RDS. Its first proof was given in [109] in the particular case when Ω is compact, $\mathcal{E}$ is a compact subset of $\Omega \times X$, Θ is continuous and only continuous functions on $\Omega \times X$ are considered. Later the result was extended in [31] to the RDS with the trivial bundle where $\mathcal{E} = \Omega \times X$ though the proof there contains some inaccuracies and unnecessary conditions. First, we state the following result parts of which appeared already in several places [109,32,49,7].

LEMMA 1.2.12. *For $\mu, \mu_n \in \mathcal{P}_{\mathbb{P}}(\mathcal{E})$, $n = 1, 2, \ldots$, write $\mu_n \Rightarrow \mu$ if $\int f\, d\mu_n \to \int f\, d\mu$ as $n \to \infty$ for any $f \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$ which introduces a weak* topology in $\mathcal{P}_{\mathbb{P}}(\mathcal{E})$. Then*

(i) *The space $\mathcal{P}_{\mathbb{P}}(\mathcal{E})$ is compact in this weak* topology;*

(ii) *For any sequence* $\nu_k \in \mathcal{P}_{\mathbb{P}}(\mathcal{E})$, $k = 0, 1, 2, \dots$, *the set of limit points in the above weak* topology of the sequence* $\mu^{(n)} = \frac{1}{n}\sum_{k=0}^{n-1} \Theta^k \nu_n$ *as* $n \to \infty$ *is not empty and it is contained in* $\mathcal{I}_{\mathbb{P}}(\mathcal{E})$;

(iii) *Let* $\mu, \mu_n \in \mathcal{P}_{\mathbb{P}}(\mathcal{E})$, $n = 1, 2, \dots$, *and* $\mu_n \Rightarrow \mu$ *as* $n \to \infty$. *Let* $\mathcal{P} = \{P_1, \dots, P_k\}$ *be a finite partition of* X *satisfying* $\int \mu_\omega(\partial\mathcal{P}_\omega)\, d\mathbb{P}(\omega) = 0$, *where* μ_ω *are disintegrations of* μ *and* $\partial\mathcal{P}_\omega = \bigcup_{i=1}^k \partial(P_i \cap \mathcal{E}_\omega)$ *is the boundary of* $\mathcal{P}_\omega = \{P_1 \cap \mathcal{E}_\omega, \dots, P_k \cap \mathcal{E}_\omega\}$. *Denote by* $\mathcal{R}$ *the partition of* $\Omega \times X$ *into sets* $\Omega \times P_i$. *Then*

$$\limsup_{n\to\infty} H_{\mu_n}(\mathcal{R}|\mathcal{F}_{\mathcal{E}}) \leqslant H_\mu(\mathcal{R}|\mathcal{F}_{\mathcal{E}}).$$

Observe that though in [99] this Lemma 1.2.12(iii) has been proved under the assumption that $(\Omega, \mathcal{F}, \mathbb{P})$ is a Lebesgue space, in fact, employing Theorem 3.17 from [49] all assertions remain true (with, essentially, the same proof) also without this assumption. Now we can state the main result of this section called the fiber (relative) variational principle whose proof can be found in [99].

THEOREM 1.2.13. *Let* F *be a continuous bundle RDS on* $\mathcal{E}$ *and* $f \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$. *Then*

$$\pi_F(f) = \sup\left\{h_\mu(F) + \int f\, d\mu\colon\ \mu \in \mathcal{I}_{\mathbb{P}}(\mathcal{E})\right\}. \tag{1.2.11}$$

In order to obtain the fiber topological pressure $\pi_F(f)$ via covers it will be convenient to use the notion of a random set S which is just the collection S of fibers $S(\omega) = \{x \in \mathcal{E}_\omega\colon\ (\omega, x) \in S\}$ of a measurable with respect to $\mathcal{F} \times \mathcal{B}$ set $S \subset \mathcal{E}$ which we denote by the same letter and which, in fact, is the graph of the map $S\colon \Omega \to 2^X$. If all fibers $S(\omega)$ are open (closed) subsets of $\mathcal{E}_\omega$ in its induced from X topology we call S an open (closed) random set. A collection $A = \{a\}$ of random sets will be called a random cover of $\mathcal{E}$ if $\mathcal{E}_\omega = \bigcup_{a\in A} a(\omega)$ for all $\omega \in \Omega$ where, again, $a(\omega) = \{x\colon\ (\omega, x) \in a\}$ and it will be called an open random cover if all $a \in A$ are open random sets. For each $f \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$ and a finite open random cover $A = \{a\}$ of $\mathcal{E}$ set

$$\begin{aligned}&\tilde\pi_F(f)(\omega, A, n)\\ &\quad = \inf\left\{\sum_{b\in B(\omega)} \sup_{x\in b} e^{(S_n f)(\omega,x)}\colon\ B(\omega) \text{ is a finite subcover of } A^{(n)}(\omega)\right\}\end{aligned} \tag{1.2.12}$$

where $A(\omega) = A \cap \mathcal{E}_\omega = \{a(\omega),\ a \in A\}$, $A^{(n)}(\omega) = \bigvee_{k=0}^{n-1}(F^k_\omega)^{-1}A(\vartheta^k\omega)$, and $C \vee D$ for covers C and D is the cover whose elements are all nonempty intersections $c \cap d$, $c \in C$, $d \in D$. Clearly, $A^{(n)}(\omega)$ is an open cover of $\mathcal{E}_\omega$ and the infimum in (1.2.12) is taken over all its finite subcovers which exist since $\mathcal{E}_\omega$ is compact. It turns out that $\tilde\pi_F(f)(\omega, A, n)$ is measurable in ω. Indeed, $A^{(n)}(\omega)$ is the ω-fiber of the open random cover $A^{(n)} = \bigvee_{k=0}^{n-1}(\Theta^k)^{-1}A$. Let $A = \{a_1, \dots, a_m\}$. For each string $\xi = (\xi_0, \dots, \xi_{n-1})$, $1 \leqslant \xi_i \leqslant m$ set $a^\xi(\omega) = \bigcap_{k=0}^{n-1}(F^k_\omega)^{-1}a_{\xi_k}(\vartheta^k\omega) \in A^{(n)}(\omega)$ which is an ω-fiber of the random set $\bigcap_{k=0}^{n-1}(\Theta^k)^{-1}a_{\xi_k}$. For each collection Ψ of such strings denote by Ω_Ψ the subset of

Ω such that for each $\omega \in \Omega_\Psi$, $\bigcup_{\xi\in\Psi} a^\xi(\omega) = \mathcal{E}_\omega$. Since $\Omega_\Psi = \{\omega\colon \mathcal{E}_\omega \setminus \bigcup_{\xi\in\Psi} a^\xi(\omega) = \emptyset\}$ it follows from Proposition 2.4 in [49] (see also [44, p. 80]) that Ω_Ψ is measurable. Intersections of Ω_Ψ for different Ψ yield a finite measurable partition of Ω. On each element of this partition the infimum in (1.2.12) is, in fact, the minimum taken over a finite number of subcovers determined by corresponding string collections Ψ which yields immediately that $\tilde\pi_F(f)(\omega, A, n)$ is measurable. Furthermore, it is easy to see that

$$\tilde\pi_F(f)(\omega, A, n+k) \leqslant \tilde\pi_F(f)(\omega, A, n)\tilde\pi_F(f)(\vartheta^n\omega, A, k).$$

Hence by the subadditive ergodic theorem (see, for instance, [82]) $\mathbb{P}$-a.s. the limit

$$\tilde\pi_F(f)(A) = \lim_{n\to\infty} \frac{1}{n} \log \tilde\pi_F(\omega, A, n)$$

exists though it can be random if $\mathbb{P}$ is not assumed to be ergodic. Next, we set

$$\tilde\pi_F(f) = \sup_A \int \tilde\pi_F(f)(A)\, d\mathbb{P},$$

where the supremum is taken over all finite open random covers of $\mathcal{E}$. By modifying slightly the arguments of Section 2.3 in [82] we derive that $\tilde\pi_F(f) = \pi_F(f)$. It can be shown also that we arrive at the same quantity $\tilde\pi_F(f)$ if we start with (nonrandom) open covers A of X in place of random covers as above, define $\tilde\pi_F(f)(\omega, A, n)$ by (1.2.12) with $A(\omega) = A \cap \mathcal{E}_\omega$, and proceed as above.

In conclusion observe that the relations between the fiber entropy, the volume growth and the growth in the homology group under the actions of random smooth maps were obtained in [103].

1.3. *Expansivity and topological generators*

In this section we will assume that $\mathbb{P}$ is ergodic invariant measure of ϑ and will deal with the same setup as in Section 1.2, i.e., when random transformations F_ω are continuous and fibers $\mathcal{E}_\omega$ are compact, but some of the results will require all $F_\omega : \mathcal{E}_\omega \to \mathcal{E}_{\vartheta\omega}$ to be homeomorphisms. By this reason we will need here a larger class of metrics than in Section 1.2. Namely, for any positive random variable $\varepsilon = \varepsilon(\omega)$, $x, y \in \mathcal{E}_\omega$, and $m, n \in \mathbb{Z}$, $m < n$ introduce a family of metrics on each $\mathcal{E}_\omega$ by

$$d^\omega_{\varepsilon;m,n}(x, y) = \max_{m\leqslant k<n} \left(d\left(F^k_\omega x, F^k_\omega y\right)\left(\varepsilon\left(\vartheta^k\omega\right)\right)^{-1}\right)$$

allowing m to be negative if F_ω's are invertible (i.e., homeomorphisms in our case). Denote also $d^\omega_{\varepsilon,n} = d^\omega_{\varepsilon;0,n}$, $d^\omega_{\varepsilon,\pm n} = d^\omega_{\varepsilon;-n,n}$. For $Q, E \subset \mathcal{E}_\omega$ we say that E (ω, ε, n)-spans Q if for any $y \in Q$ there is $x \in E$ so that $d^\omega_{\varepsilon,n}(x, y) \leqslant 1$. Denote by $r_n(Q, \omega, \varepsilon)$ the minimum cardinality of a set which (ω, ε, n)-spans Q. If $K \subset \mathcal{E}_\omega$ is compact then $r_n(K, \omega, \varepsilon) < \infty$.

For each compact $K \subset \mathcal{E}_\omega$ and any positive random variable $\varepsilon = \varepsilon(\omega) \in \mathcal{N}$ define

$$r(K, \omega, \varepsilon) = \limsup_{n\to\infty} \frac{1}{n} \log r_n(K, \omega, \varepsilon) \tag{1.3.1}$$

and

$$h_F(K, \omega) = \lim_{\varepsilon \overset{\mathcal{N}}{\to} 0} r(K, \omega, \varepsilon), \tag{1.3.2}$$

where the limit is taken over the directed set $\mathcal{N}$ introduced in Definition 1.2.9 and it exists in view of monotonicity of $r(K, \omega, \varepsilon)$ in ε. Similarly to [99] it follows that the limit remains the same if we take it over only constant $\varepsilon \to 0$. Let $B_x(\omega, \varepsilon, n)$, $B_x(\omega, \varepsilon, \pm n)$, and $B_x(\omega, \varepsilon; m, n)$ be the closed ball in $\mathcal{E}_\omega$ centered at x and having the radius 1 with respect to the metrics $d^\omega_{\varepsilon,n}$, $d^\omega_{\varepsilon,\pm n}$, and $d^\omega_{\varepsilon;m,n}$, respectively. For $d^\omega_{\varepsilon,1}$ and $B_x(\omega, \varepsilon, 1)$ we write simply d^ω_ε and $B_x(\omega, \varepsilon)$.

Set

$$\Gamma_\varepsilon(x, \omega) = \bigcap_{n\geqslant 1} B_x(\omega, \varepsilon, \pm n) \quad \text{or} \quad \Gamma_\varepsilon(x, \omega) = \bigcap_{n\geqslant 1} B_x(\omega, \varepsilon, n)$$

provided we assume or do not assume the invertibility of F_ω's, respectively. Define

$$h^*_F(\omega, \varepsilon) = \sup_{x\in\mathcal{E}_\omega} h_F\big(\Gamma_\varepsilon(x, \omega), \omega\big). \tag{1.3.3}$$

Observe that

$$F^{-1}_\omega \Gamma_\varepsilon(F_\omega x, \vartheta\omega) = \Gamma_\varepsilon(x, \omega),$$

and so

$$r_n\big(\Gamma_\varepsilon(x, \omega), \omega, \delta\big) \geqslant r_{n-1}\big(\Gamma_\varepsilon(F_\omega x, \vartheta\omega), \vartheta\omega, \delta\big)$$

for any positive random variable δ. Hence

$$h_F\big(\Gamma_\varepsilon(x, \omega), \omega\big) \geqslant h_F\big(\Gamma_\varepsilon(F_\omega x, \vartheta\omega), \vartheta\omega\big)$$

and since $F_\omega \mathcal{E}_\omega = \mathcal{E}_{\vartheta\omega}$ we conclude that $\mathbb{P}$-a.s.

$$h^*_F(\omega, \varepsilon) \geqslant h^*_F(\vartheta\omega, \varepsilon)$$

and by ergodicity of ϑ it follows that $h^*_F(\omega, \varepsilon) = \text{const}$ $\mathbb{P}$-a.s. and from now on it will be denoted by $h^*_F(\varepsilon)$. Observe that $h^*_F(\varepsilon)$ is the same for all ε such that $\varepsilon(\omega) \geqslant \operatorname{diam} \mathcal{E}_\omega$ $\mathbb{P}$-a.s. and it coincides then with the fiber topological entropy $h_{\text{top}}(F)$ of F introduced in Definition 1.2.3 via separated sets but the same quantity can be easily obtained employing

spanning sets as above (see [82, Sections 2.2 and 2.3]). In view of monotonicity of $h_F^*(\varepsilon)$ in ε the limit

$$h_F^* = \lim_{\varepsilon \overset{\mathcal{N}}{\to} 0} h_F^*(\varepsilon) \tag{1.3.4}$$

exists and $h_F^* \leqslant h_{\text{top}}(F)$. Similarly to [99] it follows that this limit remains the same if it is taken only over constant $\varepsilon \to 0$.

DEFINITION 1.3.1. F will be called fiber (relative) expansive with an expansivity characteristic $\varepsilon \in \mathcal{N}$ if $\mathbb{P}$-a.s. $\Gamma_\varepsilon(x, \omega) = x$ for all $x \in \mathcal{E}_\omega$. F will be called fiber (relative) h-expansive if $h_F^*(\varepsilon) = 0$ for some $\varepsilon \in \mathcal{N}$ and F will be called fiber (relative) asymptotically h-expansive if $h_F^* = 0$.

DEFINITION 1.3.2. A finite or countable partition $\mathcal{Q} = \{Q_i\}$ of $\mathcal{E}$ into measurable sets Q_i will be called a fiber (relative) topological generator provided $x = y$ whenever $\mathbb{P}$-a.s. $\Theta^n(\omega, x)$ and $\Theta^n(\omega, y)$ belong to the same element of $\mathcal{Q}$ for all $n \in \mathbb{Z}$ or for all $n \in \mathbb{N}$ in the invertible or in the noninvertible case, respectively. In other words, $\mathbb{P}$-a.s. any points $x, y \in \mathcal{E}_\omega$, $x \neq y$ have different $(\omega, \mathcal{Q})$-names.

The emphasis here is that all points $x \neq y$ on the same fiber $\mathcal{E}_\omega$ must be separated. This implies that $\mathcal{Q}$ is a generator for any invariant measure whose marginal on Ω is $\mathbb{P}$ but, in fact, it is a much more stringent requirement which cannot be satisfied in many nonrelative situations. The following result has been proved in [102].

THEOREM 1.3.3. *Suppose that ϑ is aperiodic, F_ω are homeomorphisms, $h_F^* = 0$ and $h_{\text{top}}(F) < \log \ell$ for some integer $\ell > 1$ then there exists a fiber topological generator consisting of ℓ sets.*

The ideologies of the proofs of Theorems 1.1.3 and 1.3.3 are somewhat similar, only in the case of Theorem 1.3.3 we employ ε-entropies $h_F^*(\varepsilon)$ (and not conditional entropies as in Theorem 1.1.3) in order to estimate the number of levels of Rohlin's towers needed to encode points more and more precisely on subsequent steps of our inductive construction. We stress that Theorem 1.3.3 does not have a counterpart in the usual deterministic theory without strong additional assumptions since, for instance, the only generator for the identity transformation is the partition into points. Observe also that the fiber generating partitions constructed in Theorems 1.1.3 and 1.3.3 enable us to represent the system as a random symbolic shift (see Section 4.1).

Next, we proceed to the further study of fiber (relative) topological entropy-like characteristics (introduced in the deterministic case in [129]) which both clarify the condition $h_F^* = 0$ from Theorem 1.3.3 and enable us to derive another important result about the upper semi-continuity of the fiber metric entropy of fiber expansive random continuous transformations. Let $\mathcal{P}(X)$ denote the set of all covers of the space X containing a finite subcover and let $\mathcal{U}(X)$ be the set of all open finite covers of X. For $A \in \mathcal{P}(X)$ we write $A_F^n(\omega) = \bigvee_{k=0}^{n-1} (F_\omega^k)^{-1} A(\vartheta^k \omega)$ where $A(\omega) = A \cap \mathcal{E}_\omega$ is the induced cover of $\mathcal{E}_\omega$

by elements $a \cap \mathcal{E}_\omega$, $a \in A$. For any nonempty set $Y \subset X$ and a cover $A \in \mathcal{P}(X)$ write $N(Y, A) = \min\{\mathrm{card}(C)\colon\ C \subset A,\ Y \subset \bigcup C\}$ and $N(\emptyset, A) = 1$. For $A, B \in \mathcal{P}(X)$ we set also $N(A(\omega)|B(\omega)) = \max_{b\in B} N(b \cap \mathcal{E}_\omega, A)$. As usual, for $A, B \in \mathcal{P}(X)$ we write also $A \succ B$ if A is a refinement of B, i.e., elements of B are unions of some elements of A. Similarly to (1.2)–(1.10) in [129] we have that for any $A, B, C, D \in \mathcal{P}(X)$,

$$N\big(A(\omega)|B(\omega)\big) \leqslant N\big(C(\omega)|D(\omega)\big) \tag{1.3.5}$$

provided $C \succ A$ and $B \succ D$,

$$N\big((F_\omega)^{-1}A(\vartheta\omega)|(F_\omega)^{-1}B(\vartheta\omega)\big) \leqslant N\big(A(\vartheta\omega)|B(\vartheta\omega)\big), \tag{1.3.6}$$

$$N\big(A(\omega) \vee B(\omega)|C(\omega)\big) \leqslant N\big(A(\omega)|C(\omega)\big)N\big(B(\omega)|A(\omega) \vee C(\omega)\big), \tag{1.3.7}$$

$$N\big(A(\omega) \vee B(\omega)|C(\omega) \vee D(\omega)\big) \leqslant N\big(A(\omega)|C(\omega)\big)N\big(B(\omega)|D(\omega)\big), \tag{1.3.8}$$

$$N\big(A(\omega)\big) \leqslant N\big(B(\omega)\big)N\big(A(\omega)|B(\omega)\big), \tag{1.3.9}$$

writing $N(A(\omega))$ in place of $N(A(\omega)|\alpha(\omega))$ provided $\alpha(\omega)$ is the trivial cover of $\mathcal{E}_\omega$, and

$$N\big(A(\omega)|B(\omega)\big) \leqslant N\big(A(\omega)|C(\omega)\big)N\big(C(\omega)|B(\omega)\big). \tag{1.3.10}$$

By (1.3.6) and (1.3.8) it is easy to see that the sequence $a_n(\omega) = \log N(A_F^n(\omega)|B_F^n(\omega))$ is subadditive, and so the subadditive ergodic theorem (see, for instance, [82]) implies that $\mathbb{P}$-a.s. the limit

$$h_F(A|B) = \lim_{n\to\infty} \frac{1}{n} \log N\big(A_F^n(\omega)|B_F^n(\omega)\big) \tag{1.3.11}$$

exists, it is not random and

$$h_F(A|B) \leqslant \int \log N\big(A(\omega)|B(\omega)\big)\, d\mathbb{P}(\omega). \tag{1.3.12}$$

The number $h_F(A|B)$ will be called the fiber conditional entropy of F on a cover A with respect to a cover B. If B is the trivial cover and A is an open cover of X then $h_F(A|B)$ is the fiber topological entropy $h_{\mathrm{top}}(F, A)$ of F with respect to the cover A, and so by (1.3.5),

$$h_{\mathrm{top}}(F, A) \geqslant h_F(A|B) \tag{1.3.13}$$

for any $B \in \mathcal{P}(X)$.

Observe that by (1.3.5),

$$h_F(A|B) \leqslant h_F(C|D) \quad \text{if } C \succ A \text{ and } B \succ D. \tag{1.3.14}$$

Thus there exists a limit (which may be finite or infinite) over the directed set $\mathcal{U}(X)$,

$$h_F(B) = \lim_{A\in\mathcal{U}(X)} h_F(A|B) = \sup_{A\in\mathcal{U}(X)} h_F(A|B) \tag{1.3.15}$$

which will be called the fiber conditional entropy of F with respect to a cover B. By (1.3.14),

$$h_F(A) \leqslant h_F(B) \quad \text{for } A \succ B, \tag{1.3.16}$$

and so we can take the limit once more

$$h_F = \lim_{B \in \mathcal{U}(X)} h_F(B) = \inf_{B \in \mathcal{U}(X)} h_F(B) \tag{1.3.17}$$

which is called the fiber conditional entropy of F. In view of (1.3.13),

$$h_{\text{top}}(F) \geqslant h_F. \tag{1.3.18}$$

Observe also that writing $A_F^n(\omega)$ and $B_F^n(\omega)$ in (1.3.9) in place of $A(\omega)$ and $B(\omega)$, respectively, then taking log, dividing by n and passing to the limit as $n \to \infty$ we obtain

$$h_{\text{top}}(F, A) \leqslant h_{\text{top}}(F, B) + h_F(A|B). \tag{1.3.19}$$

Taking here supremum over $A \in \mathcal{U}(X)$ we have

$$h_{\text{top}}(F) \leqslant h_{\text{top}}(F, B) + h_F(B). \tag{1.3.20}$$

Similarly, we obtain also from (1.3.10) that

$$h_F(A|B) \leqslant h_F(A|C) + h_F(C|B). \tag{1.3.21}$$

The following result (proved in [102]) connects $h_F^*(\varepsilon)$ introduced in the first part of this section with fiber conditional entropies $h_F(A)$.

PROPOSITION 1.3.4. *Let $\varepsilon \in \mathcal{N}$ and $A, E \in \mathcal{U}(X)$ satisfy* $\operatorname{diam} A(\omega) \leqslant \varepsilon(\omega) < \frac{1}{2} L(\omega)$ *where $L = L(\omega)$ is a random variable such that $L(\omega)$ is a Lebesgue number* (*see, for instance*, [82]) *of the cover $E(\omega)$. Then*

$$h_F(A) \leqslant h_F^*(\varepsilon) \leqslant h_F(E). \tag{1.3.22}$$

It follows that F is asymptotically fiber h-expansive if and only if $h_F = 0$.

In view of (1.3.18) it follows, in particular, that the asymptotic fiber h-expansivity assumption of Theorem 1.3.3 holds true if the fiber topological entropy of F is zero.

Let $A = \{a_1, \dots, a_r\}$ be a finite Borel partition of X then we can choose compact sets $b_i \subset a_i$, $i = 1, \dots, r$, such that

$$\int \mu_\omega\big(a_i(\omega) \setminus b_i(\omega)\big)\, d\mathbb{P}(\omega) < \varepsilon,$$

where $a_i(\omega) = a_i \cap \mathcal{E}_\omega$ and $b_i(\omega) = b_i \cap \mathcal{E}_\omega$. Then setting $b_0 = X \setminus \bigcup_{i=1}^r b_i$ we obtain a partition $B = \{b_0, b_1, \dots, b_r\}$ of X and (see the proof of (2.5) in [99]),

$$h_\mu(F, \tilde{A}) \leqslant h_\mu(F, \tilde{B}) + 1, \tag{1.3.23}$$

where $\tilde{A} = \{\tilde{a}_1, \dots, \tilde{a}_r\}$ and $\tilde{B} = \{\tilde{b}_0, \tilde{b}_1, \dots, \tilde{b}_r\}$ are partitions of $\mathcal{E}$ into sets $\tilde{a}_i = \mathcal{E} \cap (\Omega \times a_i)$ for $i = 1, \dots, r$ and $\tilde{b}_i = \mathcal{E} \cap (\Omega \times b_i)$ for $i = 0, 1, \dots, r$. Observe that $C = \{c_1, \dots, c_r\}$ with $c_i = b_0 \cup b_i$, $i = 1, \dots, r$, is an open cover of X and

$$N\big(B(\omega)|C(\omega)\big) \leqslant 2, \tag{1.3.24}$$

where $B(\omega) = B \cap \mathcal{E}_\omega$ and $C(\omega) = C \cap \mathcal{E}_\omega$ are corresponding induced covers of $\mathcal{E}_\omega$. Employing (1.3.12), (1.3.21), (1.3.23) and (1.3.24) together with standard properties of the conditional entropy (see, for instance, [82, Section 2.1]) we derive similarly to Proposition 4.2 in [129] that

$$h_\mu(F) \leqslant h_\mu(F, D) + h_F(D) \tag{1.3.25}$$

for any $\mu \in \mathcal{I}_\mathbb{P}(\mathcal{E})$ and each finite partition $D = \{d_1, \dots, d_r\}$ of $\mathcal{E}$ obtained from a finite Borel partition $\tilde{D} = \{\tilde{d}_1, \dots, \tilde{d}_r\}$ of X by taking $d_i = \mathcal{E} \cap (\Omega \times \tilde{d}_i)$.

Now we are ready to derive the upper semi-continuity of the metric fiber entropy $h_\cdot(F)$: $\mathcal{I}_\mathbb{P}(\mathcal{E}) \to \bar{\mathbb{R}}$ with respect to the narrow topology on $\mathcal{P}_\mathbb{P}(\mathcal{E})$ where a convergence $\nu \to \mu$ means that $\int f\, d\nu \to \int f\, d\mu$ for any $f \in \mathbb{L}^1_\mathcal{E}(\Omega, \mathcal{C}(X))$ (see Chapter 3 in [49]). Set $h^*_\mu(F) = \limsup_{\nu \to \mu} h_\nu(F) - h_\mu(F)$.

THEOREM 1.3.5. *Let F be a continuous bundle RDS* (*i.e., all F_ω are continuous maps*) *and $\mu \in \mathcal{I}_\mathbb{P}(\mathcal{E})$. Then*

$$h^*_\mu(F) \leqslant h^*_F. \tag{1.3.26}$$

In particular, if F is fiber asymptotically h-expansive (*i.e., $h^*_F = 0$*) *then $h_\mu(F)$ as a function of $\mu \in \mathcal{P}_\mathbb{P}(\mathcal{E})$ is upper semi-continuous in the narrow topology on $\mathcal{I}_\mathbb{P}(\mathcal{E})$. Hence, in this case there exists a maximizing measure in the variational principle* (1.2.11) (*which is called an equilibrium state*).

PROOF. Assume, first, that $h_{\text{top}}(F) = \infty$ and show that in this case $h^*_F = \infty$, as well, implying (1.3.26). Indeed, by (1.3.20) this would follow if $h_{\text{top}}(F, B) < \infty$ for $B \in \mathcal{U}(X)$. But by the subadditivity arguments

$$\begin{aligned} h_{\text{top}}(F, B) &= \lim_{n\to\infty} \frac{1}{n} \int \log N\big(B^n_F(\omega)\big)\, d\mathbb{P}(\omega) = \inf_{n \geqslant 1} \frac{1}{n} \int \log N\big(B^n_F(\omega)\big)\, d\mathbb{P}(\omega) \\ &\leqslant \int \log N\big(B(\omega)\big)\, d\mathbb{P}(\omega) < \infty \end{aligned}$$

as $N(B(\omega))$ is bounded by the number of elements in $B \in \mathcal{U}(X)$. Now assume that $h_{\text{top}}(F) < \infty$, and so by the variational principle Theorem 1.2.13 all fiber metric entropies

are finite, as well. Let $A = \{a_1, \dots, a_r\} \in \mathcal{U}(X)$. Take a cover $B = \{b_1, \dots, b_r\} \in \mathcal{U}(X)$ such that $\bar{b}_i \subset a_i$ for $i = 1, \dots, r$. Then take continuous functions $\varphi_i : X \to [0, 1]$ such that $\varphi_i(x) = 1$ for $x \in \bar{b}_i$ and $\varphi_i(x) = 0$ for $x \notin a_i$, $i = 1, \dots, r$. For some $\alpha \in [0, 1]$ we have

$$\int \mu_\omega \left(\mathcal{E}_\omega \cap \bigcup_{i=1}^{r} \varphi_i^{-1}(\alpha) \right) d\mathbb{P}(\omega) = 0.$$

Hence the finite Borel partition $C = \{c_1, \dots, c_r\}$ of X consisting of sets $c_1 = \varphi_1^{-1}[\alpha, 1]$, $c_2 = \varphi_2^{-1}[\alpha, 1] \setminus \varphi_1^{-1}[\alpha, 1], \dots, c_r = \varphi_r^{-1}[\alpha, 1] \setminus \bigcup_{i=1}^{r-1} \varphi_i^{-1}[\alpha, 1]$ satisfies

$$\int \mu_\omega \left(\mathcal{E}_\omega \cap \bigcup_{i=1}^{r} \partial c_i \right) d\mathbb{P}(\omega) = 0.$$

We also have that $a_i \supset c_i$, $i = 1, \dots, r$. For each fixed n we have also

$$\int \mu_\omega \big(\partial C_F^{(n)}(\omega) \big) \, d\mathbb{P}(\omega) = 0.$$

It follows from (1.1.1) and Theorem 3.17 in [49] that there exists an open in the narrow topology neighborhood $\mathcal{U}_n \subset \mathcal{I}_{\mathbb{P}}(\mathcal{E})$ of μ such that

$$\frac{1}{n} H_\mu \big(C_F^{(n)} | \mathcal{F}_\mathcal{E} \big) \geqslant \frac{1}{n} H_\nu \big(C_F^{(n)} | \mathcal{F}_\mathcal{E} \big) - \varepsilon$$

whenever $\nu \in \mathcal{U}_n$. Thus by (1.3.16), (1.3.25) and the subadditivity of the conditional entropy (see [142, §5] or [82, §2.1]),

$$\begin{aligned} h_\mu(F) \geqslant h_\mu(F, C) &= \lim_{n \to \infty} \frac{1}{n} H_\mu \big(C_F^{(n)} | \mathcal{F}_\mathcal{E} \big) \\ &\geqslant \limsup_{n \to \infty} \sup_{\nu \in \mathcal{U}_n} \frac{1}{n} H_\nu \big(C_F^{(n)} | \mathcal{F}_\mathcal{E} \big) - \varepsilon \geqslant \limsup_{n \to \infty} \sup_{\nu \in \mathcal{U}_n} h_\nu(F, C) - \varepsilon \\ &\geqslant \limsup_{n \to \infty} \sup_{\nu \in \mathcal{U}_n} h_\nu(F) - h_F(C) - \varepsilon \geqslant \limsup_{\nu \to \mu} h_\nu(F) - \varepsilon \end{aligned}$$

taking into account that $C \succ A$. Since ε is arbitrary we obtain $h_\mu^*(F) \leqslant h_F(A)$ and since A is arbitrary, as well, (1.3.26) follows. □

2. Smooth RDS: Invariant manifolds

2.1. *Smooth RDS*

Set-up. In this section we discuss local, differential properties of a smooth RDS. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\vartheta : (\Omega, \mathcal{F}, \mathbb{P}) \hookleftarrow$ a measure-preserving transformation which represents the time evolution of the driving noise system. Note that in this

section ϑ *will not be assumed invertible* except indicated otherwise. We will always assume that M is a smooth compact Riemannian manifold without boundary. Let $\mathrm{Diff}^r(M)$ ($r \geqslant 1$ is an integer) be the space of C^r diffeomorphisms of M, endowed with the usual C^r topology (see [70]). Let now

$$F : \Omega \to \mathrm{Diff}^r(M), \quad \omega \mapsto F_\omega$$

be a Borel measurable map and consider the RDS, denoted by the same notation F, over $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ generated by the random transformation $\{F_\omega\colon\ \omega \in \Omega\}$. To remark, F is a bundle RDS, as defined in Section 1.1, with $\mathcal{E} = \Omega \times M$. The Borel σ-algebra of M will be denoted by $\mathcal{B}$, and

$$\Theta : \Omega \times M \to \Omega \times M, \quad (\omega, x) \mapsto (\vartheta\omega, F_\omega x)$$

is the corresponding skew product transformation.

Polish noise systems. Unlike in the other sections, in this and the next one we sometimes have to be a little more careful about the measurable structure of $(\Omega, \mathcal{F}, \mathbb{P})$. Specifically, we will assume for some cases that Ω is a Polish space (i.e., a separable topological space with a complete metric) and $\mathcal{F}$ is its Borel σ-algebra (when this assumption is made, we will indicate explicitly). In such a case $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ will be called a *Polish system*. This assumption arises due partially to the need to work with Lebesgue spaces and their measurable partitions (see [142] for a detailed treatment of this topic) for many of our purposes. A measure-theoretic result states that a Polish space with a Borel probability measure and with the completion of its Borel σ-algebra with respect to this measure constitutes a Lebesgue space (see also [142]). Another nice property of Polish spaces is Lusin's theorem which says that a measurable function is in fact continuous outside of a set of arbitrarily small measure. More precisely, let X be a Polish space, E a topological space with a countable basis, μ a Borel probability measure on X, and $f : X \to E$ a Borel map. Then for each $\varepsilon > 0$ there is a compact set $K \subset X$ such that $\mu(X \setminus K) \leqslant \varepsilon$ and $f|_K$ is continuous (see [21]). This property will enable us to prove some kind of continuous dependence of local stable or unstable manifolds of an RDS F on $(\omega, x) \in \Omega \times M$, which is necessary for constructing suitable measurable partitions needed in the treatment of entropy formulae and SRB measures. We further remark that, when only the Borel σ-algebra of a Polish space is concerned, one can just use part of the topology and may treat it as a compact metric space. In fact, if $(\Omega, \mathcal{O})$ is a Polish space with topology $\mathcal{O}$, then there exists a topology $\mathcal{O}' \subset \mathcal{O}$ such that $(\Omega, \mathcal{O}')$ is compact metrizable and has the same Borel σ-algebra as $(\Omega, \mathcal{O})$ [71]. This fact is a fundamental structural property of Polish spaces and it is very helpful to overcome difficulties brought about by the noncompactness of Ω (see [12,121]).

Examples. The following are some examples of Polish measure-preserving systems which often serve as canonical models of driving noise in the theory of RDS.

(i) Here $\Omega = \mathrm{Diff}^r(M)^{\mathbb{Z}^+}$ or $\Omega = \mathrm{Diff}^r(M)^{\mathbb{Z}}$ ($r \geqslant 1$ is an integer, cf. Remark 2.2.19), endowed with the product topology. It is well known that Ω with this topology is a Polish space (see, e.g., [118, Chapter V]). Now $\vartheta : \Omega \to \Omega$ is the left shift operator, i.e.,

$(\vartheta\omega)_n = \omega_{n+1}$ for $\omega = (\omega_n) \in \Omega$, $n \in \mathbb{Z}^+$ or $\mathbb{Z}$, and $\mathbb{P}$ is a Borel probability measure on Ω which is invariant with respect to ϑ. This corresponds to the usual product of random diffeomorphisms by considering $F : \Omega \to \mathrm{Diff}^r(M)$, $\omega \mapsto \omega_0$. Specifically, the case $\mathbb{P} = \nu^{\mathbb{Z}^+}$ or $\mathbb{P} = \nu^{\mathbb{Z}}$, where ν is a Borel probability measure on $\mathrm{Diff}^r(M)$, corresponds to the classical product of i.i.d. (independent and identically distributed) random diffeomorphisms [82].

(ii) Consider the classical Wiener space $(\Omega, \mathcal{F}, \mathbb{P})$. Here $\Omega = \{\omega\colon\ \omega(\cdot) \in C(\mathbb{R}^+, \mathbb{R}^d),\ \omega(0) = 0\}$ (for the one-sided time case) or $\Omega = \{\omega\colon\ \omega(\cdot) \in C(\mathbb{R}, \mathbb{R}^d),\ \omega(0) = 0\}$ (for the two-sided time case), endowed with the metric

$$d(\omega, \omega') = \sum_{n=1}^{+\infty} 2^{-n} \Big(\sup_{|t| \leqslant n} \big|\omega(t) - \omega'(t)\big| \wedge 1 \Big), \quad \omega, \omega' \in \Omega,$$

which makes Ω a Polish space, $\mathcal{F}$ is its Borel σ-algebra and $\mathbb{P}$ is the Wiener measure on Ω. Now for each fixed $t \neq 0$, $\vartheta^t : \Omega \to \Omega$ defined by

$$\big(\vartheta^t \omega\big)(\cdot) = \omega(t + \cdot) - \omega(t), \quad \omega \in \Omega,$$

preserves $\mathbb{P}$ and it is ergodic. The coordinate process on $(\Omega, \mathcal{F}, \mathbb{P})$ describes the classical Brownian motion and $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta^1)$ serves as a canonical noise model driving the RDS generated by the time discretization of a stochastic flow of diffeomorphisms arising from the solution of a stochastic differential equation (see [7, Chapter 2] for details).

STANDING HYPOTHESES FOR SECTION 2.

- F is a C^r ($r \geqslant 1$ integer) RDS over $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$, as defined just above;
- μ is an invariant measure of F, i.e., μ is a probability on $(\Omega \times M, \mathcal{F} \times \mathcal{B})$ which is Θ-invariant and which has marginal $\mathbb{P}$ on Ω. Its disintegration μ_ω, $\omega \in \Omega$ will be called the *sample measures* of μ.

2.2. *Stable invariant manifolds*

Lyapunov exponents. By a trivialization argument on the measure-preserving system $\Theta : (\Omega \times M, \mu) \hookleftarrow$, one can easily have the following reformulation of the celebrated Oseledec multiplicative ergodic theorem (MET) [131,147].

THEOREM 2.2.1. *Assume that F is of class C^1 (i.e., $r = 1$) and assume the integrability condition*

$$\int \log^+ |D_x F_\omega| \, d\mu(\omega, x) < +\infty \tag{2.2.1}$$

(*where* $\log^+ a := \max\{\log a, 0\}$). *Then there exists a measurable set $\Delta_0 \subset \Omega \times M$ such that $\mu(\Delta_0) = 1$ and for each $(\omega, x) \in \Delta_0$ there are numbers*

$$-\infty \leqslant \lambda^{(1)}(\omega, x) < \lambda^{(2)}(\omega, x) < \cdots < \lambda^{(r(\omega,x))}(\omega, x) < +\infty$$

(*measurable in* (ω, x)) *and an associated nested sequence of subspaces of* $T_x M$,

$$\{0\} = V^{(0)}(\omega, x) \subset V^{(1)}(\omega, x) \subset \cdots \subset V^{(r(\omega,x))}(\omega, x) = T_x M$$

(*also measurable in* (ω, x)) *satisfying*

$$\lim_{n\to+\infty} \frac{1}{n} \log |D_x F_\omega^n \xi| = \lambda^{(i)}(\omega, x)$$

for $\xi \in V^{(i)}(\omega, x) \setminus V^{(i-1)}(\omega, x)$, $1 \leqslant i \leqslant r(\omega, x)$.

Each $\lambda^{(i)}(\omega, x)$ is called a *Lyapunov exponent* of F at (ω, x),

$$m^{(i)}(\omega, x) := \dim V^{(i)}(\omega, x) - \dim V^{(i-1)}(\omega, x)$$

is called its *multiplicity*, and $\{(\lambda^{(i)}(\omega, x), m^{(i)}(\omega, x))\colon 1 \leqslant i \leqslant r(\omega, x)\}$ the *Lyapunov spectrum* of F at (ω, x). The reader is referred to [7,8,23] and the survey [117] for many other interesting results on Lyapunov exponents of RDS, especially of i.i.d. RDS.

Stable manifolds: introductory remarks. Lyapunov exponents describe the exponential growth rates of the norms of vectors under successive actions of derivatives of the random diffeomorphisms. The invariant manifold theory is a nonlinear counterpart of the linear theory of Lyapunov exponents. We first give a rough description of this theory. Corresponding to each negative Lyapunov exponent $\lambda^{(i)}(\omega, x) < 0$ at a typical point (ω, x), one can consider the collection of points of the manifold whose orbits under successive actions of the random maps will cluster with that of x at least with the same exponential rate as the exponent $\lambda^{(i)}(\omega, x)$, i.e.,

$$W^{s,i}(\omega, x) := \left\{ y \in M\colon \limsup_{n\to+\infty} \frac{1}{n} \log d(F_\omega^n x, F_\omega^n y) \leqslant \lambda^{(i)}(\omega, x) \right\}, \tag{2.2.2}$$

where $d(\cdot, \cdot)$ is the metric on M induced by the Riemannian norm. This collection of points is clearly invariant in the sense that

$$F_\omega W^{s,i}(\omega, x) = W^{s,i}(\vartheta\omega, F_\omega x).$$

However, the crucial thing here is that, if F is of class C^2 (or $C^{1+\alpha}$ with $\alpha \in (0, 1]$) and some integrability condition on the C^2 (or $C^{1+\alpha}$)-norm of the random maps F_ω is satisfied, the set $W^{s,i}(\omega, x)$, as a subset of the manifold M, can be shown to have a nice submanifold structure (it is actually the image of $V^{(i)}(\omega, x)$ under an injective immersion of at least class C^1) and hence will be called the *stable manifold* of F at (ω, x) associated with the exponent $\lambda^{(i)}(\omega, x)$. The stable manifolds have another important property, i.e., the so-called "absolute continuity property" which says that maps defined by sliding along the $W^{s,i}$-manifolds between two discs transverse to these submanifolds take Lebesgue zero sets to Lebesgue zero sets. In the case of a deterministic C^2 (or $C^{1+\alpha}$) diffeomorphism preserving an absolutely continuous (with respect to the Lebesgue on the manifold) reference

measure, construction of such invariant families of submanifolds, via the crucial technical devices of the Lyapunov metric (or norm) and regular neighborhoods, together with the treatment of the absolute continuity property constitutes a substantial part of what is often called Pesin theory, due to the landmark works of Pesin in the mid-seventies (see [134] for a systematic account of this theory). This invariant manifold treatment paved the way to a series of deep ergodic–theoretic results of the system, and it (especially the absolute continuity property) serves as a bridge which allows us to pass from local differential properties of the dynamics to the study of its global metric behaviors. For example, one of the most striking of these results is the so-called Pesin entropy formula which expresses the entropy of the dynamical system through its Lyapunov exponents (see [134] and see also [110] for a substantial development of this theory). What we described above is a partial version for the RDS F of Pesin's stable manifold treatment, and in this section we will present the related results in a more detailed way. Before doing this, we first remark that our presentation below is more orientated towards dealing with entropy formula of Pesin type for the RDS F (in Section 3) and is adopted from [118, Chapter III] (see the references therein and [51] for an account of related works), which goes mainly along Pesin's original scheme (a combination of ergodic theory, contraction principle of Banach spaces and graph transformation method). We indicate that [118, Chapter III] deals with only the i.i.d. case but the treatment is identical for the general stationary case. Ruelle's alternative approach [147], which makes use of a perturbation method and is more analytic in nature, can also be adapted to the random case (see [43]). In fact, this latter approach has the advantage of an easier adaptability to the random case as well as to Hilbert spaces [149,153], but an adaption of it suitable for the purpose of dealing with entropy formula does not yet exist. We also remark that invariant manifolds have been constructed for RDS under various conditions and for various purposes by many authors, we refer the reader to [7, Chapter 7] for a comprehensive account.

We now proceed to the more precise statement of the related results. In the rest of Section 2.2 we will always assume that F is of class C^2 (i.e., $r = 2$) and satisfies

$$\int \log^+ |F_\omega|_{C^2}\, d\mathbb{P}(\omega) < +\infty, \tag{2.2.3}$$

where $|f|_{C^2}$ is the usual C^2-norm of $f \in \mathrm{Diff}^2(M)$ (see, e.g., [118, Chapter II] for the definition). See Remark 2.2.18 for a discussion about the case of F being $C^{1+\alpha}$.

Lyapunov norms and regular neighborhoods. Let $\Delta_0 \subset \Omega \times M$ be the Oseledec forward regular set of (F, μ) as introduced in Theorem 2.2.1 (note that (2.2.3) implies (2.2.1)), and put

$$\Delta = \big\{(\omega, x) \in \Delta_0 \colon \lambda^{(1)}(\omega, x) > -\infty\big\}. \tag{2.2.4}$$

Let $[a, b]$, $a < b \leqslant 0$, be a closed interval of $\mathbb{R}$. Define

$$\Delta_{a,b} = \big\{(\omega, x) \in \Delta \colon \lambda^{(1)}(\omega, x) < a \text{ and } \lambda^{(i)}(\omega, x) \notin [a,\ b] \text{ for all } i\big\}$$

and assume that $\Delta_{a,b} \neq \emptyset$. For $(\omega, x) \in \Delta_{a,b}$ and $n, l \in \mathbb{Z}^+$, set

$$E_0(\omega, x) = \bigcup_{\lambda^{(i)}(\omega,x)<a} V^{(i)}(\omega, x), \qquad H_0(\omega, x) = E_0(\omega, x)^{\perp},$$

$$E_n(\omega, x) = D_x F_\omega^n E_0(\omega, x), \quad H_n(\omega, x) = D_x F_\omega^n H_0(\omega, x), \quad n > 0,$$

$$F_\omega^{n,0} = \mathrm{id}, \qquad F_\omega^{n,l} = F_{\vartheta^{n+l-1}\omega} \circ \cdots \circ F_{\vartheta^n \omega}, \quad l > 0,$$

$$S_{(\omega,x)}^{n,l} = \left[D_{F_\omega^n x} F_\omega^{n,l}\right]\big|_{E_n(\omega,x)}, \qquad U_{(\omega,x)}^{n,l} = \left[D_{F_\omega^n x} F_\omega^{n,l}\right]\big|_{H_n(\omega,x)}$$

(note that in general $H_0(\omega, x)$ may not be invariant, i.e., $D_x F_\omega H_0(\omega, x) \neq H_0(\Theta(\omega, x))$). Fix arbitrarily $0 < \varepsilon < (b-a)/200$ which is sufficiently small as compared with $b - a$ (which is something like spectral gap).

Let now $(\omega, x) \in \Delta_{a,b}$. The vectors in $E_0(\omega, x)$ and $H_0(\omega, x)$ have respectively exponential growth rates smaller than a and bigger than b under actions of $D_x F_\omega^n$, this implies that the forward orbit of x under actions of F_ω^n, $n \geqslant 0$, has a kind of nonuniform partial hyperbolicity. Actually, there is the following

LEMMA 2.2.2 (Nonuniform partial hyperbolicity). *There exists a measurable function $l : \Delta_{a,b} \times \mathbb{Z}^+ \to (0, +\infty)$ such that for μ-a.e. $(\omega, x) \in \Delta_{a,b}$ and for all $n, l \in \mathbb{Z}^+$ one has*

$$l(\omega, x, n + l) \leqslant l(\omega, x, n) e^{\varepsilon l} \tag{2.2.5}$$

and

(1) $|S_{(\omega,x)}^{n,l} \xi| \leqslant l(\omega, x, n) e^{(a+\varepsilon)l} |\xi|$ *for* $\xi \in E_n(\omega, x)$;

(2) $|U_{(\omega,x)}^{n,l} \eta| \geqslant l(\omega, x, n)^{-1} e^{(b-\varepsilon)l} |\eta|$ *for* $\eta \in H_n(\omega, x)$;

(3) $\gamma(E_{n+l}(\omega, x), H_{n+l}(\omega, x)) \geqslant l(\omega, x, n)^{-1} e^{-\varepsilon l}$, *where $\gamma(\cdot, \cdot)$ denotes the angle between the two related subspaces.*

The inequality (2.2.5) implies that the estimates (1)–(3) of Lemma 2.2.2 may get worse along the trajectory (when n, l increase) but relatively very slowly (i.e., with an exponential rate which is very small as compared with the spectral gap $b - a$). Such things play a basic role in the study of nonuniform (partial) hyperbolicity and they are actually consequences of ergodic theorems.

The next important step is to introduce a sequence of new norms $\|\cdot\|_{(\omega,x),n}$, $n \geqslant 0$, on $T_{F_\omega^n x} M = E_n(\omega, x) \oplus H_n(\omega, x)$, called *Lyapunov norms* (*or metric*), along the forward orbit of μ-a.e. $(\omega, x) \in \Delta_{a,b}$ so that these norms do not change "seriously" the original Riemannian norm (i.e., the difference increases with n at most with an exponential rate which is very small as compared with $b - a$) and some kind of uniform (partial) hyperbolicity can be read by means of them. This is done in the following lemma.

LEMMA 2.2.3 (Lyapunov norms). *For μ-a.e. $(\omega, x) \in \Delta_{a,b}$ one can define a sequence of norms $\|\cdot\|_{(\omega,x),n}$, $n \geqslant 0$ (depending measurably on (ω, x)), on $T_{F_\omega^n x} M = E_n(\omega, x) \oplus H_n(\omega, x)$ such that*

(1) $\|S^{n,l}_{(\omega,x)}\xi\|_{(\omega,x),n+1} \leqslant e^{a+2\varepsilon}\|\xi\|_{(\omega,x),n}$ *for* $\xi \in E_n(\omega,x)$;
(2) $\|U^{n,l}_{(\omega,x)}\eta\|_{(\omega,x),n+1} \geqslant e^{b-2\varepsilon}\|\eta\|_{(\omega,x),n}$ *for* $\eta \in H_n(\omega,x)$;
(3) $\frac{1}{2}|\zeta| \leqslant \|\zeta\|_{(\omega,x),n} \leqslant A_\varepsilon l(\omega,x,0)^2 e^{2\varepsilon n}|\zeta|$ *for* $\zeta \in T_{F^n_\omega x}M$, *where* A_ε *is a constant depending only on* ε.

To construct stable manifolds, another important device is the following regular neighborhoods lemma, which says that, for μ-a.e. $(\omega,x) \in \Delta_{a,b}$, along the forward orbit $\{F^n_\omega x\}_{n=0}^{+\infty}$ there exists a sequence of neighborhoods $N_{(\omega,x),n}$ of $F^n_\omega x$ whose sizes vary with n relatively very slowly (in a similar sense as before) and on which $F_{\vartheta^n\omega}$ act like very much their derivatives at $F^n_\omega x$ (with respect to the Lyapunov norms). A condition like (2.2.3) rather than one on the C^1-norm of the random maps is perhaps necessary for this lemma.

LEMMA 2.2.4 (Regular neighborhoods). *There exists a measurable function* $r : \Delta_{a,b} \to (0,+\infty)$ *such that for* μ*-a.e.* $(\omega,x) \in \Delta_{a,b}$, *all* $n \in \mathbb{Z}^+$ *and any given* $0 < \delta \leqslant 1$ *the map*

$$\Phi_{(\omega,x),n} := \exp^{-1}_{F^{n+1}_\omega x} \circ F_{\vartheta^n\omega} \circ \exp_{F^n_\omega x} : \left\{\xi \in T_{F^n_\omega x}M : \|\xi\|_{(\omega,x),n} \leqslant \delta r(\omega,x)e^{-3\varepsilon n}\right\} \to T_{F^{n+1}_\omega x}M$$

is well defined and

$$\mathrm{Lip}_{\|\cdot\|}(\Phi_{(\omega,x),n} - D_{F^n_\omega x}F_{\vartheta^n\omega}) \leqslant \delta,$$

where $\mathrm{Lip}_{\|\cdot\|}(\cdot)$ *denotes the Lipschitz constant taken with respect to* $\|\cdot\|_{(\omega,x),n}$ *and* $\|\cdot\|_{(\omega,x),n+1}$.

Local stable manifolds. Then, using the standard graph transformation methods (see, e.g., [61, Appendix] or [118, Chapter III]), one can construct for each typical $(\omega,x) \in \Delta_{a,b}$ a sequence of local stable manifolds associated with $\lambda^{(i_a)}(\omega,x)$, the largest Lyapunov exponent smaller than a, along the forward orbit of x under actions of F^n_ω. The results are summarized in the following theorem.

THEOREM 2.2.5 (Local stable manifolds). *For* μ*-a.e.* $(\omega,x) \in \Delta_{a,b}$ *there exists a sequence of* $C^{1,1}$ *embedded* $\dim E_0(\omega,x)$*-dimensional discs* $\{W^{s,i_a}_{\mathrm{loc}}((\omega,x),n)\}_{n=0}^{+\infty}$ *in* M *together with positive numbers* $\alpha(\omega,x)$, $\beta(\omega,x)$ *and* $\gamma(\omega,x)$ (*all measurable in* (ω,x), *possibly depending on* a, b *and* ε) *such that the following hold true for each* $n \in \mathbb{Z}^+$:

(1) *There exists a* $C^{1,1}$ *map*

$$h_{(\omega,x),n} : O_n(\omega,x) \to H_n(\omega,x),$$

where $O_n(\omega,x)$ *is an open subset of* $E_n(\omega,x)$ *which contains* $\{\xi \in E_n(\omega,x) : |\xi| < \alpha(\omega,x)e^{-5\varepsilon n}\}$, *satisfying*

(i) $h_{(\omega,x),n}(0) = 0$;

(ii) $\mathrm{Lip}(h_{(\omega,x),n}) \leqslant \beta(\omega,x)e^{7\varepsilon n}$, $\mathrm{Lip}(D.h_{(\omega,x),n}) \leqslant \beta(\omega,x)e^{7\varepsilon n}$ *(with respect to* $|\cdot|$*)*;

(iii) $W^{s,i_a}_{\mathrm{loc}}((\omega,x),n) = \exp_{F^n_\omega x}\mathrm{Graph}(h_{(\omega,x),n})$ *and it is tangent to* $E_n(\omega,x)$ *at point* $F^n_\omega x$.

(2) $F_{\vartheta^n\omega}W^{s,i_a}_{\mathrm{loc}}((\omega,x),n) \subset W^{s,i_a}_{\mathrm{loc}}((\omega,x),n+1)$.

(3) $d^s(F^{n,l}_\omega y, F^{n,l}_\omega z) \leqslant \gamma(\omega,x)e^{2\varepsilon n}e^{(a+4\varepsilon)l}d^s(y,z)$, $y,z \in W^{s,i_a}_{\mathrm{loc}}((\omega,x),n)$, $l \in \mathbb{Z}^+$, *where* $d^s(\,,\,)$ *denotes the distance along* $W^{s,i_a}_{\mathrm{loc}}((\omega,x),m)$ *for* $m \in \mathbb{Z}^+$ *induced by their inherited Riemannian metric as submanifolds of* M.

From now on we will use the following notations:

$$\begin{aligned} W^{s,i_a}_{\mathrm{loc}}(\omega,x) &:= W^{s,i_a}_{\mathrm{loc}}\big((\omega,x),0\big), \\ E^{s,i_a}(\omega,x) &:= E_0(\omega,x) = V^{(i_a)}(\omega,x). \end{aligned} \tag{2.2.6}$$

In case of $(\Omega,\mathcal{F},\mathbb{P},\vartheta)$ being a Polish system, some kind of continuous dependence of the disc $W^{s,i_a}_{\mathrm{loc}}(\omega,x)$ on (ω,x) can be obtained and it will be needed to deal with entropy formula problems (i.e., to construct suitable measurable partitions of $\Omega \times M$ subordinate to stable or unstable manifolds of the RDS (F,μ)). To describe precisely this kind of continuity property, we recall the definition of a continuous family of C^1 embedded k-dimensional discs.

DEFINITION 2.2.6. Let X be a metric space and let $\{D_x\}_{x\in X}$ be a collection of subsets of M. We call $\{D_x\}_{x\in X}$ a *continuous family of* C^1 *embedded* k*-dimensional discs in* M if there is a finite open cover $\{U_i\}_{i=1}^l$ of X such that for each U_i there exists a continuous map $\theta_i : U_i \to \mathrm{Emb}^1(D^k, M)$ such that $\theta_i(x)D^k = D_x$ for each $x \in U_i$, where $D^k = \{\xi \in \mathbb{R}^k\colon \|\xi\| < 1\}$.

With the help of Lusin's theorem, one further obtains (see [12])

COMPLEMENT TO THEOREM 2.2.5 (Continuous dependence). *Assume moreover that* $(\Omega,\mathcal{F},\mathbb{P},\vartheta)$ *is a Polish system and endow* $\Omega \times M$ *with the product metric. Given* k, $1 \leqslant k \leqslant \dim M$, *put* $\Delta_{a,b,k} = \{(\omega,x) \in \Delta_{a,b}\colon \dim E^{s,i_a}(\omega,x) = k\}$. *Then the submanifolds* $W^{s,i_a}_{\mathrm{loc}}(\omega,x)$ *given in Theorem* 2.2.5 *for* μ *almost all* $(\omega,x) \in \Delta_{a,b,k}$ *can be chosen to have the following additional property*: *There are a countable number of measurable subsets* Λ_m, $m = 1,2,\ldots$, *of* $\Delta_{a,b,k}$ *such that* $\mu(\bigcup_m \Lambda_m) = \mu(\Delta_{a,b,k})$ *and* $\{W^{s,i_a}_{\mathrm{loc}}(\omega,x)\}_{(\omega,x)\in\Lambda_m}$ *is a continuous family of* C^1 *embedded* k*-dimensional discs in* M *(in fact, on each* Λ_m *the Lyapunov norm* $\|\cdot\|_{(\omega,x),0}$ *depends continuously on* (ω,x) *and there exists* $r_m > 0$ *such that* $W^{s,i_a}_{\mathrm{loc}}(\omega,x) = \exp_x \mathrm{Graph}(h_{(\omega,x),0}|_{\{\xi\in E^{s,i_a}(\omega,x)\colon \|\xi\|_{(\omega,x),0}<r_m\}})$).

For μ-a.e. $(\omega,x) \in \Delta_{a,b}$ it can be shown (see also [118]) that

$$W^{s,i_a}(\omega,x) = \bigcup_{n=0}^{+\infty}(F^n_\omega)^{-1}W^{s,i_a}_{\mathrm{loc}}\big((\omega,x),n\big), \tag{2.2.7}$$

where $W^{s,i_a}(\omega,x)$ is defined by (2.2.2) (note that $W^{s,i_a}(\omega,x)$ is determined only by the exponent $\lambda^{(i_a)}(\omega,x)$), and hence it is the image of $V^{(i_a)}(\omega,x)$ under an injective immersion of class $C^{1,1}$. This also implies that $W^{s,i_a}_{\mathrm{loc}}((\omega,x),0)$ is an open piece of $W^{s,i_a}(\omega,x)$ which contains x and on which the exponential contraction can be read "immediately" (Theorem 2.2.5(3)). Hence it can be regarded as a *local stable manifold* associated with $\lambda^{(i_a)}(\omega,x)$ and this justifies our notations W^{s,i_a}_{loc} and (2.2.6). Now, by considering a countable number of sets of the type $\Delta_{a,b}$ (e.g., considering $\Delta_{a,b}$ for all rational $a<b\leqslant 0$), one obtains the following

THEOREM 2.2.7 (Global stable manifolds). *Let Δ be as defined by* (2.2.4). *For $(\omega,x)\in\Delta$ with $\lambda^{(1)}(\omega,x)<0$, let $\lambda^{(1)}(\omega,x)<\cdots<\lambda^{(p(\omega,x))}(\omega,x)$ be the strictly negative Lyapunov exponents of F at (ω,x), and let*

$$W^{s,1}(\omega,x)\subset\cdots\subset W^{s,p(\omega,x)}(\omega,x)$$

be defined by (2.2.2). *Then for μ-a.e. $(\omega,x)\in\Delta\setminus\{(\omega,x)\in\Delta\colon\ \lambda^{(1)}(\omega,x)\geqslant 0\}$ and for each $1\leqslant i\leqslant p(\omega,x)$, $W^{s,i}(\omega,x)$ is the image of $V^{(i)}(\omega,x)$ under an injective immersion of class $C^{1,1}$ and is tangent to $V^{(i)}(\omega,x)$ at point x; in addition, for $y\in W^{s,i}(\omega,x)$ one has*

$$\limsup_{n\to+\infty}\frac{1}{n}\log d^s(F^n_\omega x,F^n_\omega y)\leqslant\lambda^{(i)}(\omega,x),$$

where $d^s(\ ,\ )$ denotes the distance along $F^n_\omega W^{s,i}(\omega,x)=W^{s,i}(\Theta^n(\omega,x))$ induced by its inherited Riemannian metric as a submanifold of M.

REMARK 2.2.8. Given $\lambda<0$, let $\Delta_\lambda=\{(\omega,x)\in\Delta\colon\ \lambda^{(1)}(\omega,x)<\lambda$ and $\lambda^{(i)}(\omega,x)\neq\lambda$ for all $i\}$. Then for μ-a.e. $(\omega,x)\in\Delta_\lambda$, letting $\lambda^{(j)}(\omega,x)$ be the largest Lyapunov exponent which is smaller than λ, one has

$$W^{s,j}(\omega,x)=\left\{y\in M\colon\ \limsup_{n\to+\infty}\frac{1}{n}\log d(F^n_\omega x,F^n_\omega y)\leqslant\lambda\right\}.$$

From this it follows that for μ-a.e. $(\omega,x)\in\Delta\setminus\{(\omega,x)\in\Delta\colon\ \lambda^{(1)}(\omega,x)\geqslant 0\}$ the (*global*) *stable manifold* $W^s(\omega,x)$ of F at (ω,x), defined by

$$W^s(\omega,x)=\left\{y\in M\colon\ \limsup_{n\to+\infty}\frac{1}{n}\log d(F^n_\omega x,F^n_\omega y)<0\right\},$$

satisfies

$$W^s(\omega,x)=W^{s,p(\omega,x)}(\omega,x)$$

and hence is the image of $E^s(\omega,x):=\bigcup_{\lambda^{(i)}(\omega,x)<0}V^{(i)}(\omega,x)$ under an injective immersion of class $C^{1,1}$ and is tangent to $E^s(\omega,x)$ at point x.

Hölder continuity of subbundles. Now we touch briefly upon the absolute continuity property of the stable manifolds $W^{s,i}(\omega, x)$. Here F being C^2 (or $C^{1+\alpha}$) with condition (2.2.3) and the consequent Hölder continuity (in x) of the subbundles of TM consisting of tangent spaces of these submanifolds are essential for this kind of absolute continuity property. We first describe the notion of Hölder continuity of a subbundle of TM. Let $\Lambda \subset M$ be a set. A family $\{E_x\}_{x\in\Lambda}$ of subspaces $E_x \subset T_xM$ is said to be *locally Hölder continuous* with exponent $\sigma > 0$ and constant $L > 0$ if for some $\rho > 0$ one has

$$d(E_x, E_y) \leqslant L d(x, y)^\sigma \quad \text{for all } x, y \in \Lambda \text{ with } d(x, y) < \rho,$$

where $d(E_x, E_y)$ is the distance between E_x and E_y (see [39] or [118, Chapter III] for the definition).

For a $C^{1+\alpha}$ $(0 < \alpha \leqslant 1)$ map $f : M \to M$ we set

$$|Df|_{H^\alpha} = \sup_{x\in M} |D_x f| + \sup_{x,y\in M} \frac{d(D_x f, D_y f)}{d(x, y)^\alpha},$$

where $d(D_x f, D_y f)$ denotes the distance between $D_x f$ and $D_y f$ (see also [39] or [118, Chapter III] for the definition, and admit here $0/0 = 1$). The following result is adopted from [39].

LEMMA 2.2.9. *Let $\{f_k : M \to M\}_{k=1}^{+\infty}$ be a sequence of $C^{1+\alpha}$ maps which satisfy*

$$\prod_{k=1}^{n} |Df_k|_{H^\alpha} \leqslant \hat{K} e^{\hat{c}n} \quad \text{for all } n \in \mathbb{N},$$

where $\hat{K} > 0, \hat{c} > 0$. Fix $\hat{C} \geqslant 1$, $\hat{a} < \hat{b}$, $\hat{\gamma} > 0$ and let $\Lambda_{\hat{C},\hat{a},\hat{b},\hat{\gamma}}$ be the (maybe empty) set of points $x \in M$ for which there exists a splitting

$$T_xM = E_x \oplus F_x$$

such that the angle between E_x and F_x satisfies

$$\gamma(E_x, F_x) \geqslant \hat{\gamma}$$

and for all $n \in \mathbb{N}$ one has

$$\left|D_x(f_n \circ \cdots \circ f_1)\xi\right| \leqslant \hat{C} e^{\hat{a}n} |\xi| \quad \text{for } \xi \in E_x,$$

$$\left|D_x(f_n \circ \cdots \circ f_1)\eta\right| \geqslant \hat{C}^{-1} e^{\hat{b}n} |\eta| \quad \text{for } \eta \in F_x.$$

Then the family $\{E_x\}_{x\in\Lambda_{\hat{C},\hat{a},\hat{b},\hat{\gamma}}}$ is locally Hölder continuous with the Hölder exponent and constant depending only on the related constants appearing above.

The integrability condition (2.2.3) implies the following (see [118])

LEMMA 2.2.10. *For the C^2 RDS F satisfying* (2.2.3), *the following hold true*:

(1) $\lim_{n\to+\infty} \frac{1}{n} \log \prod_{k=0}^{n-1} |DF_{\vartheta^k\omega}|_{H^1} =: c(\omega) < +\infty$ *exists for $\mathbb{P}$-a.e. ω.*

(2) *There exists a measurable set $\Gamma_0 \subset \Omega$ with $\mathbb{P}(\Gamma_0) = 1$ and there is a measurable function $K : \Gamma_0 \to (0, +\infty)$ such that for every $\omega \in \Gamma_0$,*

$$\prod_{k=0}^{n-1} |DF_{\vartheta^k\omega}|_{H^1} \leqslant K(\omega) e^{2c(\omega)n} \quad \text{for all } n \in \mathbb{N}.$$

Let Λ_m be a set as given in Complement to Theorem 2.2.5. Fix arbitrarily $K > 0$, $c > 0$ and put

$$\Lambda_{m,K,c} = \big\{(\omega, x) \in \Lambda_m \colon\ \omega \in \Gamma_0 \text{ and } K(\omega) \leqslant K,\ c(\omega) \leqslant c\big\}.$$

Then, by Lemmas 2.2.9 and 2.2.10, one obtains the following (see [118])

PROPOSITION 2.2.11 (Local Hölder continuity). *In the circumstances of Complement to Theorem* 2.2.5. *Let $\Lambda_{m,K,c}$ be as defined above and put $\Lambda_{m,K,c}(\omega) = \{x \in M\colon\ (\omega, x) \in \Lambda_{m,K,c}\}$ for $\omega \in \Omega$. Then for each $\omega \in \Omega$ the map $z \mapsto E^{s,i_a}(\omega, z)$ is locally Hölder continuous on the set $\bigcup_{x\in\Lambda_{m,K,c}(\omega)} W^{s,i_a}_{\mathrm{loc}}(\omega, x)$ with the Hölder exponent and constant depending only on the related numbers appearing in the definition of $\Lambda_{m,K,c}$.*

Absolute continuity. We are now in the right position to discuss the absolute continuity property. To be more precise, we first explain the idea of an absolutely continuous family of C^1 embedded k-dimensional discs in M. Let $\Lambda \subset M$ be a set and let $\{D_x\}_{x\in\Lambda}$ be a continuous family of C^1 embedded k-dimensional discs in M such that $D_y \cap D_z = \emptyset$ if $y, z \in \Lambda$ and $y \neq z$. Let $x_0 \in \Lambda$ and $p, q \in D_{x_0}$, and let W_p, W_q be two C^1 embedded $(\dim M - k)$-dimensional discs transversal to D_{x_0} at p and q, respectively. Then there exist two open pieces $\hat{W}_p$ and $\hat{W}_q$ of W_p and W_q, respectively, such that the so-called Poincaré map

$$P_{\hat{W}_p,\hat{W}_q} : \hat{W}_p \cap \Big(\bigcup_{x\in\Lambda} D_x\Big) \to \hat{W}_q \cap \Big(\bigcup_{x\in\Lambda} D_x\Big), \quad \hat{W}_p \cap D_x \mapsto \hat{W}_q \cap D_x$$

is a homeomorphism between $\hat{W}_p \cap (\bigcup_{x\in\Lambda} D_x)$ and $\hat{W}_q \cap (\bigcup_{x\in\Lambda} D_x)$. The family of C^1 embedded discs $\{D_x\}_{x\in\Lambda}$ is said to be *absolutely continuous* if each of its Poincaré maps $P_{\hat{W}_p,\hat{W}_q}$ takes Lebesgue zero sets to Lebesgue zero sets (here Lebesgue refers to the Lebesgue measure on W_p or W_q induced by the inherited Riemannian structure).

With the help of the previous technical lemmas and Proposition 2.2.11, one can adopt the proof of the absolute continuity property of stable manifolds for a deterministic map [133,19,77]) to the random case, obtaining

THEOREM 2.2.12 (Absolute continuity). *Let $\Lambda_{m,K,c}$ be a set as introduced above. Then for $\mathbb{P}$-a.e. $\omega \in \Omega$ the family of C^1 embedded k-dimensional discs $\{W^{s,i_a}_{\mathrm{loc}}(\omega, x)\}_{x\in\Lambda_{m,K,c}(\omega)}$ is absolutely continuous.*

We refer the reader to [112,111] and [118, Sections III.5 and III.6] for alternative (and more precise) statements and related estimates on the Jacobian of $P_{\hat{W}_p,\hat{W}_q}$, etc.

Partitions subordinate to stable manifolds. Take the assumption of $(\Omega,\mathcal{F},\mathbb{P},\vartheta)$ being Polish. As in the deterministic case, the absolute continuity property Theorem 2.2.12 has some important consequences. One of them is the genericity with respect to Lebesgue or the physical relevance of SRB measures of F (for the deterministic case see [145,137]; for the case of a hyperbolic RDS see [166,115]). Another one is the fact that smooth (i.e., absolutely continuous with respect to the Lebesgue) measures on M have smooth conditional measures on the stable manifolds of F (this fact is used mainly for the purpose of dealing with entropy formula of Pesin type, but physically more important is the SRB property, i.e., the smoothness of conditional measures on unstable manifolds, see Section 3). Here we present a formulation of the latter. Intuitively speaking, let $\omega\in\Omega$ and let $V\subset M$ be a Borel set. Assume that $V=\bigcup V_x$ is the disjoint union of a continuous family of embedded discs (V_x denotes the disc containing x) and each V_x is an open piece of $W^{s,i}(\omega,x)$ for a fixed $1\leqslant i\leqslant\dim M$. Let ν be a Borel probability measure on M with $\nu(V)>0$, and let ν_x be the conditional probability measure of ν on V_x. The fact we alluded to above means that, if $\nu\ll\text{Leb}$, then $\nu_x\ll\lambda_x$ for ν-a.e. $x\in V$, where λ_x is the Lebesgue measure on $W^{s,i}(\omega,x)$ induced by its inherited Riemannian structure as a submanifold of M. To describe this property precisely, we need to appeal to the theory of conditional measures associated with measurable partitions of Lebesgue spaces, for which the reader is referred to [141]. It is for this reason that we assume Ω being Polish: $(\Omega\times M,\mathcal{B}_\mu(\Omega\times M),\mu)$ defines a Lebesgue space allowing for "nice" partitions, where $\mathcal{B}_\mu(\Omega\times M)$ denotes the completion of the Borel σ-algebra of $\Omega\times M$ with respect to μ. Note that the partition of $\Omega\times M$ into the global $W^{s,i}$-manifolds $\{\omega\}\times W^{s,i}(\omega,x)$ is in general not measurable. However, one can use (and construct) a finer but measurable partition of $\Omega\times M$ into pieces of local $W^{s,i}$-manifolds and one can take the pieces to be open in the submanifold topology (neglecting sets of μ-measure 0). This leads to the following two definitions. Here we will confine ourselves to the W^s-manifolds case for simplicity of presentation. The general $W^{s,i}$-manifolds case can be considered analogously by restricting the partition to the subset $\{(\omega,x)\colon\ i\leqslant r(\omega,x)$ and $\lambda^{(i)}(\omega,x)<0\}$ of $\Omega\times M$ (which is, neglecting a μ-null set, the union of some $W^{s,i}$-manifolds, i.e., a set of the form $\bigcup[\{\omega\}\times W^{s,i}(\omega,x)]$, see [118, Lemma IV.2.2]). In what follows we will use $\xi(x)$ to denote the element of a partition ξ of a space X which contains the point $x\in X$.

DEFINITION 2.2.13. Let F be over a Polish system $(\Omega,\mathcal{F},\mathbb{P},\vartheta)$ and assume $\lambda^{(1)}(\omega,x)>-\infty$ μ almost everywhere. A measurable partition η of $(\Omega\times M,\mathcal{B}_\mu(\Omega\times M),\mu)$ is said to be *subordinate to W^s-manifolds* of (F,μ) if for μ-a.e. $(\omega,x)\in\Omega\times M$, $\eta(\omega,x)\subset\{\omega\}\times M$, $\eta_\omega(x):=\{y\colon\ (\omega,y)\in\eta(\omega,x)\}\subset W^s(\omega,x)$ and it contains an open neighborhood of x in $W^s(\omega,x)$, this neighborhood being taken in the submanifold topology of $W^s(\omega,x)$ ($W^s(\omega,x):=\{x\}$ if $\lambda^{(i)}(\omega,x)\geqslant 0$ for all i).

Such partitions always exist (see Section 3).

DEFINITION 2.2.14. Take the assumptions in Definition 2.2.13. We say that μ *has smooth conditional measures on* W^s*-manifolds* if for every measurable partition η subordinate to W^s-manifolds of (F, μ) one has

$$\mu^{\eta}_{(\omega,x)} \ll \lambda^{s}_{(\omega,x)}, \quad \mu\text{-a.e. } (\omega, x),$$

where $\{\mu^{\eta}_{(\omega,x)}\}_{(\omega,x)\in\Omega\times M}$ is a canonical system of conditional measures of μ associated with η, $\mu^{\eta}_{(\omega,x)}$ is regarded as a measure on $\eta_\omega(x)$ by identifying $\eta(\omega, x) = \{\omega\} \times \eta_\omega(x)$ with $\eta_\omega(x)$, and $\lambda^{s}_{(\omega,x)}$ is the Lebesgue measure on $W^s(\omega, x)$ induced by its inherited Riemannian structure as a submanifold of M ($\lambda^{s}_{(\omega,x)} := \delta_x$ if $W^s(\omega, x) = \{x\}$).

The consequence of Theorem 2.2.12 we discussed just above can now be formulated as follows (see [118, Proposition IV.2.1] for an analogous proof).

COROLLARY 2.2.15. *Take the assumptions in Definition* 2.2.13. *If* $\mu \ll \mathbb{P} \times$ Leb, *then* μ *has smooth conditional measures on* W^s*-manifolds*.

REMARK 2.2.16. It is easy to show that $\mu \ll \mathbb{P} \times$ Leb is equivalent to $\mu_\omega \ll$ Leb $\mathbb{P}$-a.e. ω, where μ_ω, $\omega \in \Omega$, are the sample measures (or the disintegration) of μ.

REMARK 2.2.17. If $\lambda^{(1)}(\omega, x) = -\infty$ holds on a set of positive μ-measure, then Corollary 2.2.15 is true with the arguments being restricted to the set $\{(\omega, x)\colon \lambda^{(1)}(\omega, x) > -\infty\}$, which is the union of some W^s-manifolds (neglecting a μ-null set).

REMARK 2.2.18 (*Random endomorphisms*). Let $C^r(M, M)$ be the space of all C^r ($r \geqslant 1$ is an integer) maps from M to itself, endowed with the usual C^r topology. Let the RDS F over $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ be defined by replacing $\mathrm{Diff}^r(M)$ with $C^r(M, M)$. Assume that μ is an invariant measure of F. Under the same integrability condition as (2.2.1), the Lyapunov spectrum $(\lambda^{(i)}(\omega, x), m^{(i)}(\omega, x))$, $1 \leqslant i \leqslant r(\omega, x)$, can be defined for μ-a.e. (ω, x) in the same way as in the random diffeomorphisms case. When $r \geqslant 2$ and the same integrability condition as (2.2.3) is satisfied, almost all the stable manifold arguments presented in this section hold with possibly a slight modification. For example, Theorem 2.2.5 and Complement to Theorem 2.2.5 hold verbatim, one still has the expression (2.2.7) for the global stable manifolds, but they may lose the property of being a nice submanifold of M since now F^n_ω may be noninvertible and may have *singularities* (i.e., points at which the derivatives of F^n_ω are degenerate). We refer the reader to [154,149,116] for more details about properties of stable manifolds of endomorphisms.

REMARK 2.2.19 ($C^{1+\alpha}$ *hypothesis*). In the arguments of this and next sections we often make the specific assumption that F is C^2 (as opposed to $C^{1+\alpha}$ for some $\alpha \in (0, 1]$), with the integrability condition (2.2.3) being supposed on the C^2-norms. One reason is that this assumption will be needed for the treatment of relations between entropy, Lyapunov exponents and dimension, where we often require the map space to be Polish or at least separable so that Lusin's theorem can be applied (note that $C^{1,\alpha}(M, M)$ is in general not separable, see [106]; it is not clear to the author if this assumption can be reduced to a

$C^{1+\alpha}$ one by a suitable trick for these purposes). But under analogous $C^{1+\alpha}$ assumptions invariant manifolds can still be constructed along Pesin's line (possibly with a bit loss of order of smoothness) or by adapting Ruelle's approach. Note however that C^1 assumptions are usually not sufficient for the invariant manifold theory, especially for the absolute continuity property (see [136] for a counterexample).

2.3. *Unstable invariant manifolds, Oseledec manifolds*

In this section we assume that $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ is measurably invertible. In addition to F_ω^n, $n \geqslant 0$, $\omega \in \Omega$, one can also consider the backward compositions of random maps

$$F_\omega^n := F_{\vartheta^n\omega}^{-1} \circ \cdots \circ F_{\vartheta^{-1}\omega}^{-1}, \quad n < 0,\ \omega \in \Omega.$$

Under the integrability condition

$$\int \left(\log^+ |D_x F_\omega| + \log^+ \left|D_x (F_\omega)^{-1}\right|\right) d\mu(\omega, x) < +\infty, \tag{2.3.1}$$

the Oseledec MET, applied to the invertible system $\Theta : (\Omega \times M, \mu) \hookleftarrow$, tells that there exists a measurable set Δ such that $\mu(\Delta) = 1$, $\Theta\Delta = \Delta$ and for each $(\omega, x) \in \Delta$ one has the Lyapunov exponents of F at (ω, x),

$$-\infty < \lambda^{(1)}(\omega, x) < \lambda^{(2)}(\omega, x) < \cdots < \lambda^{(r(\omega,x))}(\omega, x) < +\infty$$

and an associated measurable (in (ω, x)) splitting

$$T_x M = E^{(1)}(\omega, x) \oplus \cdots \oplus E^{(r(\omega,x))}(\omega, x) \tag{2.3.2}$$

which satisfy

$$\lim_{n\to+\infty} \frac{1}{n} \log |D_x F_\omega^{\pm} \xi| = \pm\lambda^{(i)}(\omega, x) \quad \text{for } 0 \neq \xi \in E^{(i)}(\omega, x),$$

$1 \leqslant i \leqslant r(\omega, x)$. (2.3.2) will be called the *Oseledec splitting* of F at (ω, x) and $E^{(i)}(\omega, x)$ will be called an *Oseledec space*. From basic properties of Lyapunov exponents it follows that $V^{(i)}(\omega, x) = \bigoplus_{j \leqslant i} E^{(j)}(\omega, x)$ and $m^{(i)}(\omega, x) = \dim E^{(i)}(\omega, x)$ for μ-a.e. (ω, x).

Unstable manifolds. In a way analogous to (2.2.2), we define the *unstable manifold* of F at (ω, x) associated with a positive exponent $\lambda^{(j)}(\omega, x) > 0$ as

$$W^{u,j}(\omega, x) = \left\{ y \in M\colon \limsup_{n\to+\infty} \frac{1}{n} \log d(F_\omega^{-n} x, F_\omega^{-n} y) \leqslant -\lambda^{(j)}(\omega, x) \right\}. \tag{2.3.3}$$

As can be easily seen, the unstable manifolds of F are just the stable manifolds of the RDS over $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta^{-1})$ generated by the measurable map $\omega \mapsto F_{\vartheta^{-1}\omega}^{-1}$ from Ω to $\mathrm{Diff}^r(M)$.

Thus, *the verbatim counterpart of all the stable manifold arguments of Section* 2.2 *can be obtained for the unstable manifolds of* F*, provided that it is of class* C^2 *and satisfies* (2.3.1) *together with*

$$\int \log^+ \left|(F_\omega)^{-1}\right|_{C^2} d\mathbb{P}(\omega) < +\infty. \tag{2.3.4}$$

For example, for a closed interval $[c,\ d]$, $0 \leqslant c < d$, one defines

$$\Delta_{c,d} = \left\{(\omega, x) \in \Delta\colon\ \lambda^{(i)}(\omega, x) \notin [c,\ d] \text{ for all } i \text{ and } d < \lambda^{(r(\omega,x))}(\omega, x)\right\}.$$

Setting

$$E_0(\omega, x) = \bigoplus_{d<\lambda^{(j)}(\omega,x)} E^{(j)}(\omega, x), \qquad H_0(\omega, x) = E_0(\omega, x)^\perp$$

and considering F_ω^{-n}, $n \geqslant 0$, one can obtain for μ-a.e. $(\omega, x) \in \Delta_{c,d}$ a *local unstable manifold* $W^{u,j_d}_{\text{loc}}(\omega, x)$ associated with $\lambda^{(j_d)}(\omega, x)$, the smallest Lyapunov exponent larger than d. It is the $\exp_x$-image of the graph of a $C^{1,1}$ map $h_{(\omega,x),0}$ from an open neighborhood of 0 in $E_0(\omega, x)$ to $H_0(\omega, x)$, and it has the property

$$d^u(F_\omega^{-n} y, F_\omega^{-n} z) \leqslant \gamma_\varepsilon(\omega, x) e^{(-d+4\varepsilon)n}\, d^u(y, z)$$

for $y, z \in W^{u,j_d}_{\text{loc}}(\omega, x)$ and $n \geqslant 0$, where $d^u(\ ,\ )$ denotes the distance along $F_\omega^{-m} W^{u,j_d}_{\text{loc}}(\omega, x)$ for $m \geqslant 0$, ε is a sufficiently small (as compared with $d - c$) positive number and $\gamma_\varepsilon(\cdot)$ is a measurable function on $\Delta_{c,d}$ related to ε. For each $(\omega, x) \in \Delta$ minus a μ-null set, if $\lambda^{(j)}(\omega, x) > 0$ for some j and $\lambda^{(q(\omega,x))}(\omega, x) < \cdots < \lambda^{(r(\omega,x))}(\omega, x)$ are all the strictly positive exponents of F at (ω, x), letting

$$W^{u,r(\omega,x)}(\omega, x) \subset \cdots \subset W^{s,q(\omega,x)}(\omega, x) \tag{2.3.5}$$

be the corresponding nested family of unstable manifolds of F associated with these exponents, then each $W^{u,j}(\omega, x)$, $q(\omega, x) \leqslant j \leqslant r(\omega, x)$ is the image of $\bigoplus_{j\leqslant k\leqslant r(\omega,x)} E^{(k)}(\omega, x)$ under an injective immersion of class $C^{1,1}$, is tangent to this subspace at point x and has the additional property

$$\limsup_{n\to+\infty} \frac{1}{n} \log d^u(F_\omega^{-n} x, F_\omega^{-n} y) \leqslant -\lambda^{(j)}(\omega, x)$$

for $y \in W^{u,j}(\omega, x)$, where $d^u(\ ,\ )$ denotes the distance along $F_\omega^{-n} W^{u,j}(\omega, x) = W^{u,j}(\Theta^{-n}(\omega, x))$, moreover, the *(global) unstable manifold* of F at (ω, x) defined by

$$W^u(\omega, x) = \left\{y \in M\colon\ \limsup_{n\to+\infty} \frac{1}{n} \log d(F_\omega^{-n} x, F_\omega^{-n} y) < 0\right\} \tag{2.3.6}$$

satisfies

$$W^u(\omega,x)=W^{u,q(\omega,x)}(\omega,x).$$

We leave to the reader all the other statements analogous to those in Section 2.2.

We remark that measurable partitions of $(\Omega\times M,\mathcal{B}_\mu(\Omega\times M),\mu)$ subordinate to W^u-manifolds of F and smoothness of conditional measures of μ on the W^u-manifolds give rise to the definition of a very important property—SRB property—of μ which can also be characterized by an entropy formula of Pesin type. We will deal with this topic in Section 3.

As indicated in Section 2.2, there can be various types of results concerning the stable and unstable manifolds of an RDS for various purposes and under various conditions. For example, under the integrability condition

$$\int\left(\log^+|F_\omega|_{C^2}+\log^+\left|(F_\omega)^{-1}\right|_{C^2}\right)d\mathbb{P}(\omega)<+\infty, \tag{2.3.7}$$

one can develop for (F,μ), due to its invertibility, a completely analogous counterpart of [134, Theorems 4.1–4.3]. This counterpart uses the splitting $T_xM=E_0(\omega,x)\oplus H_0(\omega,x)$, where $E_0(\omega,x)=\bigoplus_{1\leqslant i\leqslant k}E^{(i)}(\omega,x)$ and $H_0(\omega,x)=\bigoplus_{k+1\leqslant i\leqslant r(\omega,x)}E^{(i)}(\omega,x)$ which are D_xF_ω-invariant, i.e., $D_xF_\omega E_0(\omega,x)=E_0(\Theta(\omega,x))$ and similarly for $H_0(\omega,x)$. In this way it can avoid constructing a sequence of local stable or unstable manifolds at one point (as opposed to our Theorem 2.2.5), and it can give dynamical characterizations as well as some additional interesting properties of the local invariant manifolds. However, such local invariant manifolds are essentially unique since they are open subsets of the corresponding global ones. There have also been some results on higher order smoothness of the invariant manifolds, for which we refer to [7, Part III], [153] and [43].

Oseledec manifolds. We end this subsection with a remark on Oseledec manifolds. Assume the integrability condition (2.3.7). For simplicity of presentation we assume (F,μ) being ergodic and let $\lambda^{(1)}<\cdots<\lambda^{(r)}$ be its Lyapunov exponents. Note that, if we take the closed interval $[a,b]$ with $\lambda^{(i)}\notin[a,b]$ for all i but without the assumption $b\leqslant 0$, techniques similar to those for the proof of our Theorem 2.2.5 yield for μ-a.e. (ω,x) the existence of a C^1 embedded disc in M, called a *pseudo local stable manifold* of F at (ω,x) and denoted by $\hat W^{ps}_{[a,\,b]}(\omega,x)$, which is the $\exp_x$-image of the graph of a C^1 map from an open subset of $E^{ps}_a(\omega,x):=\bigoplus_{\lambda^{(i)}<a}E^{(i)}(\omega,x)$ to $E^{pu}_a(\omega,x):=\bigoplus_{\lambda^{(j)}>a}E^{(j)}(\omega,x)$ and which is locally invariant and tangent to $E^{ps}_a(\omega,x)$ at x. Here the *local invariance* means that F_ω maps an open neighborhood of x in $\hat W^{ps}_{[a,b]}(\omega,x)$ into $\hat W^{ps}_{[a,b]}(\Theta(\omega,x))$. Similarly, considering the backward application of the random maps gives for μ-a.e. (ω,x) the existence of another C^1 embedded disc in M, called a *pseudo local unstable manifold* of F at (ω,x) and denoted by $\hat W^{pu}_{[a,b]}(\omega,x)$, which is the $\exp_x$-image of the graph of a C^1 map from an open subset of $E^{pu}_a(\omega,x)$ to $E^{ps}_a(\omega,x)$ and which is locally invariant and tangent to $E^{pu}_a(\omega,x)$ at x. Let now $[a,b]$ and $[c,d]$ be two intervals such that they contain no $\lambda^{(i)}$, $b<c$ and $[b,c]\cap\{\lambda^{(i)}\colon\ 1\leqslant i\leqslant r\}\neq\emptyset$. For μ-a.e. (ω,x) define $\hat W_{[b,c]}(\omega,x)$ to be the intersection $\hat W^{pu}_{[a,b]}(\omega,x)\cap\hat W^{ps}_{[c,d]}(\omega,x)$. It can be shown that $\hat W_{[b,c]}(\omega,x)$ is a C^1 embedded disc in M which is locally invariant and tangent to $E_{[b,c]}(\omega,x):=\bigoplus_{\lambda^{(i)}\in[b,c]}E^{(i)}(\omega,x)$

at x (see [51] for a detailed treatment). When $[b, c]$ contains only one exponent $\lambda^{(i)}$, $\hat{W}_{[b,c]}(\omega, x)$ is called a *local Oseledec manifold* of F at (ω, x) associated with $\lambda^{(i)}$; and when $[b, c]$ contains more exponents $\hat{W}_{[b,c]}(\omega, x)$ is called a *generalized local Oseledec manifold* of F at (ω, x) associated with the exponents in $[b, c]$. We remark that such manifolds usually depend on the construction process and are not essentially unique. When $[b, c]$ contains, respectively, only the zero exponent, all the negative or all the positive exponents, the corresponding (suitably constructed) local Oseledec manifold defines a classical *local central*, stable or unstable manifold. The Oseledec manifolds theory has applications in the linearization (or Hartmann–Grobman) theory and bifurcation theory of random dynamical systems. We again refer the reader to [7, Part III] for a comprehensive treatment of this topic and for an account of previous works [38,137,51], etc.

2.4. *Invariant manifolds for continuous time RDS*

RDS with continuous time are of great interest due to their generation by random and stochastic differential equations (see [7, Chapter 2] for a comprehensive treatment of this generation theory). Here the time set will be $T = \mathbb{R}^+$ or $T = \mathbb{R}$ and the underlying noise system will be modelled by a measure-preserving flow $(\Omega, \mathcal{F}, \mathbb{P}, \{\vartheta^t\}_{t\in T})$. Let the manifold M be as given previously and let $r \geqslant 1$ be an integer. In this section, by a C^r RDS F on M over $(\Omega, \mathcal{F}, \mathbb{P}, \{\vartheta^t\}_{t\in T})$ we will mean a family of maps $\{F_\omega^t \in \mathrm{Diff}^r(M)\colon t \in T,\ \omega \in \Omega\}$ such that $F_\omega^t x$ is measurable in (t, ω, x), $F_\omega^0 = \mathrm{id}$, $F_\omega^{t+s} = F_{\vartheta^s\omega}^t \circ F_\omega^s$ for all $s, t \in T$ and $\omega \in \Omega$, and the derivatives, up to r orders, of $F_\omega^t x$ with respect to x are continuous in (t, x). Note that we require F_ω^t being a diffeomorphism even when $T = \mathbb{R}^+$ (this holds automatically when $T = \mathbb{R}$), but the RDS with $T = \mathbb{R}^+$ defined here appear typically in the setting of the classical stochastic differential equations (see, e.g., [107]). As one can easily expect, under suitable integrability conditions on the norms of the random maps, almost all the previous results in this section can be carried over to a continuous time RDS by reducing the problem to its time-one system.

Let F be as defined just above and μ an F-invariant measure, i.e., a probability on $(\Omega \times M, \mathcal{F} \times \mathcal{B})$ which has marginal $\mathbb{P}$ on Ω and is invariant under $\Theta^t : \Omega \times M \to \Omega \times M$, $(\omega, x) \mapsto (\vartheta^t\omega, F_\omega^t x)$ for all $t \in T$ (such measures always exist, see [7, Chapter 1]). We first consider forward applications of the random maps. If the integrability condition

$$\int \Big(\sup_{0\leqslant t\leqslant 1} \log^+ \big|D_x F_\omega^t\big| + \sup_{0\leqslant t\leqslant 1} \log^+ \big|\big(D_x F_\omega^t\big)^{-1}\big| \Big)\, d\mu(\omega, x) < +\infty \qquad (2.4.1)$$

is satisfied, then Theorem 2.2.1 holds for (F, μ) with n being replaced by t and with $\lambda^{(1)}(\omega, x) > -\infty$ for μ-a.e. (ω, x) (see [7]). Define $W^{s,i}(\omega, x)$, the *stable manifold* of F at (ω, x) associated with exponent $\lambda^{(i)}(\omega, x) < 0$, in a way analogous to (2.2.2) by replacing n with t. Then we have the following (see [43] or [118, Chapter V] for an analogous proof):

THEOREM 2.4.1 (Global stable manifolds). *If F is of class C^2 and satisfies*

$$\int \Big(\sup_{0\leqslant t\leqslant 1} \log^+ \big|F_\omega^t\big|_{C^2} + \sup_{0\leqslant t\leqslant 1} \log^+ \big|\big(F_\omega^t\big)^{-1}\big|_{C^2} \Big)\, d\mathbb{P}(\omega) < +\infty, \qquad (2.4.2)$$

then the statements of Theorem 2.2.7 *hold for* (F, μ) *with* n *being replaced by* t *and with* Δ *being replaced by a measurable set of full* μ *measure.*

A local stable manifold theorem can be stated as follows (see [118, Theorem V.2.3] for a proof and see [43] for a similar result).

THEOREM 2.4.2 (Local stable manifolds). *Take the assumptions of Theorem* 2.4.1. *Given* $\lambda < 0$, *put* $\Delta_\lambda = \{(\omega, x)$: (ω, x) *is forward regular in the sense of Oseledec,* $\lambda^{(i)}(\omega, x) \neq \lambda$ *for all* i *and* $\lambda^{(1)}(\omega, x) < \lambda\}$. *Then for* μ*-a.e.* $(\omega, x) \in \Delta_\lambda$ *there exists a* $C^{1,1}$ *embedded disc* $W^{s,\lambda}_{\mathrm{loc}}(\omega, x)$ *in* M *together with positive numbers* $\alpha(\omega, x)$, $\beta(\omega, x)$ *and* $\gamma(\omega, x)$, *all measurable in* (ω, x) *and possibly depending on* λ, *such that*:

(1) $W^{s,\lambda}_{\mathrm{loc}}(\omega, x)$ *contains* x *and is tangent to* $V^{(i_\lambda)}(\omega, x)$ *at* x, *where* $\lambda^{(i_\lambda)}(\omega, x)$ *is the largest exponent smaller than* λ, *and* $W^{s,\lambda}_{\mathrm{loc}}(\omega, x) = \exp_x \mathrm{Graph}(h_{(\omega,x)})$ *where* $h_{(\omega,x)} : \{\xi \in V^{(i_\lambda)}(\omega, x)$: $|\xi| < \alpha(\omega, x)\} \to V^{(i_\lambda)}(\omega, x)^\perp$ *is a* $C^{1,1}$ *map with* $h_{(\omega,x)}(0) = 0$ *and* $\mathrm{Lip}(h_{(\omega,x)})$, $\mathrm{Lip}(D.h_{(\omega,x)}) \leqslant \beta(\omega, x)$.

(2) *For* $y, z \in W^{s,\lambda}_{\mathrm{loc}}(\omega, x)$ *and for all* $t \geqslant 0$ *one has*

$$d^s\left(F^t_\omega y, F^t_\omega z\right) \leqslant \gamma(\omega, x) e^{\lambda t} d^s(y, z),$$

where $d^s(\,,\,)$ *denotes the distance along the submanifolds* $F^t_\omega W^{s,\lambda}_{\mathrm{loc}}(\omega, x)$.

If it is assumed moreover that $(\Omega, \mathcal{F}, \mathbb{P}, \{\vartheta^t\}_{t\in T})$ is a Polish system and $\omega \mapsto F^1_\omega$ is a measurable map from Ω to $\mathrm{Diff}^2(M)$, then results similar to Complement to Theorem 2.2.5 etc. can clearly be formulated in the present setting.

For the case of $T = \mathbb{R}$, it is also clear that results analogous to those in Section 2.3 concerning unstable manifolds hold true for a C^2 RDS F over $(\Omega, \mathcal{F}, \mathbb{P}, \{\vartheta^t\}_{t\in\mathbb{R}})$ which satisfies the integrability condition (2.4.2) (see [7] for results concerning Oseledec manifolds).

Stochastic flows. A classical model of continuous time RDS which is of particular interest is a stochastic flow of diffeomorphisms (or a Brownian motion in the diffeomorphisms group), and such stochastic flows are basically in a one-to-one correspondence to stochastic differential equations (generally with vector field valued driving Brownian motions, see [22] and [107]). It is usually defined as follows.

DEFINITION 2.4.3. Let $(\hat\Omega, \hat{\mathcal{F}}, \hat{\mathbb{P}})$ be a probability space. A random process $\{\psi_t : (\hat\Omega, \hat{\mathcal{F}}, \hat{\mathbb{P}}) \to \mathrm{Diff}^r(M)\}_{t\geqslant 0}$ is called *a stochastic flow of* C^r *diffeomorphisms* if it has the following properties:

(1) for any $0 \leqslant t_0 \leqslant t_1 \leqslant \cdots \leqslant t_n$, $\psi_{t_i} \circ \psi^{-1}_{t_{i-1}}$, $1 \leqslant i \leqslant n$, are independent;
(2) for any $0 \leqslant s \leqslant t$, the distribution of $\psi_t \circ \psi_s^{-1}$ depends only on $t - s$;
(3) with probability one $\{\psi_t\}_{t\geqslant 0}$ has continuous sample paths, i.e., the map $t \mapsto \psi_t(\hat\omega)$ is a continuous map from $\mathbb{R}^+$ to $\mathrm{Diff}^r(M)$ for $\hat{\mathbb{P}}$-a.e. $\hat\omega$;
(4) $\psi_0 = \mathrm{id}$ $\hat{\mathbb{P}}$-a.e.

Note that the probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$ may not be born with a $\hat{\mathbb{P}}$-preserving transformation group $\{\hat{\vartheta}^t\}_{t\geqslant 0}$. One can however consider the path space $(\Omega, \mathcal{F})$, where $\Omega = \{\omega = \{\omega(t)\}_{t\geqslant 0}\colon\ \omega(\cdot)$ is a continuous map from $\mathbb{R}^+$ to $\mathrm{Diff}^r(M)$ with $\omega(0) = \mathrm{id}\}$ and $\mathcal{F}$ is its natural Borel σ-algebra. Let $\mathbb{P}$ be the probability on $(\Omega, \mathcal{F})$ induced by $\hat{\mathbb{P}}$ via the measurable map

$$\Sigma : (\hat{\Omega}, \hat{\mathcal{F}}) \to (\Omega, \mathcal{F}), \quad \hat{\omega} \mapsto \big\{\psi_t(\hat{\omega})\big\}_{t\geqslant 0}.$$

Then it is preserved by the transformation group $\{\vartheta^t\}_{t\geqslant 0}$ on Ω defined by $(\vartheta^t\omega)(\cdot) = \omega(t+\cdot)\circ\omega(t)^{-1}$ (see, e.g., [118, p. 121]). The coordinate process on $(\Omega, \mathcal{F}, \mathbb{P})$, i.e., $F^t_\omega := \omega(t)$ for $\omega \in \Omega$ and $t \geqslant 0$ defines then an RDS F over $(\Omega, \mathcal{F}, \mathbb{P}, \{\vartheta^t\}_{t\geqslant 0})$. As can be seen below, measure-theoretic results obtained for F can usually be carried back to $\{\psi_t\}_{t\geqslant 0}$ via the map Σ, so a stochastic flow defined above can be essentially included in our framework of continuous time RDS.

Let now $\{\psi_t\}_{t\geqslant 0}$ be a stochastic flow as given in Definition 2.4.3. It holds automatically that

$$\int \Big(\sup_{0\leqslant t\leqslant T} \log^+ \big|\psi_t(\hat{\omega})\big|_{C^r} + \sup_{0\leqslant t\leqslant T} \log^+ \big|\psi_t(\hat{\omega})^{-1}\big|_{C^r} \Big)\, d\hat{\mathbb{P}}(\hat{\omega}) < +\infty \qquad (2.4.3)$$

for any $T > 0$, where $|f|_{C^r}$ denotes the C^r-norm of $f \in \mathrm{Diff}^r(M)$ (see [118, Proposition V.1.2] for a proof which is a slight modification of the proof of an even stronger statement in [84]). When $r = 2$, (2.4.3) clearly implies (2.4.2) for the RDS F defined just above.

Let $P_t(x, \cdot)$, $t \geqslant 0$, $x \in M$, be the transition probabilities of the one-point Markov process associated with $\{\psi_t\}_{t\geqslant 0}$, i.e., $P_t(x, A) = \hat{\mathbb{P}}\{\hat{\omega}\colon\ \psi_t(\hat{\omega})x \in A\}$ for Borel $A \subset M$. Let ρ be a Borel probability measure on M which is stationary for this Markov process (i.e., $\rho(A) = \int P_t(x, A)\, d\rho(x)$ for all Borel $A \subset M$ and all $t \geqslant 0$; such a measure always exists, see [82, Chapter V]). Then $\mathbb{P}\times\rho$ constitutes an invariant measure of the continuous time RDS F defined just above [82, Lemma I.2.3]. Thus, by the measure-preserving property of the map $(\hat{\Omega}\times M, \hat{\mathbb{P}}\times\rho) \to (\Omega\times M, \mathbb{P}\times\rho)$, $(\hat{\omega}, x) \mapsto (\Sigma\hat{\omega}, x)$, one easily knows that, if the stochastic flow $\{\psi_t\}_{t\geqslant 0}$ is of class C^2, the stable manifolds statements given by Theorems 2.4.1 and 2.4.2 hold true with μ, (ω, x) and F^t_ω being replaced respectively by $\hat{\mathbb{P}}\times\rho$, $(\hat{\omega}, x)$ and $\psi_t(\hat{\omega})$ (the Lyapunov exponents together with their multiplicities are essentially nonrandom in this case, see, e.g., [118, Chapter V]). The reader is referred to [7] for an account of various works on stochastic flows of diffeomorphisms by Baxendale and others and to Section 3.3 for entropy formula of Pesin type for such flows.

3. Relations between entropy, exponents and dimension

In this section we present some results concerning the relationship between the (Kolmogorov–Sinai) entropy, Lyapunov exponents and some dimensional characteristics of an RDS.

3.1. *Entropy formula of Pesin type*

In this section we assume that $\vartheta : (\Omega, \mathcal{F}, \mathbb{P}) \hookleftarrow$ is a (possibly noninvertible) measure-preserving transformation of a probability space and the RDS F over $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ is generated by a measurable map

$$F : \Omega \to C^r(M, M), \quad \omega \mapsto F_\omega.$$

Entropy revisited. In Section 1.1 the entropy was defined for a bundle RDS over an invertible system $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$. When $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ is possibly noninvertible, entropy can be defined in the same way. Namely, let in our present setting ξ be a finite Borel partition of M, the limit

$$h_\mu(F, \xi) := \lim_{n \to +\infty} \frac{1}{n} \int H_{\mu_\omega}\left(\bigvee_{k=0}^{n-1} \left(F_\omega^k\right)^{-1} \xi \right) d\mathbb{P}(\omega) \tag{3.1.1}$$

always exists since

$$\frac{1}{n} \int H_{\mu_\omega}\left(\bigvee_{k=0}^{n-1} \left(F_\omega^k\right)^{-1} \xi \right) d\mathbb{P}(\omega) = \frac{1}{n} H_\mu\left(\bigvee_{k=0}^{n-1} \Theta^{-k} \pi_2^{-1} \xi \,\Big|\, \pi_1^{-1} \mathcal{F} \right), \tag{3.1.2}$$

where $\pi_1 : \Omega \times M \to \Omega$, $(\omega, x) \mapsto \omega$ is the projection to the first factor, $\pi_2 : \Omega \times M \to M$, $(\omega, x) \mapsto x$ is the projection to the second, and the limit of the right-hand side of (3.1.2) exists as $n \to +\infty$ [82, Theorem II.1.1]; the *entropy* of (F, μ) is defined by

$$h_\mu(F) = \sup\left\{ h_\mu(F, \xi) \colon \xi \text{ a finite Borel partition of } M \right\}.$$

This notion describes the average information creation rate by actions on M of the random sequence of maps and it coincides with the relative (or conditional) entropy of Θ: $(\Omega \times M, \mu) \hookleftarrow$ (with respect to $\pi_1^{-1}\mathcal{F} = \mathcal{F}_\mathcal{E}$ where $\mathcal{E} = \Omega \times M$) defined by (1.1.2) and (1.1.3), namely,

$$h_\mu(F) = h_\mu^{(\mathrm{r})}(\Theta). \tag{3.1.3}$$

In many cases this coincidence enables us to investigate the entropy of an RDS via the classical (conditional) entropy theory.

As with a deterministic transformation, just considering finite or countable partitions of the manifold is not enough to investigate various properties of the entropy of an RDS, in some cases one has to appeal to finer (noncountable) partitions (for example, partitions into pieces of stable or unstable manifolds). This was made possible by the theory of conditional entropies associated with measurable partitions (i.e., partitions which can be approximated by finite or countable ones) of Lebesgue spaces (see [142] for the theory). When $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ is Polish, $(\Omega \times M, \mathcal{B}_\mu(\Omega \times M), \mu)$ defines a Lebesgue space, which will be written as $(\Omega \times M, \mu)$ for simplicity of notation. In this case with ϑ being

assumed measurably invertible, one can obtain the entropy $h_\mu(F)$ by the formula [118, Theorem 0.5.1]

$$h_\mu^{(\mathrm{r})}(\Theta) = \sup_\xi H_\mu\left(\xi \,\Big|\, \bigvee_{k=1}^{+\infty} \Theta^{-k}\xi \vee \sigma_0\right), \tag{3.1.4}$$

where the supremum is taken over the set of all measurable partitions of $(\Omega \times M, \mu)$ and $\sigma_0 = \{\{\omega\} \times M\colon\ \omega \in \Omega\}$. This means that, if one takes for each $\omega \in \Omega$ a partition ξ_ω of M such that $\xi := \{\{\omega\} \times \xi_\omega(x)\colon\ \omega \in \Omega,\ x \in M\}$ constitutes a measurable partition of $(\Omega \times M, \mu)$, then (3.1.1) can be extended by

$$\begin{aligned} h_\mu(F, \xi) &:= H_\mu\left(\xi \,\Big|\, \bigvee_{k=1}^{+\infty} \Theta^{-k}\xi\right) \\ &= \int H_{\mu_\omega}\left(\xi_\omega \,\Big|\, \bigvee_{k=1}^{+\infty} \left(F_\omega^k\right)^{-1}\xi_{\vartheta^k\omega}\right) d\mathbb{P}(\omega) \end{aligned} \tag{3.1.5}$$

and one can obtain the entropy $h_\mu(F)$ by taking the supremum of $h_\mu(F, \xi)$ over the set of all such measurable partitions ξ. For more information about relative (or conditional) entropy of measure-preserving transformations of Lebesgue spaces see [118, Chapter 0] which is a partial modification of the usual entropy arguments given in [142].

Ruelle inequality. Roughly speaking, Lyapunov exponents and entropy provide two different ways of measuring the dynamical complexity of an RDS: The Lyapunov exponents measure geometrically how fast nearby orbits diverge (due to the corresponding invariant manifolds theory) or how fast volume elements are expanded (by sums of the exponents), and the entropy measures from the pointview of information the complexity related to such dynamical behaviors. In general, there is the following inequality (3.1.6) relating these two kinds of quantities.

THEOREM 3.1.1 (Ruelle inequality). *Let F be of class C^1 (i.e., $r = 1$) and assume $\log^+ |F_\omega|_{C^1} \in L^1(\Omega, \mathbb{P})$ where $|f|_{C^1} = \sup_{\xi \in TM,\, |\xi|=1} |Df\xi|$ for $f \in C^1(M, M)$. Then for any F-invariant measure μ one has*

$$h_\mu(F) \leqslant \int \sum_i \lambda^{(i)}(\omega, x)^+ m^{(i)}(\omega, x)\, d\mu. \tag{3.1.6}$$

The inequality (3.1.6) was first proved by Ruelle [146] (also by Margulis in an unpublished work) for a deterministic C^1 map, and it was first extended to i.i.d. RDS by Kifer [82]. When $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ is measurably invertible, Theorem 3.1.1 was proved in [11]. We include below a proof of the result in its general form stated above. The reader is referred to Ruelle [150] for an elaboration of the inequality for infinite dimensional RDS, which is suitable for applications to the Navier–Stokes time evolution.

LEMMA 3.1.2. *Let ξ_n, $n = 1, 2, \dots$, be a sequence of finite measurable partitions of M such that* $\operatorname{diam}\xi_n \to 0$ *as* $n \to +\infty$ (*where* $\operatorname{diam}\xi_n := \sup_{C\in\xi_n} \operatorname{diam} C$). *Then*

$$h_\mu(F) = \lim_{n\to+\infty} h_\mu(F, \xi_n).$$

PROOF. Fix $\varepsilon > 0$ arbitrarily small and choose a finite measurable partition α of M such that $h_\mu(F) \leqslant h_\mu(F, \alpha) + \varepsilon$. Identifying $\pi_2^{-1}\xi_n$ with the σ-algebra generated by it, one has, by [118, Theorem 0.4.2],

$$\begin{aligned} h_\mu(F, \alpha) &= h_\mu^{\pi_1^{-1}\mathcal{F}}\big(\Theta, \pi_2^{-1}\alpha\big) \\ &\leqslant h_\mu^{\pi_1^{-1}\mathcal{F}}\big(\Theta, \pi_2^{-1}\xi_n\big) + H_\mu\big(\pi_2^{-1}\alpha|\pi_2^{-1}\xi_n \vee \pi_1^{-1}\mathcal{F}\big) \\ &= h_\mu(F, \xi_n) + \int H_{\mu_\omega}(\alpha|\xi_n)\, d\mathbb{P}(\omega). \end{aligned}$$

This implies that for sufficiently large n,

$$h_\mu(F) \leqslant h_\mu(F, \xi_n) + 2\varepsilon,$$

since $H_{\mu_\omega}(\alpha|\xi_n) \to 0$ for every ω as $n \to +\infty$. This proves the lemma. □

PROOF OF THEOREM 3.1.1. We first make the following additional assumptions: M is a compact subset of $(\mathbb{R}^d, \|\cdot\|)$, U is an open $2r_0$-neighborhood of M for some $r_0 > 0$, $F : \Omega \to C^1(U, U)$, $\omega \to F_\omega$ is a measurable map such that $F_\omega M \subset M$ for each ω and

$$\int_\Omega \log^+ \sup_{x\in U} \|D_x F_\omega\|\, d\mathbb{P}(\omega) < +\infty. \tag{3.1.7}$$

This generates an RDS F on U over $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ and let μ be an F-invariant measure concentrated on $\Omega \times M$.

For a $d \times d$ matrix A with real entries we will use a singular value decomposition of it, $A = Q_1 \Delta Q_2$, where Q_1, Q_2 are unitary matrices and Δ is a diagonal matrix whose diagonal elements will be denoted as $0 \leqslant \delta_1(A) \leqslant \delta_2(A) \leqslant \cdots \leqslant \delta_d(A)$.

Let $k \in \mathbb{N}$. Write $\omega \in \Omega_k$ if for all $x, y \in B(M, r_0)$ with $\|x - y\| \leqslant \frac{r_0}{k}$ we have

$$\big\|F_\omega y - F_\omega x - D_x F_\omega(y - x)\big\| \leqslant \|y - x\| \tag{3.1.8}$$

and

$$\big|\delta_i(D_y F_\omega) - \delta_i(D_x F_\omega)\big| \leqslant \frac{1}{2}, \quad 1 \leqslant i \leqslant d,$$

which clearly implies that

$$\frac{1}{2} \leqslant \frac{\max\{1, \delta_i(D_y F_\omega)\}}{\max\{1, \delta_i(D_x F_\omega)\}} \leqslant 2, \quad 1 \leqslant i \leqslant d. \tag{3.1.9}$$

One can check that each Ω_k is measurable and $\mathbb{P}(\Omega_k) \to 1$ as $k \to +\infty$.

For each $k \in \mathbb{N}$ take a maximal $\frac{r_0}{k}$-separated set E_k of M, i.e., a subset E_k of M such that $d(x, y) > \frac{r_0}{k}$ for any $x, y \in E_k$ with $x \neq y$ and for any $z \in M$ there is an element $x \in E_k$ satisfying $d(x, z) \leqslant \frac{r_0}{k}$. We then define a finite measurable partition $\xi_k = \{\xi_k(x)\colon\ x \in E_k\}$ of M such that, with respect to the inherited topology of M as a subset of $\mathbb{R}^d$, $\xi_k(x) \subset \overline{\operatorname{int}(\xi_k(x))}$ and $\operatorname{int}(\xi_k(x)) = \{z \in M\colon\ \|z - x\| < \|z - y\| \text{ for all } y \in E_k \text{ with } y \neq x\}$. Clearly $\xi_k(x) \subset B(x, \frac{r_0}{k})$ for all $x \in E_k$ and $\operatorname{diam} \xi_k \leqslant \frac{r_0}{k}$. Then, by Lemma 3.1.2,

$$h_\mu(F) = \lim_{k \to +\infty} h_\mu(F, \xi_k). \tag{3.1.10}$$

For each ξ_k,

$$\begin{aligned}
&h_\mu(F, \xi_k)\\
&\quad= \lim_{n \to +\infty} \frac{1}{n} H_\mu\left(\bigvee_{l=0}^{n-1} \Theta^{-l} \pi_2^{-1} \xi_k \,\Big|\, \pi_1^{-1} \mathcal{F} \right)\\
&\quad= \lim_{n \to +\infty} \frac{1}{n} \sum_{l=1}^{n-1} H_\mu\left(\Theta^{-l}\left(\pi_2^{-1} \xi_k\right) \,\Big|\, \bigvee_{i=0}^{l-1} \Theta^{-i}\left(\pi_2^{-1} \xi_k\right) \vee \left(\pi_1^{-1} \mathcal{F}\right) \right)\\
&\qquad+ \lim_{n \to +\infty} \frac{1}{n} H_\mu\left(\pi_2^{-1} \xi_k | \pi_1^{-1} \mathcal{F}\right)\\
&\quad\leqslant \lim_{n \to +\infty} \frac{1}{n} \sum_{l=1}^{n-1} H_\mu\left(\Theta^{-l}\left(\pi_2^{-1} \xi_k\right) \big| \Theta^{-(l-1)}\left(\pi_2^{-1} \xi_k\right) \vee \Theta^{-(l-1)}\left(\pi_1^{-1} \mathcal{F}\right)\right)\\
&\quad= H_\mu\left(\Theta^{-1}\left(\pi_2^{-1} \xi_k\right) \big| \left(\pi_2^{-1} \xi_k\right) \vee \pi_1^{-1} \mathcal{F}\right)\\
&\quad= \int H_{\mu_\omega}\left((F_\omega)^{-1} \xi_k | \xi_k\right) d\mathbb{P}(\omega)\\
&\quad= \int_{\Omega_k} H_{\mu_\omega}\left((F_\omega)^{-1} \xi_k | \xi_k\right) d\mathbb{P}(\omega) + \int_{\Omega \setminus \Omega_k} H_{\mu_\omega}\left((F_\omega)^{-1} \xi_k | \xi_k\right) d\mathbb{P}(\omega)\\
&\quad=: M_k + m_k.
\end{aligned}$$

In what follows we estimate M_k and m_k. Let $\omega \in \Omega$. Let $N_k(\omega, x)$ be the number of elements of ξ_k which intersect $F_\omega \xi_k(x)$ for $x \in E_k$. By (3.1.8),

$$\begin{aligned}
F_\omega \xi_k(x) &\subset F_\omega B\left(x, \frac{r_0}{k}\right) \subset B\left(F_\omega x + D_x F_\omega B\left(0, \frac{r_0}{k}\right), \frac{r_0}{k}\right)\\
&= F_\omega x + B\left(D_x F_\omega B\left(0, \frac{r_0}{k}\right), \frac{r_0}{k}\right).
\end{aligned}$$

Hence, if $\xi_k(x') \cap F_\omega \xi_k(x) \neq \emptyset$, then

$$B\left(x', \frac{r_0}{2k}\right) \cap \left[F_\omega x + B\left(D_x F_\omega B\left(0, \frac{r_0}{k}\right), \frac{2r_0}{k}\right)\right] \neq \emptyset.$$

Since $B(x', \frac{r_0}{2k})$, $x' \in E_k$ are disjoint, one has

$$N_k(\omega, x) \leqslant K(\omega, x) := C(d) \prod_{j=1}^{d} \max\{\delta_j(D_x F_\omega), 1\},$$

where $C(d)$ is a constant depending only on d. Thus

$$\begin{aligned} H_{\mu_\omega}\big((F_\omega)^{-1}\xi_k|\xi_k\big) &\leqslant \sum_{x\in E_k} \mu_\omega\big(\xi_k(x)\big) \log K(\omega, x) \\ &= \sum_{x\in E_k} \int_{\xi_k(x)} \log K(\omega, x)\, d\mu_\omega(y) \\ &\leqslant \log C(d) + d\log 2 + \int_M \sum_{j=1}^{d} \log^+ \delta_j(D_y F_\omega)\, d\mu_\omega(y) \\ &\qquad \text{(by (3.1.9))} \end{aligned}$$

and we then have

$$M_k \leqslant \log C(d) + d\log 2 + \int_\Omega \int_M \sum_{j=1}^{d} \log^+ \delta_j(D_y F_\omega)\, d\mu_\omega(y)\, d\mathbb{P}(\omega). \quad (3.1.11)$$

We next estimate $N_k(\omega, x)$ for $\omega \in \Omega \setminus \Omega_k$. We now cannot make use of (3.1.8) and instead we will use the following estimate:

$$\|F_\omega y - F_\omega z\| \leqslant L_k(\omega)\|y - z\|$$

for all $k \geqslant 2$ and $y, z \in B(M, \frac{r_0}{k})$ with $\|y - z\| \leqslant \frac{r_0}{k}$, where

$$L_k(\omega) := \sup_{x\in B(M, \frac{r_0}{k})} \|D_x F_\omega\|.$$

Let $k \geqslant 2$. Then for each $x \in E_k$,

$$F_\omega \xi_k(x) \subset F_\omega B\left(x, \frac{r_0}{k}\right) \subset B\left(F_\omega x, L_k(\omega)\frac{r_0}{k}\right).$$

Thus, $N_k(\omega, x)$ cannot exceed the maximal number of disjoint balls with radius $\frac{r_0}{k}$ which intersect $B(F_\omega x, (L_k(\omega)+1)\frac{r_0}{k})$, and hence

$$N_k(\omega, x) \leqslant C'(d) \max\{L_k(\omega), 1\}^d,$$

where $C'(d)$ is a constant depending only on d. It then follows that

$$\begin{aligned} H_{\mu_\omega}\big((F_\omega)^{-1}\xi_k|\xi_k\big) &\leqslant \sum_{x\in E_k} \mu_\omega\big(\xi_k(x)\big)\log N_k(\omega, x) \\ &\leqslant \log C'(d) + d\log^+ L_k(\omega) \\ &\leqslant \log C'(d) + d\log^+ L_2(\omega) \end{aligned}$$

and hence

$$m_k \leqslant \log C'(d) + d\int_{\Omega\setminus\Omega_k} \log^+ L_2(\omega)\, d\mathbb{P}(\omega). \tag{3.1.12}$$

(3.1.7) implies that $\log^+ L_2(\omega) \in L^1(\Omega, \mathbb{P})$. This together with (3.1.12) yields

$$\limsup_{k\to+\infty} m_k \leqslant \log C'(d) \tag{3.1.13}$$

since $\mathbb{P}(\Omega \setminus \Omega_k) \to 0$ as $k \to +\infty$. By (3.1.10), (3.1.11) and (3.1.13) one has

$$\begin{aligned} h_\mu(F) &\leqslant \limsup_{k\to+\infty} M_k + \limsup_{k\to+\infty} m_k \\ &\leqslant C'' + \int_\Omega\int_M \sum_{j=1}^d \log^+ \delta_j(D_y F_\omega)\, d\mu_\omega(y)\, d\mathbb{P}(\omega), \end{aligned}$$

where $C'' = \log C(d) + d\log 2 + \log C'(d)$.

Considering the RDS F^n generated by the map $\omega \mapsto F_\omega^n$ for any $n \geqslant 1$ in the arguments above yields

$$\begin{aligned} h_\mu(F) &= \lim_{n\to+\infty} \frac{1}{n} h_\mu(F^n) \\ &\leqslant \lim_{n\to+\infty} \frac{1}{n}\left[C'' + \int_{\Omega\times M} \sum_{j=1}^d \log^+ \delta_j(D_y F_\omega^n)\, d\mu(\omega, y)\right] \\ &= \int \sum_j \lambda^{(j)}(\omega, y)^+ m^{(j)}(\omega, y)\, d\mu(\omega, y). \end{aligned}$$

Finally we show why we can make the additional assumptions introduced at the beginning of the proof to prove Theorem 3.1.1. Let M and F be as given in the statement of Theorem 3.1.1 and let $m_0 = \dim M$. Let h be a C^∞ embedding of M into

$\mathbb{R}^{2m_0+1}$. We will identify M and TM, respectively, with their images under h and Dh. Let $\nu(M) = \{(x, v) \in M \times \mathbb{R}^{2m_0+1}\colon\ v \perp T_x M\}$ be the normal bundle of M. Then there is $r_0 > 0$ such that $N_{r_0}(M) := \{(x, v) \in \nu(M)\colon\ \|v\| < r_0\}$ is equal to $B(M, r_0)$ ($N_{r_0}(M)$ is called a tubular neighborhood of M, see, e.g., [60]). Define $\hat{F}_\omega = F_\omega \circ \pi : B(M, r_0) \to B(M, r_0)$, $\omega \in \Omega$, where $\pi : \nu(M) \to M$ is the natural projection, and let $\hat{F}$ be the RDS generated by these maps. Clearly $\hat{F}$ satisfies the additional assumptions alluded to above. Since all Riemannian metrics on M are equivalent (duo to M being compact) and $D_x \hat{F}_\omega v = 0$ for $v \in T_x N_{r_0}(M)$ with $v \perp T_x M$, all the positive Lyapunov exponents of $(\hat{F}, \mu)$ coincide with those of (F, μ). $h_\mu(\hat{F})$ is clearly equal to $h_\mu(F)$. This completes the proof. □

Pesin formula. For a deterministic C^2 (or $C^{1+\alpha}$) diffeomorphism f on M and an f-invariant measure ρ, Pesin [134] proved that (3.1.6) will be an equality if ρ is smooth (i.e., absolutely continuous with respect to the Lebesgue measure on M) and this equality is known as the *Pesin* (*entropy*) *formula*. The following theorem extends the result of Pesin to RDS composed of random endomorphisms (maybe noninvertible and with singularities, i.e., points at which the derivative is degenerate).

THEOREM 3.1.3 (Pesin formula). *Let F be of class C^2 (i.e., $r = 2$) with $\log^+ |F_\omega|_{C^2} \in L^1(\Omega, \mathbb{P})$ and $\log \hat{D}(F_\omega) \in L^1(\Omega, \mathbb{P})$ where $\hat{D}(F_\omega) := \inf_{x \in M} |\det D_x F_\omega|$. Assume that $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ is Polish. Let μ be an F-invariant measure. If $\mu_\omega \ll \mathrm{Leb}$ for $\mathbb{P}$-a.e. ω, then there holds the equality*

$$h_\mu(F) = \int \sum_i \lambda^{(i)}(\omega, x)^+ m^{(i)}(\omega, x)\, d\mu. \tag{3.1.14}$$

REMARK 3.1.4. When $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ is assumed moreover measurably invertible, the integrability condition $\log \hat{D}(F_\omega) \in L^1(\Omega, \mathbb{P})$ can be suppressed in Theorem 3.1.3 [116].

REMARK 3.1.5. When $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ is Polish and noninvertible, a probability μ on $(\Omega \times M, \mathcal{F} \times \mathcal{B}(M))$ with marginal $\mathbb{P}$ on Ω is Θ-invariant if and only if

$$\int_{\vartheta^{-1}\{\omega\}} F_{\omega'} \mu_{\omega'}\, d\mathbb{P}_\omega(\omega') = \mu_\omega, \quad \mathbb{P}\text{-a.e. } \omega,$$

where $\{\mathbb{P}_\omega\}_{\omega \in \Omega}$ is a canonical system of conditional measures of $\mathbb{P}$ associated with the partition $\{\vartheta^{-1}\{\omega\}\colon\ \omega \in \Omega\}$ of Ω.

Theorem 3.1.3 was first proved by Ledrappier and Young [112] for i.i.d. random diffeomorphisms and then in [116,119] in the present form. See Section 3.3 for reformulation of this theorem for i.i.d. RDS which are of particular interest from the point of view of Markov processes. We present below a description of the main ingredients of the proof.

Jacobian. One ingredient is to overcome the difficulty caused by the fact that the random maps here are in general not one-to-one but the usual inverse limit space method does

not seem helpful. The notion of the Jacobian of a measure-preserving transformation introduced by [132] turns out to be very useful for dealing with local homeomorphisms. The definition is as follows.

Let $f : X \to Y$ be a measure-preserving transformation between two probability spaces $(X, \mathcal{A}, \nu)$ and $(Y, \mathcal{B}, \rho)$. Assume that there is a countable measurable partition $\alpha = \{A_i\}$ of X (ν-mod 0) such that for each A_i the map $f_i := f|_{A_i} : A_i \to Y$ is absolutely continuous, that is,

(i) f_i is injective;
(ii) $f_i(A)$ is measurable if A is a measurable subset of A_i;
(iii) $\rho(f_i(A)) = 0$ if $A \subset A_i$ is measurable and $\nu(A) = 0$.

By (i) and (ii) we can define a measure ν_{f_i} on each A_i by $\nu_{f_i}(A) = \rho(f_i(A))$ for measurable set $A \subset A_i$. By (iii), ν_{f_i} is absolutely continuous with respect to $\nu_i := \nu|_{A_i}$. Define a measurable function $J(f) : X \to \mathbb{R}^+$ by

$$J(f)(x) = \frac{d\nu_{f_i}}{d\nu_i}(x) \quad \text{if } x \in A_i.$$

It is easy to see that the definition of $J(f)$ is independent of the choice of partition α, and we will call $J(f)$ the *Jacobian* of f. Clearly, $J(f)(x) \geqslant 1$ for μ-a.e. $x \in X$. As an exercise, consider the following simple situation: if $f : M \to M$ is a C^1 map with no singularities and ν is a Borel probability on M such that $\nu \ll \text{Leb}$, then $f : (M, \nu) \to (M, f\nu)$ admits a Jacobian $J(f)$ given by

$$J(f)(x) = \frac{(\mathcal{L}_f l)(fx)}{l(x)} |\det D_x f|, \quad \nu\text{-a.e. } x \in M, \tag{3.1.15}$$

where l is the density of ν with respect to the Lebesgue and $\mathcal{L}_f l$ is defined by

$$(\mathcal{L}_f l)(x) = \sum_{y : fy = x} \frac{l(y)}{|\det D_y f|} \tag{3.1.16}$$

(note that $\mathcal{L}_f l = l$ ν-a.e. is equivalent to $f\nu = \nu$).

We now state a very useful property of this notion. Assume moreover that $(X, \mathcal{A}, \nu)$ and $(Y, \mathcal{B}, \rho)$ are both Lebesgue spaces. If ξ is a measurable partition of Y, by $\{\nu_x^{f^{-1}\xi}\}_{x \in X}$ and $\{\rho_y^{\xi}\}_{y \in Y}$ we will denote, respectively, a canonical system of conditional measures of ν and ρ associated with $f^{-1}\xi$ and ξ.

LEMMA 3.1.6 [116]. *Take the assumptions just above. Let ξ be a measurable partition of Y. Assume that $A \subset X$ is a measurable set such that $\nu(A) > 0$ and $f_A := f|_A : A \to fA$ is injective. Then for ν-a.e. $x \in A$ one has*

$$\nu_x^{f^{-1}\xi}(B) = \int_{fB} \frac{1}{J(f) \circ f_A^{-1}} \, d\rho_{f(x)}^{\xi}$$

for all measurable sets $B \subset (f^{-1}\xi)(x) \cap A$.

Let ϵ be the partition of Y into single points. Let $\alpha = \{A_i\}$ be the partition introduced just above for f. By taking $A = A_i$ and $B = \{x\}$ in Lemma 3.1.6, we obtain [132, Lemma 10.5] which says that

$$\log J(f)(x) = -\log \nu_x^{f^{-1}\epsilon}\big(\{x\}\big) \quad \text{for } \nu\text{-a.e. } x. \tag{3.1.17}$$

If $f : M \to M$ is a C^2 expanding map and ν is the unique smooth invariant measure of f, then, by (3.1.17), one has

$$h_\nu(f) \geqslant H_\nu\big(\epsilon | f^{-1}\epsilon\big) = -\int \log \nu_x^{f^{-1}\epsilon}\big(\{x\}\big)\, d\nu(x) = \int \log J(f)(x)\, d\nu(x)$$

which together with (3.1.16) and the Ruelle inequality gives the Pesin formula for (f, ν). This simple proof is due to Ledrappier and Young. For a general C^2 map f on M preserving a smooth measure ν, one can think of replacing each single point in the above arguments with a piece of stable manifold. Then Lemma 3.1.6 allows us to analyze the contribution to the information creation made by actions on such pieces of f restricted to each area on which the map is injective. The idea is similar for the random case, see outline of the proof of Theorem 3.1.3 given below.

We remark on another component of the proof of Theorem 3.1.3. It consists in analysis along the stable manifolds of (F, μ), only for which the smoothness of μ_ω can be used since $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ and F_ω, $\omega \in \Omega$, may be noninvertible. The smoothness of μ_ω, $\omega \in \Omega$, implies that μ has smooth (i.e., absolutely continuous) conditional measures on the stable manifolds (Corollary 2.2.15) and this allows one to measure information creation along these manifolds in terms of the Lebesgue measures on them. This statement is seemingly unreasonable since intuitively it has something to do only with the negative exponents, but now the positive and the negative exponents do not play a symmetric role (consider, e.g., an expanding map and see also Proposition 3.3.7(1)). Here is an intuitive explanation which is very incomplete and technically inaccurate, but it would tell roughly how the use of stable manifolds can lead to the relation of the entropy and the positive exponents. Let ξ be a finite Borel partition of M. For notations to make sense we assume that F_ω is invertible for $\mathbb{P}$-a.e. ω. One can rewrite the term in (3.1.1) under the integration signal in the following way:

$$H_{\mu_\omega}\left(\bigvee_{k=0}^{n-1} \big(F_\omega^k\big)^{-1}\xi\right) = H_{F_\omega^n \mu_\omega}\left(\bigvee_{k=0}^{n-1} F_{\vartheta^k\omega}^{n-k}\xi\right).$$

In this way the contraction goes into the picture of information creation. Replacing ξ by a partition into pieces of the stable manifolds and using the conditional entropy theory, the determinants $|\det D_x F_\omega^n|$ and $|\det(D_x F_\omega^n|_{E^s(\omega,x)})|$ would play a major role since the former roughly determines the *local* volume change rate of $F_\omega^n : (M, \mu_\omega) \to (M, F_\omega^n \mu_\omega)$ (see (3.1.15)) and the letter roughly determines the volume change rate under the action of F_ω along the stable manifold. Finally the entropy $h_\mu(F)$ turns out to be determined by $\frac{1}{n}\log(|\det D_x F_\omega^n|/|\det(D_x F_\omega^n|_{E^s(\omega,x)})|)$ (the quotient after the logarithm signal may be

regarded as the volume expansion rate in the expanding direction), whose integration gives that of the sum of the positive exponents.

SKETCH OF THE PROOF OF THEOREM 3.1.3. By Theorem 3.1.1, it remains to prove

$$h_\mu(F) \geqslant \int \sum_i \lambda^{(i)}(\omega, x)^+ m^{(i)}(\omega, x)\, d\mu. \tag{3.1.18}$$

Let $\Gamma \subset \Omega \times M$ be a μ-full set such that $\Theta\Gamma \subset \Gamma$ and each point in Γ is regular in the sense of Oseledec. Set $I = \{(\omega, x) \in \Gamma\colon\ \lambda^{(i)}(\omega, x) \geqslant 0,\ 1 \leqslant i \leqslant r(\omega, x)\}$ and $\Delta = \Gamma \backslash I$. Define $W^s(\omega, x) = \{x\}$ for $(\omega, x) \in I$.

In what follows we write $\lambda = \text{Leb}$. Recall that, by assumption, $\mu_\omega \ll \lambda$ for $\mathbb{P}$-a.e. ω. Construct then a measurable partition η of $(\Omega \times M, \mu)$ which has the following properties:

(i) $\Theta^{-1}\eta \leqslant \eta$ (i.e., $(\Theta^{-1}\eta)(\omega, x) \supset \eta(\omega, x)$ for μ-a.e. (ω, x)), $\sigma := \{\{\omega\} \times M\colon\ \omega \in \Omega\} \leqslant \eta$;

(ii) η is subordinate to W^s-manifolds of (F, μ);

(iii) for every $B \in \mathcal{B}(\Omega \times M)$ the function $(\omega, x) \mapsto \lambda^s_{(\omega,x)}(\eta_\omega(x) \cap B_\omega)$ is measurable and μ-a.e. finite, where $B_\omega = \{y\colon\ (\omega, y) \in B\}$ and $\lambda^s_{(\omega,x)}$ is the Lebesgue measure on $W^s(\omega, x)$ induced by its inherited Riemannian structure as a submanifold of M ($\lambda^s_{(\omega,x)} = \delta_x$ if $W^s(\omega, x) = \{x\}$);

(iv) $(\mu_\omega)^{\eta_\omega}_x \ll \lambda^s_{(\omega,x)}$ for μ-a.e. (ω, x), where $(\mu_\omega)^{\eta_\omega}_x$ is the conditional measure of μ_ω on $\eta_\omega(x)$.

Let η be as given above. By a computation similar to [116, (4.8)] one has

$$h_\mu(F) \geqslant \lim_{n\to+\infty} \frac{1}{n} H_\mu(\eta | \Theta^{-n}\eta \vee \sigma)$$

if

$$H_\mu(\eta | \Theta^{-n}\eta \vee \sigma) < +\infty \tag{3.1.19}$$

for all $n \geqslant 1$. Hence, in order to prove (3.1.18), it is sufficient to prove that for every $n \geqslant 1$ there hold (3.1.19) and

$$\frac{1}{n} H_\mu(\eta | \Theta^{-n}\eta \vee \sigma) \geqslant \int \sum_i \lambda^{(i)}(\omega, x)^+ m_i(\omega, x)\, d\mu. \tag{3.1.20}$$

Fix $n \geqslant 1$ arbitrarily. By the definition of conditional entropies one has

$$\begin{aligned} H_\mu(\eta | \Theta^{-n}\eta \vee \sigma) &= -\int \log \mu^{\Theta^{-n}\eta \vee \sigma}_{(\omega,x)}\big(\eta(\omega, x)\big)\, d\mu(\omega, x) \\ &= -\int_\Omega \int_M \log(\mu_\omega)^{(F^n_\omega)^{-1}\eta_{\vartheta^n\omega}}_x\big(\eta_\omega(x)\big)\, d\mu_\omega(x)\, d\mathbb{P}(\omega), \end{aligned}$$

where $\{\nu_z^\xi\}_{z\in X}$ denotes a canonical system of conditional measures of ν associated with a measurable partition ξ of a Lebesgue space $(X, \mathcal{A}, \nu)$.

Since $\mu \ll \lambda \times \mathbb{P}$ we can define

$$\phi = \frac{d\mu}{d(\lambda \times \mathbb{P})}$$

which implies

$$\phi_\omega(\cdot) := \phi(\omega, \cdot) = \frac{d\mu_\omega}{d\lambda}(\cdot)$$

for $\mathbb{P}$-a.e. ω. Put $\Lambda = \{(\omega, x)\colon\ \phi(\omega, x) > 0\}$. By Remark 3.1.5, $F_\omega^n \mu_\omega(M \setminus \Lambda_{\vartheta^n\omega}) = 0$ for $\mathbb{P}$-a.e. ω. Then it can be checked that $F_\omega^n : (\Lambda_\omega \cap (F_\omega^n)^{-1}\Lambda_{\vartheta^n\omega}, \mu_\omega) \to (\Lambda_{\vartheta^n\omega}, F_\omega^n\mu_\omega)$ admits the Jacobian

$$J(F_\omega^n)(x) = \frac{(\mathcal{L}_{F_\omega^n}\phi_\omega)(F_\omega^n x)}{\phi_\omega(x)} |\det D_x F_\omega^n|$$

(see (3.1.15)).

Define a Borel measure λ^* on $\Omega \times M$ by

$$\lambda^*(B) = \int \lambda^s_{(\omega,x)}\big(\eta_\omega(x) \cap B_\omega\big)\, d\mu(\omega, x) \quad \text{for Borel } B \in \Omega \times M.$$

Since $\mu(B) = \int (\mu_\omega)_x^{\eta_\omega}(\eta_\omega(x) \cap B_\omega)\, d\mu(\omega, x)$ and $(\mu_\omega)_x^{\eta_\omega} \ll \lambda^s_{(\omega,x)}$, we have $\mu \ll \lambda^*$. Put

$$g(\omega, x) = \frac{d\mu}{d\lambda^*}(\omega, x).$$

We then have for μ-a.e. (ω, x),

$$g(\omega, y) = \frac{d(\mu_\omega)_x^{\eta_\omega}}{d\lambda^s_{(\omega,x)}}(y), \quad \lambda^s_{(\omega,x)}\text{-a.e. } y \in \eta_\omega(x).$$

Define for μ-a.e. $(\omega, x) \in \Omega \times M$,

$$W_n(\omega, x) = (\mu_\omega)_x^{(F_\omega^n)^{-1}\eta_{\vartheta^n\omega}}\big(\eta_\omega(x)\big),$$

$$X_n(\omega, x) = \frac{\phi(\omega, x)}{\phi \circ \Theta^n(\omega, x)} \frac{g \circ \Theta^n(\omega, x)}{g(\omega, x)},$$

$$Y_n(\omega, x) = \begin{cases} \frac{|\det D_x F_\omega^n|_{E^s(\omega,x)}|}{|\det D_x F_\omega^n|} & \text{if } (\omega, x) \in \Delta, \\ \frac{1}{|\det D_x F_\omega^n|} & \text{if } (\omega, x) \in I, \end{cases}$$

$$Z_n(\omega, x) = \int_{\eta_{\vartheta^n\omega}(F_\omega^n x)} \Phi_n(\omega, y)\, d(\mu_{\vartheta^n\omega})^{\eta_{\vartheta^n\omega}}_{F_\omega^n x}(y),$$

where

$$\Phi_n(\omega, y) := \frac{dF_\omega^n \mu_\omega}{d\mu_{\vartheta^n\omega}}(y) = \frac{(\mathcal{L}_{F_\omega^n}\phi_\omega)(y)}{\phi_{\vartheta^n\omega}(y)}.$$

W_n, X_n, Y_n, Z_n are all measurable and μ-a.e. finite. The proof will be completed after one proves the four claims below.

Claim 1. $W_n = \frac{X_n Y_n}{Z_n}$ μ-a.e. on $\Omega \times M$;

Claim 2. $\log Y_n \in L^1(\Omega \times M, \mu)$ and

$$-\int \frac{1}{n} \log Y_n \, d\mu = \int \sum_i \lambda^{(i)}(\omega, x)^+ m^{(i)}(\omega, x) \, d\mu.$$

Claim 3. $\log Z_n \in L^1(\Omega \times M, \mu)$ and $\int \log Z_n \, d\mu \geqslant 0$.

Claim 4. $\log X_n \in L^1(\Omega \times M, \mu)$ and $\int \log X_n \, d\mu = 0$.

Claim 1 is the key point and we present its proof below. In order to prove the claim, it suffices to prove that for μ-a.e. (ω, x) one has

$$W_n(\omega, y) = \frac{X_n(\omega, y) Y_n(\omega, y)}{Z_n(\omega, y)}, \quad (\mu_\omega)_x^{\eta_\omega}\text{-a.e. } y \in \eta_\omega(x). \tag{3.1.21}$$

For $\mathbb{P}$-a.e. ω choose a countable Borel partition $\{A_{\omega,i}^n\}_{i=1}^{+\infty}$ of M such that $F_{\omega,i}^n := F_\omega^n|_{A_{\omega,i}^n}$ is injective for each i. Then, for μ-a.e. (ω, x), we have for any measurable $B \subset \eta_\omega(x)$,

$$\begin{aligned}
(\mu_\omega)_x^{\eta_\omega}(B) &= \frac{1}{W_n(\omega, x)} (\mu_\omega)_x^{(F_\omega^n)^{-1}\eta_{\vartheta^n\omega}}(B) \\
&= \frac{1}{W_n(\omega, x)} \sum_i (\mu_\omega)_x^{(F_\omega^n)^{-1}\eta_{\vartheta^n\omega}}(B_i) \quad \left(B_i = B \cap A_{\omega,i}^n\right) \\
&= \frac{1}{W_n(\omega, x)} \sum_i \int_{F_\omega^n B_i} \frac{1}{J(F_\omega^n) \circ (F_{\omega,i}^n)^{-1}} d(F_\omega^n \mu_\omega)_{F_\omega^n x}^{\eta_{\vartheta^n\omega}} \\
&\qquad \text{(by Lemma 3.1.6)} \\
&= \frac{1}{W_n(\omega, x) Z_n(\omega, x)} \sum_i \int_{F_\omega^n B_i} \frac{\Phi_n(\omega, y)}{J(F_\omega^n) \circ (F_{\omega,i}^n)^{-1}(y)} d(\mu_{\vartheta^n\omega})_{F_\omega^n x}^{\eta_{\vartheta^n\omega}} \\
&= \frac{1}{W_n(\omega, x) Z_n(\omega, x)} \\
&\quad \times \sum_i \int_{F_\omega^n B_i} \frac{\Phi_n(\omega, y)}{J(F_\omega^n) \circ (F_{\omega,i}^n)^{-1}(y)} g(\vartheta^n\omega, y) d\lambda_{\Theta^n(\omega,x)}^s \\
&= \frac{1}{W_n(\omega, x) Z_n(\omega, x)}
\end{aligned}$$

$$\times \sum_i \int_{B_i} \frac{\Phi_n(\omega, F_\omega^n y)}{J(F_\omega^n)(y)} g\big(\Theta^n(\omega, y)\big) \big|\det(D_y F_\omega^n|_{E^s(\omega,y)})\big| \, d\lambda^s_{(\omega,x)}$$

$$= \frac{1}{W_n(\omega, x) Z_n(\omega, x)} \times \int_B \frac{\phi(\omega, y)}{\phi(\Theta^n(\omega, y))} Y_n(\omega, y) g\big(\Theta^n(\omega, y)\big) \, d\lambda^s_{(\omega,x)}.$$

On the other hand,

$$(\mu_\omega)_x^{\eta_\omega}(B) = \int_B g(\omega, y) \, d\lambda^s_{(\omega,x)}.$$

From the arbitrariness of B it follows that for μ-a.e. (ω, x),

$$\frac{\phi(\omega, y)}{\phi(\Theta^n(\omega, y))} Y_n(\omega, y) g\big(\Theta^n(\omega, y)\big) = g(\omega, y)$$

for $\lambda^s_{(\omega,x)}$-a.e. $y \in \eta_\omega(x)$. This proves (3.1.21) since $(\mu_\omega)_x^{\eta_\omega} \ll \lambda^s_{(\omega,x)}$ and $W_n(\omega, y) = W_n(\omega, x)$, $Z_n(\omega, y) = Z_n(\omega, x)$ for any $y \in \eta_\omega(x)$. This proves Claim 4. □

In the rest of this section we consider Pesin entropy formula for some particular RDS.

Let f be a (deterministic) C^2 map (maybe noninvertible and with singularities) on M preserving a smooth (i.e., absolutely continuous) probability measure μ. Then, by considering the one-point space $\Omega = \{f\}$, Theorem 3.1.3 together with Remark 3.1.4 implies that the Pesin formula holds true for the system $f : (M, \mu) \hookleftarrow$. If f is a C^2 expanding map on M (i.e., there exists $a > 1$ such that $|Df v| \geqslant a|v|$ for all $v \in TM$) and μ is the unique smooth invariant measure of f, then

$$h_\mu(f) = \int \sum_i \lambda^{(i)}(x) m^{(i)}(x) \, d\mu = \int \log|\det D_x f| \, d\mu;$$

if $A : K^p \to K^p$ is a surjective group endomorphism of a p-dimensional torus and m is the Haar measure on K^p, then

$$h_m(A) = \sum_{|\lambda_i|>1} \log|\lambda_i|,$$

where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of the linear transformation $\hat{A} : \mathbb{R}^p \to \mathbb{R}^p$ that covers A. The latter two results were proved previously in other ways (see [125, Section IV.5] and [162, Section 8.4], respectively), and now they can be obtained as corollaries of the general result Theorem 3.1.3.

Entropy formula for expanding in average RDS's. As applications of Theorem 3.1.3 we consider expanding RDS. Put $\Omega = C^r(M, M)^{\mathbb{Z}}$ (endowed with the product topology)

and let $\mathbb{P}$ be a Borel probability distribution on Ω which is invariant under the left shift operator ϑ on Ω. Let F be the RDS generated by the coordinate process on $(\Omega, \mathbb{P})$, i.e., by the map $\omega \to F_\omega := \omega_0$, $\omega = (\omega_i)_{i\in\mathbb{Z}} \in \Omega$. Assume that for $\mathbb{P}$-a.e. ω the map F_ω has no singularities. Define $|f|_{C^1}^- = \inf_{\xi\in TM,|\xi|=1} |Df\xi|$ for $f \in C^1(M, M)$. If $\log |F_\omega|_{C^1}^- \in L^1(\Omega, \mathbb{P})$ and with probability one we have

$$\lim_{n\to+\infty} \frac{1}{n} \sum_{k=0}^{n-1} \log |F_{\vartheta^k\omega}|_{C^1}^- =: a(\omega) > 0 \tag{3.1.22}$$

(the limit exists for $\mathbb{P}$-a.e. ω by Birkhoff ergodic theorem), F is then said to be *expanding in average*. This model was introduced in [79] (see also Section 4.2) and thermodynamic formalism for this model was developed there. Now let F be such an RDS and assume moreover that it is of class C^2 (i.e., $r = 2$) and $\log^+ |F_\omega|_{C^2} \in L^1(\Omega, \mathbb{P})$. Then there exists a unique F-invariant measure μ whose sample measures are almost all equivalent to the Lebesgue (see [79, Theorem B]; one can verify the conditions required there under the present assumptions). By Theorem 3.1.3, Pesin formula holds true for (F, μ). It is easy to see that the condition (3.1.22) implies that the smallest Lyapunov exponent of F is positive μ almost everywhere and hence

$$h_\mu(F) = \int \log |\det D_x F_\omega| \, d\mu(\omega, x).$$

Small random perturbations of expanding maps. Let F be as given in the last paragraph and let $r = 2$. If $\mathbb{P}$ is concentrated on $\Omega_\varepsilon := B_\varepsilon(f)^{\mathbb{Z}}$, where $\varepsilon > 0$ and $B_\varepsilon(f)$ is the ε-neighborhood of a C^2 expanding map $f : M \to M$ in $C^2(M, M)$ (endowed with a C^2-metric), the RDS F can be regarded as a small random perturbation of the expanding map f. Suppose that for each sufficiently small $\varepsilon > 0$ we are given such an RDS, which will be denoted by F_ε, and the corresponding distribution $\mathbb{P}_\varepsilon$ on Ω is ergodic (if it is not ergodic, the extension of the following results is straightforward by using the ergodic decomposition method). When ε is small enough, there is a unique F_ε-invariant measure μ_ε whose sample measures can be chosen so that, for each $\omega \in \Omega_\varepsilon$, $\mu_{\varepsilon,\omega}$ is equivalent to the Lebesgue with a C^1 density $l_{\varepsilon,\omega} = d\mu_{\varepsilon,\omega}/d\,\mathrm{Leb}$. Let l_0 be a C^1 density of the unique smooth f-invariant measure μ_0 with respect to the Lebesgue. Then one has

$$\lim_{\varepsilon\to 0} \sup_{\omega\in\Omega_\varepsilon} \sup_{x\in M} \left| l_{\varepsilon,\omega}(x) - l_0(x) \right| = 0. \tag{3.1.23}$$

Moreover, the measure μ_ε is ergodic. These results were proved in [91] and [16] by using random transfer operator method (see those papers for detailed treatments and for more information). Let $(\lambda_\varepsilon^{(i)}, m_\varepsilon^{(i)})$, $1 \leqslant i \leqslant r_\varepsilon$, be the Lyapunov spectrum of $(F_\varepsilon, \mu_\varepsilon)$. Then, from Theorem 3.1.3 and (3.1.23) there follows

COROLLARY 3.1.7. *For sufficiently small $\varepsilon > 0$ one has*

$$h_{\mu_\varepsilon}(F_\varepsilon) = \sum_i \lambda_\varepsilon^{(i)} m_\varepsilon^{(i)}$$

and

$$\lim_{\varepsilon\to 0} h_{\mu_\varepsilon}(F_\varepsilon) = h_{\mu_0}(f).$$

We remark that the second statement of this corollary can also be obtained as a special case of [33, Proposition 3.5].

3.2. *Relationship between entropy, exponents and dimension*

Set-up. In this section we assume that $\vartheta : (\Omega, \mathcal{F}, \mathbb{P}) \hookleftarrow$ is Polish and measurably invertible. The RDS F over $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ is generated by a measurable map

$$F : \Omega \to \mathrm{Diff}^2(M), \quad \omega \mapsto F_\omega$$

satisfying (2.3.7), i.e.,

$$\int \left(\log^+ |F_\omega|_{C^2} + \log^+ \left|(F_\omega)^{-1}\right|_{C^2}\right) d\mathbb{P}(\omega) < +\infty.$$

Let μ be an F-invariant measure.

Pesin formula and SRB measures. Careful observation makes it plausible that the information creation is caused by the expansion in the positive iterations of the random maps and the contracting and neutral directions in the positive iterations do not contribute to the entropy. This is made precise by the following proposition which tells that the entropy is determined by the action of the random maps on the unstable manifolds. For a measurable partition ξ of a Lebesgue space $(X, \mathcal{A}, \nu)$ we will denote by $\mathcal{B}(\xi)$ the σ-algebra generated by ξ (see [142] for the definition).

PROPOSITION 3.2.1. *Let (F, μ) be given. Then there exists a measurable partition η, finer than $\{\{\omega\} \times M\colon\ \omega \in \Omega\}$, of $(\omega \times M, \mu)$ which has the following properties*:

(1) η *is* subordinate to W^u-manifolds *of (F, μ), i.e., for μ-a.e. (ω, x), $\eta_\omega(x) \subset W^u(\omega, x)$ and it contains an open neighborhood of x in $W^u(\omega, x)$ (with the submanifold topology) ($W^u(\omega, x) := \{x\}$ if $\lambda^{(i)}(\omega, x) \leqslant 0$ for all i)*;
(2) $\Theta^{-1}\eta \geqslant \eta$;
(3) $\bigvee_{n=0}^{+\infty} \Theta^{-n}\eta$ *is equal to the partition into single points, and $\mathcal{B}(\bigwedge_{n=0}^{+\infty} \Theta^n \eta) = \mathcal{B}^u$ (μ-mod 0), where $\mathcal{B}^u := \{B \in \mathcal{B}_\mu(\Omega \times M)\colon\ B = \bigcup_{(\omega,x)\in B}\{\omega\} \times W^u(\omega, x)\}$*;
(4) $h_\mu(F) = H_\mu(\Theta^{-1}\eta|\eta) = \int H_{\mu_\omega}((F_\omega)^{-1}\eta_{\vartheta\omega}|\eta_\omega)\, d\mathbb{P}(\omega)$.

Proof of the existence of such a partition is relatively easy when μ is an SRB measure (whose definition will be given below) because the hardest part (4) can be obtained with the help of estimating the entropy via the positive exponents (for a deterministic map see [108]). For a general invariant measure μ it is necessary to consider explicitly the role

played by the zero exponent and the proof is much harder. For a deterministic diffeomorphism the result is due to [110, Part I]. For the extension to (F, μ) see [12].

Noting that

$$\int \sum_i \lambda^{(i)}(\omega, x)^+ m^{(i)}(\omega, x)\, d\mu = \int \log\left|\det(D_x F_\omega|_{E^u(\omega,x)})\right| d\mu,$$

where $E^u(\omega, x) = \bigoplus_{\lambda^{(i)}(\omega,x)>0} E^{(i)}(\omega, x)$, by Proposition 3.2.1(4) one can expect an estimate sharper than Ruelle inequality (3.1.6) if the conditional measures of μ on the unstable manifolds are compatible with the Lebesgue measures on these manifolds. This leads to the notion of SRB (Sinai–Ruelle–Bowen) measures.

DEFINITION 3.2.2. An invariant measure μ of F is called an *SRB measure* if it has absolutely continuous conditional measures on W^u-manifolds, i.e., for every measurable partition η of $(\Omega \times M, \mu)$ subordinate to W^u-manifolds of (F, μ) one has

$$\mu^\eta_{(\omega,x)} \ll \lambda^u_{(\omega,x)} \quad \text{for } \mu\text{-a.e. } (\omega, x), \tag{3.2.1}$$

where $\mu^\eta_{(\omega,x)}$ is the conditional measure of μ on $\eta(\omega, x) = \{\omega\} \times \eta_\omega(x)$, which is identified with $\eta_\omega(x)$, and $\lambda^u_{(\omega,x)}$ is the Lebesgue measure on $W^u(\omega, x)$ ($\lambda^u_{(\omega,x)} := \delta_x$ if $W^u(\omega, x) = \{x\}$).

REMARK 3.2.3. Let η be a measurable partition subordinate to W^u-manifolds of (F, μ). By the transitivity of conditional measures, for $\mathbb{P}$-a.e. ω one has

$$(\mu_\omega)^{\eta_\omega}_x = \mu^\eta_{(\omega,x)} \quad \text{for } \mu_\omega\text{-a.e. } x,$$

where $\{(\mu_\omega)^{\eta_\omega}_x\}_{x\in M}$ is a canonical system of conditional measures of μ_ω associated with the partition η_ω of M. So the SRB property of μ implies that μ_ω has absolutely continuous conditional measures on the unstable manifolds for $\mathbb{P}$-a.e. ω.

It turns out that the SRB property of μ is equivalent to the equation

$$H_\mu\left(\Theta^{-1}\eta|\eta\right) = \int \log\left|\det(D_x F_\omega|_{E^u(\omega,x)})\right| d\mu,$$

where η is given in Proposition 3.2.1. This can be stated as the following

THEOREM 3.2.4. *Let (F, μ) be given. Then*

$$h_\mu(F) = \int \sum_i \lambda^{(i)}(\omega, x)^+ m^{(i)}(\omega, x)\, d\mu \tag{3.2.2}$$

holds if and only if μ is an SRB measure.

REMARK 3.2.5. (3.2.1) actually implies that, for μ-a.e. (ω, x), $\mu^{\eta}_{(\omega,x)}$ is equivalent to $\lambda^{u}_{(\omega,x)}$, more precisely, there exist a countable number of open sets $U_n(\omega, x)$ of $W^u(\omega, x)$ such that $\bigcup_n U_n(\omega, x) \subset \eta_\omega(x)$, $\lambda^{u}_{(\omega,x)}(\eta_\omega(x) \setminus \bigcup_n U_n(\omega, x)) = 0$ and on each $U_n(\omega, x)$ the density $\rho = \mu^{\eta}_{(\omega,x)}/\lambda^{u}_{(\omega,x)}$ is strictly positive and satisfies

$$\frac{\rho(y)}{\rho(z)} = \prod_{i=1}^{+\infty} \frac{J^u(\Theta^{-i}(\omega, z))}{J^u(\Theta^{-i}(\omega, y))}, \quad y, z \in U_n(\omega, x),$$

where $J^u(\omega, x) = |\det(D_x F_\omega|_{E^u(\omega,x)})|$. This is an analog of [110, Corollary 6.2], a proof was given in [118] for an i.i.d. RDS and the proof is the same for the present (F, μ).

REMARK 3.2.6. If $\mu_\omega \ll$ Leb for $\mathbb{P}$-a.e. ω, then μ is an SRB measure. This follows from the absolute continuity property of the unstable manifolds of (F, μ) and is the analog of Corollary 2.2.15 for unstable manifolds. However, $\mu_\omega \ll$ Leb for $\mathbb{P}$-a.e. ω turns out to be a singular phenomenon when F is a canonical i.i.d. RDS and μ is a Markov measure (see Section 3.3).

SRB measures were first defined by Sinai [156] for Anosov diffeomorphisms, then by Ruelle [145] for Axiom A attractors and by Bowen and Ruelle [37] for Axiom A flows. In his original study, Sinai [156] showed (using Markov partitions and symbolic dynamics) that, if μ satisfies the absolute continuity condition defining the SRB measures for an Anosov diffeomorphism, then μ is in fact a Gibbs state in the sense of equilibrium statistical mechanics. As noted by Ruelle [145], the Gibbs property is equivalent to a variational principle which is in turn the same thing as the Pesin entropy formula. Using Pesin theory, the SRB property was later extended to nonuniformly (even partially) hyperbolic dynamical systems by Ledrappier and Strelcyn [108], which proves that the SRB property of an invariant measure implies Pesin formula for a C^2 (or $C^{1+\alpha}$) diffeomorphism. The inverse implication is much harder to establish and was proved by Ledrappier and Young [110, Part I]. The result for i.i.d. diffeomorphisms was then treated in [112] and [118]. Finally Theorem 3.2.4 was confirmed in [12]. When $\mu_\omega \ll$ Leb for $\mathbb{P}$-a.e. ω, it is also possible to derive (3.2.2) along the line of [124], see [10].

Besides being characterized as those measures satisfying Pesin entropy formula, SRB measures have their physical importance: they represent the asymptotic (long time) behavior of initial points in a set of positive Lebesgue measure. This was elaborated by Ruelle [145] for Axiom A attractors and by Pugh and Shub [137] for nonuniformly hyperbolic diffeomorphisms. We refer to the brief and excellent lecture notes [153] and [167] for more information about SRB measures. See [166,115] for the physical relevance of SRB measures for small random perturbations of Axiom A attractors. One can expect that Pugh and Shub's result alluded to above could be extended to (F, μ) (by using the absolute continuity property of stable manifolds).

REMARK 3.2.7. See Ruelle [152] for applications of Theorems 3.1.1 and 3.2.4 in the study of positivity of (statistical mechanics) entropy production in nonequilibrium statistical mechanics when a thermostat acting by random forces is present.

Generalized entropy formula. As stated in Young [167], one way to understand the results Proposition 3.2.1, Theorems 3.1.1 and 3.2.4 is as follows: information creation is caused exactly by the expansion in the random maps; when the measure is SRB, all (in the topological sense) the expansion is employed to make entropy, hence we have the Pesin formula; a strict inequality signifies some "wasted" expansion, which can happen only if the invariant measure has "holes" on the unstable manifolds. This naturally leads to the following question: for a non-SRB measure, how to describe the part of the expansion which contributes to the entropy, and what is the precise relation between the entropy and the expansion? This problem is answered by the next theorem. It involves dimensions of conditional measures of the invariant measure on various layers of the unstable manifolds, which allow one to distinguish between contributions to the entropy made by expansions in different directions. For simplicity of presentation we introduce the definition for ergodic (F, μ). For our purpose it is more convenient to write the Lyapunov exponents of (F, μ) in the following way:

$$+\infty > \lambda_1 > \lambda_2 > \cdots > \lambda_r > -\infty. \tag{3.2.3}$$

Assume $\lambda_1 > 0$. Let $\lambda_1 > \cdots > \lambda_q$ be all the positive exponents and denote simply by

$$W^1(\omega, x) \subset W^2(\omega, x) \subset \cdots \subset W^q(\omega, x)$$

the corresponding nested family of unstable manifolds of F at (ω, x). Fix arbitrarily $1 \leqslant i \leqslant q$ and let η be a measurable partition of $(\Omega \times M, \mu)$ subordinate to the W^i-manifolds of (F, μ) (the definition is analogous to the W^u-manifolds case). For a typical ω, let $\{(\mu_\omega)_x^{\eta_\omega}\}_{x \in M}$ be a canonical system of conditional measures of μ_ω associated with η_ω (see Remark 3.2.3). Define $B_\omega^i(x, r) = \{y \in W^i(\omega, x)\colon\ d^i(x, y) < r\}$ where $d^i(\,,\,)$ denotes the distance along $W^i(\omega, x)$. Then, the limit

$$\delta_i := \lim_{r \to 0} \frac{\log (\mu_\omega)_x^{\eta_\omega}(B_\omega^i(x, r))}{\log r} \tag{3.2.4}$$

can be shown to exist and to be constant for μ-a.e. (ω, x). The number δ_i is clearly independent of the choice of η and will be called the *dimension of* μ_ω, $\omega \in \Omega$ *on the* W^i*-manifolds*. Number γ_i, defined by

$$\gamma_i = \begin{cases} \delta_1 & \text{for } i = 1, \\ \delta_i - \delta_{i-1} & \text{for } 2 \leqslant i \leqslant q, \end{cases}$$

is called a *partial dimension*, it can be regarded as the dimension of μ_ω, $\omega \in \Omega$, "in the direction of the Oseledec space $E^{(i)}$ corresponding to λ_i" or as the "transverse dimension" of μ_ω, $\omega \in \Omega$, on W^i / W^{i-1}. When μ is not ergodic, the definitions are analogous.

THEOREM 3.2.8. *For* (F, μ) *the following generalized entropy formula*:

$$h_\mu(F) = \int \sum_{i\colon \lambda_i(\omega, x) > 0} \lambda_i(\omega, x) \gamma_i(\omega, x)\, d\mu \tag{3.2.5}$$

holds true.

The fundamental formula (3.2.5), as well as the existence of the limit (3.2.4), was elaborated by Ledrappier and Young [110, Part II] for a deterministic C^2 diffeomorphism $f: M \to M$. The extension to (F, μ) follows essentially the same line of [110, Part II] and is due to [138] (this paper deals with an i.i.d. RDS, but the argument remains almost the same for the present (F, μ)).

Note that (3.2.5) implies the Ruelle inequality for (F, μ) because $\gamma_i(\omega, x) \leqslant m_i(\omega, x)$ ($m_i(\omega, x)$ is the multiplicity of $\lambda_i(\omega, x)$) for $1 \leqslant i \leqslant q(\omega, x)$. When μ is SRB one has $\gamma_i(\omega, x) = m_i(\omega, x)$ and hence (3.2.5) gives the Pesin formula.

Dimension of hyperbolic measures of F. Another related important problem concerns dimension of the sample measures μ_ω, $\omega \in \Omega$, themselves. Let ν be a Borel probability measure on a finite-dimensional manifold N. If the limit

$$d_\nu(x) := \lim_{r \to 0} \frac{\log \nu(B(x, r))}{\log r} \tag{3.2.6}$$

exists at ν-a.e. $x \in N$, ν is then said to be *exact dimensional* and $d_\nu(x)$ is called the *pointwise dimension* of ν at x. The existence of the limit (3.2.6) signifies its importance by the following crucial fact due to Young [165]: If $d_\nu(x)$ exists and is equal to a constant α for ν-a.e. x, then the Hausdorff dimension, the lower and the upper box dimension as well as several other dimension type characteristics of the measure ν coincide and the common value is α. Let now $f: M \to M$ be a C^2 (or $C^{1+\alpha}$) diffeomorphism and μ an ergodic f-invariant measure. Though the pointwise dimensions of conditional measures of μ on various layers of the stable and unstable manifolds can be shown well defined (as mentioned in the last paragraph), the existence of the limit (3.2.6) defining the pointwise dimension $d_\mu(x)$ turns out to be more subtle. It was conjectured by Eckmann and Ruelle that, if μ is hyperbolic (i.e., if f has no zero Lyapunov exponent μ almost everywhere), then the pointwise dimension $d_\mu(x)$ exists and is constant for μ-a.e. x, the constant d_μ is equal to the sum of the pointwise dimensions, denoted by d^s_μ and d^u_μ, respectively, of conditional measures of μ on the stable and unstable manifolds, i.e.,

$$d_\mu = d^s_\mu + d^u_\mu.$$

This conjecture had remained a long-standing open problem in the interface of dimension theory and dynamical systems, and it was finally proved by Barreira, Pesin and Schmeling [20] (see that paper for previous substantial progress by Ledrappier, Young and others). The following theorem is an extension to the RDS (F, μ) of the main result from [20] (see [120]).

THEOREM 3.2.9. *Let the RDS (F, μ) be given. Assume that μ is ergodic* (*i.e., ergodic with respect to Θ*) *and* hyperbolic (*i.e., F has no zero exponent μ almost everywhere*). *Let η^s* (*respectively η^u*) *be a measurable partition of $(\Omega \times M, \mu)$ subordinate to W^s-manifolds* (*respectively W^u-manifolds*) *of (F, μ). Let $\mu^s_{(\omega,x)}$* (*respectively $\mu^u_{(\omega,x)}$*) *be the conditional measure of μ on $\eta^s(\omega, x)$* (*respectively $\eta^u(\omega, x)$*). *Then one has the following*:

(1) *for every $\varepsilon > 0$ there exists a set $\Lambda' \subset \Omega \times M$ with $\mu(\Lambda') > 1 - \varepsilon$ and constants $\kappa \geqslant 1$, $\rho_0 > 0$ such that for every $(\omega, x) \in \Lambda'$ and every $\rho \in (0, \rho_0)$,*

$$\begin{aligned}
&\rho^{\varepsilon} \mu^s_{(\omega,x)}\left(B^s_\omega\left(x, \frac{\rho}{\kappa}\right)\right) \mu^u_{(\omega,x)}\left(B^u_\omega\left(x, \frac{\rho}{\kappa}\right)\right) \\
&\quad \leqslant \mu_\omega\big(B(x,\rho)\big) \\
&\quad \leqslant \rho^{-\varepsilon} \mu^s_{(\omega,x)}\big(B^s_\omega(x, \kappa\rho)\big) \mu^u_{(\omega,x)}\big(B^u_\omega(x, \kappa\rho)\big);
\end{aligned}$$

(2) *for $\mathbb{P}$-a.e. ω, μ_ω is exact dimensional and at μ_ω-a.e. x,*

$$d_{\mu_\omega}(x) = d^s_{\mu_\omega}(x) + d^u_{\mu_\omega}(x), \tag{3.2.7}$$

the three quantities in (3.2.7) *are constant for μ-a.e. (ω, x).*

When μ is not ergodic, μ_ω is still exact dimensional for $\mathbb{P}$-a.e. ω and (3.2.7) *is true at μ-a.e. (ω, x).*

3.3. *I.i.d. RDS*

Set-up. In this section we consider a particular class of RDS which is of great interest from the point of view of Markov processes and which is closely related to solutions of classical stochastic differential equations (see [117] for a brief review). This class of RDS is defined by assuming that the random maps chosen at different time steps are independent and identically distributed. More precisely, let $\vartheta : (\Omega, \mathcal{F}, \mathbb{P}) \hookleftarrow$ be a measure-preserving map of a probability space and let F be an RDS over $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ generated by a measurable map

$$F : \Omega \to C^r(M, M), \quad \omega \mapsto F_\omega$$

($r \geqslant 0$). If $F_{\vartheta^n \cdot}$, $n \geqslant 0$, and $n < 0$ when ϑ is measurably invertible, are independent (they are clearly identically distributed), then F is said to be *i.i.d.* In the rest of this section F will always be such an i.i.d. RDS.

We recall the following canonical model which is of particular interest. Take $\Omega = C^r(M, M)^{\mathbb{Z}^+}$, endowed with the product topology and the Borel σ-algebra (note that Ω is Polish for integer $0 \leqslant r < \infty$). Let $\vartheta : \Omega \to \Omega$ be the left shift operator, i.e., $(\vartheta\omega)_n = \omega_{n+1}$ for $\omega = (\omega_n) \in \Omega$, and let $\mathbb{P} = \nu^{\mathbb{Z}^+}$ for some Borel probability ν on $C^r(M, M)$. The coordinate map $F : \Omega \to C^r(M, M)$, $\omega \mapsto \omega_0$ defines then an i.i.d. RDS F over $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$. We will call it a *one-sided canonical i.i.d. RDS*; this model corresponds to that introduced and systematically studied by Kifer [82]. A *two-sided canonical i.i.d. RDS* is defined similarly by taking $\Omega = C^r(M, M)^{\mathbb{Z}}$ and $\mathbb{P} = \nu^{\mathbb{Z}}$.

Stationary measures. A specific feature of an i.i.d. RDS F is that it naturally induces a family of Markov processes (i.e., its one-point motions) on M whose transition probabilities $P(x,\cdot)$, $x \in M$, are defined by

$$P(x, A) = \mathbb{P}\{\omega\colon\ F_\omega x \in A\} \quad \text{for Borel } A \subset M.$$

A Borel probability measure ρ on M is called a *stationary measure* of F if it is stationary for the Markov kernel $P(x,\cdot)$, $x \in M$, i.e.,

$$\rho(A) = \int P(x, A)\, d\rho(x) \quad \text{for Borel } A \subset M$$

which is equivalent to

$$\rho(A) = \int (F_\omega \rho)(A)\, d\mathbb{P}(\omega) \quad \text{for Borel } A \subset M.$$

Such measures always exist and a systematic study of their ergodic properties was presented in [82]. We present below two properties of them. Let $\Theta : \Omega \times M \hookleftarrow$ be the skew product transformation of F.

PROPOSITION 3.3.1 [130]. *Let F be a one-sided canonical i.i.d. C^r $(0 \leqslant r \leqslant \infty)$ RDS. Then*:

(1) *a probability ρ on M is stationary for F if and only if $\mathbb{P} \times \rho$ is Θ-invariant, i.e., it is an invariant measure of F;*
(2) *a stationary measure ρ of F is ergodic from the viewpoint of Markov processes* (*see, e.g.,* [82]) *if and only if $(F, \mathbb{P} \times \rho)$, or equivalently $(\Theta, \mathbb{P} \times \rho)$, is ergodic.*

Proposition 3.3.1 shows the relation between stationary and invariant measures for a one-sided canonical i.i.d. RDS. For the two-sided case one has

PROPOSITION 3.3.2. *Let F be a two-sided canonical i.i.d. C^r RDS* $(0 \leqslant r < \infty$ *since Ω is assumed being Polish here*). *Then the set of stationary measures of F corresponds in a one-to-one way to the set of forward Markov invariant measures of F* (*that is, the set of those F-invariant measures ρ^* whose almost every sample measure ρ_ω depends only on the past $(\ldots, \omega_{-2}, \omega_{-1})$ of ω*), *with the correspondence being given by*

$$\rho \mapsto \rho^* \quad \text{with } \rho_\omega := \lim_{n \to +\infty} F^n_{\vartheta^{-n}\omega}\rho,\ \mathbb{P}\text{-a.e. } \omega$$

and

$$\rho^* \mapsto \rho := \int \rho_\omega\, d\mathbb{P}(\omega).$$

Moreover, for a given stationary measure ρ of F, the corresponding ρ^* is the unique F-invariant measure whose natural projection on $C^r(M, M)^{\mathbb{Z}^+} \times M$ is $\nu^{\mathbb{Z}^+} \times \rho$, and ρ is ergodic from the pointview of Markov processes if and only if (F, ρ^*) is ergodic.

See [7, Theorems 1.7.2 and 2.1.8] for a proof and for more information.

Entropy formula for i.i.d. RDS. An interesting phenomenon due to presence of noise is that for an i.i.d. RDS and a stationary measure the Pesin formula "always" holds true and the corresponding sample measures are almost surely SRB if the RDS is "sufficiently" random [112]. To state the results precisely, we first reformulate the definition of the entropy in this particular setting. Assume F is i.i.d. and ρ is a stationary measure of F. For a finite Borel partition ξ of M, let

$$h_\rho(F,\xi) = \lim_{n\to+\infty} \frac{1}{n} \int H_\rho\left(\bigvee_{k=0}^{n-1} \left(F_\omega^k\right)^{-1}\xi\right) d\mathbb{P}(\omega) \tag{3.3.1}$$

(the limit exists, see [82, Theorem II.1.3] for a similar proof), and define the *entropy* of (F,ρ) as $h_\rho(F) := \sup h_\rho(F,\xi)$, with the supremum being taken over the set of all finite Borel partitions of M. Let ν be the probability on $C^r(M,M)$ induced by the map $\Omega \to C^r(M,M)$, $\omega \mapsto F_\omega$. Denote by $\hat{F}$ the one-sided canonical i.i.d. RDS with $\hat{\mathbb{P}} = \nu^{\mathbb{Z}^+}$, and by $\tilde{F}$ the two-sided one with $\tilde{\mathbb{P}} = \nu^{\mathbb{Z}}$. Then $h_\rho(F)$ can be related to the previous notion of entropy by

$$h_\rho(F) = h_{\hat{\mathbb{P}}\times\rho}(\hat{F}) = h_{\rho^*}(\tilde{F}), \tag{3.3.2}$$

where $\hat{\mathbb{P}} \times \rho$ and ρ^* are given by Propositions 3.3.1 and 3.3.2, respectively (see [116]).

Assume that F is C^1 and

$$\int \log^+ |D_x F_\omega| \, d(\mathbb{P}\times\rho)(\omega,x) < +\infty.$$

Then for ρ-a.e. x there exist numbers (depending only and measurably on x) $r(x)$,

$$-\infty \leqslant \lambda^{(1)}(x) < \lambda^{(2)}(x) < \cdots < \lambda^{(r(x))}(x) < +\infty$$

such that for $\mathbb{P}$-a.e. ω there are subspaces

$$\{0\} = V^{(0)}(\omega,x) \subset V^{(1)}(\omega,x) \subset \cdots \subset V^{(r(x))}(\omega,x) = T_x M$$

satisfying

$$\lim_{n\to+\infty} \frac{1}{n} \log |D_x F_\omega^n \xi| = \lambda^{(i)}(x)$$

for $\xi \in V^{(i)}(\omega,x) \setminus V^{(i-1)}(\omega,x)$, $1 \leqslant i \leqslant r(x)$, with $m^{(i)}(x) := \dim V^{(i)}(\omega,x) - \dim V^{(i-1)}(\omega,x)$ also depending only and measurably on x. We will call $(\lambda^{(i)}(x), m^{(i)}(x))$, $1 \leqslant i \leqslant r(x)$, the *Lyapunov spectrum* of (F,ρ) at x. This result is shown in [82] for a

one-sided canonical i.i.d. RDS. The extension to (F, ρ) is straightforward by defining the Lyapunov spectrum of (F, ρ) at ρ-a.e. x as that of $(\hat{F}, \rho)$ via the measure-preserving map

$$\Sigma : (\Omega, \mathbb{P}) \to \left(C^r(M, M)^{\mathbb{Z}^+}, \nu^{\mathbb{Z}^+}\right), \quad \omega \mapsto (F_\omega, F_{\vartheta\omega}, \ldots).$$

Applying Theorems 3.1.1 and 3.1.3 to the RDS $\hat{F}$ with invariant measure $\nu^{\mathbb{Z}^+} \times \rho$, we have

COROLLARY 3.3.3 (Ruelle inequality for i.i.d. RDS). *Assume that F is C^1 (i.e., $r = 1$) and $\log^+ |F_\omega|_{C^1} \in L^1(\Omega, \mathbb{P})$. Then for any stationary measure ρ of F,*

$$h_\rho(F) \leqslant \int \sum_i \lambda^{(i)}(x)^+ m^{(i)}(x)\, d\rho(x).$$

COROLLARY 3.3.4 (Pesin formula for i.i.d. RDS). *Assume that F is C^2 and $\log^+ |F_\omega|_{C^2} \in L^1(\Omega, \mathbb{P})$, $\log D(F_\omega) \in L^1(\Omega, \mathbb{P})$. Let ρ be a stationary measure of F. If $\rho \ll$ Leb, then one has*

$$h_\rho(F) = \int \sum_i \lambda^{(i)}(x)^+ m^{(i)}(x)\, d\rho(x).$$

If the i.i.d. RDS F is sufficiently random in the sense that its transition probabilities have a density with respect to the Lebesgue, i.e., if there is a Borel function $p : M \times M \to \mathbb{R}^+$ such that for every $x \in M$ one has

$$P(x, A) = \int_A p(x, y)\, d\,\mathrm{Leb}(y) \quad \text{for Borel } A \subset M,$$

then all its stationary measures ρ satisfy $\rho \ll$ Leb (see, e.g., [118, Section IV.1]).

From Corollary 3.3.4 and Theorem 3.2.4 there follows

COROLLARY 3.3.5 (SRB property of sample measures). *Let F be a two-sided canonical i.i.d. RDS such that $F_\omega \in \mathrm{Diff}^2(M)$ for $\mathbb{P}$-a.e. ω. Assume the integrability conditions $\log^+ |F_\omega|_{C^2}, \log^+ |(F_\omega)^{-1}|_{C^2} \in L^1(\Omega, \mathbb{P})$. Let ρ be a stationary measure of F. If $\rho \ll$ Leb, then almost all the sample measures ρ_ω, $\omega \in \Omega$, are SRB, or, precisely speaking, the measure ρ^* given by Proposition* 3.3.2 *is SRB.*

REMARK 3.3.6 (*Nonsmoothness of sample measures*). Take the assumptions of Corollary 3.3.5. Then the SRB sample measures ρ_ω, $\omega \in \Omega$, are almost all smooth (i.e., $\rho_\omega \ll$ Leb for $\mathbb{P}$-a.e. ω) if and only if $F_\omega \rho = \rho$ for $\mathbb{P}$-a.e. ω. Equivalently (and roughly speaking), if the latter condition does not hold, the time evolution on the phase space will destroy with full probability (note that $\mathbb{P}$ is now ergodic) the smoothness of ρ in some contracting directions but improves it in the stretching directions, ensuring the SRB property of the sample measures (recall the pictures in Proposition 3.3.2 and Corollary 3.3.5; see also

Theorem 3.3.11 below). The fact stated above can be proved as follows. Note that, if $\rho_\omega \ll$ Leb for $\mathbb{P}$-a.e. ω, the absolute continuity property of the stable and unstable manifolds implies that ρ^* is SRB both for F and its time reversal, yielding $\int \sum_i \lambda^{(i)}(x) m^{(i)}(x)\, d\rho(x) = 0$ by the Pesin formulae. By using the interesting result Proposition 3.3.7 below, one proves $F_\omega \rho = \rho$ for $\mathbb{P}$-a.e. ω. The inverse implication is obvious.

PROPOSITION 3.3.7 [24,82]. *Let F be an i.i.d. RDS with $F_\omega \in \mathrm{Diff}^1(M)$ for $\mathbb{P}$-a.e. ω and $\log^+ |F_\omega|_{C^1} \in L^1(\Omega, \mathbb{P})$. Let ρ be a stationary measure of F. If $\rho \ll$ Leb, then*

(1) $\sum_i \lambda^{(i)}(x) m^{(i)}(x) \leqslant 0$, *$\rho$-a.e. x;*
(2) $\sum_i \lambda^{(i)}(x) m^{(i)}(x) = 0$, *$\rho$-a.e. x if and only if $F_\omega \rho = \rho$ for $\mathbb{P}$-a.e. ω.*

Ruelle inequality and Pesin formula for stochastic flows of diffeomorphisms. Let $\{\psi_t\}_{t \geqslant 0}$ be a stochastic flow of C^r ($r \in \mathbb{N}$) diffeomorphisms over a probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$ as introduced by Definition 2.4.3. For the discrete time case the definition is analogous and the discussion below will be similar (but the integrability of $\log^+ |\psi_1(\hat{\omega})|_{C^r} + \log^+ |\psi_1(\hat{\omega})^{-1}|_{C^r}$ with respect to $\hat{\mathbb{P}}$ does not hold automatically, as opposed to the continuous time case, and has to be assumed). So here we only deal with the continuous time case. Let ρ be a stationary measure of $\{\psi_t\}_{t \geqslant 0}$. Let $(\lambda^{(i)}(x), m^{(i)}(x))$, $1 \leqslant i \leqslant r(x)$, ρ-a.e. $x \in M$ be the Lyapunov spectra of $(\{\psi_t\}_{t \geqslant 0}, \rho)$ (see [118, Chapter V]). For an arbitrarily fixed $t_0 > 0$, one can define the *entropy of $(\{\psi_t\}_{t \geqslant 0}, \rho)$ with respect to t_0*, written $h_\rho^{t_0}(\{\psi_t\}_{t \geqslant 0})$, in a way similar to (3.3.1) by replacing F_ω^k with $\psi_{kt_0}(\hat{\omega})$. It turns out that

$$h_\rho^{t_0}\big(\{\psi_t\}_{t \geqslant 0}\big) = t_0 h_\rho^1\big(\{\psi_t\}_{t \geqslant 0}\big)$$

[82] and hence the particular choice of the time step length t_0 does not matter for the notion of the entropy of $(\{\psi_t\}_{t \geqslant 0}, \rho)$. Then, by Corollaries 3.3.3 and 3.3.4, one has

COROLLARY 3.3.8. *For every stationary measure ρ of a C^1 stochastic flow $\{\psi_t\}_{t \geqslant 0}$ one has the Ruelle inequality*

$$h_\rho^1\big(\{\psi_t\}_{t \geqslant 0}\big) \leqslant \int \sum_i \lambda^{(i)}(x)^+ m^{(i)}(x)\, d\rho. \tag{3.3.3}$$

If $\{\psi_t\}_{t \geqslant 0}$ is C^2 and $\rho \ll$ Leb, then the Pesin formula holds, that is, (3.3.3) *is an equality.*

If $\{\psi_t\}_{t \geqslant 0}$ arises from a smooth nondegenerate Stratonovich stochastic differential equation, then it has a unique stationary measure ρ and this measure has smooth density with respect to the Lebesgue (see [72]). So in this case one has the Pesin formula.

Endomorphisms followed by time-ε-maps of stochastic flows. As an application of Corollary 3.3.4 we consider the following random perturbation model introduced in [18]. Suppose that $f : M \to M$ is a C^2 map with no singularities. We consider the situation that a particle $x \in M$ jumps to fx and it then performs a diffusion for the time $\varepsilon > 0$ (see

also [83] for a systematic treatment of this set-up). More precisely, let $X_0, X_1, \dots, X_d$ be C^r ($r \geqslant 4$, for example) vector fields on M, and consider the SDE of Stratonovich type

$$d\xi_t = X_0(\xi_t)\,dt + \sum_{i=1}^{d} X_i(\xi_t) \circ dB_t^i, \tag{3.3.4}$$

where $\{(B_t^1, \dots, B_t^d)\}_{t\geqslant 0}$ is a standard d-dimensional Brownian motion defined on a probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$. Realizing the solution of (3.3.4) as a stochastic flow of C^2 diffeomorphisms $\{\psi_t : (\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}}) \to \mathrm{Diff}^2(M)\}_{t\geqslant 0}$, we consider the randomly perturbed process generated by compositions of random maps

$$\cdots \circ g(\hat{\omega}_2) \circ g(\hat{\omega}_1),$$

where $\hat{\omega}_1, \hat{\omega}_2, \dots \in (\hat{\Omega}, \hat{\mathbb{P}})$ are chosen independently and

$$g(\hat{\omega}_i) = \psi_\varepsilon(\hat{\omega}_i) \circ f.$$

Now we are only concerned with the distribution of the random sequence of maps, hence the randomly perturbed process introduced above is just the RDS F_ε generated by the coordinate process on the canonical probability space $(\Omega, \mathbb{P}_\varepsilon)$, where $\Omega = C^2(M, M)^{\mathbb{Z}^+}$, $\mathbb{P}_\varepsilon = \nu_\varepsilon^{\mathbb{Z}^+}$ and ν_ε is the distribution on $C^2(M, M)$ induced by the map

$$\Sigma : (\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}}) \to C^2(M, M), \quad \hat{\omega} \mapsto \psi_\varepsilon(\hat{\omega}) \circ f.$$

For any $\varepsilon > 0$ the probability $\mathbb{P}_\varepsilon$ satisfies the integrability conditions in Corollary 3.3.4 (see [116]). If SDE (3.3.4) is nondegenerate, namely $X_1, \dots, X_d$ span the tangent space of M, then the transition probabilities of F_ε have a density with respect to the Lebesgue and hence every stationary measure ρ of F_ε satisfies $\rho \ll \mathrm{Leb}$. Then, by Corollary 3.3.4, we have the following

COROLLARY 3.3.9. *Let $f \in C^2(M, M)$ with no singularities and assume that SDE* (3.3.4) *is nondegenerate. Let $\varepsilon > 0$ and let ρ be a stationary measure of F_ε. Then the Pesin formula holds true for (F_ε, ρ).*

In what follows we assume that f is a C^2 expanding map. When ε is small, F_ε is expanding in average [116] and it has a unique absolutely continuous stationary measure $d\rho_\varepsilon = l_\varepsilon d\,\mathrm{Leb}$ (see [117] for a review). l_ε can be obtained as the eigenfunction with $\int l_\varepsilon d\,\mathrm{Leb} = 1$ of the simple eigenvalue 1 of the *integrated* transfer operator $\mathcal{L}_{F_\varepsilon} : C^1(M) \to C^1(M)$ defined by

$$(\mathcal{L}_{F_\varepsilon} l)(x) = \int (\mathcal{L}_g l)(x)\, d\nu_\varepsilon(g) \quad \text{for } l \in C^1(M),$$

where $\mathcal{L}_g$ is defined by (3.1.16), and it satisfies $|l_\varepsilon - l_0|_{C^1} \to 0$ as $\varepsilon \to 0$, where $l_0 = d\mu_0/d\,\mathrm{Leb}$ and μ_0 is the unique smooth invariant measure of f (see [18] and [151] for these results). One then has the following conclusion [116]:

PROPOSITION 3.3.10. *For sufficiently small $\varepsilon > 0$, $(F_\varepsilon, \rho_\varepsilon)$ has no negative Lyapunov exponents and hence*

$$h_{\rho_\varepsilon}(F_\varepsilon) = \int \sum_i \lambda^{(i)}(x) m^{(i)}(x)\, d\rho_\varepsilon(x)$$
$$= \int\int \log\left|\det D_x\left(\psi_\varepsilon(\hat{\omega}) \circ f\right)\right| d\rho_\varepsilon(x)\, d\hat{\mathbb{P}}(\hat{\omega}).$$

Moreover, one has

$$\lim_{\varepsilon \to 0} h_{\rho_\varepsilon}(F_\varepsilon) = h_{\mu_0}(f).$$

A dimension formula. In the deterministic case, it was conjectured by Yorke et al. [63] that the Hausdorff dimension of an ergodic SRB measure is "generically" equal to its Lyapunov dimension; this was proved to be true by Young [165] for surface diffeomorphisms, but when the phase space has dimension greater than 2 the conjecture has not been mathematically verified (see [59] for a detailed review of this topic). However, Ledrappier and Young [111] showed that the conjectured scenario is indeed mathematically true in *any* dimension if the dynamical system is subjected to certain types of sufficiently random noise. This means that for an i.i.d. RDS composed of random diffeomorphisms with some conditions to guarantee genuine randomness, the Hausdorff dimension of the sample measures corresponding to an ergodic stationary measure is almost surely equal to the Lyapunov dimension. Below we recall briefly the related notions and formulate this notable result in a more precise way.

Let F be a two-sided canonical i.i.d. RDS such that $F_\omega \in \mathrm{Diff}^2(M)$ for $\mathbb{P}$-a.e. ω and $\log^+ |F_\omega|_{C^2}$, $\log^+ |(F_\omega)^{-1}|_{C^2} \in L^1(\Omega, \mathbb{P})$. Let ρ be an ergodic stationary measure of F and ρ^* the corresponding F-invariant measure (see Proposition 3.3.2). Now it is more convenient to write all the Lyapunov exponents of (F, ρ), or equivalently of (F, ρ^*), as

$$+\infty > \lambda_1 > \lambda_2 > \cdots > \lambda_r > -\infty.$$

Assume $\lambda_1 > 0$ since otherwise the statements below will hold rather trivially. By Theorem 3.2.9, if ρ^* is hyperbolic, then for $\mathbb{P}$-a.e. ω the pointwise dimension $d_{\rho_\omega}(x)$ is well defined and is constant for ρ_ω-a.e. x. We will denote this constant by $\dim(\rho_\omega)$ (which is in fact constant for $\mathbb{P}$-a.e. ω) and call it the *dimension* of ρ_ω. The *Lyapunov dimension* of the family of sample measures $\{\rho_\omega\}$ is defined as

$$\mathcal{D}(\lambda_1, \ldots, \lambda_r) := \begin{cases} \dim M & \text{if } \sum_{j=1}^K m_j = \dim M, \\ \sum_{j=1}^K m_j - \frac{1}{\lambda_{K+1}} \sum_{j=1}^K \lambda_j m_j & \text{otherwise,} \end{cases} \tag{3.3.5}$$

with K being the largest integer so that $\sum_{j=1}^K \lambda_j m_j \geqslant 0$, where $m_i = \dim E^{(i)}(\omega, x)$. If $\rho \ll \mathrm{Leb}$, then, by Proposition 3.3.7, one always has $\sum_i \lambda_i m_i \leqslant 0$ with $\sum_i \lambda_i m_i = 0$ if

and only if $\rho_\omega = \rho$, $\mathbb{P}$-a.e. Clearly for the special case $\sum_i \lambda_i m_i = 0$ one has $\dim(\rho_\omega) = \mathcal{D}(\lambda_1, \dots, \lambda_r) = \dim M$, $\mathbb{P}$-a.e. Under further conditions the coincidence of $\dim(\rho_\omega)$ and $\mathcal{D}(\lambda_1, \dots, \lambda_r)$ is obtained for the general case in the following theorem, where we will employ the "backward derivative process" naturally induced by F (with $\mathbb{P} = \nu^{\mathbb{Z}}$) on the Grassmannian manifold $\mathrm{Gr}(M)$ whose transition probabilities are given by

$$Q(v, \Gamma) = \nu\{\omega\colon\ D(F_\omega)^{-1} v \in \Gamma\}, \quad v \in \mathrm{Gr}(M),\ \text{Borel } \Gamma \subset \mathrm{Gr}(M)$$

(recall $\mathrm{Gr}(M) = \bigcup_{k=1}^{\dim M} \mathrm{Gr}(M, k)$ and $\mathrm{Gr}(M, k)$ is the manifold of k-dimensional subspaces of TM).

THEOREM 3.3.11 (A dimension formula due to Ledrappier and Young [111]). *Let F be as given above and let ρ be an ergodic stationary measure of F such that $\rho \ll$ Leb and $\lambda_i \neq 0$ for all i. Take the hypothesis that for all $v \in \mathrm{Gr}(M)$ the transition probability $Q(v, \cdot)$ is absolutely continuous with respect to the Lebesgue on $\mathrm{Gr}(M)$. Then for $\mathbb{P}$-a.e. ω,*

$$\dim(\rho_\omega) = \mathcal{D}(\lambda_1, \dots, \lambda_r). \tag{3.3.6}$$

The hypothesis in the theorem is a sufficient randomness condition which appears quite naturally, especially in the setting of stochastic flows arising from SDE. If F is generated by the time-1 maps of the solution flow of C^∞ SDE (3.3.4), then the hypothesis is satisfied if the operator $L = -\tilde{X}_0 + \sum_{k=1}^d \tilde{X}_k^2$ on $C^\infty(\mathrm{Gr}(M))$ is hypoelliptic, where $\tilde{X}_k$ is the natural lifting of X_k to $\mathrm{Gr}(M)$, $0 \leqslant k \leqslant d$. Particularly, if $d \geqslant \dim M + (\dim M)^2$, then there is an open and dense subset in the space of $(d+1)$-tuples of vector fields on M on which the hypothesis is satisfied (see [111] for details and for further references). To obtain the desired result (3.3.6), the above hypothesis can be replaced by the weaker one that, for ρ-a.e. x and $j = K+1,\ K+2$ (K being given in (3.3.5)), the distribution of $\omega \mapsto \bigoplus_{i \geqslant j} E^{(i)}(\omega, x)$ is smooth on the Grassmannian manifold of $\sum_{i \geqslant j} m_i$-dimensional subspaces of $T_x M$, or by a nonlinear version formulated in terms of two-point processes on M. In fact, the latter ones are more directly related to the key ideas of the proof (which consist in estimating the transverse dimension of ρ_ω with respect to the $W^{s,j}$-manifolds).

4. Thermodynamic formalism and its applications

In this section we will describe the construction of Gibbs measures and equilibrium states for certain classes of random transformations in the way usually called the thermodynamic formalism and will apply this machinery to limit theorems and fractal dimensions.

4.1. *Random subshifts*

In this section we describe an important class of random transformations called one- and two-sided random subshifts of finite type. This setup is generated by a positive integer valued random variable $\ell = \ell(\omega)$ on $(\Omega, \mathbb{P})$ satisfying

$$\int \log \ell \, d\mathbb{P} < \infty \tag{4.1.1}$$

together with a measurable family of $\ell(\omega) \times \ell(\vartheta\omega)$-matrices $A(\omega) = (a_{ij}(\omega))$ with 0 and 1 entries having no zero row. We define $\mathcal{E}_\omega = \{x = (x_0, x_1, \ldots)\colon\ x_i \in \{1, \ldots, \ell(\vartheta^i\omega)\}$ and $a_{x_i x_{i+1}}(\vartheta^i\omega) = 1\ \forall i = 0, 1, \ldots\}$ in the one-sided case and $\mathcal{E}_\omega = \{x = (\ldots, x_{-1}, x_0, x_1, \ldots)\colon$ $x_i \in \{1, \ldots, \ell(\vartheta^i\omega)\}$ and $a_{x_i x_{i+1}}(\vartheta^i\omega) = 1\ \forall i = \ldots, -1, 0, 1, \ldots\}$ in the two-sided case and, as before, $\mathcal{E} = \{(\omega, x)\colon\ \omega \in \Omega,\ x \in \mathcal{E}_\omega\}$. In both one- and two-sided cases $F_\omega : \mathcal{E}_\omega \to \mathcal{E}_{\vartheta\omega}$ is defined as the left shift $(F_\omega x)_i = x_{i+1}$ and the corresponding random transformation F is called in these circumstances a random subshift of finite type. If $A(\omega)$ has no zeros then we call F a random full shift. This setup can also be described in terms of random graphs or networks (see [96]). Observe that multidimensional random subshifts of finite type motivated by some statistical mechanics models (such as spin-glasses) were studied in [93].

The spaces $\mathcal{E}_\omega$ are closed imbedded subsets of the compact space $X = \bar{\mathbb{Z}}^+ \times \bar{\mathbb{Z}}^+ \times \cdots$, in the one-sided case, or of $X = \cdots \times \bar{\mathbb{Z}}^+ \times \bar{\mathbb{Z}}^+ \times \cdots$, in the two-sided case which are infinite products of the one-point compactification $\bar{\mathbb{Z}} = \mathbb{Z} \cup \{\infty\}$ of $\mathbb{Z}$ with the metric on X given by

$$d(x, \tilde{x}) = \sum_i 2^{-|i|} \left| \frac{1}{x_i} - \frac{1}{\tilde{x}_i} \right|, \tag{4.1.2}$$

where we sum in $i \in \mathbb{N}$ in the one-sided case or in $i \in \mathbb{Z}$ in the two-sided case. Clearly, the shifts $F_\omega : \mathcal{E}_\omega \to \mathcal{E}_{\vartheta\omega}$ are continuous and the corresponding bundle RDS is fiber expansive (see Definition 1.3.1) with the expansivity characteristic $\varepsilon(\omega) = (\ell(\omega))^{-2}$. Hence, not only the variational principle (1.2.11) holds true in this setup for any function $f \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$ but also according to Theorem 1.3.5 there exists an equilibrium state for such an f, i.e., a measure $\mu \in \mathcal{I}_{\mathbb{P}}(\mathcal{E})$ maximizing in (1.2.11). Observe that (4.1.1) implies that the fiber topological pressure $\pi_F(f)$ of any $f \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$ is finite and, in particular, the fiber topological entropy is finite, as well. Next, we will be interested in conditions which ensure the uniqueness of equilibrium states.

Both in the one- and two-sided subshift cases we set for any measurable function $f = f(\omega, x)$ on $\mathcal{E}$,

$$\operatorname{var}_n f(\omega) = \sup\bigl\{\bigl|f(\omega, x) - f(\omega, \tilde{x})\bigr|\colon\ x_i = \tilde{x}_i\ \forall |i| < n\bigr\}.$$

We will assume that

$$\int \sup_x \bigl|f(\omega, x)\bigr| \, d\mathbb{P}(\omega) < \infty \tag{4.1.3}$$

and

$$\operatorname{var}_n f(\omega) \leqslant K_f(\omega) e^{-\kappa n} \tag{4.1.4}$$

for some constant $\kappa > 0$ and a random variable $K_f = K_f(\omega) > 0$ such that

$$\int \ln K_f(\omega)\, d\mathbb{P}(\omega) < \infty \quad \text{in the one-sided case} \tag{4.1.5}$$

and

$$\int K_f(\omega)\, d\mathbb{P}(\omega) < \infty \quad \text{in the two-sided case.} \tag{4.1.6}$$

A random subshift of finite type F is called topologically mixing if there exists a $\mathbb{Z}^+ = \{1, 2, \ldots\}$-valued random variable $\tilde{N} = \tilde{N}(\omega) < \infty$ on $(\Omega, \mathcal{F}, \mathbb{P})$ so that for $\mathbb{P}$-a.a. ω, $A(\vartheta^{-\tilde{N}}\omega) \ldots A(\vartheta^{-2}\omega) A(\vartheta^{-1}\omega)$ is a matrix with positive entries. This property is equivalent to existence of random variables $N = N(\omega) < \infty$ and $\tilde{N} = \tilde{N}(\omega) < \infty$ such that for any $n \geqslant N(\omega)$ and $k \geqslant \tilde{N}(\omega)$ the matrices $A(\omega)A(\vartheta\omega)\ldots A(\vartheta^n\omega)$ and $A(\vartheta^{-k}\omega)A(\vartheta^{-k+1}\omega)\ldots A(\vartheta^{-1}\omega)$ have only positive entries. The main goal of this section is to establish the following result.

THEOREM 4.1.1. *Assume that $\mathcal{F}$ is countably generated* (*separable*) *σ-algebra, F is a topologically mixing random* (*one- or two-sided*) *subshift of finite type and that a measurable function $f : \mathcal{E} \to \mathbb{R}$ satisfies* (4.1.3)–(4.1.6). *Then there exists a unique probability measure $\mu = \mu_f \in \mathcal{P}_{\mathbb{P}}(\mathcal{E})$, called a fiber* (*or relative*) *Gibbs measure* (*or state*) *for a potential f, such that for some random variables $C_f = C_f(\omega) > 0$ and $\lambda_f = \lambda_f(\omega) > 0$ satisfying*

$$\int |\ln C_f|\, d\mathbb{P} < \infty \quad \textit{and} \quad \int |\ln \lambda_f|\, d\mathbb{P} < \infty \tag{4.1.7}$$

we have

$$C_f^{-1} \leqslant \frac{\mu_\omega\{\tilde{x} \in \mathcal{E}_\omega \colon \tilde{x}_i = x_i \ \forall i = 0, 1, \ldots, n-1\}}{\exp\left(\sum_{i=0}^{n-1} (f \circ \Theta^i(\omega, x) - \ln \lambda(\vartheta^i \omega))\right)} \leqslant C_f \tag{4.1.8}$$

for all $x \in \mathcal{E}_\omega$ and $\mathbb{P}$-a.a. ω, where $\{\mu_\omega\}$ are disintegrations of μ. Moreover, μ is Θ-invariant and it is the unique Θ-invariant probability measure maximizing in the variational principle (1.2.11) *so that*

$$\pi_F(f) = \int \ln \lambda_f(\omega)\, d\mathbb{P} = h_\mu(F) + \int f\, d\mu, \tag{4.1.9}$$

i.e., μ is the unique equilibrium state for f. Furthermore, if $\mathbb{P}$ is ergodic (*mixing*) *with respect to ϑ then μ is ergodic* (*mixing*) *with respect to Θ. The disintegrations satisfy certain nonuniform ω-wise decay of correlations property* (*see Lemma* 6.3 *in* [98]).

There exist several approaches to the proof of Theorem 4.1.1. The first approach appeared in [91] and extended in [79] works for a wide class of random expanding transformations including random subshifts of finite type. This method is an extension to the random (relative) situation of arguments from [161]. The paper [79] contains also a statistical mechanics motivated treatment of a part of Theorem 4.1.1. A proof based on projective metrics of a part of Theorem 4.1.1 suggested in [35] imposes more restrictive conditions requiring random variables ℓ and K_f to be constants. It is possible also to follow the strategy described in the classical exposition [36] adjusting it to the random setup though this way is longer than the method of [91] and [79]. Observe also that in [69] Gibbs measures were constructed for more general random transformations (under the condition of random (relative) specification) than just topologically mixing subshifts of finite type.

First, we will assume that $\mathbb{P}$ is ergodic with respect to ϑ since for otherwise we can restrict our attention to its ergodic components in the ergodic decomposition. Next, we will reduce the two-sided case to the one-sided one. Following the standard definition we will say that two functions $f, g \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$ (with $\mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$ defined at the beginning of Section 1.2) are cohomologous if there exists another function $u \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$ such that for all $x \in \mathcal{E}_\omega$ and $\mathbb{P}$-a.a. ω,

$$f(\omega, x) = g(\omega, x) - u(\omega, x) + u \circ \Theta(\omega, x). \tag{4.1.10}$$

Clearly, if $f, g \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$ are cohomologous and Theorem 4.1.1 holds true for f then it holds true for g, as well. Next, we will show that in the case of a two-sided subshift for any f satisfying (4.1.3), (4.1.4), and (4.1.6) there exists g cohomologous to f satisfying (4.1.3), (4.1.4), and (4.1.6) and such that $f(\omega, x) = f(\omega, \tilde{x})$ whenever $x, \tilde{x} \in \mathcal{E}_\omega$ and $x_i = \tilde{x}_i$ for all $i \geqslant 0$. This will imply that we can restrict our attention to the proof of Theorem 4.1.1 in the case of one-sided subshifts. Namely, for each t and ω, $1 \leqslant t \leqslant \ell(\omega)$ choose measurably in ω a sequence $\{a^\omega_{k,t}\}_{k=-\infty}^{\infty} \in \mathcal{E}_\omega$ with $a^\omega_{0,t} = t$. Define $r_\omega : \mathcal{E}_\omega \to \mathcal{E}_\omega$ by $r_\omega(x) = x^*$ where $x^*_k = x_k$ for $k \geqslant 0$ and $x^*_k = a^\omega_{k,x_0}$ for $k < 0$. Then

$$u(\omega, x) = \sum_{j=0}^{\infty} \big(f \circ \Theta^j(\omega, x) - f \circ \Theta^j\big(\omega, r_\omega(x)\big)\big). \tag{4.1.11}$$

Since $(F^j_\omega x)_i = (F^j_\omega r_\omega(x))_i$ provided $-j \leqslant i < \infty$ we obtain by (4.1.4) that

$$\big| f \circ \Theta^j(\omega, x) - f \circ \Theta^j\big(\omega, r_\omega(x)\big)\big| \leqslant K_f\big(\vartheta^j \omega\big) e^{-\kappa j}. \tag{4.1.12}$$

By (4.1.6) (in fact, here (4.1.5) is enough) the series (4.1.11) converges for all $x \in \mathcal{E}_\omega$ and $\mathbb{P}$-a.a. ω and $u \in \mathbb{L}^1_{\mathcal{E}}(\Omega, \mathcal{C}(X))$. Now it is easy to check directly (cf. [36]) that $g = f - u + u \circ \Theta$ satisfies the above requirements.

In what follows we will deal exclusively with one-sided random subshifts assuming that f satisfies the conditions (4.1.3)–(4.1.5). The random Ruelle–Perron–Frobenius (RPF)

operator $\mathcal{L}_f^\omega$ corresponding to a function f maps functions on $\mathcal{E}_\omega$ to functions on $\mathcal{E}_{\vartheta\omega}$ by the formula

$$\mathcal{L}_f^\omega q(x) = \sum_{z\in F_\omega^{-1}x} e^{f(\omega,z)} q(z), \quad x \in \mathcal{E}_{\vartheta\omega}. \tag{4.1.13}$$

The main step in the construction of Gibbs measures of Theorem 4.1.1 is the following fiber or relative version of the Ruelle–Perron–Frobenius (RPF) theorem.

THEOREM 4.1.2. *Let F be a topologically mixing one-sided random subshift of finite type and a function f satisfies the conditions* (4.1.3)–(4.1.5). *Then there exists a triple consisting of a positive random variable $\lambda = \lambda(\omega)$, of a positive measurable in (ω, x) and continuous in x function $h = h(\omega) = h(\omega, x)$, and of a measurable in ω family of probability measures ν_ω on $\mathcal{E}_\omega$ such that*

$$\mathcal{L}_f^\omega h(\omega)(x) = \lambda(\omega) h(\vartheta\omega, x), \quad (\mathcal{L}_f^\omega)^* \nu_{\vartheta\omega} = \lambda(\omega)\nu_\omega, \quad \textit{and}$$
$$\int_{\mathcal{E}_\omega} h(\omega, x)\, d\nu_\omega(x) = 1. \tag{4.1.14}$$

PROOF. The existence of ν_ω and $\lambda = \lambda(\omega)$ satisfying $(\mathcal{L}_f^\omega)^* \nu_{\vartheta\omega} = \lambda(\omega)\nu_\omega$ is just a consequence of the Schauder–Tichonoff fixed point theorem (see [91]). The next step is to show the existence of $h = h(\omega, x)$. Set

$$R_\omega = \sum_{l=1}^{\infty} K_f\big(\vartheta^{-l}\omega\big) e^{-\kappa l}, \ \rho(x, \tilde{x}) = e^{-\min\{j\geqslant 0:\, x_j \neq \tilde{x}_j\}}, \tag{4.1.15}$$

provided $x, \tilde{x} \in \mathcal{E}_\omega$. It follows by (4.1.5) and the Borel–Cantelli lemma that $R_\omega < \infty$ $\mathbb{P}$-a.s. Denote also $R_\omega(x, \tilde{x}) = R_\omega(\rho(x, \tilde{x}))^\kappa$ and define a family of cones of continuous functions on $\mathcal{E}_\omega$ by

$$\Lambda^\omega = \left\{ q:\ q \geqslant 0,\ \int q\, d\nu_\omega = 1 \text{ and } q(x) \leqslant e^{\gamma R_\omega(x,\tilde{x})} q(\tilde{x}) \text{ if } x_0 = \tilde{x}_0 \right\}, \tag{4.1.16}$$

where $\gamma \geqslant 1$ is a constant. Clearly, $1 \in \Lambda^\omega$ and we claim that (cf. [91]),

$$\big(\lambda(\omega)\big)^{-1} \mathcal{L}_f^\omega \Lambda^\omega \subset \Lambda^{\vartheta\omega} \quad \text{for any } \gamma \geqslant 1. \tag{4.1.17}$$

Indeed, if $q \in \Lambda^\omega$ then by the choice of ν_ω,

$$\big(\lambda(\omega)\big)^{-1} \int q\, d(\mathcal{L}_f^\omega)^* \nu_{\vartheta\omega} = \int q\, d\nu_\omega = 1.$$

Next, let $x, x' \in \mathcal{E}_{\vartheta\omega}$ and $x_0 = x_0'$. For each $y \in F_\omega^{-1}x$ there is exactly one $y' \in F_\omega^{-1}x'$ such that $y_0' = y_0$. Thus by (4.1.4), (4.1.13), (4.1.15) and (4.1.16) for any $q \in \Lambda^\omega$ and $\gamma \geqslant 1$,

$$\begin{aligned}\mathcal{L}_f^\omega q(x) &= \sum_{y \in F_\omega^{-1}x} e^{f(\omega,y)} q(y) \\ &\leqslant \sum_{y' \in F_\omega^{-1}x} e^{f(\omega,y')} q(y') \exp\bigl(K_f(\omega) e^{-\kappa} \bigl(\rho(x,x')\bigr)^\kappa + \gamma e^{-\kappa} R_\omega(x,x')\bigr) \\ &\leqslant e^{\gamma R_{\vartheta\omega}(x,x')} \mathcal{L}_f^\omega q(x')\end{aligned}$$

proving (4.1.17). Furthermore, Λ^ω is, clearly, closed and convex. We assert that Λ^ω is also bounded and compact. Indeed, by (6.10) in [98] for any $q \in \Lambda^\omega$,

$$q \leqslant G_\omega e^{\gamma e^{-1}}, \tag{4.1.18}$$

where

$$G_\omega = \lambda(\omega) \dots \lambda\bigl(\vartheta^{N(\omega)}\omega\bigr) \exp\left(\sum_{j=0}^{N(\omega)-1} \bigl\| f\bigl(\vartheta^j \omega\bigr) \bigr\|_\infty \right)$$

and $N = N(\omega)$ was defined just before the statement of Theorem 4.1.1. Hence we obtain that for any $q \in \Lambda^\omega$,

$$\bigl|q(x) - q(\tilde{x})\bigr| \leqslant \bigl|e^{\gamma R_\omega(x,\tilde{x})} - 1\bigr| G_\omega e^{\gamma e^{-1}} \tag{4.1.19}$$

which implies the equicontinuity of Λ^ω and its compactness follows by the Arzela–Ascoli theorem. Thus the product

$$\Lambda_{\mathbb{Z}}^\omega = \cdots \times \Lambda^{\vartheta^{-1}\omega} \times \Lambda^\omega \times \Lambda^{\vartheta\omega} \times \cdots$$

with the product topology is a nonempty, convex and compact space.

Define the map Φ^ω on $\Lambda_{\mathbb{Z}}^\omega$ which sends a sequence $\{q_i\} \in \Lambda_{\mathbb{Z}}^\omega$ to a sequence $\{(\Phi^\omega q)_i\}$ by the formula

$$(\Phi^\omega q)_i = \bigl(\lambda\bigl(\vartheta^{i-1}\omega\bigr)\bigr)^{-1} \mathcal{L}_f^{\vartheta^{i-1}\omega} q_{i-1}. \tag{4.1.20}$$

By (4.1.17), $\Phi^\omega \Lambda_{\mathbb{Z}}^\omega \subset \Lambda_{\mathbb{Z}}^\omega$ and by the Schauder–Tychonoff theorem we derive the existence of a fixed point of Φ^ω which we denote by $h^\omega = \{h_i^\omega\}$ and put, finally, $h(\vartheta^i\omega, x) = h_i^\omega(x)$ which satisfies, clearly, (4.1.14). Observe that since ϑ is ergodic it is either aperiodic or purely periodic. In the former case there is no problem with the above definition of $h(\vartheta^i\omega)$ and in the latter case we choose h^ω to be periodic sequences which is possible since Φ^ω preserves corresponding spaces of periodic sequences of functions. Observe that we do not claim at this point that h^ω depends measurably on ω (though we could apply some theorem

on a measurable choice to ensure this) but this follows automatically from some asymptotic formulas (see [91] and [79]) which also imply the uniqueness assertions of Theorem 4.1.2.

Next we show that $h(\omega) > 0$. Indeed, set $\mathcal{L}_f^{\omega,n} = \mathcal{L}_f^{\vartheta^{n-1}\omega} \circ \cdots \circ \mathcal{L}_f^{\vartheta\omega} \circ \mathcal{L}_f^{\omega}$. If $h(\omega, x) = 0$ for some $x \in \mathcal{E}_\omega$ then in view of

$$\mathcal{L}_f^{\vartheta^{-n}\omega,n} h(\vartheta^{-n}\omega)(x) = \left(\prod_{i=-n}^{-1} \lambda\big(\vartheta^i \omega\big) \right) h(\omega, x)$$

we conclude that $h(\vartheta^{-n}\omega, y) = 0$ whenever $F^n_{\vartheta^{-n}\omega} y = x$. For $C > 0$ large enough the set $\Gamma_C = \{\omega\colon\ R_\omega \leqslant C,\ G_\omega \leqslant C\}$ has positive $\mathbb{P}$-measure. Then for $\mathbb{P}$-almost all ω we can define successive times $n_i = n_i(\omega) \to \infty$ as $i \to \infty$ when $\vartheta^{-n_i}\omega \in \Gamma_C$. It follows by (4.1.19) that the family $h(\vartheta^{-n_i}\omega)$, $i = 1, 2, \ldots$, is equi-continuous and so the assertion that $h(\vartheta^{-n_i}\omega, y) = 0$ whenever $F^{n_i}_{\vartheta^{-n_i}\omega} y = x$ would yield that $\|h(\vartheta^{-n_i}\omega)\|_\infty \to 0$ as $i \to \infty$ contradicting the equality $\int h(\vartheta^{-n_i}\omega)\, d\nu_{\vartheta^{-n_i}\omega} = 1$, which completes the proof of Theorem 4.1.2. □

In fact, the triple $\lambda(\omega), h(\omega), \nu_\omega$ satisfying (4.1.4) is unique but this requires further arguments. Define

$$\mu_\omega = h_\omega \nu_\omega. \tag{4.1.21}$$

Let q be a bounded measurable function on $\mathcal{E}_{\vartheta\omega}$. Then

$$\begin{aligned}
\int q\, d\mu_{\vartheta\omega} &= \int q h(\vartheta\omega)\, d\nu_{\vartheta\omega} = \big(\lambda(\omega)\big)^{-1} \int q\big(\mathcal{L}_f^\omega h(\omega)\big)\, d\nu_{\vartheta\omega} \\
&= \big(\lambda(\omega)\big)^{-1} \int \mathcal{L}_f^\omega\big(h(\omega) \cdot (q \circ F_\omega)\big)\, d\nu_{\vartheta\omega} = \int h(\omega)(q \circ F_\omega)\, d\nu_\omega \\
&= \int q\, dF_\omega \mu_\omega.
\end{aligned}$$

Hence

$$F_\omega \mu_\omega = \mu_{\vartheta\omega}. \tag{4.1.22}$$

It turns out that $\{\mu_\omega\}$ are disintegrations of the unique Gibbs measure μ for f which has the properties described in Theorem 4.1.1. This can be proved following the strategy of [36] via a series of lemmas such as Lemmas 6.1–6.4 from [98] together with the variational principle (1.2.11) and Theorem 3.2 from [79] which gives ergodicity and mixing of μ. A somewhat shorter proof based on the (deterministic) approach from [161] had been exhibited in [91] and extended in [79]. If the random variables K_f and N introduced before Theorem 4.1.1 were, in fact, constants then already the method of [91] does the job but in the general case of random K_f and N some additional arguments should be employed as explained in Section 3 of [79] in order to overcome nonuniformity in ω.

The key point of the approach from [161] extended for random subshifts in [91] and [79] is the characterization of μ as the unique probability measure whose disintegrations satisfy (see Proposition 2.3 in [91] and the item (d) in Theorem 3.1 from [79]),

$$(\mathcal{L}^{\omega}_{\log g(\omega)})^{*}\mu_{\vartheta\omega} = \mu_{\omega} \tag{4.1.23}$$

for $\mathbb{P}$-a.a. ω where $g(\omega) = e^{f(\omega)}h(\omega)(\lambda(\omega)h(\vartheta\omega)\circ F_{\omega})^{-1}$. In fact, μ_{ω} for $\mathbb{P}$-a.a. ω is obtained via the limit

$$\int q\,d\mu_{\omega} = \lim_{n\to\infty}\mathcal{L}^{\omega,n}_{\log g(\omega)}q \tag{4.1.24}$$

for any continuous function q on $\mathcal{E}_{\omega}$. It remains to verify (4.1.21). Indeed, let $m_{\omega} = h(\omega)\nu_{\omega}$. Then

$$\begin{aligned}\int q\,d(\mathcal{L}^{\omega}_{\log g})^{*}m_{\vartheta\omega} &= \int h(\vartheta\omega)\mathcal{L}^{\omega}_{\log g}q\,d\nu_{\vartheta\omega}\\ &= \big(\lambda(\omega)\big)^{-1}\int \mathcal{L}^{\omega}_{f}\big(h(\omega)q\big)\,d\nu_{\vartheta\omega} = \int h(\omega)q\,d\nu_{\omega}\\ &= \int q\,dm_{\omega}.\end{aligned}$$

Hence $(\mathcal{L}^{\omega}_{\log g})^{*}m_{\vartheta\omega} = m_{\omega}$ and by uniqueness we obtain $m_{\omega} = \mu_{\omega}$. Observe that

$$\big(\lambda\big(\vartheta^{n-1}\omega\big)\ldots\lambda(\vartheta\omega)\lambda(\omega)\big)^{-1}\mathcal{L}^{\omega,n}_{f}q = h(\vartheta^{n}\omega)\mathcal{L}^{\omega,n}_{\log g}\left(\frac{q}{h(\omega)}\right).$$

This together with (4.1.21) and (4.1.24) yield

$$\lim_{n\to\infty}\big(\lambda\big(\vartheta^{n-1}\omega\big)\ldots\lambda(\vartheta\omega)\lambda(\omega)\big)^{-1}\big(h(\vartheta^{n}\omega)\big)^{-1}\mathcal{L}^{\omega,n}_{f}q = \int q\,d\nu_{\omega}. \tag{4.1.25}$$

Employing (4.1.25) for an arbitrary continuous q and for $q = 1$ and dividing one formula by the other we obtain

$$\lim_{n\to\infty}\frac{\mathcal{L}^{\omega,n}_{f}q}{\mathcal{L}^{\omega,n}_{f}1} = \int q\,d\nu_{\omega}, \tag{4.1.26}$$

and so the measures ν_{ω} are determined uniquely $\mathbb{P}$-a.s. Now $\lambda(\omega) = ((\mathcal{L}^{\omega}_{f})^{*}\nu_{\vartheta\omega})(\mathcal{E}_{\omega})$ is determined uniquely, as well. In order to prove the uniqueness of h observe that by Theorem 3.2 from [79] the measure μ is an ergodic invariant measure of the skew product transformation Θ and since all such measures satisfying (4.1.23) (constructed may be with different functions h) must be equivalent to ν by (4.1.21) we conclude that all such μ should coincide. Finally, since now both μ_{ω} and ν_{ω} are uniquely defined then $h(\omega) = \frac{d\mu_{\omega}}{d\nu_{\omega}}$

is also determined uniquely $\mathbb{P}$-a.s. It follows from the assertion (f) of Theorem 3.1 in [79] (see also [35]) that μ satisfies (4.1.8), i.e., it is a Gibbs measure for the potential f and, moreover, it is the unique equilibrium state for f (see Theorem 3.2(iii) from [91] and Theorem A in [79]). By Theorem 3.2 from [79] the measure μ is mixing with respect to Θ whenever $\mathbb{P}$ is mixing with respect to ϑ. Certain ω-wise mixing of μ is obtained in Lemma 6.3 from [98].

Among important examples of fiber Gibbs measures are random Markov and random Bernoulli measures. Their construction is the following. Let $p_i(\omega) \geqslant 0$, $i = 1, \dots, \ell(\omega)$, and $p_{ij}(\omega) \geqslant 0$, $i = 1, \dots, \ell(\omega)$, $j = 1, \dots, \ell(\vartheta\omega)$, be measurable in ω families of probability vectors and probability matrices, i.e., $\sum_{i=1}^{\ell(\omega)} p_i(\omega) = 1$ and $\sum_{j=1}^{\ell(\vartheta\omega)} p_{ij}(\omega) = 1$ $\forall i$, such that $\sum_{i=1}^{\ell(\omega)} p_i(\omega) p_{ij}(\omega) = p_j(\vartheta\omega)$. For any cylinder set $C^{\omega}_{\alpha_0,\dots,\alpha_n} = \{x \in \mathcal{E}_\omega\colon\ x_i = \alpha_i$ $\forall i = 0, 1, \dots, n\}$ put

$$\mu_\omega(C^{\omega}_{\alpha_0,\dots,\alpha_n}) = p_{\alpha_0}(\omega) p_{\alpha_0\alpha_1}(\omega) \dots p_{\alpha_{n-1}\alpha_n}\big(\vartheta^{n-1}\omega\big).$$

Employing the Kolmogorov extension theorem we conclude that μ_ω can be extended uniquely to the whole Borel σ-algebra on $\mathcal{E}_\omega$ and this construction produces a measurably depending on ω family of probability measures μ_ω on $\mathcal{E}_\omega$ such that $F_\omega \mu_\omega = \mu_{\vartheta\omega}$ where F_ω is the left shift. Hence, the probability measure μ given by $d\mu(\omega, x) = d\mu_\omega(x)\, d\mathbb{P}(\omega)$ is Θ-invariant and it is called a random (fiber) Markov measure and the corresponding pair (F, μ) is called a random Markov shift. Set $a_{ij}(\omega) = 1$ if $p_{ij}(\omega) > 0$ and $a_{ij}(\omega) = 0$ if $p_{ij}(\omega) = 0$. Then the matrices $A(\omega) = (a_{ij}(\omega))$ determine a subshift of finite type and we assume that it is topologically mixing as defined at the beginning of this section. Of course, the same condition can be expressed in terms of products of matrices $(p_{ij}(\omega))$. Then it follows from Theorem 4.1.1 and also can be verified directly (cf. [94] and Section 4.3 below) that μ is the Gibbs measure for the potential $f(\omega, x) = \ln p_{x_0 x_1}(\omega)$. If $p_{ij}(\omega) = p_j(\omega)$ does not depend on i then the corresponding measure μ obtained as above is called a random (fiber) Bernoulli measure and the pair (F, μ) is called in this case a random Bernoulli shift.

4.2. *Random expanding and hyperbolic transformations*

We start with the simpler case of random expanding in average endomorphisms and again assume that the σ-algebra $\mathcal{F}$ is countably generated. In this setup $\mathcal{E}_\omega$'s for all ω coincide with one compact connected d-dimensional C^2 Riemannian manifold M (though a more general case of manifolds M_ω depending on ω may be considered, as well) and all $F_\omega\colon M \to M$ are C^2 endomorphisms of M such that

$$\log\big\|DF_\omega^{-1}\big\|, \log\|DF_\omega\| \in L^1(\Omega, \mathbb{P}), \tag{4.2.1}$$

and

$$\alpha = \int \log\big\|DF_\omega^{-1}\big\|\, d\mathbb{P}(\omega) < 0, \tag{4.2.2}$$

where DF_ω is the differential of F_ω and $\|Df\| = \sup_{x\in M} \|D_x f\|$ for any C^1 map f of M. Again, we define a random RPF operator $\mathcal{L}_f^\omega$ by (4.1.13) and assume that $f(\omega)$'s are Hölder continuous, i.e.,

$$\big|f(\omega,x) - f(\omega,y)\big| \leqslant K_f(\omega)\big(d(x,y)\big)^\kappa \quad \forall x, y \in M \tag{4.2.3}$$

(where $d(\cdot,\cdot)$ is the distance function on M) for some $\kappa > 0$ and a random variable $K_f(\omega) > 0$ satisfying (4.1.5). Then the corresponding analogy of Theorems 4.1.1 and 4.1.2 holds true.

THEOREM 4.2.1. *Let F be a random C^2 endomorphism of M satisfying* (4.2.1) *and* (4.2.2) *and let $f = f(\omega,x)$ be a function on $\Omega \times M$ satisfying* (4.2.3), (4.1.3), *and* (4.1.5). *Then there exists a unique triple consisting of a positive random variable $\lambda = \lambda(\omega)$, of a positive measurable in (ω,x) and continuous in x function $h = h(\omega) = h(\omega,x)$, and of a measurable in ω family of probability measures ν_ω on M such that* (4.1.14) *holds true. Furthermore, the measure μ whose disintegrations have the form $\mu_\omega = h(\omega)\nu_\omega$ is Θ-invariant and ergodic* (*mixing*) *whenever $\mathbb{P}$ is ergodic* (*mixing*). *Furthermore, μ is the unique equilibrium state for f, i.e., it is the unique Θ-invariant probability measure satisfying* (4.1.9).

Let m be the normalized Riemannian volume on M and

$$f(\omega,x) = -\ln \frac{dm \circ F_\omega}{dm}(x) = -\ln |\operatorname{Jac} D_x F_\omega|,$$

where Jac *denotes the Jacobian with respect to Riemannian inner products in tangent spaces. Then f satisfies the above conditions and the corresponding objects $\lambda(\omega)$, h, ν_ω, μ_ω constructed for such f have the form $\lambda(\omega) = 1$, $\nu_\omega = m$, and $\mu_\omega = h(\omega)m$ which yield the unique Θ-invariant measure μ whose disintegrations μ_ω are equivalent to the Riemannian volume m on M.*

This result has been proved in [91] and [79]. Observe that such results for both random subshifts of finite type and random expanding transformations can be dealt with simultaneously as explained in Section 6 of [98]. Note also that under excessively restrictive conditions random Markov partitions for random expanding transformations were constructed in [34] which provides a longer way to prove Theorem 4.2.1 reducing it to the symbolic setup of random subshifts of finite type considered in Section 4.1. We note that equilibrium states were constructed in [100] also for one-dimensional random piecewise expanding in average maps which has applications to random f-expansions, in particular, to continued fractions with random digits. Transformations here may have unbounded derivatives and infinitely many branches which result in infinite entropy for some measures and the definition of equilibrium states should use the information instead. Another difficulty is the possible presence of neutral points such as 1 for the Gauss map $x \to \{\frac{1}{x}\}$ which is related to continued fractions and the study of the fiber (relative) thermodynamic formalism for it yields, in particular, a computation of the Hausdorff dimension of distributions of continued fractions with random digits (see [100]). Observe also that Θ-invariant measures with absolutely continuous disintegrations were studied also for multidimensional random

piecewise smooth expanding maps in [41] and [40]. Some results about equilibrium states for random nonuniformly expanding maps were obtained in [6]. Observe that without the invertibility assumption on the base transformation ϑ on Ω the equilibrium states theory of even expanding RDS encounters certain difficulties and corresponding results should undergo substantial modifications (see [14,52,53]) in comparison with the case of an invertible ϑ considered above.

Next, we turn our attention to random (uniformly) hyperbolic diffeomorphisms following [68]. The proofs of the results below are indicated in [68] but we admit that they are not written yet with all necessary details. The corresponding results for random hyperbolic diffeomorphism obtained via small perturbations of a deterministic hyperbolic diffeomorphism can be found in [115]. Let M be a smooth d-dimensional Riemannian manifold. We say that $\Lambda = \{\Lambda(\omega)\colon\ \omega \in \Omega\}$ is a random compact set if each $\Lambda(\omega) \subset M$ is compact and the map $(x, \omega) \to d(x, \Lambda(\omega))$ is measurable, where d is the Riemannian distance on M. A random variable $g : \Omega \to \mathbb{R}_+$ will be called tempered, if it satisfies $\lim_{n\to\pm\infty} \frac{1}{n} \log g(\vartheta^n\omega) = 0$ $\mathbb{P}$-a.s.

DEFINITION 4.2.2. A random compact nonempty set $\Lambda = \{\Lambda(\omega)\colon\ \omega \in \Omega\}$ is called invariant under F, if $F_\omega\Lambda(\omega) = \Lambda(\vartheta\omega)$ for $\mathbb{P}$-almost all $\omega \in \Omega$. Such a Λ is called a random hyperbolic set for F if there exist an open set V with a compact closure $\bar{V}$, tempered random variables $\lambda > 0$, $\alpha > 0$, $C > 0$, and subbundles $\Gamma^u(\omega)$ and $\Gamma^s(\omega)$ of the tangent bundle $T\Lambda(\omega)$, depending measurably on ω, such that

(1) For $\mathbb{P}$-almost all (a.a.) $\omega \in \Omega$ there exist a measurable in ω family of open sets $U(\omega)$ such that $\{x\colon\ d(x, \Lambda(\omega)) < \alpha(\omega)\} \subset U(\omega) \subset V$, $F_\omega U(\omega) \subset V$, and F_ω restricted to $U(\omega)$ is a diffeomorphism and both $\log^+ \sup_{x\in U(\omega)} \|D_x F_\omega\|$ and $\log^+ \sup_{x\in U(\omega)} \|(D_x F)^{-1}_\omega\|$ belong to $\mathbb{L}^1(\Omega, \mathbb{P})$;
(2) $T\Lambda(\omega) = \Gamma^u(\omega) \oplus \Gamma^s(\omega)$, $DF_\omega\Gamma^u(\omega) = \Gamma^u(\vartheta\omega)$, $DF_\omega\Gamma^s(\omega) = \Gamma^s(\vartheta\omega)$, $\angle(\Gamma^u(\omega), \Gamma^s(\omega)) \geqslant \alpha(\omega)$ $\mathbb{P}$-a.s., where $\angle(\Gamma^u(\omega), \Gamma^s(\omega))$ denotes the minimal angle between $\Gamma^u(\omega)$ and $\Gamma^s(\omega)$;
(3) for $n \in \mathbb{N}$ and $\lambda(n, \omega) = \lambda(\omega) \dots \lambda(\vartheta^{n-1}\omega)$ and $\mathbb{P}$-a.a. ω,

$$\|DF^n_\omega \xi\| \leqslant C(\omega)\lambda(n, \omega)\|\xi\| \quad \text{for } \xi \in \Gamma^s(\omega)$$

and

$$\|DF^{-n}_\omega \eta\| \leqslant C(\vartheta^{-n}\omega)\lambda(n, \vartheta^{-n}\omega)\|\eta\| \quad \text{for } \eta \in \Gamma^u(\omega);$$

(4) $\int \log\lambda \, d\mathbb{P} < 0$;
(5) $\log\alpha \in \mathbb{L}^1(\Omega, \mathbb{P})$.

If, in addition, $F_\omega U(\omega) \subset U(\vartheta\omega)$ $\mathbb{P}$-a.s. $\bigcap_{n=0}^\infty F^n_{\vartheta^{-n}\omega} U(\vartheta^{-n}\omega) = \Lambda(\omega)$ then we call Λ a random hyperbolic attractor of F. If M is compact and all $\Lambda(\omega)$ coincide with M and satisfy assumptions above then we will call F a random Anosov diffeomorphism.

Observe that the subbundles $\Gamma^u(\omega) = \{\Gamma^u_x(\omega)\colon\ x \in \Lambda(\omega)\}$, $\Gamma^s(\omega) = \{\Gamma^s_x(\omega)\colon\ x \in \Lambda(\omega)\}$ are necessarily continuous in $x \in \Lambda(\omega)$, since the inequalities in (3) being true for sequences $\xi_n \in \Gamma^s_{x_n}(\omega)$, $\eta_n \in \Gamma^u_{x_n}(\omega)$ such that $\xi = \lim_{n\to\infty} \xi_n \in \Gamma^s_x(\omega)$, $\eta = \lim_{n\to\infty} \eta_n \in$

$\Gamma^u_x(\omega)$, $x = \lim_{n\to\infty} x_n$ remain true for ξ and η. By ergodicity of ϑ, $\dim \Gamma^u(\omega)$ and $\dim \Gamma^s(\omega)$ are constant P-a.s.

Actually, (3) can be replaced by the following weaker condition.

(3′) There exists $n \in \mathbb{N}$ such that

$$\int \log \|DF^n_\omega|_{\Gamma^s(\omega)}\|\, d\mathbb{P}(\omega) < 0, \qquad \int \log \|DF^{-n}_\omega|_{\Gamma^u(\omega)}\|\, d\mathbb{P}(\omega) < 0.$$

This property already provides the needed contracting/expanding splitting relying on the ergodic theorem. For the same reason we could replace the random variable λ in Definition 4.2.2 by a constant via a change of the tempered random variable C. We arrive here at a nonuniform in $\omega \in \Omega$ but uniform in $x \in \Lambda(\omega)$ kind of hyperbolicity: due to the random variable C the time of the onset of expansion and contraction of the linear map restricted to the subbundles depends on chance. This problem can be resolved with the help of random Lyapunov norms (see Lemma 2.2.3 above and [7, Section 4.3]). Characterizations of random hyperbolic sets via random norms and random cones can be found in [68]. The equivalence of definitions of random hyperbolic sets via a splitting of the tangent bundle as above and via random cones follows from general results exhibited in Section 6.6 from [75].

The following example of a random hyperbolic set was discussed in [115].

EXAMPLE 4.2.3. Assume that a deterministic (local) diffeomorphism f has a hyperbolic invariant set Λ_f. Let $\mathcal{U}(f)$ be a small C^1 neighborhood of f so that any $g \in \mathcal{U}(f)$ has a hyperbolic invariant set Λ_g close to Λ_f. Now any measurable map $F : \Omega \to \mathcal{U}(f)$ so that $F(\omega) = F_\omega \in \mathcal{U}(f)$ generates an RDS F having a random hyperbolic set $\Lambda(\omega)$. Moreover, by the structural stability theorem there exist random Hölder continuous homeomorphisms $h_\omega : \Lambda \to \Lambda(\omega)$ such that $h_{\vartheta\omega} \circ f = F_\omega \circ h_\omega$.

Another more explicit example of a random Anosov diffeomorphism appeared originally in [9] (in the preprint form in 1995).

EXAMPLE 4.2.4. Let $\sigma : \Omega \to \Omega$ be a $\mathbb{P}$-preserving ergodic invertible map and assume that $\vartheta = \sigma^2$ is also ergodic. Then we define for a random variable $n : \Omega \to \mathbb{N}$ with $\log n \in \mathbb{L}^1(\Omega, \mathbb{P})$ a torus automorphism F_A on the 2-torus $M = \mathbb{T}^2$ via linear lifts on $\mathbb{R}^2$ of the form

$$A(\omega) = \begin{pmatrix} 1 + n(\omega)n(\sigma\omega) & n(\sigma\omega) \\ n(\omega) & 1 \end{pmatrix}.$$

Denote by $[k_1, k_2, \ldots]$ the continued fraction

$$\cfrac{1}{k_1 + \cfrac{1}{k_2 + \cdots}}$$

and set $a(\omega) = [n(\omega), n(\sigma\omega), n(\sigma^2\omega), \ldots]$, $b(\omega) = [n(\sigma^{-1}\omega), n(\sigma^{-2}\omega), \ldots]$. Define

$$\xi(\omega) = \begin{pmatrix} a(\omega) \\ -1 \end{pmatrix}, \qquad \eta(\omega) = \begin{pmatrix} 1 \\ b(\omega) \end{pmatrix},$$

$$\lambda(\omega) = a(\omega)a(\sigma\omega), \qquad \gamma(\omega) = \frac{1}{b(\sigma\omega)b(\sigma^2\omega)}.$$

Then $A(\omega)\xi(\omega) = \lambda(\omega)\xi(\vartheta\omega)$, $A(\omega)\eta(\omega) = \gamma(\omega)\eta(\vartheta\omega)$, $\lambda(\omega) < 1, \gamma(\omega) > 1$, and so, ξ and η span the contracting and expanding (in average) directions, respectively, to make the whole torus a random hyperbolic set for F_A.

As in the deterministic case it is possible to integrate the hyperbolic splitting in order to obtain random local stable and unstable manifolds (cf. Section 2.2).

THEOREM 4.2.5. *For any $x \in \Lambda(\omega)$ there exist embedded C^1 manifolds $V_x^s(\omega)$ and $V_x^u(\omega)$* (*called* local stable *and* unstable manifold, *respectively*) *tangent to $\Gamma_x^s(\omega)$ and $\Gamma_x^u(\omega)$, respectively, at x such that $V_x^s(\omega)$ and $V_x^u(\omega)$ depend measurably on ω and for fixed $\omega \in \Omega$ continuously on $x \in \Lambda(\omega)$, $F_\omega V_x^s(\omega) \subset V^s_{F_\omega x}(\vartheta\omega)$, $F_\omega^{-1} V_x^u(\vartheta\omega) \subset V^u_{F_\omega^{-1}x}(\omega)$,*

$$V_x^s(\omega) \cap V_x^u(\omega) = \{x\} \tag{4.2.4}$$

and there exist tempered random variables C' and λ' such that for any $y \in V_x^s(\omega)$, $z \in V_x^u(\omega)$ and $\mathbb{P}$-a.a. ω,

$$\operatorname{dist}_{V^s_{F_\omega^n x}(\vartheta^n\omega)}(F_\omega^n x, F_\omega^n y) \leqslant C'(\omega)\lambda'(n,\omega)\operatorname{dist}_{V_x^s(\omega)}(x,y),$$

$$\operatorname{dist}_{V^u_{F_\omega^{-n}x}(\vartheta^{-n}\omega)}(F_\omega^{-n} x, F_\omega^{-n} z) \leqslant C'(\vartheta^{-n}\omega)\lambda'(n,\vartheta^{-n}\omega)\operatorname{dist}_{V_x^u(\omega)}(x,z),$$

where $\lambda'(n,\omega) = \lambda'(\omega)\ldots\lambda'(\vartheta^{n-1}\omega)$ and $\operatorname{dist}_{V_x^s(\omega)}$ is the distance in $V_x^s(\omega)$ induced by the random norms on the tangent bundle, extended to some neighborhood of $\Lambda(\omega)$ and used for the description of the hyperbolicity properties. Moreover, the angle between $V_x^s(\omega)$ and $V_x^u(\omega)$ at x is not less than $c(\omega)$ for some random variable $c > 0$ with $\log c \in \mathbb{L}^1(\Omega, \mathbb{P})$.

Observe that in the case of Example 4.2.3 taking into account Hölder continuity of conjugating homeomorphisms h_ω random stable and unstable manifolds are obtained directly as images under h_ω of deterministic stable and unstable manifolds for the unperturbed diffeomorphism f. As a consequence of a continuous dependence of $V_x^s(\omega)$, $V_x^u(\omega)$ on $x \in \Lambda(\omega)$, of a measurable dependence on ω, and (4.2.4) we obtain the following result.

COROLLARY 4.2.6. *There exists a small tempered random variable $\gamma > 0$ such that for all $x, y \in \Lambda(\omega)$ with $d(x,y) \leqslant \gamma(\omega)$ the intersection of $V_x^s(\omega)$ and $V_y^u(\omega)$ consists precisely of one point in M which is denoted by $[x,y]_\omega$. The mapping $[.,.]_\omega : \{(x,y) \in \Lambda(\omega) \times \Lambda(\omega)\colon d(x,y) \leqslant \gamma(\omega)\} \to M$ depends measurably on ω and is continuous for each fixed $\omega \in \Omega$.*

Observe that a transversal intersection of random stable and unstable manifolds of the same point which is called a random homoclinic point leads also to a random hyperbolic set (see [66]). Next, we will introduce the following important notions.

DEFINITION 4.2.7. We say that the random hyperbolic set Λ has a local product structure, if for $\mathbb{P}$-almost all $\omega \in \Omega$, $x, y \in \Lambda(\omega)$ with $d(x, y) \leqslant \gamma(\omega)$ we have that $[x, y]_\omega \in \Lambda(\omega)$.

It is easy to see that if a deterministic hyperbolic set Λ in Example 4.2.3 has the local product structure then the random one $\Lambda(\omega)$ will also have it in the sense of the above definition provided the neighborhood $\mathcal{U}(f)$ in Example 4.2.3 is small enough.

DEFINITION 4.2.8. Let δ be a strictly positive random variable. Then for any $\omega \in \Omega$ a sequence $\{y_n\}_{n\in\mathbb{Z}}$ in M is called an (ω, δ) pseudo-orbit of F if

$$d(y_{n+1}, F_{\vartheta^n\omega} y_n) \leqslant \delta\big(\vartheta^{n+1}\omega\big) \quad \text{for all } n \in \mathbb{Z}.$$

For a strictly positive random variable ε and any $\omega \in \Omega$ the orbit of a point $x \in M$ is said to (ω, ε)-shadow the (ω, δ) pseudo-orbit $\{y_n\}_{n\in\mathbb{Z}}$ if

$$d(F^n_\omega x, y_n) \leqslant \varepsilon(\vartheta^n\omega) \quad \text{for all } n \in \mathbb{Z}.$$

The following result is important, in particular, in a construction of random Markov partitions.

PROPOSITION 4.2.9 (Random Shadowing Lemma). *Assume that the random hyperbolic set Λ has a local product structure. Then for every tempered random variable $\varepsilon > 0$ there exists a tempered random variable $\beta > 0$ such that $\mathbb{P}$-a.s. every (ω, β) pseudo-orbit $\{y_n\}_{n\in\mathbb{Z}}$ with $y_n \in \Lambda(\vartheta^n\omega)$ can be (ω, ε)-shadowed by the orbit of a point $x \in \Lambda(\omega)$. If 2ε is chosen as an expansivity characteristic* (*see Definition* 1.3.1), *then the shadowing point x is unique. Moreover, if the y_n's are chosen to be random variables such that for $\mathbb{P}$-almost all $\omega \in \Omega$ the sequence $\{y_n(\omega)\}_{n\in\mathbb{Z}}$ is an (ω, β) pseudo-orbit, then the starting point $x(\omega)$ of the corresponding (ω, ε)-shadowing orbit depends measurably on ω.*

Again, in the case of Example 4.2.3 the proof is simple as we can employ the deterministic shadowing lemma for f and obtain its random version using the conjugation by the random homeomorphisms h_ω. Next, we introduce random Markov partitions which serve as an important tool for the study of random hyperbolic sets.

DEFINITION 4.2.10. Assume that the random hyperbolic set Λ has a local product structure (with a corresponding random variable γ). A nonempty subset R of some $\Lambda(\omega)$ is called a rectangle, if it has diameter less than $\gamma(\omega)$ and $x, y \in R$ implies that $[x, y]_\omega \in R$. Moreover such a rectangle R is called proper, if it is closed in $\Lambda(\omega)$ and if it is the closure of the interior of R as a subset of $\Lambda(\omega)$.

A (random) Markov partition of Λ is a family of finite covers $\mathcal{R}(\omega) = \{R_1(\omega), \ldots, R_{\ell(\omega)}(\omega)\}$ of $\Lambda(\omega)$ which depends measurably on $\omega \in \Omega$ and satisfies $\mathbb{P}$-a.s.

(1) each $R_i(\omega)$ is a proper rectangle;
(2) $\operatorname{int} R_i(\omega) \cap \operatorname{int} R_j(\omega) = \emptyset$, if $i \neq j$;
(3) $F_\omega(V^s_x(\omega) \cap R_i(\omega)) \subset V^s_{F_\omega x}(\vartheta\omega) \cap R_j(\vartheta\omega)$ for $x \in \operatorname{int} R_i(\omega)$, $F_\omega x \in \operatorname{int} R_j(\vartheta\omega)$;
(4) $F_\omega(V^u_x(\omega) \cap R_i(\omega)) \supset V^u_{F_\omega x}(\vartheta\omega) \cap R_j(\vartheta\omega)$ for $x \in \operatorname{int} R_i(\omega)$, $F(\omega)x \in \operatorname{int} R_j(\vartheta\omega)$.

Here local invariant manifolds are used with a size that is given by an expansivity characteristic. We refer to conditions (3) and (4) as Markov properties.

THEOREM 4.2.11. *If the random hyperbolic set Λ has local product structure, then there exists a random Markov partition of Λ for F. If F has the $C^{1+\gamma}$, $\gamma > 0$, norm whose logarithm belongs to $\mathbb{L}^1(\Omega, \mathbb{P})$ then the number of rectangles in the partition is a random variable $\ell = \ell(\omega) \in \mathbb{N}$ with $\ln \ell \in \mathbb{L}^1(\Omega, \mathbb{P})$.*

Clearly, in the case of Example 4.2.3 random Markov partitions can be obtained as images under the conjugating maps h_ω of deterministic Markov partitions for f Let $A(\omega) = (a_{ij}(\omega),\ i = 1, \dots, \ell(\omega),\ j = 1, \dots, \ell(\vartheta\omega))$ be matrices with 0 and 1 entries such that $a_{ij}(\omega) = 1$ if $F_\omega R_i(\omega) \cap R_j(\vartheta\omega) \neq 0$ and $a_{ij}(\omega) = 0$, otherwise. Define

$$\mathcal{E}_\omega = \left\{\xi = (\dots, \xi_{-1}, \xi_0, \xi_1, \dots)\colon\ \xi_i \in \left\{1, \dots, \ell\left(\vartheta^i\omega\right)\right\} \text{ and } a_{\xi_i\xi_{i+1}}\left(\vartheta^i\omega\right) = 1\ \forall i\right\}$$

and let σ be the left shift. It follows from the definition of random hyperbolic sets and of random Markov partitions that for any $\xi \in \mathcal{E}_\omega$ there exists a unique $\pi_\omega(\xi) \in \Lambda(\omega)$ such that

$$\pi_\omega(\xi) = \bigcap_{n \geqslant 0} \left((F^n_\omega)^{-1} R_{\xi_n}(\vartheta^n\omega) \cap F^n_{\vartheta^{-n}\omega} R_{\xi_{-n}}(\vartheta^{-n}\omega)\right)$$

and $\pi_\omega : \mathcal{E}_\omega \to \Lambda(\omega)$ is a continuous map semi-conjugating σ and F, i.e.,

$$\pi_{\vartheta\omega}\sigma = F_\omega \pi_\omega.$$

In fact, π_ω is one-to-one except for points of the set

$$\partial_{\mathcal{R}}(\omega) = \bigcup_{n \geqslant 0} \left((F^n_\omega)^{-1}\left(\bigcup_i \partial R_i(\vartheta^n\omega)\right) \cup F^n_{\vartheta^{-n}\omega}\left(\bigcup_i \partial R_i(\vartheta^{-n}\omega)\right)\right)$$

which satisfies

$$F_\omega \partial_{\mathcal{R}}(\omega) \subset \partial_{\mathcal{R}}(\vartheta\omega).$$

Set $D(\omega) = \pi^{-1}_\omega \partial_{\mathcal{R}}(\omega)$; then

$$\sigma D(\omega) \subset D(\vartheta\omega). \tag{4.2.5}$$

It is not clear how to introduce general verifiable conditions on F which ensure topological mixing of the random subshift of finite type σ constructed above. Still, for some specific models, in particular, for Examples 4.2.4 and 4.2.3 it is possible to construct random

Markov partitions such that the corresponding random subshift of finite type σ will be topologically mixing. In the case of Example 4.2.3 it suffices to assume that the unperturbed diffeomorphism f is topologically mixing on a basic hyperbolic set Λ and that the perturbation neighborhood $\mathcal{U}(f)$ is C^1 small enough. Let $f = f(\omega, x)$ be a function on $\Omega \times M$ satisfying (4.2.3) when $x, y \in \Lambda(\omega)$ with K_f satisfying (4.2.3). Then we can define a function $\tilde{f}(\omega, \xi) = f(\omega, \pi_\omega \xi)$ which will satisfy conditions of Theorem 4.1.1. Hence there exists a unique probability measure (equilibrium state) such that $\pi_\sigma(\tilde{f}) = h_{\tilde{\mu}}(\sigma) + \int \tilde{f}\, d\tilde{\mu}$. Since $\tilde{\mu}$ is ergodic and positive on cylinder sets we can derive from (4.2.5) similarly to [36] that $\int \tilde{\mu}_\omega(D(\omega))\, d\mathbb{P}(\omega) = 0$. This enables us to define disintegrations $\mu_\omega = \pi_\omega \tilde{\mu}_\omega$ of a Θ-invariant measure μ satisfying (4.1.9). In particular, taking

$$f(\omega, x) = -\ln |\det D_x F_\omega|_{\Gamma_x^u(\omega)}|$$

and assuming that the $C^{1+\gamma}$, $\gamma > 0$, norm of F belongs to $\mathbb{L}^1(\Omega, \mathbb{P})$ we obtain random Sinai–Ruelle–Bowen (SRB) measures μ whose disintegrations μ_ω have absolutely continuous conditional measures on unstable manifolds (see Theorem 4.3 in [68] and [115]). Suppose that $\mathbb{P}$-a.s. F_ω are C^2 perturbations (not necessarily very small) of one (deterministic) diffeomorphism F_0 having a hyperbolic set Λ so that F_ω, F_0 and F_ω^{-1}, F_0^{-1} preserve corresponding expanding nonrandom cones. Then random SRB measures can be constructed more directly without random Markov partitions employing the method exhibited in [13].

An attempt to deal with continuous time hyperbolic RDS, i.e., with stochastic flows, has been made in Section 5 of [68]. Still, at present there are neither final definitions nor substantial results concerning this situation. In the continuous time case RDS can be generated either by stochastic or by random differential equations (see [7, Chapter 2]). There are no examples of (spatially uniform) hyperbolic stochastic flows given by nondegenerate stochastic differential equations. It is easier to generate what could be called hyperbolic stochastic flows by random differential equations. Even then there is an important difference from the deterministic case, namely, there is no flow direction, i.e., the one-dimensional invariant subbundle tangent to orbits. Observe that the structural stability of hyperbolic flows theorem does not help much in constructing easy workable examples as in the discrete time case since perturbations of hyperbolic flows conjugate to it with time change and we still do not have much control in the random one-dimensional "flow" direction. Random hyperbolic type splittings yield random stable and unstable manifolds (see [68]) but it is not clear how to proceed any further. So we restrict ourselves by considering the following model example from [68]. Suppose that vector fields $B_1, \dots, B_k$ are taken from a small C^2 neighborhood of one vector field B which generates an Anosov flow. Let $\nu_t = \nu_t(\omega)$ be a continuous time Markov chain with the finite state space $\{1, \dots, k\}$ and set $B(\vartheta^t \omega, x) = B_{\nu_t(\omega)}(x)$. This defines a cocycle $F(t, \omega)$ for $\omega \in \Omega$ given by

$$\frac{dF(t, \omega)x}{dt} = B\big(\vartheta^t \omega, F(t, \omega)x\big).$$

One can consider, for instance, a "random geodesic flow" when several close metrics of negative curvature are switched at random times (when ν_t moves from state to state). Clearly, some expanding and contracting cones will be preserved for forward and backward actions, respectively, and it is not difficult to see that this model admits an invariant

random hyperbolic splitting. The sum of the dimensions of stable and unstable subbundles is one less than the dimension of the manifold as in the deterministic case but now there is no good control of the action on the additional one-dimensional subbundle. Though we cannot provide thermodynamic formalism constructions of equilibrium states for the above model in its full generality one very particular case of it can be dealt with. Namely, assume that $B_j = q_j B$ where q_j are positive functions on M and B is a (nonrandom) C^2 vector field generating on M a transitive Anosov flow f^t. Then $F(t,\omega)$ is obtained from f^t by a random time change and both flows have the same orbits. Using a Markov partition for f^t we can represent $F(t,\omega)$ as a suspension over a nonrandom subshift of finite type with a random ceiling function (see [98]). A more general random time change will lead to a suspension over a random subshift of finite type if we require that the corresponding random ceiling function is uniformly bounded away from zero and infinity in order to satisfy the assumptions from [98].

4.3. *Markov chains with random transitions*

In this section we will discuss Markov chains with random transition probabilities which are both ideologically close to RDS and also emerge directly in the study of random Markov subshifts of finite type described at the end of Section 4.1. Our exposition will follow mainly [94] though we will mention some other related papers, as well. Let $(\Omega, \mathcal{F}, \mathbb{P})$ and ϑ be as before, Y be a compact space and $\mathcal{G}$ be a measurable subset of $\Omega \times Y$ with compact fibers $\mathcal{G}_\omega = \{x\colon\ (\omega, x) \in \mathcal{G}\}$. Suppose that for each $\omega \in \Omega$ and $x \in \mathcal{G}_\omega$ we are given a Borel probability measure $P^\omega(x,\cdot)$ on $\mathcal{G}_{\vartheta\omega}$ Borel measurably depending on the pair (ω, x). By the Kolmogorov extension theorem, for each $x \in \mathcal{G}_\omega$, there is a unique probability measure $\mathbf{P}_x^\omega$ on the product space $\mathbf{G}_\omega = \mathcal{G}_\omega \times \mathcal{G}_{\vartheta\omega} \times \mathcal{G}_{\vartheta^2\omega} \times \cdots$ such that under $\mathbf{P}_x^\omega$ the coordinate maps $\{Z_n^\omega,\ n \geqslant 0\}$ become a time inhomogeneous Markov chain starting at x and evolving according to $\{P^{\vartheta^n\omega}\}$, i.e., $Z_{n+1}^\omega \in \Gamma \subset \mathcal{G}_{\vartheta^{n+1}\omega}$ with probability $P^{\vartheta^n\omega}(y,\Gamma)$ provided $Z_n^\omega = y \in \mathcal{G}_{\vartheta^n\omega}$. In this way, one can think of $(\Omega, \mathcal{F}, \mathbb{P}, \vartheta)$ as a random stationary environment for the Markov chains Z_n^ω. Note, that in the physical literature parameters considered for each specific state of the environment (ω-wise) are often called quenched (which we call here fiber or relative) and parameters averaged in the environment or parameters related to the Markov chains with the averaged transition probabilities $\int P^\omega(x,\cdot)\,d\mathbb{P}(\omega)$ are called annealed. In the case when $\mathcal{G}_\omega \equiv Y$ is a countable set the ergodic theory for such Markov chains was studied in a series of papers (see references in [94]) with rather different motivations. Observe that we have changed the notations here with respect to other sections denoting random spaces by $\mathcal{G}$ and not by $\mathcal{E}$ as before by the reason that in an application of such Markov chains to random Markov subshifts of finite type discussed below these subshifts act on the product space $\mathcal{E}_\omega = \mathbf{G}_\omega$ and not on $\mathcal{G}_\omega$.

For $x \in \mathcal{G}_\omega$ and a Borel $\Gamma \subset \mathcal{G}_{\vartheta^n\omega}$, set

$$P^\omega(n,x,\Gamma) = \int \dots \int P^\omega(x,dy_1)P^{\vartheta\omega}(y_1,dy_2)\dots P^{\vartheta^{n-1}\omega}(y_{n-1},\Gamma).$$

Thus, $P^\omega(n,x,\cdot)$ is the n-step transition probability of the Markov chain Z_n^ω. Our main assumption here is that there exist random variables $N = N_\omega \in \mathbb{Z}_+ = \{1,2,\dots\}$ and $\gamma_\omega > 0$

and a family of measures m_ω from the space $\mathcal{P}(\mathcal{G}_\omega)$ of probability measures on $\mathcal{G}_\omega$ such that for $\mathbb{P}$-a.a. $\omega \in \Omega$, any $x \in \mathcal{G}_{\vartheta^{-N}\omega}$, and each Borel $\Gamma \subset \mathcal{G}_\omega$,

$$P^{\vartheta^{-N}\omega}(N, x, \Gamma) \geqslant \gamma_\omega m_\omega(\Gamma). \tag{4.3.1}$$

Clearly, (4.3.1) is a randomized version of the classical Doeblin condition and it turns out that it implies also a randomized version of Doeblin's conclusion.

THEOREM 4.3.1. *Suppose that* (4.3.1) *holds true but no topological assumptions on Y and $\mathcal{G}_\omega$ are made. Then for $\mathbb{P}$-a.a.ω there exists a unique family of probability measures μ_ω on $\mathcal{G}_\omega$ satisfying*

$$\mu_\omega P^\omega = \mu_{\vartheta\omega} \tag{4.3.2}$$

measurably depending on ω and such that for some $\kappa = \kappa_\omega \in (0, 1)$, $C = C(\omega) \in (0, \infty)$, any $x \in \mathcal{G}_{\vartheta^{-n}\omega}$, and a Borel set $\Gamma \subset \mathcal{G}_\omega$,

$$\left|P^{\vartheta^{-n}\omega}(n, x, \Gamma) - \mu_\omega(\Gamma)\right| \leqslant C(\omega)(1 - \kappa_\omega)^n. \tag{4.3.3}$$

Furthermore, if in addition to (4.3.1) *we have also the upper bound*

$$P^{\vartheta^{-N}\omega}(N, x, \Gamma) \leqslant \gamma_\omega^{-1} m_\omega(\Gamma), \quad N = N_\omega, \tag{4.3.4}$$

then for any $n \geqslant N_\omega$ there exist densities

$$\begin{aligned} p^{\vartheta^{-n}\omega}(n, x, y) &= \frac{dP^{\vartheta^{-n}\omega}(n, x, \cdot)}{m_\omega}(y), \\ p^\omega(y) &= \frac{d\mu_\omega}{dm_\omega}(y) \quad \forall x \in \mathcal{G}_{\vartheta^{-n}\omega},\ \forall y \in \mathcal{G}_\omega, \end{aligned} \tag{4.3.5}$$

and we have

$$\sup_{x,y}\left|p^{\vartheta^{-n}\omega}(n, x, y) - p^\omega(y)\right| \leqslant C(\omega)(1 - \kappa_\omega)^n. \tag{4.3.6}$$

The above setup includes finite Markov chains with random transition probabilities where $\mathcal{G}_\omega = \{1, \ldots, \ell(\omega)\}$ with a random variable $\ell(\omega) \in \mathbb{Z}_+$ satisfying $\int \ell\, d\mathbb{P} < \infty$. The transition probabilities are given here by $\ell(\omega) \times \ell(\vartheta\omega)$ matrices $P^\omega = (P_{ij}^\omega)$, $i = 1, \ldots, \ell(\omega)$, $j = 1, \ldots, \ell(\vartheta\omega)$. Consider the products $P^\omega(n) = P^\omega P^{\vartheta\omega} \ldots P^{\vartheta^{n-1}\omega}$ and assume that there exists a random variable $N = N_\omega$ such that for P-a.a. ω the matrix $P^{\vartheta^{-N}\omega}(N)$ has positive entries. Then (4.3.1) holds true where m_ω is the uniform

measure on $\mathcal{G}_\omega$. By Theorem 4.3.1 there exist unique probability vectors $p^\omega = (p_i^\omega)$, $i = 1, \dots, \ell(\omega)$, such that

$$p^\omega P^\omega = p^{\vartheta\omega}. \tag{4.3.7}$$

Next consider the space of sequences $\mathcal{E}_\omega = \mathcal{G}_\omega \times \mathcal{G}_{\vartheta\omega} \times \cdots = \{x = (x_0, x_1, \dots),\ x_i \in \mathcal{G}_{\vartheta^i\omega}\}$ together with the left shift $F_\omega : \mathcal{E}_\omega \to \mathcal{E}_{\vartheta\omega}$. Let $C^\omega_{\xi_0,\dots,\xi_n} = \{x \in \mathcal{E}_\omega\colon\ x_i = \xi_i\ \forall i = 0, 1, \dots, n\}$ be a cylinder set. Define the probability measures $\nu_\omega^{p,P}$ on $\mathcal{E}_\omega$ setting

$$\nu_\omega^{p,P}(C^\omega_{\xi_0,\dots,\xi_n}) = p^\omega_{\xi_0} P^\omega_{\xi_0\xi_1} \dots P^{\vartheta^{n-1}\omega}_{\xi_{n-1}\xi_n}$$

which is legitimate in view of Kolmogorov's extension theorem. By (4.3.7) it follows easily that

$$F_\omega \nu_\omega^{p,P} = \nu_{\vartheta\omega}^{p,P}. \tag{4.3.8}$$

The measure ν having disintegrations $\nu_\omega^{p,P}$ is the invariant measure of the RDS F determined by F_ω above and we arrive again at a Markov subshift of finite type considered at the end of Section 4.1. Theorem 4.3.1 implies in this case exponentially fast ω-wise mixing of disintegrations and the ergodicity (mixing) of ν whenever $\mathbb{P}$ is ergodic (mixing) with respect to ϑ (see Proposition 2.2 in [94]) though we know these facts already for general Gibbs measures by Section 4.1. Next, let $\mathcal{G}$ be as above. We exhibit a random Perron–Frobenius type theorem for random operators which resembles Theorem 4.1.2 considered in the case of random subshifts of finite type. Let R^ω, $\omega \in \Omega$, be a family of bounded, nonnegative operators mapping a bounded Borel function g on $\mathcal{G}_{\vartheta\omega}$ to a function $R^\omega g$ on $\mathcal{G}_\omega$ by the formula

$$R^\omega g(x) = \int R^\omega(x, dy) g(y). \tag{4.3.9}$$

Assume that there exist random variables $c_\omega > 0$ and $N = N_\omega \in \mathbb{Z}_+$ and a family of probability measures $m_\omega \in \mathcal{P}(\mathcal{G}_\omega)$ such that for $\mathbb{P}$-a.a. ω, any $x \in \mathcal{G}_\omega$, and each Borel $\Gamma \subset \mathcal{G}_\omega$,

$$R^{\vartheta^{-N}\omega}(N, x, \Gamma) \geqslant c_\omega m_\omega(\Gamma) \tag{4.3.10}$$

and

$$\operatorname{supp} R^{\vartheta^{-N}\omega}(N, x, \cdot) = \operatorname{supp} m_\omega. \tag{4.3.11}$$

Here $R^\omega(n, x, \Gamma) = \int \dots \int R^\omega(x, dy_1) \dots R^{\vartheta^{n-1}\omega}(y_{n-1}, \Gamma)$. When in addition to (4.3.10) we assume that

$$R^{\vartheta^{-N}\omega}(N, x, \Gamma) \leqslant c_\omega^{-1} m_\omega(\Gamma) \tag{4.3.12}$$

then (4.3.11) is automatically satisfied.

THEOREM 4.3.2. *There exist uniquely defined numbers* $\lambda(\omega) > 0$, *positive functions* $h(\omega)$ *on* $\mathcal{G}_\omega$, *and measures* $\nu_\omega \in \mathcal{P}(\mathcal{G}_\omega)$ (*dependent measurably on* ω *if* $\mathcal{F}$ *is countably generated*) *such that for any* $x \in \mathcal{G}_\omega$ *and* $\Gamma \subset \mathcal{G}_{\vartheta\omega}$,

$$\int d\nu_\omega(x) R^\omega(x, \Gamma) = \lambda(\omega)\nu_{\vartheta\omega}(\Gamma) \tag{4.3.13}$$

and

$$\int h(\vartheta\omega, y) R^\omega(x, dy) = \lambda(\omega)h(\omega, x). \tag{4.3.14}$$

Furthermore, for some $C(\omega) > 0$, $\kappa = \kappa_\omega \in (0, 1)$, *any* $x \in \mathcal{G}_{\vartheta^{-n}\omega}$, *and a Borel* $\Gamma \subset \mathcal{G}_\omega$,

$$\begin{aligned} &\left|\left(\lambda(\vartheta^{-n}\omega)\dots\lambda\left(\vartheta^{-1}\omega\right)\right)^{-1}\left(h(\vartheta^{-n}\omega, x)\right)^{-1} R^{\vartheta^{-n}\omega}(n, x, \Gamma) - \nu_\omega(\Gamma)\right| \\ &\quad \leqslant C(\omega)(1-\kappa)^n. \end{aligned} \tag{4.3.15}$$

If also the upper bound (4.3.12) *holds true then for any* $n \geqslant N_\omega$ *there are densities*

$$\begin{aligned} &r^{\vartheta^{-n}\omega}(n, x, y) = \frac{dR^{\vartheta^{-n}\omega}(n, x, \cdot)}{dm_\omega}(y), \\ &r^\omega(y) = \frac{d\nu_\omega}{dm_\omega}(y) \quad \forall x \in \mathcal{G}_{\vartheta^{-n}\omega},\ y \in \mathcal{G}_\omega, \end{aligned} \tag{4.3.16}$$

and

$$\begin{aligned} &\sup_{x,y}\left|\left(\lambda(\vartheta^{-(n-1)}\omega)\dots\lambda(\omega)\right)^{-1}\left(h(\vartheta^{-n}\omega, x)\right)^{-1} r^{\vartheta^{-n}\omega}(n, x, y)h(\omega, y) - r^\omega(y)\right| \\ &\quad \leqslant C(\omega)(1-\kappa)^n. \end{aligned} \tag{4.3.17}$$

The above result enables us to obtain a Donsker–Varadhan type formula in a random environment which in certain sense is similar to the relative variational principle for random transformations. Namely, for any measurable in ω family of continuous functions $\varphi = \{\varphi_\omega\}$ on $\mathcal{G}_\omega$ let $R^\omega = R^\omega_\varphi$ be the operator given by the formula

$$R^\omega g(x) = \int g(y) e^{\varphi_{\vartheta\omega}(y)} P^\omega(x, dy); \tag{4.3.18}$$

and let $\lambda(\omega)_\varphi$ be the corresponding random variable appearing in Theorem 4.3.2. In addition to (4.3.10), assume that

$$\int \sup_{x \in \mathcal{G}_\omega} \left|\varphi_\omega(x)\right| d\mathbb{P}(\omega) < \infty. \tag{4.3.19}$$

Then $\ln \lambda(\omega)_\varphi$ is integrable and the following variational formula holds true:

$$Q(\varphi) \stackrel{\text{def}}{=} \int \log \lambda(\omega)_\varphi \, d\mathbb{P}(\omega) = \sup_\eta \left(\int \varphi \, d\eta - I(\eta) \right), \tag{4.3.20}$$

where the supremum is taken over the space $\mathcal{P}_{\mathbb{P}}(\Omega \times Y)$ of probability measures on $\Omega \times Y$ with the disintegration $d\eta(\omega, x) = d\eta_\omega(x)\, d\mathbb{P}(\omega)$ and

$$I(\eta) \equiv -\inf_u \int \log \left(\frac{P^\omega u(\vartheta\omega)}{u(\omega)} \right) d\eta_\omega \, d\mathbb{P}(\omega), \tag{4.3.21}$$

the infimum being taken over measurable families $u = \{u(\omega)\}$ of positive continuous functions on $\mathcal{G}_\omega$ such that $\int \sup_x |\log u(\omega)(x)| \, d\mathbb{P}(\omega) < \infty$. It turns out that for a wide class of families $\varphi = \{\varphi(\omega)\}$ of continuous functions $\varphi(\omega)$ on $\mathcal{G}_\omega$ there exists a unique $\mu = \mu^\varphi \in \mathcal{P}_{\mathbb{P}}(\Omega \times Y)$ such that

$$Q(\varphi) = \int \varphi \, d\mu - I(\mu). \tag{4.3.22}$$

If each $\mathcal{G}_\omega$ is a finite set $\{1, \dots, \ell(\omega)\}$ then it is natural to take m_ω to be the uniform measure on $\mathcal{G}_\omega$ and Theorem 4.3.2 yields a Perron–Frobenius type theorem for random positive matrices. Observe that the first construction of Gibbs measures for Anosov diffeomorphisms in [156] used the classical Perron–Frobenius theorem for random matrices but later the Ruelle–Perron–Frobenius theorem enabled a more direct construction as described in [36]. In the random case we used from the beginning the latter approach but Theorem 4.3.2 reveals a close connection between thermodynamic formalism type constructions for random transformations and related results for Markov chains with random transitions and random positive operators.

The construction of the triple $\lambda(\omega), h(\omega), \nu_\omega$ in Theorem 4.3.2 is similar to Theorem 4.1.2 but then we use $R^\omega, \lambda(\omega)$ and $h(\omega)$ in order to construct a random Markov operator satisfying conditions of Theorem 4.3.1 and via the latter result we derive remaining assertions of Theorem 4.3.2. Once Markov chains with random transitions are defined it is natural to proceed to other standard problems of the theory of Markov chains such as their asymptotic behavior which is related to existence of bounded and positive "harmonic" functions. It turns out that random harmonic functions with respect to Markov chains with random transitions can be defined, as well (see [101]). It is especially interesting to study such questions for random walks on groups whose increments are independent but have stationarily changing distributions (see [101] and [73]). In particular, existence of random bounded harmonic functions can be characterized via certain entropy characteristics. If the group in question is the group of invertible matrices we end up with products of independent matrices whose distributions form a stationary process and we can obtain results concerning, for instance, simplicity of the Lyapunov spectrum for such products and to develop a theory which extends many of well-known results concerning products of independent identically distributed random matrices (see [101]).

4.4. *Limit theorems*

In this section we will discuss the fiber (ω-wise) central limit theorem and the law of iterated logarithm type results for random transformations and for Markov chains with random transitions following [98], as well, as some ω-wise large deviations results. For the proof of Theorems 4.4.1–4.4.5 below we refer the reader to [98]. We employ here the standard notations and setup as in previous sections and we assume that $\mathbb{P}$ is ergodic with respect to ϑ. The setup includes here also a Θ-invariant measure $\mu \in \mathcal{P}_{\mathbb{P}}(\mathcal{E})$ with disintegrations $\{\mu_\omega\}$ and a measurable family of σ-algebras $\mathcal{F}^\omega_{m,n}$, $n \geqslant m$, $\omega \in \Omega$, of sets from $\mathcal{E}_\omega$ such that

$$\mathcal{F}^\omega_{m,n} \subset \mathcal{F}^\omega_{m',n'} \quad \text{if } m' \leqslant m \text{and } n' \geqslant n \text{ and } F_\omega^{-1}\mathcal{F}^{\vartheta\omega}_{m,n} = \mathcal{F}^\omega_{m+1,n+1}. \tag{4.4.1}$$

The uniform mixing (ϕ-mixing) coefficient is defined by

$$\phi^\omega_{i,j} = \sup_{A\in\mathcal{F}^\omega_{0,i},\, \mu_\omega(A)\neq 0,\, B\in\mathcal{F}^\omega_{j,\infty}} \left| \frac{\mu_\omega(A\cap B)}{\mu_\omega(A)} - \mu_\omega(B)\right|, \quad j > i. \tag{4.4.2}$$

Let $\varphi = \varphi(\omega, x) = \varphi_\omega(x)$ be a measurable function on $\mathcal{E}$ so that φ_ω as a function on $\mathcal{E}_\omega$ is $\mathcal{F}^\omega_{0,\infty}$-measurable, and so in view of (4.4.1), $\varphi_{\vartheta^i\omega}\circ F^i_\omega$ is $\mathcal{F}^\omega_{i,\infty}$-measurable as a function on $\mathcal{E}_\omega$. The setup includes also a measurable set $Q \subset \Omega$ with $\mathbb{P}(Q) > 0$ and the corresponding sequence of hitting times

$$k_{i+1}(\omega) = \min\bigl\{k > k_i(\omega)\colon\ \vartheta^k\omega \in Q\bigr\} \quad \text{with } k_0 \equiv 0. \tag{4.4.3}$$

Set

$$\psi_\omega = \varphi_\omega - E_{\mu_\omega}\varphi_\omega, \qquad \Psi(\omega, x) = \Psi_\omega(x) = \sum_{i=0}^{k_1(\omega)-1} \psi\circ\Theta^i(\omega, x),$$

$$c(\omega) = \bigl(E_{\mu_\omega}|\psi_\omega|^2\bigr)^{1/2}, \qquad C(\omega) = \bigl(E_{\mu_\omega}|\Psi_\omega|^2\bigr)^{1/2},$$

$$d_n(\omega) = \bigl(E_{\mu_\omega}\bigl(\psi_\omega - E_{\mu_\omega}(\psi_\omega|\mathcal{F}^\omega_{0,n})\bigr)^2\bigr)^{1/2},$$

and

$$D_n(\omega) = \bigl(E_{\mu_\omega}\bigl(\Psi_\omega - E_{\mu_\omega}(\Psi_\omega|\mathcal{F}^\omega_{0,n})\bigr)^2\bigr)^{1/2},$$

where E_ν always denotes the expectation (i.e., the integral) with respect to a probability measure ν and $E_\nu(\cdot|\cdot)$ is the corresponding conditional expectation. Observe that

$$C(\omega) \leqslant \sum_{i=0}^{k_1(\omega)-1} c\bigl(\vartheta^i\omega\bigr) \quad \text{and} \quad D_n(\omega) \leqslant \sum_{i=0}^{k_1(\omega)-1} d_{n-i}\bigl(\vartheta^i\omega\bigr). \tag{4.4.4}$$

Set $\tau = \vartheta^{k_1(\omega)}$, $\Phi_\omega = F_\omega^{k_1(\omega)}$, and $T(\omega, x) = (\tau\omega, \Phi_\omega x)$. Since we assume also that $\mathbb{P}$ is ergodic, it is well known (see, for instance, [48]) that τ is an ergodic measure preserving transformation on the space (Q, P_Q) where $P_Q(A) = \frac{\mathbb{P}(A\cap Q)}{\mathbb{P}(Q)}$. Denote by $\mathcal{E}_Q$ the restriction of $\mathcal{E}$ to $Q \times X$ and by μ_Q the normalized restriction of μ to $\mathcal{E}_Q$, i.e., $d\mu_Q(\omega, x) = d\mu_\omega(x)\, dP_Q(\omega)$. It follows that μ_Q is invariant under the action of T.

THEOREM 4.4.1. *Let* $\phi_j = \sup_{\omega, i\geqslant 1} \phi^\omega_{k_i(\omega), k_{i+j}(\omega)}$ *and* $\beta_j = (E_{P_Q} D_{k_j}^2)^{1/2}$ *and suppose that*

$$\sum_{j=1}^{\infty} \phi_j^{1/2} < \infty, \tag{4.4.5}$$

$$\sum_{j=1}^{\infty} \beta_j < \infty, \tag{4.4.6}$$

and

$$E_{P_Q}\left(\sum_{i=0}^{k_1(\omega)-1} c \circ \vartheta^i\right)^2 < \infty. \tag{4.4.7}$$

Then $\mathbb{P}$*-a.s.,*

$$\begin{aligned}\sigma^2 &= \lim_{n\to\infty} \frac{1}{n} E_{\mu_\omega}\left(\sum_{j=0}^{n-1} \psi_{\vartheta^j\omega} \circ F_\omega^j\right)^2 \\ &= \mathbb{P}(Q)\left(E_{\mu_Q}\Psi^2 + \sum_{l=1}^{\infty} E_{\mu_Q}\big(\Psi\big(\Psi \circ T^l\big)\big)\right)\end{aligned} \tag{4.4.8}$$

and the series in the right-hand side of (4.4.8) *converges. Furthermore,* $\mathbb{P}$*-a.s. for any number* a*,*

$$\begin{aligned}&\lim_{n\to\infty} \mu_\omega\left\{x \in \mathcal{E}_\omega\colon \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} \psi \circ \Theta^i(\omega, x) \leqslant a\right\} \\ &\quad = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2\sigma^2}}\, dx,\end{aligned} \tag{4.4.9}$$

i.e., for $\mathbb{P}$*-a.a.* ω *the* μ_ω*-distribution of* $n^{-1/2}\sum_{i=0}^{n-1}(\psi \circ \Theta^i)_\omega$ *converges to the normal distribution with zero mean and the variance* σ^2 *which in case* $\sigma = 0$ *is understood as the unit mass at* 0*. Finally,* $\sigma = 0$ *if and only if there exists a function* η *on* $\mathcal{E}_Q$ *from* $L^2(\mathcal{E}_Q, \mu_Q)$ *such that* μ_Q*-a.s.,*

$$\Psi \circ T = \eta \circ T - \eta. \tag{4.4.10}$$

Furthermore, assume that $\sigma > 0$ *and set* $\rho(t) = (2t \log\log t)^{1/2}$ *and for* $k = 0, 1, \dots, n-1$,

$$\eta_n(t) = \sigma^{-1} n^{-1/2} \left(\sum_{j=0}^{k-1} \psi \circ \Theta^j + (nt - k)\psi \circ \Theta^k \right) \quad \text{for } t \in \left[\frac{k}{n}, \frac{k+1}{n} \right),$$

$$\zeta_n(t) = \left(\rho\left(\sigma^2 n\right) \right)^{-1} \left(\sum_{j=0}^{k-1} \psi \circ \Theta^j + (nt - k)\psi \circ \Theta^k \right) \quad \text{for } t \in \left[\frac{k}{n}, \frac{k+1}{n} \right).$$

Then μ_Q*-a.s.* η_n *converges in distribution as* $n \to \infty$ *to the standard Wiener process on the time interval* $[0, 1]$ (*the invariance principle in the central limit theorem*) *and the invariance principle for the law of iterated logarithm* (*LIL*) *holds true, as well. Namely,* μ_Q*-a.s. the sequence of functions* $\{\zeta_n(\cdot),\ n \geqslant 3/\sigma^2\}$ *is relatively compact in the space* $C[0, 1]$ (*of continuous functions on* $[0, 1]$ *considered with the supremum norm*) *and the set of its limit points as* $n \to \infty$ *coincides with the set* K *of absolutely continuous* $x \in C[0, 1]$ *with* $\int_0^1 (\dot{x}(t))^2\, dt \leqslant 1$.

This result can be proved via the classical characteristic functions technique though another technique based on nonstationary backwards martingale differences approximation and martingale limit theorems which also works here is often more convenient since there is no need in an explicit uniform mixing condition with respect to a certain family of σ-algebras as in Theorem 4.4.1. The corresponding setup includes the same objects as above, but now in place of a family of σ-algebras $\mathcal{F}_{m,n}^{\omega}$ we consider another family of σ-algebras $\mathcal{T}_l^{\omega} = (F_{\omega}^l)^{-1} \mathcal{T}_0^{\vartheta^l \omega}$ where $\mathcal{T}_0^{\omega}$ for each ω is the restriction of the σ-algebra $\mathcal{F}$ to $\mathcal{E}_{\omega}$. Then $\mathcal{T}_l^{\omega}$, $l = 0, 1, \dots$, is a nonincreasing sequence of σ-algebras on $\mathcal{E}_{\omega}$. Let $u_{\omega} : \mathbb{L}^2(\mathcal{E}_{\vartheta\omega}, \mu_{\vartheta\omega}) \to \mathbb{L}^2(\mathcal{E}_{\omega}, \mu_{\omega})$ be an isometry operator acting by the formula $u_{\omega}\varphi(x) = \varphi(F_{\omega}x)$ and let $u_{\omega}^* : \mathbb{L}^2(\mathcal{E}_{\omega}, \mu_{\omega}) \to \mathbb{L}^2(\mathcal{E}_{\vartheta\omega}, \mu_{\vartheta\omega})$ be its conjugate defined by $\int \varphi(u_{\omega}^* \tilde{\varphi})\, d\mu_{\vartheta\omega} = \int (u_{\omega}\varphi)\tilde{\varphi}\, d\mu_{\omega}$ for any $\varphi \in \mathbb{L}^2(\mathcal{E}_{\vartheta\omega}, \mu_{\vartheta\omega})$ and $\tilde{\varphi} \in \mathbb{L}^2(\mathcal{E}_{\omega}, \mu_{\omega})$. It is easy to see that $u_{\omega}u_{\omega}^* : \mathbb{L}^2(\mathcal{E}_{\omega}, \mu_{\omega}) \to \mathbb{L}^2(\mathcal{E}_{\omega}, \mu_{\omega})$ is the orthogonal projection to $u_{\omega}\mathbb{L}^2(\mathcal{E}_{\vartheta\omega}, \mu_{\vartheta\omega})$ and the last set is exactly the set of $\mathcal{T}_1^{\omega}$-measurable functions in $\mathbb{L}^2(\mathcal{E}_{\omega}, \mu_{\omega})$. Introduce also the operator $U_{\omega} = u_{\omega}^{k_1(\omega)}$, where $u_{\omega}^n = u_{\omega} \circ u_{\vartheta\omega} \circ \cdots \circ u_{\vartheta^{n-1}\omega}$, and let U_{ω}^* be its conjugate.

THEOREM 4.4.2. *Assume that* (4.4.7) *holds true and that the following two conditions are satisfied*:

$$E_{P_Q} \sum_{n=0}^{\infty} \left| E_{\mu_{\omega}} (\Psi_{\omega} (\Psi \circ T^n)_{\omega}) \right| < \infty \tag{4.4.11}$$

and

$$E_{P_Q} E_{\mu_{\omega}} \left(\sum_{n=0}^{\infty} |U_{\tau^{-n}\omega}^{*n} \Psi_{\tau^{-n}\omega}| \right)^2 < \infty, \tag{4.4.12}$$

where $U^{*n}_\omega = U^*_{\tau^{n-1}\omega} \circ \cdots \circ U^*_{\tau\omega} U^*_\omega$. *Then* $\mathbb{P}$*-a.s.* (4.4.8) *and* (4.4.9) *hold true and the criterion* (4.4.10) *for* $\sigma = 0$ *remains valid, as well. Moreover, assuming that* $\sigma > 0$ *the invariance principles stated in Theorem* 4.4.1 *hold true.*

Observe that a similar result has been proved also in [143]. Next, we discuss specific examples of processes in random environments and random transformations for which either mixing conditions of Theorem 4.4.1 or convergence conditions of Theorem 4.4.2 needed for martingale differences approximations hold true. First, we consider the case when μ_ω's are Markov measures observing that the random Doeblin condition (4.3.1) is suffice for the uniform mixing condition of Theorem 4.4.1 to hold true. Let Z^ω_n be a Markov chain in a random stationary environment introduced at the beginning of Section 4.3 and suppose that probability measures μ_ω on $\mathcal{E}_\omega = \mathcal{G}_\omega \times \mathcal{G}_{\vartheta\omega} \times \cdots$ are determined by

$$\begin{aligned}
&\mu_\omega\big\{x \in \mathcal{E}_\omega\colon\ Z^\omega_1(x) \in \Gamma_1,\ Z^\omega_2(x) \in \Gamma_2,\ \ldots,\ Z^\omega_n(x) \in \Gamma_n\big\} \\
&\quad = \int_{\mathcal{G}_\omega} \eta_\omega(dx_0) \int_{\Gamma_1} \ldots \int_{\Gamma_{n-1}} P^\omega(x_0, dx_1) \ldots \\
&\qquad \times P^{\vartheta^{n-2}\omega}(x_{n-2}, dx_{n-1}) P^{\vartheta^{n-1}\omega}(x_{n-1}, \Gamma_n),
\end{aligned} \tag{4.4.13}$$

where Γ_i is a measurable subset of $\mathcal{G}_{\vartheta^i\omega}$ and measures $\eta_\omega \in \mathcal{P}(\mathcal{G}_\omega)$ satisfy $\eta_\omega P^\omega = \eta_{\vartheta\omega}$. Existence and uniqueness of such measures η_ω is ensured by Theorem 4.3.1 whenever the random Doeblin condition (4.3.1) holds true. Let σ-algebras $\mathcal{F}^\omega_{m,n}$ be generated by all sets of the form $\{x\colon\ Z^\omega_l(x) \in \Gamma\},\ l = m, m+1, \ldots, n$, for measurable $\Gamma \subset \mathcal{E}_{\vartheta^l\omega}$.

THEOREM 4.4.3. *Let* (4.3.1) *holds true with some* $N = N_\omega$, $\gamma_\omega > 0$. *Set* $Q = Q_L = \{\omega\colon\ \max(N_\omega, \gamma_\omega^{-1}) \leqslant L\}$ *for a sufficiently large* L *so that* $\mathbb{P}(Q) > 0$ *and suppose that* $k_i(\omega)$*'s are defined by* (4.4.3). *Then the condition* (4.4.5) *holds true. Thus if* φ *is a measurable function on* $\mathcal{E}$ *as in Theorem* 4.4.1 *satisfying the conditions* (4.4.6) *and* (4.4.7) *then* (4.4.8) *and* (4.4.9) *hold true, as well. If* $\varphi_\omega(x) = \varphi_\omega(Z^\omega_0(x))$ *is, in fact, a function on* Y (*and, as before,* $\psi_\omega = \varphi(\omega) - E_{\mu_\omega}\varphi(\omega)$) *then* (4.4.9) *can be written in the form*

$$\begin{aligned}
&\lim_{n\to\infty} \mu_\omega\left\{x \in \mathcal{E}_\omega\colon\ \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} \psi_{\vartheta^k\omega}\big(Z^\omega_k(x)\big) \leqslant a \,\middle|\, Z^\omega_0(x) = x_0\right\} \\
&\quad = \lim_{n\to\infty} \mu_\omega\left\{x \in \mathcal{E}_\omega\colon\ \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} \psi_{\vartheta^k\omega}\big(Z^\omega_k(x)\big) \leqslant a\right\} \\
&\quad = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{u^2}{2\sigma^2}}\, du
\end{aligned} \tag{4.4.14}$$

which is satisfied for any initial point $x_0 \in \mathcal{G}_\omega$. *A corresponding LIL described in Theorem* 2.1 *holds true for this case, as well, as the invariance principles for both CLT and LIL.*

Next we deal with one-sided random subshifts of finite type considered in Section 3.1.1. Let $\mathcal{F}^{\omega}_{m,n}$, $m \leqslant n < \infty$, be the finite σ-algebra generated by cylinder sets $C_{\alpha_m,\alpha_{m+1},\ldots,\alpha_n} = \{x = (x_0, x_1, \ldots)\colon x_i = \alpha_i \text{ for } i = m, m+1, \ldots, n\}$ and let $\mathcal{F}^{\omega}_{m,\infty}$ be the minimal σ-algebra containing $\bigcup_{n \geqslant m} \mathcal{F}^{\omega}_{m,n}$.

THEOREM 4.4.4. *Suppose that $\{F_\omega\}$ is a random topologically mixing subshift of finite type,* (4.1.1) *is satisfied, the σ-algebras $\mathcal{F}^{\omega}_{m,n}$ are defined as above, and a measurable family of probability measures μ_ω is constructed by a function f satisfying* (4.1.3)–(4.1.5) *so that* (4.1.14), (4.1.21) *and* (4.1.22) *hold true. Then one can choose a set Q in the form $Q = Q_L = \{\omega\colon L_\omega \leqslant L\}$ with $\mathbb{P}(Q) > 0$, for some random variable L_ω which can be constructed explicitly so that the condition* (4.4.5) *will be satisfied. Thus by Theorem* 4.4.1 *if* (4.4.6) *and* (4.4.7) *are satisfied then* (4.4.8) *and* (4.4.9) *hold true together with the characterization* (4.4.10) *of the case $\sigma = 0$. The corresponding LIL follows, as well, as the invariance principles for both CLT and LIL.*

This theorem had been applied in [115] in order to obtain an ω-wise central limit theorem as above for RDS described in Example 4.2.3. Next we will describe a similar result for random expanding in average transformations.

THEOREM 4.4.5. *Assume that F_ω, $\omega \in \Omega$, are C^2 endomorphisms of a compact connected C^2 Riemannian manifold M satisfying* (4.2.1) *and* (4.2.2). *Suppose that the measures μ_ω are constructed by Theorem* 4.2.1 *with a function f satisfying* (4.1.3), (4.1.5), *and* (4.2.3) *and that a function $\varphi = \varphi(\omega, x)$ (for which the CLT is going to be proved) is Hölder continuous in x, i.e., it also satisfies* (4.2.3) *with, say, the same exponent $\kappa > 0$ and a random variable $K_\varphi(\omega) > 0$ such that*

$$E_{P_Q}\left(\sum_{j=0}^{k_1(\omega)-1} \left(\|\varphi_{\vartheta^j\omega}\| + K_\varphi\left(\vartheta^j\omega\right)\right)\right)^2 < \infty, \tag{4.4.15}$$

with k_1 given by (4.4.3). *Then one can choose a set Q in the form $Q = Q_L = \{\omega\colon L_\omega \leqslant L\}$ with $\mathbb{P}(Q) > 0$ and a random variable L_ω which can be constructed explicitly so that the conditions* (4.4.11) *and* (4.4.12) *of Theorem* 4.4.2 *will be satisfied, and so* (4.4.8) *and* (4.4.9) *together with the characterization of the case $\sigma = 0$ hold true, as well, as the corresponding LIL.*

We will conclude this section with a description of large deviations results for random transformations. Consider the occupational measures

$$\zeta^n_\omega = \frac{1}{n}\sum_{k=0}^{n-1} \delta_{\vartheta^k\omega, Z^\omega_k} \in \mathcal{P}(\Omega \times Y) \tag{4.4.16}$$

(where δ_z denotes the unit mass at z) in the case of Markov chains with random transition probabilities and

$$\zeta_\omega^n = \frac{1}{n}\sum_{k=0}^{n-1} \delta_{(\vartheta^k\omega, F_\omega^k x)} \in \mathcal{P}(\Omega \times X) \tag{4.4.17}$$

in the case of random transformations. The following result has been proved in [94] employing general large deviations results from [87] together with Theorem 4.3.2 and the variational formula (4.3.20) considerations.

THEOREM 4.4.6. *Suppose that Markov chains with random transition probabilities Z_n^ω satisfy the random Doeblin condition* (4.3.1). *Then $\mathbb{P}$-a.s. for any $x \in Y$ and closed set $K \subset \mathcal{P}(\Omega \times Y)$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}_x^\omega\{\zeta_\omega^n \in K\} \leqslant -\inf_{\nu\in K} I_{\mathbb{P}}(\nu) \tag{4.4.18}$$

and for any open set $G \subset \mathcal{P}(\Omega \times Y)$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}_x^\omega\{\zeta_\omega^n \in G\} \geqslant -\inf_{\nu\in G} I_{\mathbb{P}}(\nu), \tag{4.4.19}$$

where $\mathbf{P}_x^\omega$ is the path space probability for the process Z_n^ω starting at x and $I_{\mathbb{P}}(\nu) = I(\nu)$ with $I(\eta)$ given by (4.3.21) *if $\eta \in \mathcal{P}(\Omega \times Y)$ and $I(\eta) = \infty$, otherwise.*

In the case of random expanding transformations and topologically mixing random subshifts of finite type the corresponding theorem has been established in [89] and [91] as a consequence of results from [87] together with Theorems 4.1.1 and 4.2.1 above. Large deviations for random transformations taken from a small C^2 neighbourhood of a diffeomorphism near a basic hyperbolic set (see Example 4.2.3) were derived in [121].

THEOREM 4.4.7. *Suppose that $\{F_\omega\}$ is either a random topologically mixing subshift of finite type as in Theorem* 4.4.4 *satisfying* (4.1.1) *or a random expanding in average C^2 endomorphism of a compact Riemannian manifold M satisfying* (4.2.1) *and* (4.2.2). *Let X be the product space $\bar{\mathbb{Z}}^+ \times \bar{\mathbb{Z}}^+ \times \cdots$ in the subshifts case and $X = M$ in the expanding transformations case. Suppose that the measures μ_ω are constructed by Theorems* 4.2.1 *or* 4.1.1, *respectively, for a function f satisfying* (4.1.3)–(4.1.5) *or* (4.1.3), (4.1.5), *and* (4.2.3), *respectively. Then $\mathbb{P}$-a.s. for closed set $K \subset \mathcal{P}(\Omega \times X)$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_\omega\{\zeta_\omega^n \in K\} \leqslant -\inf_{\nu\in K} I_f(\nu) \tag{4.4.20}$$

and for any open set $G \subset \mathcal{P}(\Omega \times X)$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_\omega\{\zeta_\omega^n \in G\} \geqslant -\inf_{\nu\in G} I_f(\nu), \tag{4.4.21}$$

where $I_f(\nu) = \pi_F(f) - \int f\,d\nu - h_\nu(F)$ *if* $\nu \in \mathcal{P}_{\mathbb{P}}(\Omega \times X)$ *is* Θ*-invariant and* $I_f(\nu) = \infty$*, otherwise. The relations* (4.4.20) *and* (4.4.21) *remain true if we replace* μ_ω *there by the normalized Riemannian volume* m *and take* $f(\omega, x) = -\ln|\operatorname{Jac} D_x F_\omega|$ *as in Theorem* 4.2.1.

4.5. *Random fractals*

In this section we apply thermodynamical formalism machinery for random transformations exhibited in Sections 4.1–4.3 to the study of fractal dimensions of random sets and random measures emerging naturally in various expansions. We start with random base expansions following [95]. The general setup includes as before an ergodic measure preserving transformation ϑ of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a countably generated σ-algebra $\mathcal{F}$ and a $\mathbb{Z}_+^d$-valued random vector $m(\omega) = (m_1(\omega), \dots, m_d(\omega))$, $\omega \in \Omega$, $m_i(\omega) \in \mathbb{Z}_+ = \{1, 2, \dots\}$ such that for all $i = 1, \dots, d$,

$$0 < \int \log m_i \, d\mathbb{P} = \int \log m_1 \, d\mathbb{P} < \infty. \tag{4.5.1}$$

For any $x = (x_1, \dots, x_d) \in [0,1)^d$ we can write

$$x = \sum_{j=0}^{\infty} x_j(\omega)\big(M\big(\vartheta^j\omega\big)\cdots M(\vartheta\omega)M(\omega)\big)^{-1}, \tag{4.5.2}$$

where $M(\omega)$ is the $d \times d$ diagonal matrix with the diagonal elements $m_1(\omega), \dots, m_d(\omega)$ and $x_j(\omega) = (x_{1j}(\omega), \dots, x_{dj}(\omega))$ are row vectors with $x_{ij}(\omega) \in \{0, 1, \dots, m_i(\vartheta^j\omega) - 1\}$. This expansion is always possible since for $x_i \in [0,1)$ we can set $x_{i0}(\omega) = [x_i m_i(\omega)]$, where $[\cdot]$ denotes the integer part, and

$$x_{ij}(\omega) = \left[m_i(\omega) m_i(\vartheta\omega) \dots m_i\big(\vartheta^j\omega\big)\left(x - \sum_{n=0}^{j-1} \frac{x_{in}(\omega)}{m_i(\omega)\dots m_i(\vartheta^n\omega)} \right) \right] \tag{4.5.3}$$

after $x_{i0}(\omega), \dots, x_{i,j-1}(\omega)$ have been already defined. Thus zero tails are permitted but the tails with $x_{ij}(\omega) = m_i(\vartheta^j\omega) - 1$ $\forall j \geqslant n$ are not. When $d = 1$ we arrive at an expansion with random digits of numbers x in the semi-open unit interval $[0,1)$ with respect to random stationarily changing bases $m(\omega), m(\vartheta\omega), \dots$, i.e.,

$$x = \sum_{n=0}^{\infty} \frac{x_n(\omega)}{m(\omega)\dots m(\vartheta^n\omega)} \tag{4.5.4}$$

(random base expansions) and when $m(\omega) = m$ is not random we are back to the usual expansion with respect to a fixed base m.

Define the transformations $F_\omega : [0,1)^d \to [0,1)^d$ and $\Theta : \Omega \times [0,1)^d \to \Omega \times [0,1)^d$ by the formulas

$$(F_\omega x)_i = m_i(\omega)x_i - \big[m_i(\omega)x_i\big]$$
$$\forall i = 1, \dots, d \text{ and } x = (x_1, \dots, x_d) \in [0,1)^d \tag{4.5.5}$$

and

$$\Theta(\omega, x) = (\vartheta\omega, F_\omega x). \tag{4.5.6}$$

It is convenient to identify $[0,1)^d$ with the d-dimensional torus $\mathbb{T}^d$, and so we assume that the expansion (4.5.2) is given for points of $\mathbb{T}^d$. Thus F_ω becomes a random smooth expanding transformation of $\mathbb{T}^d$. For any $x \in \mathbb{T}^d$ and $\omega \in \Omega$ set $\phi(\omega, x) = x_0(\omega) \in \mathbb{Z}^d_+$. Then $x_1(\omega) = (F_\omega x)_0(\vartheta\omega) = (\phi \circ \Theta)(\omega, x)$, and so by induction, assuming that $(\phi \circ \Theta^{i-1})(\omega, x) = x_{i-1}(\omega)$, we obtain

$$\big(\phi \circ \Theta^i\big)(\omega, x) = \big(\phi \circ \Theta^{i-1}\big)(\vartheta\omega, F_\omega x) = (F_\omega x)_{i-1}(\vartheta\omega) = x_i(\omega). \tag{4.5.7}$$

This connection between the expansion (4.5.2) and the skew product transformation Θ enables us to apply the machinery of previous sections in computations of fractal dimensions of random sets related to this expansion.

For any $k, l \in \mathbb{Z}^d_+$ set

$$N^\omega_{kl}(x,n) = \big|\big\{j \geqslant 0,\ j < n\colon m\big(\vartheta^j\omega\big) = k,\ x_j(\omega) = l - \mathbf{1}\big\}\big| \quad \text{and}$$
$$N^\omega_l(x,n) = \sum_{k \in \mathbb{Z}^d_+} N^\omega_{kl}(x,n), \tag{4.5.8}$$

where $\mathbf{1} = (1, \dots, 1) \in \mathbb{Z}^d_+$ and $|\{\cdot\}|$ denotes the number of elements in a set $\{\cdot\}$. Let $r = (r_k,\ k \in \mathbb{Z}^d_+)$ be an infinite probability vector and $A = (a_{kl},\ k, l \in \mathbb{Z}^d_+)$ be an infinite probability matrix such that $a_{kl} = 0$ unless $l \leqslant k$ where $l = (l_1, \dots, l_d) \leqslant k = (k_1, \dots, k_d)$ means that $l_i \leqslant k_i\ \forall i = 1, \dots, d$. Set

$$U^\omega_r = \left\{x \in [0,1)^d\colon \lim_{n\to\infty} \frac{1}{n} N^\omega_l(x,n) = r_l \text{ for all } l \in \mathbb{Z}^d_+\right\} \tag{4.5.9}$$

and

$$V^\omega_A = \left\{x \in [0,1)^d\colon \lim_{n\to\infty} \frac{1}{n} N^\omega_{kl}(x,n) = q_k a_{kl} \text{ for all } k, l \in \mathbb{Z}^d_+\right\}, \tag{4.5.10}$$

where $q_k = \mathbb{P}\{m = k\}$. It is easy to check that $\mathbb{P}$-a.s.,

$$F_\omega V^\omega_A = V^{\vartheta\omega}_A \quad \text{and} \quad F_\omega U^\omega_r = U^{\vartheta\omega}_r. \tag{4.5.11}$$

Thus, clearly,

$$HD(U_r^\omega) = HD\big(U_r^{\vartheta\omega}\big) \quad \text{and} \quad HD(V_A^\omega) = HD\big(V_A^{\vartheta\omega}\big), \tag{4.5.12}$$

where $HD(\cdot)$ denotes the Hausdorff dimension of a set in brackets, and since ϑ is ergodic and $HD(U_r^\omega)$, $HD(V_A^\omega)$ depend measurably on ω we conclude that with $\mathbb{P}$-probability one,

$$HD(U_r^\omega) = \text{const} \quad \text{and} \quad HD(V_A^\omega) = \text{const}. \tag{4.5.13}$$

The following result is proved in [95].

THEOREM 4.5.1 (see [95]). *With $\mathbb{P}$-probability one,*

$$H_A = HD(V_A^\omega) = \frac{-\sum_{k\in\mathbb{Z}_+^d} q_k \sum_{l\leqslant k} a_{kl}\log a_{kl}}{\int \log m_1\, d\mathbb{P}}, \tag{4.5.14}$$

and so $HD(V_A^\omega) = d$ if and only if $a_{kl} = \prod_{i=1}^d k_i^{-1}$ for all $l \leqslant k$ and any $k = (k_1, \dots, k_d) \in \mathbb{Z}_+^d$ such that $q_k \neq 0$. In the last case with probability one V_A^ω has also the Lebesgue measure one. The sets U_r^ω have the Lebesgue measure one for $\mathbb{P}$-a.a. ω if and only if $r_l = \sum_{k\geqslant l} q_k \prod_{i=1}^d k_i^{-1}$ for all $l \in \mathbb{Z}_+^d$. Furthermore, for P-a.a. ω,

$$HD(U_r^\omega) = \sup_{A\in\mathcal{A}_{qr}} H_A, \tag{4.5.15}$$

where the supremum in (4.5.15) *is taken over the set $\mathcal{A}_{qr}$ of all infinite probability matrices $A = (a_{kl})$ such that $a_{kl} = 0$ unless $l \leqslant k$ and $qA = r$ with q and r considered as the row vectors* (*i.e.*, $\sum_{k\in\mathbb{Z}_+^d} q_k a_{kl} = r_l \ \forall l \in \mathbb{Z}_+^d$). *The set $\mathcal{A}_{qr}$ is nonempty if and only if*

$$\sum_{l\in R} q_l \geqslant \sum_{l\in R} r_l \tag{4.5.16}$$

for any $R \in \mathcal{R}$ where $\mathcal{R}$ is the collection of all filters in $\mathbb{Z}_+^d$, i.e., the subsets $R \subset \mathbb{Z}^d$ such that if $l \in R$ and $l \leqslant k$ then $k \in R$. If (4.5.16) *does not hold true for some $R \in \mathcal{R}$ then with probability one U_r^ω is empty.*

The above theorem gives Hausdorff dimensions of everywhere dense fractal random sets. Next, we deal with Hausdorff dimensions of random compact sets and, in particular, random Cantor-like sets which are random repellers for random transformations F defined above. Let $\mathcal{E}_\omega$, $\omega \in \Omega$, be a measurable family of compact subsets of $\mathbb{T}^d$ satisfying

$$F_\omega \mathcal{E}_\omega = \mathcal{E}_{\vartheta\omega} \tag{4.5.17}$$

with F_ω given by (4.5.5). Set $\mathcal{E} = \{(\omega, x)\colon\ x \in \mathcal{E}_\omega\}$. Then we have the setup of Section 1.2 and we can speak about the fiber topological entropy $h_{\text{top}}(F, \mathcal{E})$ of F restricted to the $\mathcal{E}$ introduced in Definition 1.2.3. The following result is proved in [95].

THEOREM 4.5.2. *With $\mathbb{P}$-probability one,*

$$HD(\mathcal{E}_\omega) = \frac{h_{\text{top}}(F, \mathcal{E})}{\int \log m_1 \, d\mathbb{P}}. \tag{4.5.18}$$

A natural subclass of sets $\mathcal{E}_\omega$ satisfying the conditions of Theorem 4.5.2 consists of random Cantor-like sets obtained by choosing a measurable family of finite subsets $\Psi(\omega) \subset \{l - \mathbf{1}: l \in \mathbb{Z}^d_+,\ l \leqslant m(\omega)\} = \mathcal{L}(\omega)$ and setting $\mathcal{E}_\omega = \{x \in \mathbb{T}^d: x_j(\omega) \in \Psi(\vartheta^j\omega)\ \forall j = 0, 1, \ldots\}$. If with positive probability $\Psi(\omega) \neq \mathcal{L}(\omega)$ then for $\mathbb{P}$-a.a. ω $\mathcal{E}_\omega$ are proper closed subsets of $\mathbb{T}^d$ which are naturally to call random Cantor sets. They are statistically self similar in the sense that if $\alpha_j \in \Psi(\vartheta^j\omega)$, $j = 0, 1, \ldots, n-1$, then looking at the intersection of $\mathcal{E}_\omega$ with the cylinder $C^\omega_{\alpha_0,\ldots,\alpha_{n-1}} = \{x \in \mathbb{T}^d: x_j(\omega) = \alpha_j\ \forall j = 0, 1, \ldots, n-1\}$ and rescaling by means of $M(\vartheta^{n-1}\omega)\ldots M(\omega)$, where $M(\omega)$ is the same as in (4.5.2), we obtain $\mathcal{E}_{\vartheta^n\omega}$, namely, $F^n_\omega(\mathcal{E}_\omega \cap C^\omega_{\alpha_0,\ldots,\alpha_{n-1}}) = \mathcal{E}_{\vartheta^n\omega}$ which has the same distribution as $\mathcal{E}_\omega$. Observe that $h_{\text{top}}(F, \mathcal{E})$ for such sets $\mathcal{E}_\omega$ is equal to $\int \log|\Psi(\omega)| \, d\mathbb{P}(\omega)$. More general family of sets $\mathcal{E}_\omega$ satisfying (4.5.17) can be obtained by taking a measurable family of matrices $B(\omega) = (b_{kl}(\omega),\ k, l \in \mathbb{Z}^d_+,\ k \leqslant m(\omega),\ l \leqslant m(\vartheta\omega))$ with 0 and 1 entries and by setting $\mathcal{E}_\omega = \{x \in \mathbb{T}^d: b_{x_i(\omega)+\mathbf{1},\mathbf{x_{i+1}}(\omega)+\mathbf{1}} = 1\ \forall i = 0, 1, \ldots\}$. In this case with $\mathbb{P}$-probability one,

$$h_{\text{top}}(F, \mathcal{E}) = \lim_{n\to\infty} \frac{1}{n} \log \left\| B\left(\vartheta^{n-1}\omega\right) \ldots B(\vartheta\omega) B(\omega) \right\|. \tag{4.5.19}$$

This example is connected with random subshifts of finite type considered in Section 4.1.

Random transformations considered in Theorem 4.5.2 are particular cases of random conformal maps acting on their random repellers which are defined in the following way. Let $(\Omega, \mathbb{P})$ and ϑ be as before, M be a C^2 locally compact Riemannian manifold, and $F_\omega, \omega \in \Omega$ be a measurable family of C^2 maps of M such that there exist a compact set $M_0 \subset M$ with nonempty interior $\text{int}M_0$ and a function $\lambda_\omega(x)$ satisfying:

(i) $D_x F_\omega = \lambda_\omega(x) I^\omega_x \quad \forall x \in M_0,\ \omega \in \Omega,$

where $D_x F_\omega$ is the differential of F_ω at x and I^ω_x is an isometry of the tangent space $T_x M$ onto $T_{F_\omega x} M$;

(ii) With probability one $\inf_{x\in M_0} \lambda_\omega(x) > 1$;

(iii) If $\gamma_0(\omega) = \sup_{x\in M_0} \log \lambda_\omega(x)$ and $\gamma_1(\omega) = \sup_{x\in M_0} |\log \max_i |\frac{\partial \lambda_\omega(x)}{\partial x_i}||$ then $\int (\gamma_0 + \gamma_1) \, d\mathbb{P} < \infty$.

Assume also the mixing condition saying that for any open set $U \subset M_0$ there exists $N = N_\omega$ such that $F^N_\omega U \supset M_0$. As an example of maps satisfying the above conditions we may keep in mind algebraic endomorphisms of the torus $\mathbb{T}^d$ given by integer valued matrices $L = (l_{ij})$ such that $\sum_{j=1}^d l_{ij} l_{kj} = \delta_{ik} \lambda_L^2$ for some $\lambda_L > 1$ independent of i where $\delta_{ik} = 1$ if $i = k$ and $= 0$, otherwise. Then $\lambda_L^{-1} L$ is an orthogonal matrix and the conditions above are satisfied for F_ω taken out of this family of endomorphisms with $\lambda_\omega(x)$ depending on ω but not on x. Now let $\mathcal{E}_\omega$, $\omega \in \Omega$, be a measurable family of compact sets satisfying (4.5.17) and $\mathcal{E}_\omega \subset M_1\ \forall \omega \in \Omega$ for some compact set $M_1 \subset \text{int}M_0$. Denote by $\pi_F(\psi, \mathcal{E})$ the fiber

topological pressure of F restricted to $\mathcal{E} = \{(\omega, x)\colon\ x \in \mathcal{E}_\omega\}$ for a random function ψ (see Definition 1.2.3). The following result proved in [95] is a random version of the Bowen–Ruelle formula for the Hausdorff dimension of conformal repellers.

THEOREM 4.5.3. *Let $\varphi_\omega(x) = -\log \lambda_\omega(x)$. Then under assumptions above for $\mathbb{P}$-a.a. ω,*

$$HD(\mathcal{E}_\omega) = t_0 \quad \text{if } \pi_F(t_0\varphi, \mathcal{E}) = 0. \tag{4.5.20}$$

If $\lambda_\omega(x)$, and so $\varphi_\omega(x)$, are independent of x on $\mathcal{E}$ then

$$HD(\mathcal{E}_\omega) = \frac{h_{\text{top}}(F, \mathcal{E})}{\int \log \lambda_\omega \, d\mathbb{P}(\omega)}. \tag{4.5.21}$$

It is not difficult to see that the Minkowski (box) dimensions of random sets $\mathcal{E}_\omega$ from Theorems 4.5.2 and 4.5.3 are the same as their Hausdorff dimensions. For random compact repellers of nonconformal random maps this may already be not true (see [95]). Hausdorff dimensions of other random sets obtained via random geometric constructions and as attractors of random iterated function systems were computed in [92].

More general random f-expansions, the corresponding random transformations and their random invariant measures were studied in [100]. The main part of this paper deals with an extension of the relative thermodynamic formalism to this situation, the construction of random Gibbs measures and an application of this machinery to the computation of their Hausdorff dimensions. Observe that the study of random invariant measures, in particular, random Gibbs measures is interesting even for deterministic transformations such as the Gauss map $Fx = \{x\}$ of the interval $[0, 1]$ which emerges in the continued fraction expansion. Consider, for instance, a sequence of independent positive integer valued random variables $A_\omega, A_{\vartheta\omega}, A_{\vartheta^2\omega}, \dots$ with distributions $p_\omega, p_{\vartheta\omega}, p_{\vartheta^2\omega}, \dots$ and let X_ω be the continued fraction with digits $A_\omega, A_{\vartheta\omega}, A_{\vartheta^2\omega}, \dots$. Then the distribution μ_ω of X_ω satisfies $F\mu_\omega = \mu_{\vartheta\omega}$ $\mathbb{P}$-a.s. and we arrive at a random invariant measure of the Gauss map F. The Hausdorff dimension of this measure has been computed in [100] and it is always less than 1.

5. Random perturbations of dynamical systems

5.1. *Markov chain type perturbations*

The theory of random perturbations of dynamical systems deals either with discrete time models where each iteration of a deterministic transformation F is followed by a small noise (which can be chosen in various ways) which amounts to a Markov chain X_n^ε, $n = 0, 1, 2, \dots$, or with a continuous time setup whose main model is a diffusion Markov process X_t^ε, $t \geqslant 0$, generated by differential operators $L^\varepsilon = \varepsilon L + B$, $\varepsilon > 0$, where L is a second order elliptic differential operator and B is a vector field considered as a first order differential operator which generates a flow F^t determined by the differential equation $d(F^t x)/dt = B(F^t x)$, $t \geqslant 0$. The importance of the latter model is, of course, its connection with partial differential equations and with so-called singular perturbations problems

there. On the other hand, the discrete time model does not necessarily require a differentiable structure and it allows more flexibility in choices of transformations, noises and ambient spaces. The processes X_t^ε, $t \in \mathbb{Z}_+$ or $t \in \mathbb{R}_+$, are viewed as random perturbations of the dynamical system F^t. Not surprisingly the behavior of X_t^ε depends crucially on dynamical properties of F^t both on bounded time intervals, which is clear, and also when $t \to \infty$, which is more interesting. Of course, this behavior depends also on the noise type but we are especially interested in parameters of the process X_t^ε which approximate corresponding parameters of the dynamical system F^t for a large and natural class of perturbations. We deal mainly with the discrete time setup indicating from time to time the corresponding results for the continuous time case when they exist.

Let $F : M \to M$ be a continuous map of a compact metric space M and $\{Q_x^\varepsilon,\ x \in M,\ \varepsilon > 0\}$ be a family of probability distributions on M Borel measurably dependent on x on and such that for any continuous function g,

$$\lim_{\varepsilon \to 0} \sup_{x \in M} \left| \int_M g(y) Q_x^\varepsilon(dy) - g(x) \right| = 0. \tag{5.1.1}$$

Markov chains X_n^ε, $n = 0, 1, 2, \ldots$, with transition probabilities

$$P^\varepsilon(x, \Gamma) = \mathbb{P}\{X_{n+1}^\varepsilon \in \Gamma \mid X_n^\varepsilon = x\} = Q_{Fx}^\varepsilon(\Gamma) \tag{5.1.2}$$

(where $\mathbb{P}$ is the probability on the corresponding probability space) are called random perturbations of the dynamical system F^n, $n \in \mathbb{N}$. A probability measure μ^ε on M is called an invariant measure of the Markov chain X_n^ε if

$$\int_M d\mu^\varepsilon(x)\, P^\varepsilon(x, \Gamma) = \mu^\varepsilon(\Gamma) \tag{5.1.3}$$

for any Borel set $\Gamma \subset M$. The following simple result established originally in [80] is a starting point in the study of stability of invariant measures of dynamical systems.

THEOREM 5.1.1. *Suppose that* (5.1.1)–(5.1.3) *hold true and*

$$(\mathrm{w})\lim_{ve_i \to 0} \mu^{\varepsilon_i} = \mu \tag{5.1.4}$$

(*where* (w)lim *denotes the limit in the weak sense*) *for some subsequence* $\varepsilon_i \to 0$. *Then* μ *is an invariant measure of the map* F (*i.e.*, $\mu(F^{-1}\Gamma) = \mu(\Gamma)$ *for any Borel set* $\Gamma \subset M$).

In physical applications it is natural to think about measures μ obtained as limits of μ^ε as more stable to random perturbations and so having more physical sense than other invariant measures. Especially, this is true when μ^ε converges weakly to μ as $\varepsilon \to 0$ for a natural class of perturbations in which case μ is naturally to be called stochastically stable and physically relevant (as real systems are subject to random perturbations). In the

continuous time case of diffusions X_t^ε generated by an elliptic operator L^ε as above defined on a compact smooth manifold M the transition probabilities $P^\varepsilon(t,x,\cdot)$ of X_t^ε satisfy

$$\lim_{\varepsilon\to 0}\sup_{x\in M}\left|\int_M g(y)P^\varepsilon(t,x,dy)-g(x)\right|=0 \tag{5.1.5}$$

for any continuous function g, and so Theorem 5.1.1 yields that any limit in the weak sense of invariant measures of the process X_t^ε is an invariant measure of the flow F^t.

It is natural to believe that under general circumstances all weak limits of invariant measures μ^ε of the processes X^ε are supported by attractors. To make this precise, recall, that a finite sequence of points $x_1,\dots,x_n$ is called a δ-pseudo orbit if $\operatorname{dist}(Fx_i,x_{i+1})<\delta$ for all $i=1,\dots,n-1$. We say that y is attainable from x and write $y\succ x$ if for any $\delta>0$ there exists a δ-pseudo-orbit starting at x and ending at y. We also have an equivalence relation writing $x\sim y$ whenever $y\succ x$ and $x\succ y$. By the definition we set $x\sim x$. We call a compact set K an attractor if there exists an open set $V\supset K$ such that $F\bar{V}\subset V$ and $\bigcap_{n\geqslant 0}F^nV=K$. The proof of the following result can be found in [85].

THEOREM 5.1.2. *Suppose that there exists only a finite number of equivalence classes in M and assume that the process X^ε exits "much faster" from small neighborhoods of nonattractors than from small neighborhoods of attractors* (*for the precise condition see* [85]). *Then all weak limit points as $\varepsilon\to 0$ of invariant measures μ^ε of X^ε have supports in the union of attractors.*

Other results in this direction can be found in [148] and [65]. We refer the reader also to [126] where the behavior of certain random perturbations in a vicinity of an attractor is studied.

In general, it is necessary to take in the above result all attractors since if we take a local perturbation (i.e., each distribution Q_x^ε is supported by a small neighborhood of x) then starting in a small neighborhood of an attractor K the process X^ε will never leave the open set V appearing in the definition of an attractor, and so there is an invariant measure μ^ε of X^ε whose support is contained in V. It follows that all limit points of such μ^ε have supports in K. For more specific perturbations it is possible to indicate a subset of attractors which support limiting measures. Suppose that for any open set $U\subset M$ uniformly in $x\in M$,

$$\lim_{\varepsilon\to 0}\varepsilon\log P^\varepsilon(x,U)=-\inf_{y\in U}\rho(x,y), \tag{5.1.6}$$

where $\rho\geqslant 0$ is a continuous function on $M\times M$. Set

$$D(x,y)=\inf\left\{\sum_{i=0}^{n-1}\rho(x_i,x_{i+1})\colon\ n\geqslant 1,\ x_0=x,\ x_n=y\right\}.$$

The function measures "the difficulty" for the process X^ε to go from x to y and it induces another preorder writing $y\succ_\rho x$ if $D(x,y)=0$. This yields a ρ-equivalence relation if we write $x\sim_\rho y$ provided $y\succ_\rho x$ and $x\succ_\rho y$. It turns out that in these circumstances

it is possible to give a more precise description of attractors which support limit points of invariant measures μ^ε of X^ε (see [86]). For a continuous time diffusion process X_t^ε generated by a second order elliptic differential operator $L^\varepsilon = \varepsilon L + B$ as described above this was done much earlier in [164] (see also [64]). In this case the functional B should be defined by

$$D(x, y) = \inf_{t \geqslant 0} \inf_{\varphi_0 = x,\, \varphi_t = y} \int_0^t \left\| B(\varphi_s) - \dot{\varphi}_s \right\|^2 ds,$$

where the infimum is taken over absolutely continuous curves φ_s, $\dot{\varphi}_s$ denotes the tangent (speed) vector of this curve at φ_s, and $\| \cdot \|$ denotes the Riemannian norm in the tangent bundle constructed by means of coefficients in second derivatives of the elliptic operator L. Observe that a more difficult question about weights which limiting measures give to various attractors has been solved only in very particular cases.

As we mentioned this above, in general, there is no way to say that limiting measures prefer some attractors rather than others as their supports, i.e., that some attractors are more stable under random perturbations than others, since the answer strongly depends on a chosen type of random perturbations. On the other hand, a more delicate question about a most stable to a natural class of random perturbations invariant measure of F on a transitive attractor K makes sense. Namely, we want to study situations when the normalized restriction to K of any limit (with respect to this natural class of random perturbations) measure μ coincides with a fixed F-invariant probability measure ν on K which is naturally to be called stable under random perturbations and thus having a physical sense. This question is especially important for dynamical systems with abundance of invariant measures, in particular, for chaotic dynamical systems. The most well-understood subclass of the latter is Axiom A systems, in particular, Anosov systems for which the question on random perturbations can be answered rather satisfactory.

In order to obtain more specific results we have to deal with more restricted classes of dynamical systems and random perturbations. We assume now that M is a compact C^2 d-dimensional Riemannian manifold and F is either C^2 diffeomorphism or C^2 endomorphism of M. We assume that the distributions Q^ε in the definition of transition probabilities P^ε of processes X^ε have densities q^ε with respect to the Riemannian volume m on M, i.e., $Q_y^\varepsilon(\Gamma) = \int_\Gamma q_y^\varepsilon(z)\, dm(z)$ for any Borel set $\Gamma \subset M$. Moreover, we suppose that there exist constants $\alpha, \beta, C > 0$, $\beta < 1$ and a family of nonnegative functions $\{r_x(\xi),\ x \in M,\ \xi \in T_x M\}$, where $T_x M$ denotes the tangent space at x, such that $\int r_x(\xi)\, d\xi = 1$ and

$$q_x^\varepsilon(y) \leqslant \varepsilon^{-d} \Bigg((1 + \varepsilon^\alpha) r_x \left(\frac{1}{\varepsilon} \operatorname{Exp}_x^{-1} y \right) \mathbb{I}_{U_{\varepsilon^\beta}(x)}(y) + C e^{-\frac{\alpha}{\varepsilon} \operatorname{dist}(x,y)} \mathbb{I}_{M \setminus U_{\varepsilon^\beta}(x)}(y) \Bigg), \tag{5.1.7}$$

where the functions r_x serve as local scales, $\mathbb{I}_U(y) = 1$ if $y \in U$ and = 0, otherwise, $U_\delta(x)$ is an open δ-ball around x, and $\operatorname{Exp}_x : T_x M \to M$ denotes the exponential map. We assume

furthermore that $q_x^\varepsilon(y)$ is uniformly Lipschitz continuous in both x and y in the domain of positivity of this function and the δ-neighborhood of the boundary of this domain has the Lebesgue measure of order δ (see details in [83, Section II.2.1]). This condition is satisfied for many natural perturbations Q^ε such as Q_x^ε being the distribution of the diffusion particle starting at x and moving time ε or Q_x^ε being the (local) uniform distribution in the ε-ball around x.

Recall, that a compact F-invariant set $\Lambda \subset \mathbb{R}^d$ is called hyperbolic if F is a diffeomorphism of a neighborhood of Λ to its image and there is a continuous splitting $x + \mathbb{R}^d = \Gamma_x^s \oplus \Gamma_x^u$, $x \in \Lambda$, into linear subspaces Γ_x^s and Γ_x^u such that $DF\Gamma_x^s = \Gamma_{Fx}^s$ and $DF\Gamma_x^u = \Gamma_{Fx}^u$, where DF is the differential of F, and $\|DF^n\xi\| \leqslant Ce^{-\gamma n}\|\xi\|$ and $\|DF^{-n}\eta\| \leqslant Ce^{-\gamma n}\|\eta\|$ when $n \geqslant 0$, $\xi \in \Gamma_x^s$, $\eta \in \Gamma_x^u$, $x \in \Lambda$, where $C > 0$, $\gamma > 0$ are constants. A set Λ is called basic hyperbolic if:

(a) Λ is hyperbolic;
(b) the periodic orbits of F restricted to Λ are dense in Λ;
(c) F is topologically transitive on Λ;
(d) there exists an open set $U_\Lambda \supset \Lambda$ such that $\Lambda = \bigcap_{n\in\mathbb{Z}} F^n U_\Lambda$.

If, moreover,

$$FU_\Lambda \subset U_\Lambda \quad \text{and} \quad \Lambda = \bigcap_{n\geqslant 0} F^n U_\Lambda, \tag{5.1.8}$$

then Λ is called a hyperbolic attractor. In the continuous time case a compact F^t-invariant set Λ is called hyperbolic if there is a continuous splitting $x + \mathbb{R}^d = \Gamma_x^s \oplus \Gamma_x^0 \oplus \Gamma_x^u$, $x \in \Lambda$, with Γ^s and Γ^u satisfying the same properties as above and with Γ_x^0 being the one-dimensional linear subspace generated by a vector field B such that $\frac{dF^t x}{dt} = B(F^t x)$. Such a Λ is called basic hyperbolic if it has no fixed points and (b)–(d) above hold true. If, in addition, (5.1.8) holds true for $F = F^1$ then Λ is called a hyperbolic attractor.

For $x \in \Lambda$ let $\mathcal{J}_t(x)$ be the absolute value of the Jacobian of the linear map $DF^t : \Gamma_x^u \to \Gamma_{F^t x}^u$. Define

$$\varphi^{(u)}(x) = -\log \mathcal{J}_1(x) \quad \text{and} \quad \varphi^{(u)}(x) = -\frac{d\mathcal{J}_t(x)}{dt}\bigg|_{t=0} \tag{5.1.9}$$

in the discrete and the continuous time cases, respectively. It is known (see [37] and Chapter 20 in [75]) that both in the discrete and the continuous time cases if Λ is a hyperbolic attractor and $F = F^1$ then for any Hölder continuous function g there exists a unique F-invariant (flow invariant in the continuous time case) probability measure μ_g on Λ such that

$$R_\Lambda(g) = \sup_\mu \left(\int g\, d\mu + h_\mu(F) \right) = \int g\, d\mu_g + h_{\mu_g}(F), \tag{5.1.10}$$

where $R_\Lambda(g)$ is called the topological pressure of g, $h_\mu(F)$ is the entropy of F with respect to μ, and the supremum is taken over all F-invariant probability measures on Λ. Such μ_g is called the equilibrium state for g. If $g = \varphi^{(u)}$ then the corresponding $\mu_g = \mu_\Lambda^{\mathrm{SRB}}$ is called

the Sinai–Ruelle–Bowen (SRB) measure. If Λ is a hyperbolic attractor then $R_\Lambda(\varphi^{(u)}) = 0$ and $\mu_\Lambda^{\mathrm{SRB}}$ is characterized by the equality

$$h_{\mu_\Lambda^{\mathrm{SRB}}}(F) = -\int_\Lambda \varphi^{(u)}(x)\, d\mu_\Lambda^{\mathrm{SRB}}(x). \tag{5.1.11}$$

If F is a C^2 endomorphism of M and $\|DF^{-n}\eta\| \leqslant Ce^{-\gamma n}\|\eta\|$ when $n \geqslant 0$ for all $\eta \in TM$ then F is called an expanding map of M. It is known (see [125]) that in this case F preserves a measure μ_M^{SRB} which is equivalent to the Riemannian volume, such measure is unique, and it is also characterized by the equality similar to (5.1.11) (with $\Lambda = M$).

First, we formulate a result concerning perturbations in a neighborhood of a hyperbolic attractor.

THEOREM 5.1.3. *Let $\Lambda \subset M$ be a hyperbolic attractor and U_Λ satisfies* (5.1.8) *and, moreover, $\overline{FU_\Lambda} \subset U_\Lambda$. Assume that the distributions Q^ε of random perturbations satisfy conditions* (5.1.7) *and, in addition,*

$$q_x^\varepsilon(y) = 0 \quad \text{if } x \in \overline{FU_\Lambda} \text{ and } y \notin U_\Lambda. \tag{5.1.12}$$

Then starting in U_Λ the Markov chains X^ε stay in U_Λ forever and their invariant measures μ^ε in U_Λ converge in the weak sense to $\mu_\Lambda^{\mathrm{SRB}}$.

We observe that some regularity condition, such as the existence of densities q_y^ε of measures Q_y^ε is necessary in order to obtain Theorem 5.1.3. Indeed, consider the following example. Let $M = \mathbb{T}^2$ be the two-dimensional torus and G be its automorphism with no eigenvalues equal 1 in absolute value, for instance,

$$G = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

Now let F be a small C^2 perturbation of G so that F is an Anosov diffeomorphism of $\mathbb{T}^2$ conjugate to G by means of a Hölder continuous homeomorphism h, i.e., $h \circ G = F \circ h$. Assume that the SRB measure μ^{SRB} for F is different from its measure with maximal entropy μ_0. Observe that the Lebesgue measure ℓ on $\mathbb{T}^2$ is both the SRB measure and the measure with maximal entropy for G. It is not difficult to see that $\mu_0 = h\ell$. Now let X^ε be Markov chains with transition probabilities $P^\varepsilon(x, \Gamma)$ which are random perturbations of iterates of G satisfying conditions of Theorem 5.1.3 and let μ^ε be the invariant measure of X^ε. Then, by Theorem 5.1.3, μ^ε converges weakly as $\varepsilon \to 0$ to ℓ. Now define $\tilde{X}_n^\varepsilon = hX_n^\varepsilon$ which is a family of Markov chains whose transition probabilities $\tilde{P}^\varepsilon(x, \Gamma) = P^\varepsilon(h^{-1}x, h^{-1}\Gamma)$ satisfy the definition (5.1.1)–(5.1.2) of random perturbations of F with $\tilde{Q}_y^\varepsilon(\Gamma) = Q_{h^{-1}y}^\varepsilon(h^{-1}\Gamma)$. Then $h\mu^\varepsilon$ is the invariant measure of $\tilde{X}^\varepsilon$ and, clearly, $h\mu^\varepsilon$ converges weakly as $\varepsilon \to 0$ to $h\ell = \mu_0 \neq \mu^{\mathrm{SRB}}$.

Next we formulate a global result.

THEOREM 5.1.4. *Suppose that the limit set of F in M consists of a finite number of basic hyperbolic sets. Assume that the distributions Q^ε of random perturbations satisfy the condition* (5.1.7) *but in place of* (5.1.12) *assume that there exist $\gamma, \delta > 0$ such that*

$$q_x^\varepsilon(y) \geqslant \varepsilon^{-d}\gamma \quad \textit{whenever } \operatorname{dist}(x, y) \leqslant \delta\varepsilon. \tag{5.1.13}$$

Then conditions of Theorem 5.1.2 *hold true and any limit in the weak sense of invariant measures μ^ε of Markov chains X^ε is a linear combination of SRB measures on hyperbolic attractors. The similar result holds true in the continuous time case when X_t^ε are diffusions generated by second order nondegenerate elliptic differential operators $\varepsilon L + B$ with smooth coefficients while the vector field B generates a flow F^t whose limit set consists of a finite number of basic hyperbolic sets* (*and then no additional assumptions on perturbations are needed*). *When F is an expanding endomorphism and perturbations satisfy conditions* (5.1.7) *then all invariant measures μ^ε of Markov chains X_n^ε converge in the weak sense to μ_M^{SRB}.*

The main part of the proof of Theorems 5.1.3 and 5.1.4 is to show that all limits in the weak sense of measures μ^ε have in the hyperbolic case conditional measures on the unstable foliation absolutely continuous with respect to the Lebesgue measures there and in the expanding case these limits are absolutely continuous with respect to the Lebesgue measure on M. This identifies the limit uniquely as the SRB measure. Indeed, this property yields (5.1.11) which characterizes the SRB measure. Similar results were proved in [83] and [128] for random perturbations of model Lorenz type systems and in [76] for random perturbations of maps of the interval satisfying Misiurewicz's condition. Random perturbations of piecewise expanding maps and other related problems are discussed in [27] which contains also some counterexamples to stochastic stability of smooth invariant measures in the setup under consideration (see also [29]). The papers [30,28] and the book [27] discuss also the spectral stability of the Perron–Frobenius operator under deterministic and random perturbations. Recently, the stability of the corresponding SRB measure (in the sense of weak convergence as above) has been established in [25] with respect to certain random perturbations (in particular, parameter random perturbations) of Hènon-like maps. Observe that under the conditions of [25] the Markov chains X_n^ε can be represented (because of the two-dimensional flat phase space) as compositions of random transformations which are small C^2 perturbations of the Hènon-like map in question enabling the authors to apply crucial for their method distortion arguments for their random maps.

Next, we will discuss another approach to the stochastic stability based on variational formulas and convexity arguments. In the discrete time case assume in addition to (5.1.7) that the densities $q_x^\varepsilon(y)$ are positive and continuous in both arguments which makes the transition densities $p^\varepsilon(x, y) = q_{Fx}^\varepsilon(x)$ of Markov chains X_n^ε positive and continuous, as well. In the continuous time case transition densities $p^\varepsilon(t, x, y)$ for time t from x to y of diffusions X_t^ε are automatically positive and continuous provided the generator $L^\varepsilon = \varepsilon L + B$ of X^ε is nondegenerate. For each continuous function V on M consider the operators $T^\varepsilon(V)$ acting on continuous functions g on M by the formula

$$T^\varepsilon(V)g(x) = E_x^\varepsilon g(X_1^\varepsilon)\exp(A^\varepsilon), \tag{5.1.14}$$

where $A^\varepsilon = V(X_1^\varepsilon)$ in the discrete time case and $A^\varepsilon = \int_0^1 V(X_s^\varepsilon)\,ds$ in the continuous time case. The logarithm of the principal eigenvalue of this operator can be obtained via the limit

$$\lambda_\varepsilon(V) = \lim_{n\to\infty} \frac{1}{n} \ln \big\| \big(T_\varepsilon(V)\big)^n \big\|, \tag{5.1.15}$$

where $\|\cdot\|$ denotes the supremum norm. It follows from [55] that $\lambda_\varepsilon(V)$ satisfies the variational formula

$$\lambda_\varepsilon(V) = \sup_\mu \left(\int_M V\,d\mu - I^\varepsilon(\mu) \right), \tag{5.1.16}$$

where the supremum is taken over all probability measures on M and in the discrete time case

$$I^\varepsilon(\mu) = - \inf_{u>0,\, u \text{ is continuous}} \int_M \ln\left(\frac{P^\varepsilon u}{u} \right) d\mu,$$

where $P^\varepsilon u(x) = \int p^\varepsilon(x,y) u(y)\, dm(y)$, and in the continuous time case

$$I^\varepsilon(\mu) = - \inf_{u>0,\, u \text{ is } C^2} \int_M \frac{L^\varepsilon u}{u}.$$

The functional $I^\varepsilon(\mu)$ is lower semicontinuous, and so for any continuous function V there exists a probability measure μ_V^ε such that

$$\lambda_\varepsilon(V) = \int_M V\,d\mu_V^\varepsilon - I^\varepsilon(V). \tag{5.1.17}$$

Moreover, by Proposition 3.1 from [88] if V is Hölder continuous then such measure μ_V^ε is unique and μ_0^ε is the unique invariant measure of the Markov process X^ε.

THEOREM 5.1.5. *Suppose that the limit set of F (or of the flow F^t in the continuous time case) consists of a finite number of basic hyperbolic sets $\Lambda_1, \dots, \Lambda_\kappa$ and in the discrete time case in addition to* (5.1.7) *assume that for some $\alpha > 0$, $\beta \in (0,1)$,*

$$q_x^\varepsilon(y) \geqslant \varepsilon^{-d}\big(1-\varepsilon^\alpha\big) r_x\left(\frac{1}{\varepsilon} \operatorname{Exp}_x^{-1} y \right) \tag{5.1.18}$$

whenever $y \in U_{\varepsilon^\beta}(x)$ (while in the continuous time case we assume only the nondegeneracy of the operator L). Then

$$\lim_{\varepsilon\to 0} \lambda_\varepsilon(V) = \max_{1\leqslant i\leqslant \kappa} \pi_F^{\Lambda_i}\big(\varphi_{\Lambda_i}^{(u)} + V\big), \tag{5.1.19}$$

where $\pi_F^{\Lambda_i}$ is the topological pressure (see [162] *and Section* 1.2 *above) of F restricted to Λ_i and $\varphi_{\Lambda_i}^{(u)}$ is defined by* (5.1.9) *for F restricted to Λ_i.*

Recall (see [162] and Section 1.2 above) that the topological pressure $\pi_F^{\Lambda_i}$ satisfies the variational formula

$$\pi_F^{\Lambda_i}\big(V+\varphi_{\Lambda_i}^{(u)}\big)=\sup_{\mu}\bigg(\int V\,d\mu - I_i(\mu)\bigg), \tag{5.1.20}$$

where $I_i(\mu)=-\int \varphi_{\Lambda_i}^{(u)}\,d\mu - h_\mu(F)$ if μ is an F-invariant probability measure on Λ_i and $I_i(\mu)=\infty$, otherwise. Since here $h_\mu(F)$ is an upper semicontinuous function of μ (see [162]) then $I_i(\mu)$ is a lower semicontinuous functional, and so for any continuous function V there exists a probability measure μ^0_{V,Λ_i} on Λ_i (called an equilibrium state) such that

$$\pi_F^{\Lambda_i}\big(V+\varphi_{\Lambda_i}^{(u)}\big)=\int V\,d\mu^0_{V,\Lambda_i} - I_i\big(\mu^0_{V,\Lambda_i}\big). \tag{5.1.21}$$

Moreover, it is known that $\varphi_{\Lambda_i}^{(u)}$ is Hölder continuous and if V is also Hölder continuous then (see [37] and [75]) a probability measure μ^0_{V,Λ_i} on Λ_i satisfying (5.1.21) is unique. Let $\varphi^{(u)}$ be a continuous function on M which coincides with $\varphi_{\Lambda_i}^{(u)}$ on each Λ_i. Since any F-invariant measure is supported by $\bigcup_i \Lambda_i$ we can write now the variational principle in the form

$$\pi_F\big(V+\varphi^{(u)}\big)=\sup_{\mu}\bigg(\int V\,d\mu - I_i(\mu)\bigg)=\max_{1\leqslant i\leqslant\kappa}\pi_F^{\Lambda_i}\big(\varphi_{\Lambda_i}^{(u)}+V\big), \tag{5.1.22}$$

where $I(\mu)=-\int \varphi_M^{(u)}\,d\mu - h_\mu(F)$ if μ is an F-invariant probability measure on M and $I(\mu)=\infty$, otherwise. Now the supremum (5.1.22) is attained at several measures but all of them are linear combinations of equilibrium states μ^0_{V,Λ_i}. Since $I^\varepsilon(\mu)$ and $I(\mu)$ are convex lower semicontinuous nonnegative functionals it follows from (5.1.16), (5.1.19), and (5.1.22) by convex analysis arguments (see Proposition 3.2 in [88]) that any weak limit as $\varepsilon\to 0$ of measures μ_V^ε maximizing in (5.1.16) belongs to the set of F-invariant measures maximizing in (5.1.22). Thus we arrive at the following result.

THEOREM 5.1.6. *Suppose that the conditions of Theorem* 5.1.5 *hold true. Then for any continuous function V all limit points of measures μ_V^ε are equilibrium states μ_V^0 of F corresponding to the function $V+\varphi^{(u)}$, i.e.,*

$$\pi_F\big(V+\varphi^{(u)}\big)=\int V\,d\mu_V^0 - I\big(\mu_V^0\big). \tag{5.1.23}$$

All equilibrium states μ_V^0 are linear combinations of measures μ^0_{V,Λ_i}. If $V\equiv 0$ then the maximum $\max_{1\leqslant i\leqslant\kappa}\pi_F^{\Lambda_i}(\varphi_{\Lambda_i}^{(u)}+V)$ *is zero and it is attained at attractors only. If this maximum is attained at only one attractor Λ_j and V is Hölder continuous then μ_V^ε converges in the weak sense as $\varepsilon\to 0$ to μ^0_{V,Λ_j} (which is unique in this case). In particular, when $V=0$ we obtain that any weak limit as $\varepsilon\to 0$ of invariant measures μ_0^ε of Markov processes X^ε is a linear combination of SRB measures on hyperbolic attractors.*

Observe that Theorem 5.1.6 requires a stronger than (5.1.13) condition (5.1.18) but, on the other hand, it describes the behavior as $\varepsilon \to 0$ of a large class of measures than Theorems 5.1.3 and 5.1.4. Furthermore, Theorems 5.1.5 and 5.1.6 exhibit the similarity of variational formulas for the principal eigenvalue and for the topological pressure and they show how the former is being transformed into the latter in the random perturbations setup when the perturbation parameter tends to zero.

5.2. *Random perturbations via random transformations*

Let F be a continuous map of a compact metric space M and let ν_ε, $\varepsilon > 0$, be a family of probability measures on the space of continuous maps $C(M, M)$ of M considered with the uniform metric ρ such that for any $\delta > 0$,

$$\nu_\varepsilon\big(\{f \in C(M, M)\colon\ \rho(f, F) > \delta\}\big) \to 0 \quad \text{as } \varepsilon \to 0. \tag{5.2.1}$$

For any Borel $\Gamma \subset M$ set

$$P^\varepsilon(x, \Gamma) = Q^\varepsilon_{Fx}(\Gamma) = \nu_\varepsilon\big(\{f \in C(M, M)\colon\ fx \in \Gamma\}\big) \tag{5.2.2}$$

and for any $y \notin FM$ we can define Q^ε_y as the unit mass at y. Then, clearly, (5.1.1) holds true and the Markov chains X^ε_n with transition probabilities $P^\varepsilon(x, \cdot)$ are random perturbations of the dynamical system F^n, $n \in \mathbb{N}$. This setup can be considered also in the framework of (i.i.d.) random transformations defined on the product probability space as follows. Let $\Omega = (C(M, M))^{\mathbb{Z}}$, $\mathbb{P}^\varepsilon = \nu_\varepsilon^{\mathbb{Z}}$, $\mathcal{F}$ be the product of Borel σ-algebras on $C(M, M)$, and ϑ be the left shift on the sequence space Ω. Now for any $\omega = (\ldots, \omega_{-1}, \omega_0, \omega_1, \ldots)$ we define $F_\omega = F_{\omega,\varepsilon} = \omega_0$. Then $X^\varepsilon_n(\omega) = F^n_\omega x = F_{\vartheta^{n-1}\omega} \circ \cdots \circ F_{\vartheta\omega} \circ F_\omega x$ provided $X^\varepsilon_0 = x$. In fact, according to Section 1.1 from [82] any Markov chain can be represented as a composition of independent random transformations but this general result does not help in applications to random perturbations. In order to obtain specific results the initial map F should be chosen from certain classes of transformations and random maps F_ω are usually chosen from small neighborhoods of F. Let F be a C^2 diffeomorphism of a compact Riemannian manifold M which has a hyperbolic attractor $\Lambda \subset M$ with U_Λ satisfying (5.1.8) and, moreover, $\overline{FU_\Lambda} \subset U_\Lambda$. Let

$$\mathcal{D}_{\alpha,\beta}(f) = \big\{g\colon\ \rho_{C^1}(f, g) \leqslant \alpha \text{ and } L(Dg) \leqslant \beta\big\},$$

where $\rho_{C^1}(\cdot, \cdot)$ is the C^1 distance, $L(\cdot)$ is the Lipschitz constant, and Dg is the differential of g.

THEOREM 5.2.1 (see [166]). *Let $F, \Lambda, U_\Lambda, \mathcal{D}_{\alpha,\beta}(F)$ be as above with α small enough and $\beta > L(DF)$. Let ν_ε be a Borel probability measure with support in $\mathcal{D}_{\alpha,\beta}(F)$ such that ν_ε tends in the weak sense as $\varepsilon \to 0$ to the unit mass at F and for each $\varepsilon > 0$ and $x \in \bar{U}_\Lambda$ the probability measure $P^\varepsilon(x, \cdot)$ defined by* (5.2.2) *is absolutely continuous with respect to the Lebesgue measure on M. Then any invariant measure μ^ε of the corresponding Markov chain X^ε_n converges in the weak sense to the SRB measure $\mu^{\mathrm{SRB}}_\Lambda$.*

Random perturbations of Theorem 5.2.1 enable us to use directly diffeomorphisms close to F which are also hyperbolic and whose differentials preserve one expanding cone in the tangent bundle. It is not difficult to understand from here that applications of the corresponding random diffeomorphisms smooth out measures in the unstable direction with the result that any weak limit of μ^ε has conditional measures on the unstable foliation absolutely continuous with respect to the Lebesgue measures there and the assertion follows. Next, we will see that also in the above setup we can apply the approach based on variational formulas discussed at the end of Section 5.2 but now we have to employ the fiber variational principle of Section 1.2. Denote by $\pi_F^\varepsilon(V)$ the fiber topological pressure for random transformations $F_\omega = F_{\omega,\varepsilon}$ constructed via $\mathbb{P}^\varepsilon$ of a bounded measurable function $V_\omega(x) = V(\omega, x)$ on $\Omega \times M$ which is continuous in x (see Definition 1.2.3). By $\pi_F^0(V)$ we denote the topological pressure of V for the map F.

THEOREM 5.2.2 (see [90]). *Under the conditions of Theorem* 5.2.1 *for each function V as above,*

$$\lim_{\varepsilon\to 0} \pi_F^\varepsilon(V) = \pi_F^0(V). \tag{5.2.3}$$

By the same convex analysis arguments as in Theorem 5.1.6 *it follows that all weak limits of measures maximizing in the fiber variational principle* (1.2.11) *for $\pi_F^\varepsilon(V)$ are maximizing measures in the variational principle for $\pi_F^0(V)$. Taking here $V_\omega(x)$ to be a continuous extension from Λ to $\bar{U}_\Lambda$ of the function $\varphi_\omega^{(u)}(x) = -\ln|\det D_x F_\omega|_{\Gamma_x^u(\omega)}|$, where $\Gamma^u(\omega)$ is the unstable subbundle of F_ω, we obtain that SRB measures of random transformations $F_{\omega,\varepsilon}$ converge weakly as $\varepsilon \to 0$ to the SRB measure of F. Integrating in ω the conclusion of Theorem* 5.2.1 *follows.*

Recently stochastic stability of attractors and nonuniformly expanding maps with respect to this type of random perturbations has been studied in [5] and [3], respectively. In particular, [3] provides not only sufficient but also some necessary conditions on perturbations to have stochastic stability of nonuniformly expanding maps. Random perturbations considered in this section of nonuniformly hyperbolic diffeomorphisms with dominated splitting were studied in [4].

Considering random perturbations of a smooth map (of a diffeomorphism) by means of random smooth maps (of random diffeomorphisms) we may be interested in conditions which ensure stability of various parameters of the map (of the diffeomorphism) such as the topological pressure, the entropy, Lyapunov exponents etc. in the sense that the corresponding fiber (relative) parameters of the perturbed random transformations converge to them (similarly to Theorem 5.2.2) as the perturbation tends to zero. In particular, we mention [113] which studies the stability of Lyapunov exponents of a diffeomorphism from this point of view.

Random perturbations of random transformations type became popular, especially, in lower-dimensional dynamics and also for random perturbations of (piecewise) expanding maps. In both cases the Ruelle–Perron–Frobenius operator plays an important role in the study of an unperturbed map and the corresponding operator for such random perturbations also can be written down and its comparison with the unperturbed one leads often

to required proofs. This approach has been first exhibited in [46] in the study of random perturbations of a map F of the interval $[-1,1]$ satisfying, so-called, Collet–Eckmann conditions. In order to define perturbations close to the boundary points we have either to assume that F extends to a neighborhood of the interval $[-1,1]$, or to assume that $F[-1,1] \subset (-1,1)$, or to identify the end points of the interval and consider addition on the circle. In place of a map F we consider random maps $F_{\omega,\varepsilon}x = Fx + f_{\omega,\varepsilon}$ where $f_{\omega,\varepsilon}$ is a random variable distributed on $[-1,1]$ with the density $g_\varepsilon(x) = \varepsilon^{-1}g(\varepsilon^{-1}x)$ where $g > 0$ and $\int_{-1}^{1} g(x)\,dx = 1$. Next, we apply these random maps independently and arrive at a Markov chain X_n^ε whose transition operator P^ε acts by the simple formula $P^\varepsilon q(x) = \int g_\varepsilon(x - Fy)q(y)\,dy$. The conjugate of this operator acts on measures and applying its iterates to an initial distribution we obtain in the limit invariant measures of X_n^ε. Moreover, we can study the evolution of densities under the action of this operator which can give some information about densities of such invariant measures and lead not only to a proof of their weak convergence to some invariant measure of F but also to convergence of densities in some sense. In [18] this method has been applied to piecewise expanding one-dimensional maps. An extension of this method in [26] led to the proof of weak convergence as $\varepsilon \to 0$ of invariant measures of perturbations X_n^ε described above to the absolutely continuous invariant measure of a one-dimensional unimodal map F satisfying Benedicks–Carleson conditions and a further extension based on the analysis of the spectrum of the (averaged or annealed) Ruelle–Perron–Frobenius operator yielded the proof also of L^1-convergence of densities of invariant measures of X_n^ε in the same situation (see [17]). Observe that random tower constructions introduced in [15] may help to study random perturbations (of the above type) of nonuniformly expanding transformations, as well as random SRB measures of random nonuniformly expanding transformations.

In all works on random perturbations of one-dimensional maps described in the previous paragraph the authors dealt with additive perturbations independent of the point in the interval where they are applied. This excluded the interesting class of random parametric perturbations. Namely, let $F_\lambda(x) = 4\lambda x(1-x)$, $\lambda, x \in [0,1]$, and

$$X_n^\varepsilon = F_{\lambda(1+\varepsilon\zeta_n)} \circ \cdots \circ F_{\lambda(1+\varepsilon\zeta_1)}, \quad n \in \mathbb{N}, \tag{5.2.4}$$

where $\zeta_1, \dots, \zeta_n, \dots$ are independent identically distributed on $[-1,1]$ random variables, $\varepsilon < \lambda^{-1} - 1$ if $\lambda \in (0,1)$ and if $\lambda = 1$ the random variables ζ_n are required to be distributed on $[-1,0]$. Since $F_{\lambda(1+\varepsilon\zeta_1)}x - F_\lambda x = 4\varepsilon\zeta_1 x(1-x)$ the additive perturbation depends on the point where it is applied, and so such perturbations cannot be considered within the setup described in the previous paragraph. The following result has been proved in [45].

THEOREM 5.2.3. *Suppose that the distribution of random variables ζ_n has a Lipschitz continuous density with respect to the Lebesgue measure on $[-1,1]$ if $\lambda < 1$ and on $[-1,0]$ if $\lambda = 1$. Then there exists a set $\Lambda \subset [0,1]$ of positive Lebesgue measure such that for any $\lambda \in \Lambda$ the map F_λ has an absolutely continuous invariant measure μ_λ and as $\varepsilon \to 0$ the densities of invariant measures μ_λ^ε of Markov chains X_n^ε defined by* (5.2.4) *converge in the L^1-sense to the density of μ_λ.*

5.3. *Computations via random perturbations*

In view of the increasing role of computers in the study of dynamical systems and the growing importance of chaotic dynamical systems in modelling physical processes more and more researchers are trying to justify computations of various parameters of such systems. The difficulty here lies in the very definition of a chaotic system which is based on the notion of a "sensitive dependence on initial conditions". Thus, in view of roundoff errors a computer orbit of a chaotic system very fast loses track of a true orbit and it is not clear what one obtains as a result of computations. Some people put forward the results on stability of dynamical systems under random perturbations as a justification of these computations. Since roundoff errors are not random (unless one employs a special algorithm which uses random roundoffs) it hardly makes sense to use random perturbations results as a justification of these computations. On the other hand, it turns out that random perturbations can be used for computations themselves, i.e., instead of trying to compute a dynamical system, in question, we can make computations of its small random perturbations. We exhibit here results showing that this provides a robust method for making computations for chaotic dynamical systems though precise statements concern only the uniformly hyperbolic and (piecewise) expanding maps and the Misiurewicz type maps of the interval.

We will consider here random perturbations of dynamical systems relevant to computer simulations. In view of roundoffs any computer deals with a discrete space rather than with a continuous one and it is not difficult to understand that a computer simulation of random perturbations of a dynamical system should be represented by discretized random perturbations of an original continuous dynamical system. For instance, let $\mathcal{S}_\delta$ be a δ-lattice in $\mathbb{R}^d$ and define the transition probabilities of random perturbations by $P^\varepsilon(x, U) = P^{\varepsilon,\delta}(x, U) = |\mathcal{S}_\delta \cap B_\varepsilon(Fx)|^{-1}|\mathcal{S}_\delta \cap U \cap B_\varepsilon(Fx)|$ where $|A|$ denotes the number of points in a set A and $B_r(y)$ is a closed r-ball centered at y. Considering a hyperbolic attractor Λ (or an expanding endomorphism) and random perturbations on a substantially larger scale ε than the discretization step δ (which can be viewed as the precision of a computer) it follows from [97] that invariant measures of corresponding Markov chains weakly converge as $\delta \backsim \varepsilon^{1+c} \to 0$, $c > 0$, to the Sinai–Ruelle–Bowen (SRB) measure on Λ. Similar results hold true for maps of the interval satisfying the Misiurewicz condition and such "almost" hyperbolic sets as the geometric Lorenz attractor. Observe that discretizations of piecewise expanding maps and justifications of their computations via random perturbations can be found in [27]. In the case of a hyperbolic attractor Λ with localized random perturbations and in the case of a map of the interval I it suffices to consider lattices $\mathcal{S}_\delta$ only in a small neighborhood of Λ or on I, respectively, and so one obtains a Markov chain on $\mathcal{S}_\delta$ with a finite number of states whose invariant measures (stationary distributions) can be found by solving a system of linear equations. These results mean also that typical paths of a random perturbation visit different sets with frequencies close to their SRB measure which says in the one-dimensional case that the corresponding measure can be approximated by histograms constructed by typical paths. In practice, every path may be considered as typical when independent random errors are added on each step of iterations since for large numbers of iterations it is extremely improbable to get a nontypical path and it seems that this is a quite stable self correcting algorithm for evaluation of invariant

measures and it should work faster for the same precision than methods based on solving systems of linear equation. Observe that similar results can be obtained for continuous time dynamical systems, i.e., for flows, considering either random perturbations of time-one maps or taking space-time discretizations of corresponding diffusion perturbations of flows.

Another possible procedure for approximating the absolutely continuous invariant measure of a map F of an interval was suggested by Ulam [158, Section 4] where this measure is conjectured to be close to the invariant measure of the finite Markov chain with transition probabilities $p_{ij} = m(V_i \cap F^{-1}V_j)/m(V_i)$ where m is the Lebesgue measure and $\{V_i\}$ is a finite partition of the interval into subintervals with small length. For one-dimensional piecewise expanding maps this problem was studied in many papers (see [27,122,54] and references therein). A similar construction can be considered in any dimension with appropriate partitions of a neighborhood of an invariant set. If we set $P^\varepsilon(x, U \cap V_j) = p_{ij}m(U \cap V_j)/m(V_j)$, provided $x \in V_i$ and ε is the maximal diameter of elements of the partition then we arrive at a model of random perturbations. It turns out that Ulam's approximations can be justified also for multidimensional expanding maps but in a general hyperbolic situation it is not clear whether the Ulam approach leads to the SRB measure.

In order to formulate precise results let $\mathcal{S}_\delta$ be a cubic δ-lattice in $\mathbb{R}^d$ and F be a C^2 diffeomorphism of an open neighborhood U_Λ of its hyperbolic attractor Λ and $\Lambda \subset U_\Lambda \subset \mathbb{R}^d$. Consider random perturbations $X_n^{\varepsilon,\delta}$ of iterates F^n of F which are Markov chains with transition probabilities $P^{\varepsilon,\delta}(x,\cdot) = Q_{Fx}^{\varepsilon,\delta}(\cdot)$ where $Q_y^{\varepsilon,\delta} \in \mathcal{P}(\mathcal{S}_\delta)$, $y \in \mathbb{R}^d$, is a family of measures which can be written in the form $Q_y^{\varepsilon,\delta}(G) = \sum_{z\in G} q_y^{\varepsilon,\delta}(z)$ with $q_y^{\varepsilon,\delta}(z) = Q_y^{\varepsilon,\delta}(\{z\})$, $z \in \mathcal{S}_\delta$, satisfying the following properties:

(i) $q_y^{\varepsilon,\delta}(z) = 0$ if $y \in \overline{FU_\Lambda}$ and $z \notin U_\Lambda$;

(ii) There exist constants $\alpha, C > 0, \alpha < 1$ and a family of nonnegative functions $\{r_y(z),\ y, z \in \mathbb{R}^d\}$ such that

$$q_y^{\varepsilon,\delta}(z) \leqslant C\left(\frac{\delta}{\varepsilon}\right)^d e^{-\frac{\alpha}{\varepsilon}|z-y|} \quad \text{for all } z \in \mathcal{S}_\delta \text{ and } y \in \mathbb{R}^d \tag{5.3.1}$$

and

$$q_y^{\varepsilon,\delta}(z) \leqslant \left(1+\varepsilon^\alpha\right)\left(\frac{\delta}{\varepsilon}\right)^d r_y\left(\frac{z-y}{\varepsilon}\right) \tag{5.3.2}$$

provided $|z-y| \leqslant \varepsilon^{1-\alpha}$ and $z \in \mathcal{S}_\delta$;

(iii) For any $y, z, v, w \in \mathbb{R}^d$,

$$\int r_y(v)\,dv = 1 \quad \text{and} \quad r_y(z) \leqslant Ce^{-\alpha|z-y|}, \tag{5.3.3}$$

$$\int_{\partial V_y^+(\gamma)} r_y(v)\,dv \leqslant C(\gamma+\delta), \tag{5.3.4}$$

and

$$r_y(z) \leqslant r_v(w) + C\max(\rho,\delta) + \mathbb{I}_{\partial V_x^+(C\rho)}(z) r_y(z), \tag{5.3.5}$$

where $\rho = |y-v| + |z-w|$, $V_y^+ = \{v \in \mathbb{R}^d\colon\ r_y(v) > 0\}$, $\partial V_y^+(\gamma)$ denotes the γ-neighborhood of the boundary ∂V_y^+ of V_y^+, and $\mathbb{I}_A(x) = 1$ if $x \in A$ and $= 0$, otherwise.

THEOREM 5.3.1 (see [97]). *Let Λ be a hyperbolic attractor as above and suppose that the conditions* (i)–(iii) *hold true. Let $\mu^{\varepsilon,\delta}$ be an invariant measure of the Markov chain $X_n^{\varepsilon,\delta}$ having the support in $\bar{U}_\Lambda$. Then $\mu^{\varepsilon,\delta}$ weakly converges as $\varepsilon \to 0$ to the SRB measure $\mu_\Lambda^{\mathrm{SRB}}$ provided $\delta \leqslant \varepsilon^{1+c}$ for some fixed $c > 0$. If $\operatorname{supp} Q_y^{\varepsilon,\delta} \supset B_{\rho\varepsilon}(y) \cap \mathcal{S}_\delta$ for all y and a fixed ρ then $\mu^{\varepsilon,\delta}$ is unique provided ε, δ are small enough.*

The proof of Theorem 5.3.1 proceeds by modifying arguments of the proof of Theorem 5.1.3 in order to show that any weak limit μ of measures $\mu^{\varepsilon,\delta}$ is an F-invariant probability measure on Λ having conditional measures on unstable manifolds with bounded densities with respect to the Riemannian volume there which yields the result similarly to Theorem 5.1.3.

Next we discuss Ulam's approximations (see also [27]). Let M be a compact d-dimensional Riemannian manifold and $F : M \to M$ be a C^2 expanding endomorphism of it. Let $\mathcal{V}_\varepsilon = \{V_i,\ i = 1, \dots, N_\varepsilon\}$, $\varepsilon > 0$, be a family of finite partitions of M such that:

(i′) $\overline{\operatorname{int} V_i} = \bar{V}_i\ \forall i$ and $\exists \kappa > 0$ such that each V_i contains a ball of radius $\kappa\varepsilon$ and is contained in a ball of radius $\kappa^{-1}\varepsilon$;

(ii′) the $(d-1)$-dimensional volume of each boundary ∂V_i is bounded by $\kappa^{-1}\varepsilon^{d-1}$;

(iii′) all V_i's have the same volume $C_\varepsilon \varepsilon^d$ with $\kappa \leqslant C_\varepsilon \leqslant \kappa^{-1}$.

The Ulam Markov chains X_n^ε considered on the manifold M have transition probabilities

$$P^\varepsilon(x, U) = \sum_{j=1}^{N_\varepsilon} p_{ij} \frac{m(V_j \cap U)}{m(V_j)}, \quad p_{ij} = \frac{m(V_i \cap F^{-1}V_j)}{m(V_i)}, \tag{5.3.6}$$

provided $x \in V_i$, where m denotes the Riemannian volume on M.

THEOREM 5.3.2. *Invariant measures of "Ulam's Markov chains" X_n^ε constructed by an expanding C^2 endomorphism F of a compact manifold and a family of partitions $\mathcal{V}_\varepsilon$ satisfying* (i′)–(iii′) *weakly converge as $\varepsilon \to 0$ to the unique absolutely continuous invariant measure of F.*

The main point in the proof of Theorem 5.3.1 (as well, as of Theorem 5.1.3 from Section 5.2) concerns certain estimates of probabilities for the Markov chain $X_n^{\varepsilon,\delta}$ to stay in a tube around an orbit $\{F^i z,\ i = 0, \dots, n\}$ which are done via the linearization of the map F near this orbit, i.e., by dealing with its differential instead, and then one has to study corresponding Markov chains in linear spaces considered in Section 2.2 of [83]. A similar Markov chain representation can be done for Ulam's model with the crucial difference

that now Markov chains become sums of dependent and not independent random vectors as in [83], and so the argument there for main estimates does not go through. This is due to the fact that in Ulam's method when x moves through the boundary of an element of the partition the transition probabilities $P^\varepsilon(x,\cdot)$ jump in the way that (5.3.2) and (5.3.5) cannot be both satisfied. If F is an expanding endomorphism then it is still possible to show directly for Ulam's model that the "tube" probabilities, in question, satisfy desired estimates which yields the weak convergence of invariant measures of X_n^ε to the smooth invariant measure of F. In the hyperbolic case if for Ulam's method one takes Markov partitions then desired estimates for "tube" probabilities are valid and the weak convergence of invariant measures of X_n^ε to the SRB measure follows though this has no practical applications since it is hardly possible to compute Markov partitions for nonlinear transformations. On the other hand, if boundaries of elements of the partition are not pieces of stable and unstable manifolds then the situation becomes rather complicated and it is not clear whether Ulam's approach works in this situation.

6. Concluding remarks

6.1. *Some open problems in RDS*

We will discuss here some open problems in random dynamics which seem to us most important for further development of this subject. In Section 4.2 we gave a definition of random hyperbolic sets but the global definition of random Axiom A diffeomorphisms and their properties are not quite clear yet except for the case when they are picked at random from a small C^2 neighborhood of a deterministic Axiom A diffeomorphism as in Example 4.2.3. We note that some important general notions of dynamical systems such as nonwandering points and, in particular, periodic points may not play a natural role in the random setup. It seems that some deterministic results do not have natural extensions for the random setup and, on the other hand, some results on random transformations may have no counterparts for the deterministic case assuming that transformations are truly random, for instance, that ϑ acts on Ω aperiodically or has positive entropy, etc. As an example of such results we can mention the existence of finite relative topological generators (a relative topological version of Krieger's theorem) described in Section 1.3.

We start the discussion on open problems with basic questions concerning random Anosov diffeomorphisms F_ω of the d-dimensional torus $\mathbb{T}^d$ (see Definition 4.2.2). The deterministic version of the following is a well-known theorem (see, for instance, Section 18.6 in [75]).

CONJECTURE 6.1.1. For any random Anosov diffeomorphism F_ω, $\omega\in\Omega$, of $\mathbb{T}^d$ there exists a random Anosov automorphism A_ω, $\omega\in\Omega$, and a random homeomorphism h_ω, $\omega\in\Omega$, of $\mathbb{T}^d$ such that $A_\omega = h_{\vartheta\omega}^{-1}F_\omega h_\omega$ and A_ω is in the same homotopy class as F_ω, i.e., they induce the same action in the fundamental group $\mathbb{Z}^d$ of $\mathbb{T}^d$.

In order to prove this result, one shows, first, that the linear actions $\tilde F_\omega$ induced by F_ω in the fundamental group produce a random Anosov automorphism. Next, it is necessary

to construct a random semiconjugacy which should be similar to arguments in the random Hartman–Grobman theorem (see Section 7.4 in [7]). The final part of the proof should yield that, in fact, this semiconjugacy is the required random conjugacy. The deterministic proof employs, usually, periodic points which are rare in the random setup and, it seems, some form of random shadowing (see Proposition 4.2.9) should be used instead. Observe that if all realizations of a random Anosov (more generally, hyperbolic) diffeomorphism are close to the same deterministic one then the former is randomly conjugate to the latter which follows via the shadowing (see [115]). The existence of $\mathbb{P}$-a.s. smooth conjugations h_ω is less clear (cf. [75, pp. 640–641]). It is easy to see that if h_ω is a random conjugation and probability measures m_ω satisfy $A_\omega m_\omega = m_{\vartheta\omega}$ $\mathbb{P}$-a.s. then the measures $\mu_\omega = h_\omega m_\omega$ satisfy $F_\omega \mu_\omega = \mu_{\vartheta\omega}$ $\mathbb{P}$-a.s. and if h_ω is smooth $\mathbb{P}$-a.s. then Lyapunov exponents of A with respect to m and of F with respect to μ must be the same.

It would be interesting also to verify in the random situation the following rigidity statement which concerns atypically smooth stable and unstable foliations.

CONJECTURE 6.1.2. Any C^∞ random Anosov diffeomorphism of $\mathbb{T}^d$ with $\mathbb{P}$-a.s. C^∞ random stable and unstable foliations is C^∞ random conjugate to a random Anosov automorphism of $\mathbb{T}^d$.

The topological pressure of a Hölder continuous potential g, the corresponding equilibrium state and its entropy for an Anosov diffeomorphism F are known to depend smoothly both on g and on F in an appropriate sense. A random extension of this result has not been dealt with properly as yet which would be important for various applications.

PROBLEM 6.1.3. Show that the fiber topological pressure of a random Hölder potential g for a random Anosov diffeomorphism F depend smoothly on g and F in an appropriate sense and that the same is true for the corresponding equilibrium state and its fiber entropy.

Periodic orbits play an extremely important role in the deterministic dynamics, especially, for hyperbolic dynamical systems where already definition of some important objects, such as the ζ-function, relies on periodic orbits. In the random setup it is not quite clear what should play their role, in general. It is shown in [42] that standard results about the deterministic zeta function such as its analytic extension fail for its straightforward random version. One of problems whose solution is given in the deterministic case via periodic orbits is Livschitz's theorem (see [75, Section 19.2]). Its random counterpart can be described in the following way. Let F_ω, $\omega \in \Omega$, be a random Anosov diffeomorphism of $\mathbb{T}^d$.

CONJECTURE 6.1.4. Let $\varphi_\omega = \varphi_\omega(x)$ be a measurable function satisfying the random Hölder condition $|\varphi_\omega(x) - \varphi_\omega(y)| \leqslant K_\varphi(\omega)(d(x, y))^\alpha$ where $\alpha > 0$ is a constant and $\int \log K_\varphi \, d\mathbb{P} < \infty$. Assume that $\int \varphi \, d\mu = 0$ for any Θ-invariant measure (where Θ is the skew product transformation) having the marginal $\mathbb{P}$ on Ω. Then there exist a random variable $h = h(\omega)$ with $\int h \, d\mathbb{P} = 0$ and a measurable function $\psi_\omega = \psi_\omega(x)$ satisfying the random Hölder condition as above with the same $\alpha > 0$ and some log integrable random variable $K_\psi(\omega)$ such that $\varphi = \psi \circ \Theta - \psi + h$.

Observe also that periodicity (and a more general notion of recurrence) plays an important role also in the theory of usual Markov chains and it would be interesting to develop a corresponding notion for Markov chains with random transition probabilities considered in Section 4.3 which would substantially enhance their theory.

An interesting study of periodicity for random dynamical systems has been recently done in [104]. Suppose that the fibers $\Gamma(\omega) = \{x\colon\ (\omega, x) \in \Gamma\}$ of a measurable set $\Gamma \subset \mathcal{E}$ satisfy $F_\omega \Gamma(\omega) = \Gamma(\vartheta\omega)$ $\mathbb{P}$-a.s. Then Γ is called sometimes a random invariant set . By ergodicity of ϑ either the number $\#\Gamma(\omega)$ of elements of $\Gamma(\omega)$ is infinite $\mathbb{P}$-a.s. or it is finite $\mathbb{P}$-a.s. and in the latter case it is constant $\mathbb{P}$-a.s. A finite random invariant set Γ with $\#\Gamma(\omega) = N$ $\mathbb{P}$-a.s. has been called in [104] a random periodic orbit of period N. If such Γ does not contain a proper nonempty random invariant subset then Γ is said to have minimal period N. Assuming that ϑ^N is ergodic a random point $x(\omega)$ has been called a random periodic point of period N if $F^N_\omega x(\omega) = x(\vartheta^N\omega)$ $\mathbb{P}$-a.s. Such an N is called the minimal period if $F^l_\omega x(\omega) \neq x(\vartheta^l\omega)$ with positive probability for all $l \in \{1, \dots, N-1\}$. If random maps F_ω act on the line $\mathbb{R}$ and assuming that ϑ^N is ergodic N random points $x_i(\omega) \in \mathbb{R}$, where $x_1 < x_2 < \cdots < x_N$ $\mathbb{P}$-a.s., are called in [104] a random periodic cycle of period N, if there exists a deterministic N-permutation π such that $F_\omega x_i(\omega) = x_{\pi(i)}(\vartheta\omega)$. Such an N is called the minimal period if $F^l_\omega x_1(\omega) \neq x_1(\vartheta^l\omega)$ with positive probability for all $l \in \{1, \dots, N-1\}$. It turns out that these notions are not quite the same, in particular, the existence of a random periodic orbit of minimal period N does not imply, in general, the existence of a random periodic point of minimal period N and the latter does not imply, in general, the existence of a random periodic cycle of minimal period N (see Theorem 4.15 in [104]). The main result in [104] is the following random version of the Sharkovsky type theorem.

THEOREM 6.1.5. *Let F be a RDS with F_ω being continuous maps of $\mathbb{R}$ and assume that ϑ^n is ergodic for all $n \in \mathbb{N}$. Let $\{x_1(\omega), \dots, x_N(\omega)\}$ be a random periodic cycle of minimal period N, where $x_1 < \cdots < x_N$ $\mathbb{P}$-a.s. Then for any $l \lhd N$, where $\lhd$ denotes the Sharkovsky ordering, F has a random periodic orbit of minimal period l or $2l$. Let ϑ be weakly mixing and assume that F has a random periodic orbit of minimal period three. Then it also has a random periodic orbit of minimal period l or $2l$ for all $l \in \mathbb{N}$.*

Periodic orbits are also somewhat involved in the proof of the thermodynamic formalism constructions for β-transformations in [160] and by this reason a corresponding extension to the random β-transformations encounters difficulties and has not be done yet (see the corresponding discussion in Section 4 of [100]). Observe that existence of random absolutely continuous invariant measures for such transformations was derived in [41] and [40] shows that they have an exponential decay of correlations property.

Among other interesting questions which would be important to understand is the meaning of the fiber (relative) spectrum for random transformations. This could be important both for an extension of the classical spectral theory of dynamical systems to the random setup and, for instance, for a characterization of random Anosov diffeomorphisms via an appropriate gap in a random version of Mather's spectrum (see [127]). Observe that (fiber) relative discrete spectrum is easier to define and some results based on it appeared already in the literature (see [114] and references therein).

We mention also the basic problem of representations of Markov chains by appropriate classes of random transformations discussed in Section 1.1 of [82] which asks about conditions on transition probabilities $P(x,\cdot)$ of a Markov chain X_n which enable us to find a probability measure μ on a nice family of transformations Φ (homeomorphisms, smooth maps, diffeomorphisms, automorphisms, linear transformations, etc.) such that $P(x,\Gamma)=\mu\{F\in\Phi\colon\ Fx\in\Gamma\}$. Then X_n can be viewed as a composition of independent identically distributed random maps with the distribution μ. Apart from a theoretical interest of this problem such representations can be useful in proving various results about Markov chains. For instance, an appropriate representation of this sort has been recently employed in [25] in the study of random perturbations of the Hènnon-like maps. In spite of some progress in this problem achieved in [139] and [140] there are no sufficiently general results about such representations by families of maps mentioned above and a further research on this problem is needed.

Discrete time random hyperbolic dynamical systems are still in a relatively good shape in comparison to stochastic flows where in view of noninvariance of the flow direction no substantial theory of hyperbolic RDS has been developed as yet which remains an important standing problem.

6.2. *Remarks on random perturbations*

In conclusion we will discuss a bit the current situation with the theory of random perturbations. After a relatively complete theory of random perturbations of uniformly hyperbolic dynamical systems there were hopes to extend these theory to a wide class of nonuniformly hyperbolic dynamical systems so that in spite of their deterministic instability they could be considered stable in some probabilistic sense. This program has not been realized and it does not seem feasible nowadays. Moreover, this program has lost some of its original romantic lustre since unlike natural smoothness assumptions on deterministic perturbations there exist no widely acceptable standard on the type of random perturbations. In the continuous time case the diffusion type random perturbations is the most natural and, essentially, the only reasonable setup. On the other hand, in the discrete time case much more general random perturbations can be considered but, nevertheless, in order to obtain the stability results described in Theorem 5.1.3 certain assumptions (which are not so easy to justify from a physical point of view) are needed (as the example after Theorem 5.1.3 shows). Most of the recent results about random perturbations are just extensions of some deterministic results and they concern only some rather restricted classes of random perturbations of some families of nonuniformly hyperbolic dynamical systems (see [3] and [4]). It is important to extend the study to more general natural classes of random perturbations of different families of nonuniformly hyperbolic systems. For instance, it is not known yet whether the stochastic stability holds true under general conditions of the Pesin theory.

PROBLEM 6.2.1. Suppose that f is a C^2 diffeomorphism of a compact Riemannian manifold M preserving a probability measure μ equivalent to the Riemannian volume m on M or even suppose that $\mu=m$. Assume that μ is ergodic and that there are no zero Lyapunov exponents with respect to μ. Prove (or construct a counterexample) that μ is stable to random perturbations (in the sense of Section 5.2) satisfying the condition (5.1.7).

In order to advance the theory of random perturbations it would be very important to construct various examples of transitive dynamical systems unstable to reasonable random perturbations (i.e., for example, satisfying conditions of Section 5.2) in the sense, for instance, that invariant measures of random perturbations do not converge to SRB measures (assuming they exist) of dynamical systems under consideration.

PROBLEM 6.2.2. Construct an example of a topologically transitive C^2 diffeomorphism f of a compact Riemannian manifold M preserving an ergodic probability measure μ equivalent to the Riemannian volume m on M such that μ is not stable to random perturbations (in the sense of Section 5.2) satisfying the condition (5.1.7), i.e., that invariant measures μ^ε of the random perturbations X^ε do not converge weakly to μ.

Coupled map lattices attracted a substantial attention recently, in particular, as discrete time models describing turbulence. It would be interesting to generalize the results on stochastic stability of SRB measures with respect to (rather special) random perturbations of (weakly) coupled (expanding) map lattices obtained in [123].

References

Survey in this volume

[1] L. Barreira and Ya. Pesin, *Smooth ergodic theory and nonuniformly hyperbolic dynamics*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 57–263.

Other sources

[2] L.M. Abramov and V.A. Rohlin, *The entropy of a skew product of measure-preserving transformation*, Amer. Math. Soc. Transl. Ser. 2, **48** (1966), 255–265.
[3] J. Alves and V. Araújo, *Random perturbations of nonuniformly expanding maps*, Geometric Methods in Dynamics I, Astérisque **286** (2003), 25–62.
[4] J. Alves, V. Araújo and C. Vásquez, *Random perturbations of diffeomorphisms with dominated splitting*, Preprint.
[5] V. Araújo, *Infinitely many stochastically stable attractors*, Nonlinearity **14** (2001), 583–596.
[6] A. Arbieto, C. Matheus and K. Oliveira, *Equilibrium states for random nonuniformly expanding maps*, Preprint (2003).
[7] L. Arnold, *Random Dynamical Systems*, Springer-Verlag, New York, Berlin (1998).
[8] L. Arnold, H. Crauel and J.-P. Eckmann, eds, *Lyapunov Exponents*, Lecture Notes in Math., Vol. 1486, Springer (1991).
[9] P. Arnoux and A. Fisher, *Anosov families, renormalization and non-stationary subshifts*, Ergodic Theory Dynam. Systems **25** (2005), 661–709.
[10] J. Bahnmüller, *The Pesin formula for random dynamical systems*, PhD thesis, Institut für Dynamische Systeme, Universität Bremen (1996).
[11] J. Bahnmüller and T. Bogenschütz, *A Margulis–Ruelle inequality for random dynamical systems*, Arch. Math. **64** (1995), 246–253.
[12] J. Bahnmüller and P.-D. Liu, *Characterization of measures satisfying Pesin's entropy formula for random dynamical systems*, J. Dynam. Differential Equations **10** (3) (1998), 425–448.
[13] V.I. Bakhtin, *Random processes generated by a hyperbolic sequence of mappings I, II*, Russian Acad. Sci. Izv. Math. **44** (1995), 247–279, 617–627.

[14] V. Baladi, *Correlation spectrum of quenched and annealed equilibrium states for random expanding maps*, Comm. Math. Phys. **186** (1997), 671–700.
[15] V. Baladi, M. Benedicks and V. Maume-Deschamps, *Almost sure rates of mixing for i.i.d. unimodal maps*, Ann. Sci. Ècole Norm. Sup. **35** (2002), 77–126.
[16] V. Baladi, A. Kondah and B. Schmitt, *Random correlations for small perturbations of expanding maps*, Random Comput. Dynam. **4** (1996), 179–204.
[17] V. Baladi and M. Viana, *Strong stochastic stability and the rate of mixing for unimodal maps*, Ann. Sci. Ècole Norm. Sup. **29** (1996), 483–517.
[18] V. Baladi and L.-S. Young, *On the spectra of randomly perturbed expanding maps*, Comm. Math. Phys. **156** (1993), 355–385; Erratum: **166** (1994), 219–220.
[19] L. Barreira and Ya. Pesin, *Lectures on Lyapunov exponents and smooth ergodic theory*, Proc. Sympos. Pure Math. **69** (2001), 3–95.
[20] L. Barreira, Ya. Pesin and J. Schmeling, *Dimension and product structure of hyperbolic measures*, Ann. of Math. **149** (1999), 755–783.
[21] H. Bauer, Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie, 3rd edn, De Gruyter, Berlin (1978).
[22] P. Baxendale, *Brownian motions in the diffeomorphism group I*, Compositio Math. **53** (1984), 19–50.
[23] P. Baxendale, *The Lyapunov spectrum of a stochastic flow of diffeomorphisms*, Lecture Notes in Math., Vol. 1186, Springer (1986), 322–337.
[24] P. Baxendale, *Lyapunov exponents and relative entropy for a stochastic flow of diffeomorphisms*, Probab. Theory Related Fields, **81** (1989), 521–554.
[25] M. Benedicks and M. Viana, *Random perturbations and statistical properties of Hènnon-like maps*, Ann. Inst. H. Poincaré Anal. Non Linéare, to appear.
[26] M. Benedicks and L.-S. Young, *Absolutely continuous invariant measures and random perturbations for certain one-dimensional maps*, Ergodic Theory Dynam. Systems **12** (1992), 13–37.
[27] M. Blank, *Discreteness and Continuity in Problems of Chaotic Dynamics*, Transl. Math. Monographs, Vol. 161, Amer. Math. Soc., Providence, RI (1997).
[28] M. Blank, *Perron–Frobenius spectrum for random maps and its approximation*, *Moscow Math. J.* **1** (2001), 315–344.
[29] M. Blank and G. Keller, *Stochastic stability versus localization in one-dimensional chaotic dynamical systems*, Nonlinearity **10** (1997), 81–107.
[30] M. Blank and G. Keller, *Random perturbations of chaotic dynamical systems: stability of the spectrum*, Nonlinearity **11** (1998), 1351–1364.
[31] T. Bogenschütz, *Entropy, pressure, and a variational principle for random dynamical systems*, Random Comput. Dynam. **1** (1992), 99–116.
[32] T. Bogenschütz, *Equilibrium states for random dynamical systems*, PhD thesis, Universität Bremen (1993).
[33] T. Bogenschütz, *Stochastic stability of equilibrium states*, Random Comput. Dynam. **4** (1996), 85–98.
[34] T. Bogenschütz and V.M. Gundlach, *Symbolic dynamics for expanding random dynamical systems*, Random Comput. Dynam. **1** (1992), 219–227.
[35] T. Bogenschütz and V.M. Gundlach, *Ruelle's transfer operator for random subshifts of finite type*, Ergodic Theory Dynam. Systems **15** (1995), 413–447.
[36] R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, Lecture Notes in Math., Vol. 470, Springer-Verlag, New York (1975).
[37] R. Bowen and D. Ruelle, *The ergodic theory of Axiom A flows*, Invent. Math. **29** (1975), 181–202.
[38] P. Boxler, *A stochastic version of center manifold theory*, Probab. Theory Related Fields **83** (1989), 509–545.
[39] M. Brin and Y. Kifer, *Dynamics of Markov chains and stable manifolds for random diffeomorphisms*, Ergodic Theory Dynam. Systems **7** (1987), 351–374.
[40] J. Buzzi, *Exponential decay of correlations for random Lasota–Yorke maps*, Comm. Math. Phys. **208** (1999), 25–54.
[41] J. Buzzi, *Absolutely continuous S.R.B. measures for random Lasota–Yorke maps*, Trans. Amer. Math. Soc. **352** (2000), 3289–3303.
[42] J. Buzzi, *Some remarks on random zeta functions*, Ergodic Theory Dynam. Systems **22** (2002), 1031–1040.
[43] A. Carverhill, *Flows of stochastic dynamical systems: Ergodic theory*, Stochastics **14** (1985), 273–317.

[44] C. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math., Vol. 580, Springer-Verlag, New York (1977).
[45] G. Chakvetadze, *Parameter random perturbations of Collet–Eckmann maps of the interval*, Preprint.
[46] P. Collet, *Ergodic properties of some unimodal mappings of the interval*, Preprint, Institut Mittag-Leffler (1984).
[47] N.-D. Cong, *Topological Dynamics of Random Dynamical Systems*, Oxford Univ. Press, Oxford (1997).
[48] I.P. Cornfeld, S.V. Fomin and Ya.G. Sinai, *Ergodic Theory*, Springer-Verlag, Berlin (1982).
[49] H. Crauel, *Random Probability Measures on Polish Spaces*, Taylor & Francis, London (2002).
[50] H. Crauel and F. Flandoli, *Attractors for random dynamical systems*, Probab. Theory Related Fields **100** (1994), 365–393.
[51] S. Dahlke, *Invariant manifolds for products of random diffeomorphisms*, J. Dynam. Differential Equations **9** (2) (1997), 157–210.
[52] M. Denker and M. Gordin, *Gibbs measures for fibred systems*, Adv. Math. **148** (1999), 161–192.
[53] M. Denker, M. Gordin and S.-M. Heinemann, *On the relative variational principle for fibre expanding maps*, Ergodic Theory Dynam. Systems **22** (2002), 757–782.
[54] G. Domokos and D. Szász, *Ulam's scheme revisited: digital modeling of chaotic attractors via micro-perturbations*, Discrete Contin. Dynam. Systems **9** (2003), 859–876.
[55] M.D. Donsker and S.R.S. Varadhan, *On a variational formula for the principal eigenvalue for operators with maximum principle*, Proc. Natl. Acad. Sci. U.S.A. **72** (1975), 780–783.
[56] J. Duan, H. Gao and B. Schmalfuß, *Stochastic dynamics of a coupled atmosphere–ocean model*, Stochastics Dynamics **2** (2002), 357–380.
[57] R.M. Dudley, *Real Analysis and Probability*, Wadsworth & Brooks/Cole, Pacific Grove (1989).
[58] H.A. Dye, *On groups of measure preserving transformations I*, Amer. J. Math. **81** (1959), 119–159.
[59] J.-P. Eckmann and D. Ruelle, *Ergodic theory of chaos and strange attractors*, Rev. Modern Phys. **57** (3) (1985), 617–656.
[60] K.D. Elworthy, *Geometric aspect of diffusions on manifolds*, Lecture Notes in Math., Vol. 1362, Springer-Verlag (1988), 277–425.
[61] A. Fathi, M.R. Herman and J.C. Yoccoz, *A proof of Pesin's stable manifold theorem*, Lecture Notes in Math., Vol. 1007, Springer-Verlag (1983), 177–215.
[62] F. Flandoli, *Regularity theory and stochastic flows for parabolic SPDEs*, Gordon and Breach, Yverdon (1995).
[63] P. Frederickson, J.L. Kaplan, E.D. Yorke and J.A. Yorke, *The Lyapunov dimension of strange attractors*, J. Differential Equations **49** (1983), 183–207.
[64] M.I. Freidlin and A.D. Wentzell, *Random Perturbations of Dynamical Systems*, 2nd edn, Springer-Verlag, New York (1998).
[65] P. Gora, *Random composing of mappings, small stochastic perturbations and attractors*, Z. Warsch. Verw. Geb. **69** (1985), 137–160.
[66] V.M. Gundlach, *Random homoclinic orbits*, Random Comput. Dynam. **3** (1995), 1–33.
[67] V.M. Gundlach, *Isomorphic random Bernoulli shifts*, Colloq. Math. **84/85** (2) (2000), 327–344.
[68] V.M. Gundlach and Yu. Kifer, *Random hyperbolic systems*, Stochastic Dynamics, H. Crauel and M. Gundlach, eds, Springer-Verlag, New York (1999), 117–145.
[69] V.M. Gundlach and Yu. Kifer, *Expansiveness, specification, and equilibrium states for random bundle transformations*, Discrete Contin. Dynam. Systems **6** (2000), 89–120.
[70] M.W. Hirsch, *Differential Topology*, Springer, Berlin (1976).
[71] J. Hoffmann-Jorgensen, *The Theory of Analytic Spaces*, Lecture Notes Aarhus (1970).
[72] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, Tokyo (1981).
[73] V.A. Kaimanovich, Yu. Kifer and B.-Z. Rubshtein, *Boundaries and harmonic functions for random walks with random transition probabilities*, J. Theor. Probab. **17** (2004), 605–646.
[74] S. Kakutani, *Random ergodic theorems and Markoff processes with a stable distribution, Proc. 2nd Berkeley Sympos. on Math. Stat. and Probab.*, Univ. Calif. Press, Berkeley–Los Angeles (1951), 247–261.
[75] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge Univ. Press, Cambridge (1995).
[76] A. Katok and Yu. Kifer, *Random perturbations of transformations of an interval*, J. Anal. Math. **47** (1986), 193–237.

[77] A. Katok and J. M. Strelcyn, *Invariant Manifold, Entropy and Billiards; Smooth Maps with Singularities*, Lecture Notes in Math., Vol. 1222, Springer-Verlag (1986).
[78] Y. Katznelson and B. Weiss, *Commuting measure-preserving transformations*, Israel J. Math. **12** (1972), 161–173.
[79] K. Khanin and Y. Kifer, *Thermodynamic formalism for random transformations and statistical mechanics*, Amer. Math. Soc. Transl. Ser. 2 **171** (1996), 107–140.
[80] R.Z. Khasminskii, *The averaging principle for parabolic and elliptic differential equations and Markov processes with small diffusion*, Theor. Probab. Appl. **8** (1963), 1–21.
[81] Yu. Kifer, *On small random perturbations of some smooth dynamical systems*, Math. USSR Izv. **8** (5) (1974), 1083–1107.
[82] Yu. Kifer, *Ergodic Theory of Random Transformations*, Birkhäuser, Boston (1986).
[83] Yu. Kifer, *Random Perturbations of Dynamical Systems*, Birkhäuser, Boston (1988).
[84] Yu. Kifer, *A note on integrability of C^r-norms of stochastic flows and applications*, Lecture Notes in Math., Vol. 1325, Springer-Verlag (1988).
[85] Yu. Kifer, *Attractors via random perturbations*, Comm. Math. Phys. **121** (1989), 445–455.
[86] Yu. Kifer, *A discrete-time version of the Wentzell–Freidlin theory*, Ann. Probab. **18** (1990), 1676–1692.
[87] Yu. Kifer, *Large deviations in dynamical systems and stochastic processes*, Trans. Amer. Math. Soc. **321** (1990), 505–524.
[88] Yu. Kifer, *Principal eigenvalues, topological pressure, and stochastic stability of equilibrium states*, Israel J. Math. **70** (1990), 1–47.
[89] Yu. Kifer, *Large deviations for random expanding maps*, Lyapunov Exponents, L. Arnold, H. Crauel and J.-P. Eckmann, eds, Springer-Verlag (1991), 178–186.
[90] Yu. Kifer, *A variational approach to the random diffeomorphism type random perturbations of a hyperbolic diffeomorphism*, Mathematical Physics X, K. Schüdgen, ed., Springer-Verlag (1992), 334–340.
[91] Yu. Kifer, *Equilibrium states for random expanding transformations*, Random Comput. Dynam. **1** (1992), 1–31.
[92] Yu. Kifer, *Fractals via random iterated function systems and random geometric constructions*, Fractal Geometry and Stochastics, C. Bandt, S. Graf and M. Zähle, eds, Birkhäuser, Basel (1995), 145–164.
[93] Yu. Kifer, *Multidimensional random subshifts of finite type and their large deviations*, Probab. Theory Related Fields **102** (1995), 223–248.
[94] Yu. Kifer, *Perron–Frobenius theorem, large deviations, and random perturbations in random environments*, Math. Z. **222** (1996), 677–698.
[95] Yu. Kifer, *Fractal dimensions and random transformations*, Trans. Amer. Math. Soc. **348** (1996), 2003–2038.
[96] Yu. Kifer, *Large deviations for paths and configurations counting*, Ergodic Theory of $\mathbb{Z}^d$-Actions, M. Pollicott and K. Schmidt, eds, London Math. Soc. Lecture Notes, Vol. 228, Cambridge Univ. Press (1996), 415–432.
[97] Yu. Kifer, *Computations in dynamical systems via random perturbations*, Discrete Contin. Dynam. Systems **3** (1997), 457–476.
[98] Yu. Kifer, *Limit theorems for random transformations and processes in random environments*, Trans. Amer. Math. Soc. **350** (1998), 1481–1518.
[99] Yu. Kifer, *On the topological pressure for random bundle transformations*, Rohlin's Memorial Volume, V. Turaev and A. Vershik, eds, Amer. Math. Soc. Transl., Vol. 202 (2001), 197–214.
[100] Yu. Kifer, *Random f-expansions*, Proc. Sympos. Pure Math. **69** (2001), 385–407.
[101] Yu. Kifer, *"Random" random matrix products*, J. Anal. Math. **83** (2001), 41–88.
[102] Y. Kifer and B. Weiss, *Generating partitions for random transformations*, Ergodic Theory Dynam. Systems **22** (2002), 1813–1830.
[103] Y. Kifer and Y. Yomdin, *Volume growth and topological entropy for random transformations*, Dynamical Systems, J.C. Alexander, ed, Lecture Notes in Math., Vol. 1342, Springer-Verlag, New York (1988), 361–373.
[104] M. Klünger, *Periodicity and Sharkovsky's theorem for random dynamical systems*, Stochastics Dynamics **1** (2001), 299–338.
[105] W. Krieger, *On entropy and generators of measure-preserving transformations*, Trans. Amer. Math. Soc. **149** (1970), 453–464; Erratum: **168** (1972), 519.

[106] A. Kufner, O. John and S. Fučik, *Function Spaces*, Academia, Publishing House of the Czechoslovak Academy of sciences, Prague (1977).
[107] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge Univ. Press, Cambridge (1990).
[108] F. Ledrappier and J.-M. Strelcyn, *A proof of the estimation from below in Pesin's entropy formula*, Ergodic Theory Dynam. Systems **2** (1982), 203–219.
[109] F. Ledrappier and P. Walters, *A relativized variational principle for continuous transformations*, J. London Math. Soc. (2) **16** (1977), 568–576.
[110] F. Ledrappier and L.-S. Young, *The metric entropy of diffeomorphisms. Part I: Characterization of measures satisfying Pesin's formula. Part II: Relations between entropy, exponents and dimension*, Ann. of Math. **122** (1985), 509–574.
[111] F. Ledrappier and L.-S. Young, *Dimension formula for random transformations*, Comm. Math. Phys. **117** (1988), 529–548.
[112] F. Ledrappier and L.-S. Young, *Entropy formula for random transformations*, Probab. Theory Related Fields **80** (1988), 217–240.
[113] F. Ledrappier and L.-S. Young, *Stability of Lyapunov exponents*, Ergodic Theory Dynam. Systems **11** (1991), 469–484.
[114] M. Lemańczyk, J.-P. Thouvenot and B. Weiss, *Relative discrete spectrum and joinings*, Monatsh. Math. **137** (2002), 57–75.
[115] P.-D. Liu, *Random perturbations of Axiom A sets*, J. Statist. Phys. **90** (1998), 467–490.
[116] P.-D. Liu, *Entropy formula of Pesin type for noninvertible random dynamical systems*, Math. Z. **230** (1999), 201–239.
[117] P.-D. Liu, *Dynamics of random transformations: smooth ergodic theory*, Ergodic Theory Dynam. Systems **21** (2001), 1279–1319.
[118] P.-D. Liu and M. Qian, *Smooth Ergodic Theory of Random Dynamical Systems*, Lecture Notes in Math., Vol. 1606, Springer-Verlag, Berlin (1995).
[119] P.-D. Liu, M. Qian and F.-X. Zhang, *Entropy formula of Pesin type for one-sided stationary random maps*, Ergodic Theory Dynam. Systems **22** (2002), 1831–1844.
[120] P.-D. Liu and J.-Sh. Xie, *Dimension of hyperbolic measures for random diffeomorphisms*, Trans. Amer. Math. Soc., to appear.
[121] P.-D. Liu and Y. Zhao, *Large deviations in random perturbations of Axiom A basic sets*, J. London Math. Soc. (2) **68** (2003), 148–164.
[122] C. Liverani, *Rigorous numerical investigation of the statistical properties of piecewise expanding maps. A feasibility study*, Nonlinearity **14** (2001), 463–490.
[123] C. Maes and A. Van Moffaert, *Stochastic stability of weakly coupled lattice maps*, Nonlinearity **10** (1997), 715–730.
[124] R. Mañé, *A proof of Pesin's formula*, Ergodic Theory Dynam. Systems **1** (1981), 95–102.
[125] R. Mañé, *Ergodic Theory and Differentiable Dynamics*, Springer-Verlag (1987).
[126] F. Martinelli, L. Sbano and E. Scappola, *Small random perturbations of dynamical systems: recursive multiscale analysis*, Stochastics Stochastics Rep. **49** (1994), 253–272.
[127] J. Mather, *Characterization of Anosov diffeomorphisms*, Indag. Math. **30** (1968), 479–483.
[128] R.J. Metzger, *Stochastic stability for contracting Lorenz maps and flows*, Comm. Math. Phys. **212** (2000), 277–296.
[129] M. Misiurewicz, *Topological conditional entropy*, Studia Math. **55** (1976), 175–200.
[130] T. Ohno, *Asymptotic behaviors of dynamical systems with random parameters*, Publ. RIMS Kyoto Univ. **19** (1983), 83–98.
[131] V.I. Oseledec, *A multiplicative ergodic theorem, Liapunov characteristic numbers for dynamical systems*, Trans. Moscow Math. Soc. **19** (1968), 197–221.
[132] W. Parry, *Entropy and Generators in Ergodic Theory*, Benjamin, New York (1969).
[133] Ya.B. Pesin, *Families of invariant manifolds corresponding to nonzero characteristic exponents*, Math. USSR Izv. **40** (6) (1976), 1261–1350.
[134] Ya.B. Pesin, *Lyapunov characteristic exponents and smooth ergodic theory*, Russ. Math. Surveys **32** (4) (1977), 55–114.

[135] L.S. Pontryagin, A.A. Andronov and A.A. Vitt, *On statistical consideration of dynamical systems*, J. Experiment. Theor. Phys. **3** (1) (1933), 165–180 (in Russian).
[136] C.C. Pugh, *The $C^{1+\alpha}$ hypothesis in Pesin theory*, I.H.E.S. Publ. Math. **59** (1984), 143–161.
[137] C.C. Pugh and M. Shub, *Ergodic attractors*, Trans. Amer. Math. Soc. **312** (1) (1989), 1–54.
[138] M. Qian and J.-Sh. Xie, *Entropy formula for random dynamical systems—Relations between entropy, exponents and dimension*, Ergodic Theory Dynam. Systems, to appear.
[139] A. Quas, *On representation of Markov chains by random smooth maps*, Bull. London Math. Soc. **23** (1991), 487–492.
[140] A. Quas, *Representation of Markov chains on tori*, Random Comput. Dynam. **1** (1992–1993), 261–276.
[141] V.A. Rohlin, *On the fundamental ideas of measure theory*, Amer. Math. Soc. Transl. (1) **10** (1962), 1–52, Translated from Mat. Sbornik (N.S.) **25** (67) (1949), 107–150.
[142] V.A. Rohlin, *Lectures on the entropy theory of measure-preserving transformations*, Russian Math. Surveys **22** (5) (1967), 1–52.
[143] B.-Z. Rubshtein, *A central limit theorem for conditional distributions*, Convergence in Ergodic Theory and Probability, V. Bergelson, P. March and J. Rosenblatt, eds, de Gruyter, Berlin (1996), 373–380.
[144] D.J. Rudolph and B. Weiss, *Entropy and mixing for amenable group actions*, Ann. of Math. **151** (2000), 1119–1150.
[145] D. Ruelle, *A measure associated with Axiom A attractors*, Amer. J. Math. **98** (1976), 619–654.
[146] D. Ruelle, *An inequality for the entropy of differentiable maps*, Bol. Soc. Brasil Math. **9** (1978), 83–87.
[147] D. Ruelle, *Ergodic theory of differentiable dynamical systems*, I.H.E.S. Publ. Math. **50** (1979), 27–58.
[148] D. Ruelle, *Small random perturbations of dynamical systems and the definition of attractors*, Comm. Math. Phys. **82** (1981), 137–151.
[149] D. Ruelle, *Characteristic exponents and invariant manifolds in Hilbert space*, Ann. of Math. **115** (1982), 243–290.
[150] D. Ruelle, *Characteristic exponents for a viscous fluid subjected to time dependent forces*, Comm. Math. Phys. **93** (1984), 285–300.
[151] D. Ruelle, *An extension of the theory of Fredholm determinants*, I.H.E.S. Publ. Math. **72** (1990), 175–193.
[152] D. Ruelle, *Positivity of entropy production in the presence of a random thermostat*, J. Statist. Phys. **86** (1996), 935–952.
[153] D. Ruelle, *Random smooth dynamical systems*, Lecture Notes in Rutgers (1996).
[154] D. Ruelle and M. Shub, *Stable manifolds for maps*, Lecture Notes in Math., Vol. 819, Springer-Verlag, (1980), 389–392.
[155] K.R. Schenk-Hoppé, *Random dynamical systems in economics*, Stochastics Dynamics **1** (2001), 63–83.
[156] Ya.G. Sinai, *Gibbs measures in ergodic theory*, Russian Math. Surveys **27** (4) (1972), 21–69.
[157] J.-P. Thouvenot, *Quelques properties des systemes dynamiques que se decomposent en un produit de deux systemes dont l'un est un schema de Bernoulli*, Israel J. Math. **21** (1975), 177–207.
[158] S. Ulam, *Problems in Modern Mathematics*, Interscience, New York (1960).
[159] S.M. Ulam and J. von Neumann, *Random ergodic theorems*, Bull. Amer. Math. Soc. **51** (1945), 660.
[160] P. Walters, *Equilibrium states for β-transformations and related transformations*, Math. Z. **159** (1978), 65–88.
[161] P. Walters, *Invariant measures and equilibrium states for some mappings which expand distances*, Trans. Amer. Math. Soc. **236** (1978), 121–153.
[162] P. Walters, *An Introduction to Ergodic Theory*, Springer-Verlag, New York (1982).
[163] E. Weinan, K. Khanin, A. Mazel and Ya. Sinai, *Invariant measures for Burgers equation with stochastic forcing*, Ann. of Math. **151** (2000), 877–960.
[164] A.D. Wentzell and M.I. Freidlin, *On small random perturbations of dynamical systems*, Russian Math. Surveys **25** (1) (1970), 1–56.
[165] L.-S. Young, *Dimension, entropy and Lyapunov exponents*, Ergodic Theory Dynam. Systems **2** (1982), 109–124.
[166] L.-S. Young, *Stochastic stability of hyperbolic attractors*, Ergodic Theory Dynam. Systems **6** (1986), 311–319.
[167] L.-S. Young, *Ergodic theory of chaotic dynamical systems*, XIIth International Congress of Mathematical Physics (ICMP'97) (Brisbane), Internat. Press, Cambridge, MA (1999), 131–143.

This page intentionally left blank

CHAPTER 6

An Introduction to Veech Surfaces

Pascal Hubert

Institut de Mathématiques de Luminy, 163, avenue de Luminy, case 907, 13288 Marseille, cedex 09, France
E-mail: *hubert@iml.univ-mrs.fr*

Thomas A. Schmidt

Department of Mathematics, Kidder 362, Oregon State University, Corvallis, OR 97331-4605, USA
E-mail: *toms@math.orst.edu*

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

This page intentionally left blank

1. Introduction to Veech surfaces

We give a gentle introduction to the basics of Veech surfaces, with an emphasis on the Veech Dichotomy, followed by a sketch of the present state of the literature. These notes arose from lectures for a summer school held at the Institute de Mathématiques de Luminy in June 2003. We thank the participants, especially Jayadev Athreya who prepared an initial set of notes, and other speakers for various comments.

1.1. *From billiards to flat surfaces*

1.1.1. *Billiards* A seemingly innocuous problem is to analyze the billiard flow on rational-angle Euclidean polygons. That is, given a polygon whose angles are rational multiples of π, consider the trajectories of an ideal point mass, that moves at a constant speed without friction in the interior of the polygon and enjoys elastic collisions with the boundary—angles of incidence and reflection are equal.

For more on billiards and related matters, see [45] and [3] as well as the chapters of Eskin, Forni and Masur.

1.1.2. *Unfolding* We now describe the unfolding process for rational billiards. Given a billiard trajectory (that avoids the vertices) beginning at a side of a rational angle polygon, this yields a surface. The process has arisen in various guises, see in particular Katok and Zemlyakov [29].

Given a collision with a side we reflect the *polygon* along the side, obtaining a mirror image of the original polygon, on which the billiard now continues in its original direction, instead of reflecting off the side. Continuing this process *ad infinitum*, we would obtain a laundry line (a ray in the plane), along which various copies of the polygon are strung. But, since our polygon has rational angles, there are only finitely many possible angles of incidence of our chosen trajectory with these copies. Thus, the billiard eventually exits a copy of the polygon in a side that is parallel with the initial side. We now identify these sides by translation; we continue this process, considering any unpaired side that the billiards meets as the new initial side. The result is a new polygon with various 'opposite' sides identified; on this 'flat surface', the billiard moves along straight line segments, up to translation.

The 1-form dz on the complex plane induces a 1-form on our surface. There is a unique complex structure on the surface such that this 1-form is holomorphic. The process thus results in a Riemann surface with a distinguished Abelian differential (that is, holomorphic 1-form). There is a close relationship between the flows on the flat surface and various properties of the 1-form.

Unfolding: Two examples. First, let us consider billiards in the unit square, see Figure 1. Suppose our billiard trajectory starts near the bottom left corner (the origin) and has slope $1 > s > 0$. Thus it collides initially with the right side. We reflect about this side to get a mirror image of the square upon which our trajectory continues with this slope. The next side it hits is the top of the new (right) square; reflecting about that side we get a third square that sits above the second (bottom right) square. Continuing this procedure, we

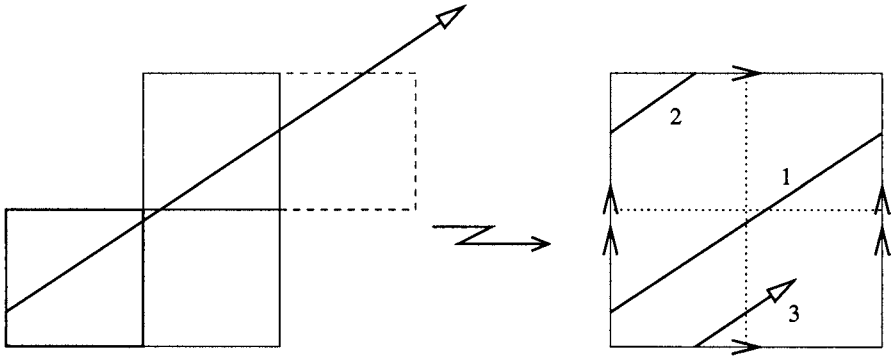


Fig. 1. Unfolding; square table to torus surface.

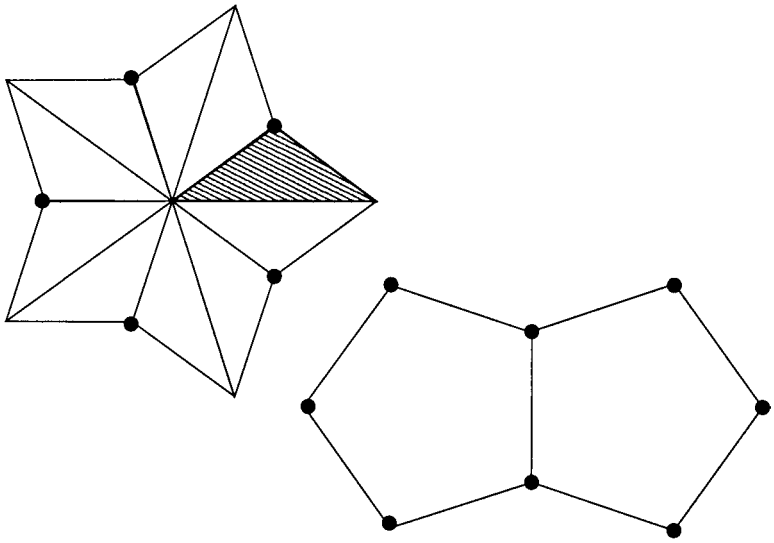


Fig. 2. Surface from triangle; same translation surface. (Identify parallel sides by translation.)

eventually end up with four copies of our original square; we can appropriately translate one of the copies so as to form a larger square. As an exercise, the reader should now check that we can follow all billiard paths within this larger square, if we identify opposite sides by translation. Thus, a torus is formed. Each trajectory of the billiard flow is mapped to a trajectory for the linear flow in the same direction on the torus.

If we now take the isosceles triangle with angles $(\pi/5, \pi/5, 3\pi/5)$ as our initial table, the unfolding process yields a star-shaped polygon with opposite sides identified, see Figure 2.

(The reader should note that differing billiard trajectories give apparently different polygons, but should show that these differences are accounted for by the translations of the various identified sides!) This is a compact, oriented topological surface. An easy Euler characteristic calculation shows that it has genus two.

The identifications of the sides lead to interesting identifications of the vertices. While the "outside" vertices of the stellated pentagon collapse to a point with angle 2π, the "inside" vertices yield a point with total angle 6π! (This phenomenon did not arise in our first example—the large square with its sides identified—as there the vertices are identified to a single point of angle 2π.) Indeed, a Gauss–Bonnet calculation will now confirm that our surface is of genus two.

This difference between our genus two and genus one examples reflects the fact that while the torus is naturally flat (its universal cover is the Euclidean plane $\mathbb{R}^2$), a genus 2 surface is naturally hyperbolic (universal cover $\mathbb{H}^2$), and *cannot be forced to be flat*.

1.1.3. *From 1-forms to surfaces* Now consider a pair (X, ω), a Riemann surface X with a holomorphic 1-form ω. Locally (i.e., in each coordinate patch) $\omega = f(w)\,dw$. Given a point $p_0 \in X$, we define new coordinates by the map

$$z(p) = \int_{p_0}^{p} \omega.$$

In these coordinates, $\omega = dz$ locally.

If we change base points in some small patch, then our coordinates change by a translation:

$$c := \int_{p_0}^{p} \omega - \int_{p_1}^{p} \omega = \int_{p_0}^{p_1} \omega.$$

Since c does not depend on p, our transition maps are of the form $z \mapsto z + c$. Thus the pair (X, ω) gives a structure which is reasonably called a *translation surface*.

We need to take care in the above discussion. At a zero of multiplicity k, locally we have $\omega = z^k\,dz$, hence $\omega = d(z^{k+1}/(k+1))$. That is, instead of the surface locally resembling the complex plane $\mathbb{C}$ (as it does away from the zeros), at a zero the surface instead locally resembles the $(k+1)$-fold cover of $\mathbb{C}$ via the map $z \mapsto z^{k+1}$. Thus, the total angle around the zero is $2\pi(k+1)$.

By your favorite general theorem about Riemann surfaces (either Gauss–Bonnet or Riemann–Roch), the total number of zeros (counting multiplicity) of the Abelian differential ω is $2g-2$, where g is the genus of the surface X.

Fixing the orders of all zeros, we call the associated subset of translation surfaces a *stratum*. Thus, we have a stratum for each integer partition of $2g-2$. See [2] for more discussion of these matters.

1.1.4. $\mathrm{SL}(2,\mathbb{R})$*-action and Veech groups* The group $\mathrm{SL}(2,\mathbb{R})$ acts on the space of translation surfaces: a pair (X, ω) is given by its charts, with coordinate functions to the complex plane (and all transition maps are translations). We will now consider $\mathbb{C}$ with its natural structure as the real plane. Given a matrix $A \in \mathrm{SL}(2,\mathbb{R})$, the new point $A \circ (X, \omega)$ is the surface whose charts are the charts for (X, ω), with coordinate functions post-composed with the linear action of A on $\mathbb{R}^2$. This action preserves orders of zeros, it thus preserves

each stratum. Note that an element of $\mathrm{SO}(2,\mathbb{R})$ acts on a translation surface as a (piecewise) rotation; this action corresponds to multiplying ω by a non-zero complex number of norm one.

We denote the stabilizer of (X,ω) under the action of $\mathrm{SL}(2,\mathbb{R})$ by $\mathrm{SL}(X,\omega)$. Recall that $\mathrm{SL}(2,\mathbb{R})$ does not act faithfully on the upper half-plane; it is the projective group $\mathrm{PSL}(2,\mathbb{R})$ that does so. We define the *Veech group*, $\mathrm{PSL}(X,\omega)$, to be the image of $\mathrm{SL}(X,\omega)$ in $\mathrm{PSL}(2,\mathbb{R})$.

Examples revisited. For the torus, we consider the maps

$$(x,y)\mapsto (x, x+y \bmod 1)$$

and

$$(x,y)\mapsto (x+y \bmod 1, y).$$

These are *Dehn twists* about the curves corresponding to the x- and y-axes, respectively. Their derivatives are given by the matrices $A_1=\left(\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right)$ and $A_2=\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$, respectively. We have that $A_i\in \mathrm{SL}(\mathbb{C}/\mathbb{Z}^2, dz)=\mathrm{SL}(2,\mathbb{Z})$. The reader should verify this last equality!

For our genus two example, we can decompose the surface into two vertical cylinders of height and width (h_1,w_1) and (h_2,w_2), see Figure 3. On each cylinder we can define a Dehn twist via

$$(x,y)\mapsto \big(x, y+\mu^{-1}x \bmod h\big),$$

where following tradition, the *modulus* of the cylinder is $\mu=w/h$. Note that each Dehn twist is constant on the vertical sides of the corresponding cylinder; we can certainly glue

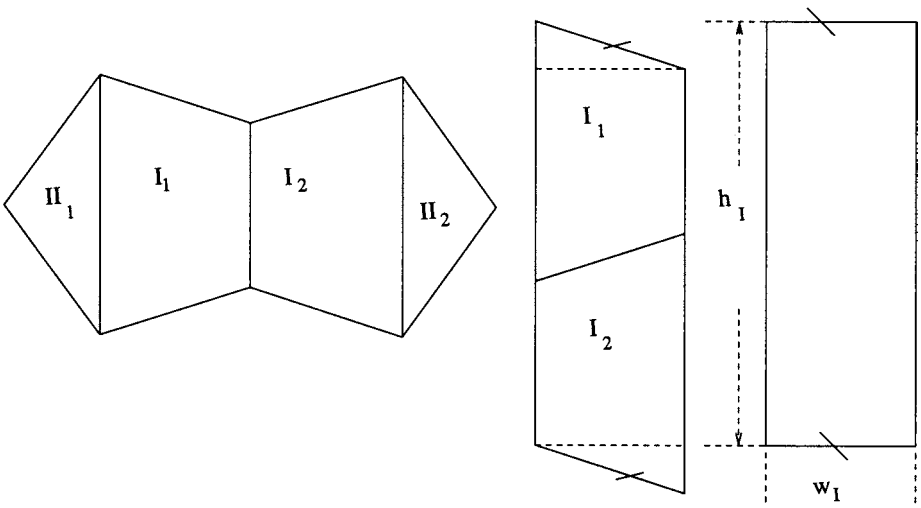


Fig. 3. Vertical cylinders.

them together to get a globally defined function. But, in order to preserve our flat structure, a diffeomorphism must have its derivative (off of the singularities) constant in our coordinates. We call such maps *affine diffeomorphisms*, and denote the group that they form by $\mathrm{Aff}(X, \omega)$.

Thus, in order to construct an affine diffeomorphism of the surface from these Dehn twists we must be able to take some power of each twist so that the resulting derivatives agree. For this, we must have $r\mu_1 = s\mu_2$ for some integers r and s; in words: *the moduli of the cylinders must be rationally related*. In this example, we get very lucky and the moduli are in fact the same. The reader is encouraged to check this trigonometry!

This stellated pentagon has its Veech group generated by an element of order five—the obvious rotation—and an element of order two. Can you find 'the' element of order two? On a related surface—the Golden Cross, see say [25] or [35]—it acts as a square root of the famous "hyperelliptic involution" of the surface.

We must emphasize that it is very rare that the Dehn twists on cylinders match up to give a global affine diffeomorphism!

1.2. *The Veech Dichotomy*

Recall the theorem of Weyl for geodesic flow on the torus: in any rational direction θ, all orbits are closed, whereas the flow in any irrational direction is uniquely ergodic: it is ergodic with respect to a unique non-atomic measure, which is (induced by) Lebesgue measure. Veech proved an analogous result for a class of particularly nice surfaces.

We can define directions θ of flow on a given translation surface (X, ω): use the coordinate charts to pull-back from the real plane the straight lines of direction θ. The directional flow F_θ is the map from $X \times \mathbb{R}^+$ to X sending pairs (x, t) to x', where x' is length t from x along a line segment in the direction θ. Of course, the true definition of F_θ recognizes that the translation surface has singularities! There is a theorem of Kerckhoff–Masur–Smillie [31] that for a fixed translation surface (X, ω), for almost every direction θ the flow F_θ is uniquely ergodic. See [2] for related discussion.

We say that F_θ is periodic if the surface decomposes into a finite number of cylinders in the direction θ, and furthermore these cylinders have pairwise commensurable moduli: $\mu_i/\mu_j \in \mathbb{Q}$. Note that it is not necessary that the actual period lengths of the cylinders be the same, nor even commensurable—as the vertical flow on our genus two example already shows!

Recall that the Veech group of (X, ω) is defined such that it acts on the hyperbolic plane. We say that such a group is a *lattice* if the quotient space under this action has finite (induced) hyperbolic area. In this setting, we also say that $\mathrm{SL}(X, \omega)$ is a lattice. (There are several ways of defining the term lattice; this definition works in our setting.)

THEOREM 1 (Veech Dichotomy[1]). *Let (X, ω) be a translation surface. Suppose $\mathrm{SL}(X, \omega)$ is a lattice in $\mathrm{SL}(2, \mathbb{R})$. Then for each direction θ, the flow F_θ is either periodic or uniquely ergodic.*

[1]The authors of [3] have asked us to point out that this clarifies their statement of the Veech Dichotomy.

If $\mathrm{SL}(X,\omega)$ is a lattice, then (X,ω) is called a Veech surface. The theorem states that a Veech surface has dynamical properties similar to the touchstone surface, the square torus. In what follows, we will sketch a proof—coming from Veech's original proof [48], especially as adapted by Vorobets [51].

1.3. *Structure of Veech groups*

A *separatrix* is a geodesic line emanating from a singularity, a *saddle connection* is a separatrix connecting singularities (with no singularities on its interior). To each saddle connection we can associate a *holonomy vector*: we 'develop' the saddle connection to the plane by using local coordinates, the difference vector defined by the planar line segment is the holonomy vector.

1.3.1. *Discreteness* The following theorem seems to be in the folklore of the subject, our proof is modeled on that of Proposition 3.1 of [51]. See [2] for a second proof of this fundamental result.

PROPOSITION 1. *Let (X,ω) be a translation surface. Then the set of holonomy vectors of saddle connections, $V_{\mathrm{sc}}(X,\omega)$, is discrete in $\mathbb{R}^2$.*

SKETCH OF PROOF. We assume that the surface does admit singularities. Since there are only finitely many of these singularities, it is clear that every point p of the surface admits some positive $\epsilon(p)$ such that there is a punctured disk of radius $\epsilon(p)$ centered at p that is void of singularities.

Choose any vector $v\in\mathbb{R}^2$. At each singularity, form every geodesic ray of holonomy v. Each ray is in general a sequence of saddle connections followed by a separatrix. Since there are only finitely many singularities and the total angle at any of these is finite, there are only finitely many of these geodesic rays. Let $\epsilon=\min(\epsilon(p))$, where p runs over the endpoints of the paths of these geodesic rays.

Clearly, there is no saddle connection ending within the punctured ϵ-disk about the end point of any of our geodesic rays. But, this means that v cannot be the limit of holonomy vectors of saddle connections. Since v was arbitrary, we find that $V_{\mathrm{sc}}(X,\omega)$ is discrete. □

1.3.2. *Non-cocompactness* Again following Vorobets, one has an easy proof of the following result, originally due to Veech [48].

LEMMA 2. *Let (X,ω) be a translation surface. Then the group* $\mathrm{SL}(X,\omega)$ *is a discrete subgroup of* $\mathrm{SL}(2,\mathbb{R})$.

SKETCH OF PROOF. Any $A\in\mathrm{SL}(2,\mathbb{R})$ acts so as to send saddle connections of (X,ω) to saddle connections of $A\circ(X,\omega)$. Let $\{A_n\}\subset\mathrm{SL}(X,\omega)$ be a sequence approaching the identity (where $\mathrm{SL}(2,\mathbb{R})$ has its usual topology), $A_n\to I$. Let $v,w,\in V_{\mathrm{sc}}(X,\omega)$ be linearly independent. Then $A_nv\to v$ and $A_nw\to w$. By discreteness of $V_{\mathrm{sc}}(X,\omega)$, for n sufficiently large, $A_nv=v$ and $A_nw=w$. But v and w are linearly independent; they form

a basis for $\mathbb{R}^2$. Hence, for all large n we have that $A_n = I$. We conclude that $\mathrm{SL}(X,\omega)$ is discrete. □

Standard terminology: a discrete subgroup of $\mathrm{SL}(2,\mathbb{R})$ is a *Fuchsian group*.

Similarly, $\mathrm{SL}(X,\omega)$ is never cocompact: $\mathrm{SL}(X,\omega)$ being *cocompact* would simply mean that in the natural quotient topology $\mathrm{SL}(X,\mathbb{R})/\mathrm{SL}(X,\omega)$ is compact. We disprove this by finding a continuous (non-negative) real valued function on $\mathrm{SL}(2,\mathbb{R})$ that is constant on cosets, but has no minimum value.

Consider the function $\Lambda : \mathrm{SL}(2,\mathbb{R}) \to \mathbb{R}^+$, given by $A \mapsto l(A \circ (X,\omega))$, where $l(X,\omega)$ denotes the length of the shortest saddle connection. If $\mathrm{SL}(X,\omega)$ were cocompact, the function Λ would have a minimum, say $\alpha > 0$.

But, take any saddle connection. We can normalize by rotating (X,ω) so that this saddle connection is in the vertical direction; we can send the length to zero via the *geodesic flow*: $g_t := \left(\begin{smallmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{smallmatrix}\right)$. Since both rotation and geodesic flow are realized in $\mathrm{SL}(2,\mathbb{R})$, we clearly have a contradiction to the minimality of α. We conclude that $\mathrm{SL}(X,\omega)$ is not cocompact.

1.3.3. *Parabolic elements* It is a well-known fact for Fuchsian groups that any non-cocompact lattice must have a parabolic element; see, say, [28]. Conjugating the group, the parabolic fixed point may be taken to be infinity, the parabolic then acts as a translation; the quotient can be informally envisioned as having a cone with missing point at infinity, a *cusp*.

The following is a restatement of Lemma 3.7 of [51].

LEMMA 3. *Let $\Gamma \subset \mathrm{SL}(2,\mathbb{R})$ be a non-cocompact lattice, such that $g_t\Gamma$ is divergent* (*i.e., leaves every compact set*) *in $\mathrm{SL}(2,\mathbb{R})/\Gamma$. Then there is an $\alpha \neq 0$ with $h_\alpha = \left(\begin{smallmatrix} 1 & 0 \\ \alpha & 1 \end{smallmatrix}\right) \in \Gamma$.*

Thus, if Γ is a lattice, the only way a trajectory of the geodesic flow on $\mathrm{SL}(2,\mathbb{R})/\Gamma$ can escape to infinity is via a cusp.

1.3.4. *Affine diffeomorphisms and Veech groups* In fact, $\mathrm{SL}(X,\omega)$ is the group of derivatives of orientation-preserving affine diffeomorphisms. To sketch a proof of this, we take (X,ω) normalized such that X has area one with respect to the area form, $d\lambda$, induced by ω. Let ϕ be an orientation-preserving affine diffeomorphism of (X,ω). The derivative of ϕ is its Jacobian derivative in the usual sense. With the real structure of the translation surface, this derivative is a constant (off of the singularities) 2×2 real matrix. Thus

$$1 = \int_X d\lambda = \int_{\phi^{-1}(X)} \big|\operatorname{Jac}(\phi)\big|\, d\lambda = \big|\operatorname{Jac}(\phi)\big|.$$

Thus, the derivative of ϕ is of determinant one. In brief: Area preserving implies determinant one. (By the way, it is a significant fact that the "derivative" map has finite kernel in $\mathrm{Aff}(X,\omega)$, [48]: any ϕ whose derivative is the identity is certainly an automorphism of the complex structure of X, in genus greater than one, there are only finitely many of these.)

1.4. *Proof of the Veech Dichotomy*

Rotations leave the underlying structure unchanged, we can thus suppose that the vertical direction is non-uniquely ergodic. This is only possible if $g_t\omega$ is divergent, that is if $g_t \mathrm{SL}(X,\omega)$ leaves every compact set of the quotient $\mathrm{SL}(2,\mathbb{R})/\mathrm{SL}(X,\omega)$; this follows from *Masur's criterion*, see Theorem 3 of [2] and the sketch of its proof, given in §3 there. This criterion is key to the proof; it is closely related to a combinatorial criterion of Boshernitzan for non-unique ergodicity of an interval exchange transformation [10,47] and the discussion in [2].

By hypothesis, $\mathrm{SL}(X,\omega)$ is a lattice; by our basic facts, it has a parabolic element. In fact, since the vertical direction is divergent, there is a parabolic element of the type given in Lemma 3. The next lemma shows that the existence of a parabolic element implies important geometric information about the translation surface (X,ω).

LEMMA 4. *Let h_α be as above. If $h_\alpha \in \mathrm{SL}(X,\omega)$, then X decomposes into a finite number of vertical cylinders of moduli $\mu_i = \frac{p_i}{q_i}\alpha$, $p_i, q_i \in \mathbb{Z}$.*

PROOF. Denote the affine map with derivative h_α by ϕ. Let Σ be the set of singular points on (X,ω). Then, ϕ acts by permutation on Σ. At each $p_i \in \Sigma$, we have outgoing *separatrices*—geodesics emanating from the singularities, see Figure 4. Let $\{L_1, L_2, \ldots, L_k\}$ denote the set of outgoing separatrices in the vertical direction. Then ϕ also acts on this set by permutation; by passing to a power $\psi = \phi^n$, we can assume that ψ fixes both every singularity and each of the L_i.

The affine diffeomorphism ψ acts up to translation exactly as its derivative; the derivative fixes the vertical direction, and hence ψ restricted to any L_i acts as a pure translation. Since a translation with a fixed point can only be the identity, we conclude that ψ fixes each vertical separatrix L_i pointwise.

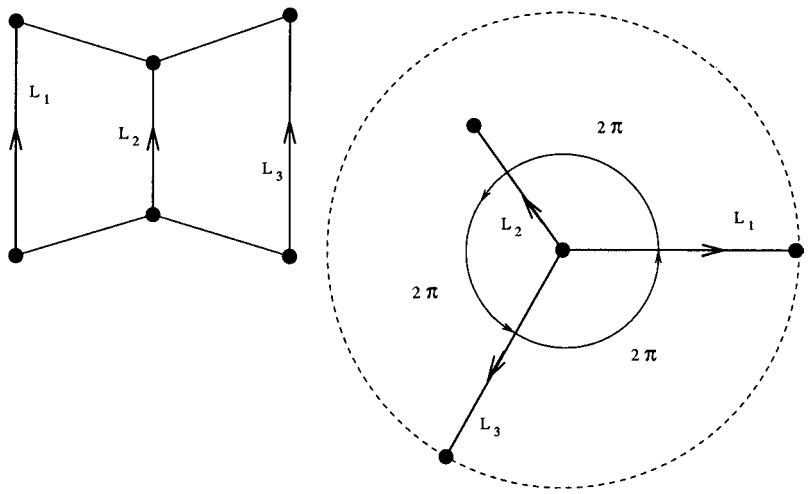


Fig. 4. Vertical saddle connections. (Three outgoing, giving also three incoming.)

We claim that each L_i is in fact an outgoing saddle connection. Indeed, if a separatrix L is not a saddle connection, then it must in fact be dense in some open subset U of X. But if L_i is dense in some U, then ψ is identity on U; since $h_\alpha \neq I$, this leads to a contradiction.

Next, we claim that ALL vertical leaves are closed. Consider an arbitrary point $p \in X$ not lying on any of our L_i. Let $\mathcal{F}_t$ denote the vertical flow on X. If $\mathcal{F}_t(p)$ is not closed, then it is dense in some minimal component—see the proof of Theorem 1.8 of [3]. On the other hand, $\mathcal{F}_t(p)$ does not encounter any singularity, as we have assumed that p is not on any of the L_i. Hence, p flows in parallel with the L_i; in particular, the distance of any $\mathcal{F}_t(p)$ to the L_i cannot be made arbitrarily small. Thus, $\mathcal{F}_t(p)$ is certainly not dense; it must be closed.

We now have a cylinder decomposition of (X, ω) in the vertical direction. The powers of the affine Dehn twist of a given vertical cylinder are of derivative $\left(\begin{smallmatrix} 1 & 0 \\ k\mu & 1 \end{smallmatrix}\right)$ where μ is the modulus. Since $d\psi = \left(\begin{smallmatrix} 1 & 0 \\ n\alpha & 1 \end{smallmatrix}\right)$ is constant, the moduli of the various vertical cylinders are all rational multiples of α. □

So we have the Veech Dichotomy: if the flow is not uniquely ergodic, it gives a divergent trajectory in $\mathbb{H}/\mathrm{PSL}(X, \omega)$, thus there is a parabolic element in $\mathrm{SL}(X, \omega)$, and we can then decompose our surface into cylinders with commensurable moduli.

REMARK 1. Note that the theorem leads to a simple necessary condition for a surface to be Veech: in each direction with a cylinder decomposition, the moduli of the cylinders must be commensurable. That is, if there are two cylinders with moduli $\mu_1, \mu_2, \mu_1/\mu_2 \notin \mathbb{Q}$, we are not on a Veech surface. In fact, a Veech surface has a cylinder decomposition in the direction of any of its saddle connections.

Consider our basic example, the square torus. In this case, $\mathrm{SL}(X, \omega) = \mathrm{SL}(2, \mathbb{Z})$; it is thus a lattice, and Veech's result recovers the result we mentioned as a theorem of Weyl.

1.5. *Arithmeticity*

1.5.1. *Theorem of Gutkin and Judge* For surfaces that can be tiled by squares—called, most simply, *square-tiled surfaces*—we have that $\mathrm{SL}(X, \omega)$ is *commensurate* to $\mathrm{SL}(2, \mathbb{Z})$ (the groups share a common finite index subgroup) and thus (X, ω) is a Veech surface. Any lattice that has a $\mathrm{SL}(2, \mathbb{R})$-conjugate commensurate to $\mathrm{SL}(2, \mathbb{Z})$ is called *arithmetic*. (This weaker type of relationship between groups is called *commensurability*.) Let us say that a surface (X, ω) is *tiled by parallelograms* if it is in the $\mathrm{SL}(2, \mathbb{R})$ orbit of a square-tiled surface.

One has the following theorem of Gutkin–Judge, for a simple proof see [22].

THEOREM 5 (Gutkin–Judge). *The surface (X, ω) is tiled by parallelograms if and only if $\mathrm{SL}(X, \omega)$ is arithmetic.*

In particular this theorem proves that all square-tiled surfaces are Veech, since any arithmetic group is a lattice. This implies that any square-tiled surface satisfies the Veech alter-

native; this difficult result had previously been shown by Veech [47] using Boshernitzan's criterion.

1.5.2. *Consequences and examples* Note that an arithmetic group need not be contained in $\mathrm{SL}(2,\mathbb{Z})$. For example, consider the surface given by two unit volume squares placed one on top of the other. This is a degree 2 cover of the torus, with a one-cylinder decomposition, of modulus $1/2$. Thus, in $\mathrm{SL}(X,\omega)$ we have the element $\left(\begin{smallmatrix}1 & 1/2\\ 0 & 1\end{smallmatrix}\right)$, that is obviously not in $\mathrm{SL}(2,\mathbb{Z})$.

Another square-tiled surface provides a cautionary example. There exist oriented affine diffeomorphisms of parabolic derivative that are not formed by taking powers of Dehn twists in the cylinder decomposition of the corresponding fixed direction. (However, as Veech [47] showed, some finite power of such an affine diffeomorphism is given in such a manner.) Consider the genus two square-tiled surface formed by 3 squares stacked one on top of the other, with top and bottom identified, and side segments identified such that there is a single singularity of total angle 6π. Then one can show that there is an affine diffeomorphism of derivative $\left(\begin{smallmatrix}1 & 0\\ 1 & 1\end{smallmatrix}\right)$; however, it is the cube of this matrix that corresponds to the fundamental vertical Dehn twist here. For more on this, see [22].

The Gutkin–Judge result implies that any surface of arithmetic Veech group is a branched cover of the torus, with branching above one sole point. In general there are surfaces that have the same (or commensurate) Veech group, but are not related by any tree of finite covers that are "balanced", see [25].

The group $\Gamma = \left\langle \left(\begin{smallmatrix}1 & 3\\ 0 & 1\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & 0\\ 3 & 1\end{smallmatrix}\right)\right\rangle$ is not commensurable to any Veech group [18]. Indeed, it is known that any Veech group with a hyperbolic element of trace in $\mathbb{Q}$ must be arithmetic [30,34], and in particular a lattice. The group Γ however, is not a lattice, but possesses hyperbolic elements. Note that any finite-index subgroup H of Γ then includes hyperbolic elements with rational trace. The same is thus true for any group commensurable to Γ, and our result follows.

In any fixed stratum, the set of square-tiled surfaces of that stratum is dense. Indeed, integration of ω along its periods relative to the singularities provides local coordinates for the stratum, see [1]; these coordinates are contained in $\mathbb{Q}+i\mathbb{Q}$ exactly when (X,ω) is square-tiled. Thus, density of $\mathbb{Q}+i\mathbb{Q}$ in $\mathbb{C}$ gives the result. On the other hand, Gutkin and Judge gave an argument showing that in any stratum the set of Veech surfaces is of measure zero (if $g \geqslant 2$)—see [2] for the definition of this measure. This is loosely analogous to the fact that the rationals are of measure zero in the real numbers.

1.5.3. *Non-arithmetic surfaces exist* Non-arithmetic lattice Veech groups exist. In fact, our other favorite example—the surface arising from the $(\pi/5, \pi/5, 3\pi/5)$-triangle—has Veech group that contains $\langle S, R\rangle$, where S is the aforementioned diffeomorphism that induces the Dehn twist on each of the two vertical cylinders, and R the order five rotation. In fact, this is the entire Veech group. This group is a lattice; moreover, it is non-arithmetic.

This Veech group is (conjugate to) a well-known group, a so-called *Hecke group*. The Hecke group of index n is $\Gamma_n = \langle z \to -1/z, z \to z + 2\cos(\pi/n)\rangle$. The group above is conjugate to Γ_5. In fact, Veech showed that each Hecke group of odd index n, as well as a subgroup of index two in each even index case, is also realized as a Veech group. All

but three of these are non-arithmetic groups, and are known to be pairwise incommensurable [33].

2. State of the art

In this new century, two perspectives on Veech groups have been fruitful. The first, of a longer tradition, employs so-called scissor invariants of linear flows on the translation surface (X, ω). The second, pioneered by McMullen [34], emphasizes the algebro-geometric aspects of the Riemann surface X imposed by characteristics of $\mathrm{SL}(X, \omega)$.

2.1. *Background: Scissor invariants*

Kenyon and Smillie [30] introduced an invariant for translation surfaces, called the *J-invariant*; this invariant is an extension of the Sah–Arnoux–Fathi invariant used for the study of interval exchange transformations. Calta [11] has recently used the J-invariant to characterize the Veech surfaces in the stratum of genus 2 surfaces with a single singularity; this stratum is denoted $\mathcal{H}(2)$, see §2 of [2].

DEFINITION 1. Let P be a planar polygon of vertices $v_1, \ldots, v_n$. We define $J(P)$ as $v_1 \wedge v_2 + v_2 \wedge v_3 + \cdots + v_{n-1} \wedge v_n + v_n \wedge v_1 \in \mathbb{R}^2 \wedge_{\mathbb{Q}} \mathbb{R}^2$.

This is indeed a scissors invariant, in the following sense.

PROPOSITION 2. *Suppose that $P = P_1 \cup \cdots \cup P_k$ is a cellular decomposition of P into polygons P_i. Then $J(P) = J(P_1) + \cdots + J(P_k)$.*

Now, any translation surface can be given as a finite union of polygons, with appropriate side identification; indeed, some authors define the notion of translation surface in this way, see Definition 4 of [2]. If (X, ω) is a translation surface, and $(X, \omega) = P_1 \cup \cdots \cup P_k$ is a cellular decomposition of Σ into polygons P_i, then we define $J(X, \omega)$ as the sum of the $J(P_i)$.

THEOREM 6 (Kenyon–Smillie). *The value $J(X, \omega)$ is independent of choice of polygonal cellular decomposition of (X, ω).*

One has the possibility of studying various projections of the J-invariant. In particular, the Sah–Arnoux–Fathi invariant can be recovered in this manner. Consider

$$\pi_{xx} : \mathbb{R}^2 \wedge \mathbb{R}^2 \to \mathbb{R} \wedge \mathbb{R},$$
$$\begin{pmatrix} a \\ b \end{pmatrix} \wedge \begin{pmatrix} c \\ d \end{pmatrix} \mapsto a \wedge c.$$

We define J_{xx} as $\pi_{xx}(J)$ and J_{yy} analogously. Let $T : I \to I$ be an interval exchange transformation on a real interval I, with the lengths of the ith subinterval denoted by λ_i,

$1 \leqslant i \leqslant n$. For $i \in \{1, \ldots, n\}$, let $t_i \in \mathbb{R}$ denote the translation applied to the ith subinterval. The Sah–Arnoux–Fathi invariant is defined as

$$\mathrm{SAF}(T) = \sum_{j=1}^{n} \lambda_j \wedge t_j \in \mathbb{R} \wedge_{\mathbb{Q}} \mathbb{R}.$$

The set of all interval exchange transformations on I forms a group under composition of functions; Arnoux [5], see also [4], showed that the SAF-invariant defines a group homomorphism to $\mathbb{R} \wedge_{\mathbb{Q}} \mathbb{R}$. Furthermore, since the commutator subgroup of the group of interval exchange transformations is a simple group, the SAF-invariant gives what is essentially the only non-trivial homomorphism defined on the group.

The fundamental property of the SAF-invariant is its invariance under induction:

PROPOSITION 3 (Arnoux). *Let T be an interval exchange transformation on an interval I, and suppose that $K \subset I$ is a subinterval that meets every orbit of T. Let S denote the interval exchange transformation induced on K by T. Then* $\mathrm{SAF}(S) = \mathrm{SAF}(T)$.

The following is crucial in the work of Calta.

REMARK 2. One easily shows that if T is periodic, then $\mathrm{SAF}(T) = 0$. Furthermore, an interval exchange transformation T of *three* subintervals is periodic if and only if $\mathrm{SAF}(T) = 0$. This last is directly related to rotations: let R_α denote the rotation of angle $\alpha \in \mathbb{R}$; this map of the circle to itself is periodic if and only if $\alpha \in \mathbb{Q}$.

Note, however, Arnoux and Yoccoz [8] constructed an interval exchange transformation T of 7 subintervals with $\mathrm{SAF}(T) = 0$, but such that T is minimal, and in fact uniquely ergodic. The geometry of this interval exchange transformation is extremely interesting, see [6].

The invariance under induction of interval exchange transformation of the SAF-invariant affords the possibility of defining an SAF-invariant for a measured foliation $\mathcal{F}$ of a surface: Choose a normalized full transversal I for $\mathcal{F}$, thus in particular this interval I meets all leaves of $\mathcal{F}$, and define $\mathrm{SAF}(\mathcal{F}) = \mathrm{SAF}(T)$, where T is the interval exchange transformation defined on I by the first return map along leaves of $\mathcal{F}$. This invariant is independent of choice of I.

Kenyon and Smillie easily show the following.

PROPOSITION 4. *Let (X, ω) be a translation surface. Then $J_{xx}(X, \omega)$ equals the* SAF-*invariant for the vertical foliation of (X, ω); similarly, $J_{yy}(X, \omega)$ equals the* SAF-*invariant for the horizontal foliation of (X, ω).*

It is deft use of the J-invariant that allows Kenyon–Smillie to reach the main result of [30], that in turn lead to the following sobering result.

THEOREM 7 (Kenyon–Smillie, Puchta). *Suppose that T is an acute, non-isosceles, rational-angled triangle, and that (X, ω) is the translation surface associated to T by the usual unfolding process. Then (X, ω) is a Veech surface if and only if T has angles*:

(a) $(\pi/4, \pi/3, 5\pi/12)$,
(b) $(\pi/5, \pi/3, 7\pi/15)$, *or*
(c) $(2\pi/9, \pi/3, 4\pi/9)$.

Kenyon and Smillie also show that an acute, isosceles, rational-angled triangle gives a Veech surface if and only if the smallest angle is of the form π/n.

2.2. *Results of Calta*

A translation surface (X, ω) is said to be *completely periodic* if for every direction whose linear flow admits a periodic orbit, and hence a cylinder, (X, ω) admits a decomposition into cylinders in this direction. Clearly, Veech surfaces are completely periodic. The converse is in general false; consider the slit torus examples of [2], see also [26,35]. However, one has the following.

THEOREM 8 (Calta). *A translation surface belonging to $\mathcal{H}(2)$ is completely periodic if and only if it is a Veech surface.*

Furthermore, in this stratum, every non-arithmetic Veech surface is "quadratic" in the sense that up a change within the $\mathrm{SL}(2, \mathbb{R})$-orbit, all of its (absolute) periods are contained in some real quadratic field. Here, the *absolute periods* of (X, ω) are the periods of ω: $p(\gamma) = \int_\gamma \omega$ with $\gamma \in H_1(X, \mathbb{Z})$; thus the result is that $p(H_1(X, \mathbb{Z})) \subset \mathbb{Q}(\sqrt{d}) \times \mathbb{Q}(\sqrt{d})$, with $d > 0$ a non-square integer. Amongst all quadratic translation surfaces, Calta gives equations distinguishing the Veech surfaces.

The main idea of the proof is to introduce the following intermediate property. Here, given a direction v, the projection J_{vv} is defined analogously to J_{xx} and J_{yy}.

DEFINITION 2. A direction is called a *homological direction* for (X, ω) if it is the direction of some absolute period of ω. A translation surface has *Property* X if for every homological direction v one has $J_{vv} = 0$.

Every periodic direction of course has a representative in $p(H_1(X, \mathbb{Z}))$; Property X may be thought of as being "virtually" completely periodic—every direction that is a candidate to be completely periodic passes the test of vanishing of the corresponding projection of the J-invariant.

Calta's proof of Theorem 8 consists of showing that for translation surfaces of $\mathcal{H}(2)$ the three properties are equivalent: Property X, completely periodic, Veech. One easily shows that Property X does imply complete periodicity here—this is an application of Remark 2, and strongly depends on the genus being 2. The converse is significantly more complicated, and Calta uses explicit quadratic equations. A number theoretic argument shows that the $\mathrm{SL}(2, \mathbb{R})$-orbit of a translation surface with Property X is closed in $\mathcal{H}(2)$; by *Smillie's Theorem*, announced in [49], the surface must then be Veech.

An analogous discussion allows Calta to show that the completely periodic surfaces of the remaining stratum of genus 2 translation surfaces, $\mathcal{H}(1, 1)$, are also quadratic, and to again give explicit equations.

One can give a geometric interpretation of Calta's work, that can be compared to the appearance of Hilbert modular surfaces in the work of McMullen, see below. Beginning with a completely periodic surface in $\mathcal{H}(1,1)$, consider the $\mathrm{SL}(2,\mathbb{R})$-orbits of the surface found by fixing the absolute periods and deforming the relative periods; here "relative" means relative to the singularities. Thus, one considers the $\mathrm{SL}(2,\mathbb{R})$-orbits of the various surfaces found by varying the position of the zeros of ω. The result, $\mathcal{M}$, is a *closed* submanifold of $\mathcal{H}(1,1) \cup \mathcal{H}(2)$ of real dimension 5. The intersection of $\mathcal{M}$ with $\mathcal{H}(2)$ is a finite union of $\mathrm{SL}(2,\mathbb{R})$-orbits of Veech surfaces.

2.3. *McMullen's approach*

The approach emphasized by McMullen [34] studies properties of the Riemann surface X implied by hypotheses on the group $\mathrm{SL}(X,\omega)$. Any affine diffeomorphism ϕ of (X,ω) is such that the pull-back map ϕ^* acts on $H^1(X,\mathbb{R})$ so as to preserve the two dimensional real subspace V generated by the real and imaginary parts of ω. If ϕ has derivative $D\phi$ hyperbolic of trace t, then $T^* := \phi^* + (\phi^*)^{-1}$ acts on V as multiplication by t. McMullen relates this to the structure of the endomorphism ring of the Jacobian of X.

2.3.1. *Algebro-geometric background* We briefly recall some standard terminology and results from algebraic geometry, see the textbooks [20,17,15]; the classic reference on Abelian varieties is [40]; for a constructive treatment of real multiplication see [9], as well as [43]. See [21] or [46] for an introduction to the study by the school of F. Hirzebruch of the geometry and arithmetic of Hilbert modular surfaces. Our discussion closely follows §4 of [36].

The Jacobian. Key to our discussion is the g-complex dimensional vector space $\Omega(X)$ of 1-forms on a Riemann surface X of genus g. Indeed, whereas the results discussed so far are related to the flat structure induced on X by integration of a single 1-form, we now fix a base point and consider integration of a vector whose entries form a basis for $\Omega(X)$. This gives a map to $\mathbb{C}^g$ that is only well defined after dividing by the lattice formed by the integrals along closed curves. The result is the famed *Abel–Jacobi map* from X to the complex torus defined as the *Jacobian variety* of X, $\mathrm{Jac}(X)$.

The celebrated *Riemann Relations* show that $\mathrm{Jac}(X)$ is a *principally polarized Abelian variety*: It is in particular a complex torus equipped with an embedding into complex projective space. Expressing $\mathrm{Jac}(X)$ as $\Omega^*(X)/H_1(X,\mathbb{Z})$, one avatar of the polarization is as a symplectic form on $H_1(X,\mathbb{Z})$. In fact, the intersection pairing on $H_1(X,\mathbb{Z})$ gives this symplectic form. Of course, as real vector spaces, $\Omega^*(X)$ and $H_1(X,\mathbb{R})$ are isomorphic; we can thus view $\Omega^*(X)$ as $H_1(X,\mathbb{R})$ with a complex structure. See Chapter 4 of [12] for a discussion of related canonical isomorphisms.

Real multiplication by a field; eigenforms. Given any principally polarized Abelian variety $A \cong \mathbb{C}^g/\Lambda$, the polarization of A equips $\Lambda \cong H_1(A,\mathbb{Z}) \cong \mathbb{Z}^{2g}$ with a symplectic form. The *endomorphism ring* $\mathrm{End}(A)$ consists of the Lie group homomorphisms of A;

each endomorphism respects the Hodge decomposition $H^1(A,\mathbb{C}) \cong H^{(1,0)} \oplus H^{(0,1)}$ and induces an endomorphism of Λ.

A field K is called *totally real* if it is a number field all of whose embeddings fixing $\mathbb{Q}$ have image in $\mathbb{R}$. Given a totally real field K with $[K:\mathbb{Q}] = g$, we say that A admits *real multiplication* by K if there is a faithful representation $\rho : K \to \mathrm{End}(A) \otimes \mathbb{Q}$ such that each $\rho(\kappa)$ is self-adjoint with respect to the induced symplectic form on $\Lambda \otimes \mathbb{Q}$. The holomorphic 1-forms on A form the g-dimensional $\mathbb{C}$-vector space $\Omega(A) \cong H^{(1,0)}$. Since $\rho(K)$ respects the Hodge decomposition, K acts on $\Omega(A)$ in a complex linear fashion. An eigenvector for this action is called an *eigenform* for the real multiplication of A. The action can always be diagonalized: $\Omega(A) = \bigoplus \mathbb{C}\omega_i$ for g eigenforms ω_i, thus there are eigenforms for any real multiplication.

In the case that $A = \mathrm{Jac}(X)$, we can speak of $\omega \in \Omega(X)$ as being an eigenform. Indeed, given real multiplication on $\mathrm{Jac}(X) \cong \Omega(X)^*/H_1(X,\mathbb{Z})$, one finds that the eigenforms are exactly the eigenvectors for the dual action on $\Omega(X)$. The *eigenform locus* in $\Omega\mathcal{M}_g$ is the space of (X,ω) with ω an eigenform.

REMARK 3. With only slight complication of the above, one can define real multiplication on an Abelian variety of complex dimension g by a product K of totally real fields K_i, with $\sum [K_i:\mathbb{Q}] = g$.

Endomorphisms to real multiplication. The integral points $\mathfrak{o} = K \cap \mathrm{End}(A)$ of elements of K which act as endomorphisms of A form an *order* of K. That is, $\mathfrak{o}$ is a finite-index subring of $\mathcal{O}_K$, where $\mathcal{O}_K$ is the product of the rings of algebraic integers of the K_i. Of course, given an order $\mathfrak{o} \subset K$, and any faithful representation of $\mathfrak{o}$ as self-adjoint endomorphisms of A, there is an induced real multiplication of A by K.

Indeed, suppose that some totally real algebraic integer t acts as an endomorphism T on an Abelian variety A. Then one finds that $\mathbb{Z}[t] \subset \mathrm{End}(A)$, by extending the map $t \mapsto T$ in the usual manner. Tensoring with $\mathbb{Q}$, one finds that A admits real multiplication by $\mathbb{Q}(t)$. Thus, a single endomorphism can induce real multiplication by a field.

Families with real multiplication by an order. The appropriate level of abstraction is obtained by fixing a symplectic form on a lattice $L \cong \mathbb{Z}^{2g}$, and considering the injective homomorphisms ρ which send $\mathfrak{o}$ to $\mathrm{End}(L)$ as self-adjoint endomorphisms. One then says that A admits *real multiplication by* $(\mathfrak{o},\rho)$ if there is a symplectic isomorphism of L with $H_1(A,\mathbb{Z})$ such that $\rho(\mathfrak{o})$ coincides with the restriction of $\mathrm{End}(A)$.

The space of all Abelian varieties admitting real multiplication by some $(\rho,\mathfrak{o})$ can be determined in the following constructive manner. Tensoring the rank two $\mathfrak{o}$-module L with $\mathbb{R}$ allows us to find a decomposition into orthogonal eigenspaces, each of real dimension two: $L \otimes \mathbb{R} \cong \bigoplus_{i=1}^{g} S_i$. Fix i, and choose some positively ordered symplectic basis (a_i, b_i) for S_i; to each $\tau_i \in \mathbb{H}$, we have an $\mathbb{R}$-linear map from $\mathbb{C}$ to S_i induced by sending 1 to a_i and τ_i to b_i. Note that in particular this map respects the orientation of $\mathbb{R}^2 \cong S_i$.

Each $\tau := (\tau_1, \ldots, \tau_g) \in \mathbb{H}^g$ thus determines an isomorphism of real vector spaces that takes $L \otimes \mathbb{R}$ to $\mathbb{C}^g$ and thus induces a symplectic structure on $\mathbb{C}^g$; the image of $L \otimes 1$ is a lattice. The quotient, A_τ, of $\mathbb{C}^g$ by this lattice has real multiplication by $(\mathfrak{o},\rho)$.

Every Abelian variety admitting real multiplication by $(\mathfrak{o}, \rho)$ arises in this fashion. Indeed, given some $A = \mathbb{C}^g/\Lambda$, take Λ as L and use the symplectic form given by the principal polarization. Choose an integral basis for Λ and a compatible splitting of $\mathbb{C}^g$; we may assume that the basis of Λ is of the form $(1, b_i)$ with $b_i \in \mathbb{H}$. With $\tau = (b_1, \ldots, b_g)$, we find that $A_\tau = A$.

Hilbert modular varieties. Given L and $(\mathfrak{o}, \rho)$ as above, let $\mathrm{Sp}(L \otimes \mathbb{R}) \cong \mathrm{Sp}(2g, \mathbb{R})$ denote the $\mathbb{R}$-linear operators on $L \otimes \mathbb{R}$ which respect the symplectic form. Those symplectic automorphisms that commute with the action of $\mathfrak{o}$ preserve the splitting $L \otimes \mathbb{R} \cong \bigoplus_{i=1}^{g} S_i$. Therefore, each such automorphism acts on the set of complex structures on $L \otimes \mathbb{R}$ that are compatible with the splitting. Since these complex structures are indexed by $\mathbb{H}^g$, one finds that the subgroup of symplectic automorphisms that commute with the action of $\mathfrak{o}$ is the image of an injective homomorphism $\iota : \mathrm{SL}(2, \mathbb{R})^g \to \mathrm{Sp}(L \otimes \mathbb{R})$. The integral points $\Gamma(\mathfrak{o}, \rho) := \iota(\mathrm{SL}(2, \mathbb{Z})^g)$ are exactly the automorphisms of the symplectic $\mathfrak{o}$-module L. The group $\Gamma(\mathfrak{o}, \rho)$ acts isometrically on $\mathbb{H}^g$ as elements of $\mathrm{SL}(2, \mathbb{Z})^g$, the finite volume quotient $X(\mathfrak{o}, \rho) := \Gamma(\mathfrak{o}, \rho)\backslash\mathbb{H}^g$, called the *Hilbert modular variety* of $(\mathfrak{o}, \rho)$, parametrizes pairs $(A, \mathfrak{o} \to \mathrm{End}(A))$ compatible with ρ. There is a natural forgetful map from $X(\mathfrak{o}, \rho)$ to $\mathcal{A}_g$, the coarse moduli space of principally polarized Abelian varieties—one forgets the maps $\mathfrak{o} \to \mathrm{End}(A)$.

Multiplication by a real quadratic order. When $g = 2$, there are two facts that simplify the above. First, it is well known that the orders $\mathfrak{o}$ in real quadratic fields are uniquely determined by their discriminants $D = D(\mathfrak{o}) \in \mathbb{Z}$; we thus write $\mathfrak{o}_D$. Second, for each such $\mathfrak{o}_D$, there is essentially a unique representation $\rho_D : \mathfrak{o}_D \to \mathbb{Z}^4$ which respects the standard symplectic form on $\mathbb{Z}^4$; see say Theorem 2 of [43]. One thus finds a single Hilbert modular surface for each discriminant, $X_D := X(\mathfrak{o}_D, \rho_D)$.

Furthermore, one can give an explicit model for each of these. Let σ denote the nontrivial element in $\mathrm{Gal}(K/\mathbb{Q})$; for $M \in \mathrm{SL}(2, K)$, let M^σ denote the matrix whose elements are the images by σ of the corresponding elements of M. Then $\mathrm{SL}(2, K)$ acts on $\mathbb{H}^2$ by $M \circ (z_1, z_2) = (Mz_1, M^\sigma z_2)$, where elements of $\mathrm{SL}(2, \mathbb{R})$ act on $\mathbb{H}$ in the usual manner. One can show that $X_D \cong \mathrm{SL}(2, \mathfrak{o}_D)\backslash\mathbb{H}^2$.

For each of these X_D, the forgetful map to $\mathcal{A}_2$ is generically 2-to-1: homomorphisms from $\mathfrak{o}_D$ to $\mathrm{End}(A)$ are conflated with their compositions with σ. This forgetful map factors through the *symmetric Hilbert modular surface* formed as the quotient of X_D by the involution induced by the standard permutation on $\mathbb{H} \times \mathbb{H}$. The image variety in $\mathcal{A}_2$ is called a *Humbert surface*, after the work of G. Humbert in the late 19th century.

2.3.2. *McMullen's action by the trace field* With ϕ an affine diffeomorphism of hyperbolic derivative $D\phi$ having trace t, consider $T = \phi_* + (\phi_*)^{-1}$ acting on $H_1(X, \mathbb{R})$. Since ϕ preserves intersections, it is easy to show that T is self-adjoint with respect to the corresponding symplectic form. Since the pull-back of any affine diffeomorphism leaves $V \subset H^1(X, \mathbb{R})$ invariant, T leaves invariant the annihilator of V, defined as the space of cycles upon which all elements of V vanish.

In genus two, the annihilator and its orthogonal complement are both of real dimension two, giving thus complex lines in $\Omega^*(X)$. The self-adjoint T acts on each of these

eigenspaces as multiplication by a real number. That is to say, T induces an endomorphism of $\mathrm{Jac}(X)$. When t is quadratic over $\mathbb{Q}$, the map $t \mapsto T$ as discussed in the treatment of real multiplication in Section 2.3.1 shows that $\mathrm{Jac}(X)$ admits real multiplication by $K = \mathbb{Q}(t)$.

The field K is independent of choice of hyperbolic element in $\mathrm{SL}(X, \omega)$; see the appendix of [30] for the following: since those ϕ with $D\phi$ hyperbolic are in fact *pseudo-Anosov* maps, earlier results allow one to prove both that

(1) K is the full *trace field* of $\mathrm{SL}(X, \omega)$, defined as the field generated by adjoining to $\mathbb{Q}$ the traces of all elements of the group; and
(2) $[K : \mathbb{Q}] \leqslant g$.

Furthermore, see say Lemma 8 on p. 167 of [16], t is an algebraic integer.

2.3.3. *Projecting orbits to $\mathcal{M}_g$ and $\mathcal{A}_g$* The projection $\pi : \Omega\mathcal{M}_g \to \mathcal{M}_g$ is constant on orbits of $\mathrm{SO}(2, \mathbb{R})$. On the other hand, the stabilizer of $z = i$ under the transitive action of $\mathrm{SL}(2, \mathbb{R})$ by Möbius transformations on the Poincaré upper half-plane, $\mathbb{H}$, is $\mathrm{SO}(2, \mathbb{R})$. There is thus a map $\mathbb{H} \to \mathcal{M}_g$ that factors through $\mathrm{SL}(X, \omega)\backslash\mathbb{H}$. In fact, it is of great importance that this image in $\mathcal{M}_g$ is isometrically immersed with respect to the so-called Teichmüller metric, see [13] for discussion of this metric in terms related to $\mathrm{SL}(2, \mathbb{R})$. The image in $\mathcal{M}_g$ is an algebraic curve if and only if $\mathrm{SL}(X, \omega)$ is a lattice, in which case this image is called a *Teichmüller curve* in $\mathcal{M}_g$.

The *Torelli map* $\tau : \mathcal{M}_g \to \mathcal{A}_g$ is defined by sending each X to $\mathrm{Jac}(X)$; for a discussion of the geometry of this map, see [41]. In dimension $g = 2$, in fact $\mathcal{A}_2 = \tau(\mathcal{M}_2) \sqcup H_1$, where H_1 is the locus of Abelian varieties that split as a product of two polarized elliptic curves. In particular, the Torelli map has dense open image in $\mathcal{A}_2$; there is thus a tendency in the literature to slur over the distinction of certain loci as being in one or the other of the spaces $\mathcal{M}_2$ and $\mathcal{A}_2$. For simplicity, call the map $\Omega\mathcal{M}_g \to \mathcal{A}_g$, given by composing the Torelli map with π, the projection to $\mathcal{A}_g$.

2.3.4. *A selection of results* The fundamental observation of McMullen is that as soon as a translation surface (X, ω) with X of genus 2 admits a hyperbolic element in $\mathrm{SL}(X, \omega)$, then $\mathrm{Jac}(X)$ admits real multiplication by the trace field of $\mathrm{SL}(X, \omega)$, with ω an eigenform for this multiplication. The following result, false in higher genus, is crucial to McMullen's study in genus two.

THEOREM 9 (McMullen). *The eigenform locus in $\Omega\mathcal{M}_2$ is $\mathrm{SL}(2, \mathbb{R})$-invariant.*

The main result of McMullen on Teichmüller curves in $\mathcal{M}_2$ is the following

THEOREM 10 (McMullen). *Suppose that $\mathrm{SL}(X, \omega)$ is a non-arithmetic lattice and X is of genus 2. Then the $\mathrm{SL}(2, \mathbb{R})$-orbit of (X, ω) projects to $\mathcal{A}_2$ to be an algebraic curve contained in some symmetric Hilbert modular surface.*

In fact, Remark 3 can be invoked to show that in genus 2 if $\mathrm{SL}(X, \omega)$ is arithmetic, then $\mathrm{Jac}(X)$ admits real multiplication by $\mathbb{Q} \times \mathbb{Q}$, and the $\mathrm{SL}(2, \mathbb{R})$-orbit then projects to an appropriate symmetric Hilbert modular surface [36].

The previous theorem easily leads to the following result, which can also be deduced from Calta's results.

THEOREM 11 (McMullen). *Suppose that* $(X, \omega) \in \mathcal{H}(2)$. *If there is a hyperbolic element in* $\mathrm{SL}(X, \omega)$, *then* (X, ω) *is a Veech surface.*

The situation is completely different for $\mathcal{H}(1, 1)$. Indeed, let $\mathcal{D}$ denote the translation surface given by identifying by translation opposite sides of the regular decagon. In [37], McMullen conjectured, and in [38] proved, the following

THEOREM 12 (McMullen). *The only non-arithmetic Veech surface of* $\mathcal{H}(1, 1)$ *is* $\mathcal{D}$.

McMullen [37] gives an algorithm for determining those (X, ω) whose $\mathrm{SL}(2, \mathbb{R})$-orbit projects to a Hilbert modular surface for a given discriminant of order. In particular, he shows that Veech's original examples of a double pentagon and a double decagon account for all lattice groups giving rise to curves on the symmetric Hilbert modular surface of real multiplication by the order with discriminant $D = 5$.

REMARK 4. For reasons of time and space, we have not discussed an important aspect of the projections of $\mathrm{SL}(2, \mathbb{R})$-orbits in $\Omega\mathcal{M}_g$ to each of $\mathcal{M}_g$ and A_g: These projections are isometries for the appropriate metrics. This result is due to Kra [32]. This isometry is in some sense what allows one to use the structure of the homogeneous space $\mathcal{A}_g$ to study Veech groups. As well, there are many curves in moduli space, but very few of them are isometrically embedded with respect to the Teichmüller metric.

Using the above, McMullen [36] proves an analog of the celebrated Ratner Theorem, see [1].

THEOREM 13 (McMullen). *The closure of the* $\mathrm{SL}(2, \mathbb{R})$*-orbit of any* $(X, \omega) \in \Omega\mathcal{M}_2$ *projects to* $\mathcal{M}_2$ *as exactly one of the following*: *an algebraic curve*; *a Hilbert modular surface*; *all of* $\mathcal{M}_2$.

In recent work, M. Möller [39] has extended McMullen's result for lattice $\mathrm{SL}(X, \omega)$. In particular, for $g > 2$, he shows that even though the action by the trace field identified by McMullen may not extend to the full Jacobian of X, it does identify special properties, which he studies in terms of variation of Hodge structures. (For an introduction to this study of splittings of bundles generalizing the study of the Hodge decomposition, see [50].) An *isogeny* of an Abelian variety is a surjective morphism of algebraic varieties to some Abelian variety, and this morphism is a group homomorphism, of finite kernel. (Isogenous Abelian varieties are thus morally equivalent.)

THEOREM 14 (Möller). *Suppose that* $\mathrm{SL}(X, \omega)$ *is a lattice. Then the* $\mathrm{SL}(2, \mathbb{R})$*-orbit of* (X, ω) *projects to* $\mathcal{A}_g$ *to be an algebraic curve contained in the locus parametrizing Abelian varieties A splitting up to isogeny to a product* $A_1 \times A_2$, *where* A_1 *admits real multiplication by the trace field of* $\mathrm{SL}(X, \omega)$.

2.4. *Infinitely generated Veech groups*

In [49], Veech asked if a $\mathrm{SL}(X,\omega)$ can ever be an infinitely generated Fuchsian group. This has recently been answered in the affirmative [26,35].

THEOREM 15 [26]. *For each genus $g \geqslant 4$, there exist $(Y,\alpha) \in \Omega\mathcal{M}_g$ with $\mathrm{SL}(Y,\alpha)$ infinitely generated. In particular, the genus four translation surface arising from the triangle of angles $(3\pi/10, 3\pi/10, 2\pi/5)$ has infinitely generated Veech group.*

THEOREM 16 (McMullen). *Suppose that $(X,\omega) \in \Omega\mathcal{M}_2$ is such that $\mathrm{SL}(X,\omega)$ admits a hyperbolic element. Then the limit set of $\mathrm{SL}(X,\omega)$ is the full boundary $\partial\mathbb{H}$. Furthermore, there exist infinitely many distinct $(X,\omega) \in \Omega\mathcal{M}_2$ with $\mathrm{SL}(X,\omega)$ infinitely generated.*

2.4.1. *Commonalities of proofs* Other than the specifics of the examples, the proofs of these two results have common logic, both beginning with the fact that a non-lattice Fuchsian group whose limit set is all of $\partial\mathbb{H}$ must be infinitely generated. Now, it is often quite easy to show that the Veech group of a given translation surface is not a lattice: simply exhibit a saddle connection in whose direction the surface does not admit a decomposition into cylinders of commensurable moduli.

To show that the limit set of the Veech groups under consideration in the two theorems have all of $\partial\mathbb{H}$ as limit sets, both proofs show that the *parabolic directions* of the corresponding translation surfaces—that is, the directions for which there is a cylinder decomposition with commensurable moduli, and thus a corresponding parabolic element in the group—form a dense set in the unit circle of all directions. In both cases, one exhibits some point $p \in X$ such that every direction in which there is a separatrix passing through p is in fact a parabolic direction. This is the difficult step in each proof.

2.4.2. *Sketch: Proof of Theorem 16* Suppose that X is of genus two and $\mathrm{SL}(X,\omega)$ admits a hyperbolic element, of trace say t. Let $K = \mathbb{Q}(t)$ be the trace field. By results of the appendix of [30], one can assume that the relative (to the singularities of ω) periods of ω on X lie in $K(i)$. Let ϕ be an affine diffeomorphism corresponding to the hyperbolic element. As in the previous section, $T^* := \phi^* + (\phi^*)^{-1}$ acts as multiplication by t on V, the real subspace spanned in $H^1(X,\mathbb{R})$ by the real and imaginary parts of ω. Once again, we let σ denote the non-trivial Galois group element. One finds that T^* thus acts as multiplication by $\sigma(t)$ on the subspace V^σ spanned by the real and imaginary parts of $\sigma(\omega)$. Since T^* is appropriately self-adjoint, V and V^σ are orthogonal. One thus has that the integral over X of each of $\omega \wedge \sigma(\omega)$ and $\omega \wedge \overline{\sigma(\omega)}$ is zero, where the bar here denotes complex conjugation. From this, $\int_X \rho \wedge \sigma(\rho) = 0$ when ρ is the closed real form associated to any directional flow of slope in $\mathbb{P}^1(K) = K \cup \{\infty\}$.

However, if f is the interval exchange transformation on a transversal of the measured foliation associated to ρ, then $\int_X \rho \wedge \sigma(\rho) = \mathrm{flux}(f)$, where $\mathrm{flux}(f)$ is a version of the SAF-invariant introduced by McMullen, the *Galois flux*: Suppose that all the translations

for some interval exchange transformation T are contained in some quadratic number field K, then one defines

$$\text{flux}(T) = \sum_{j=1}^{n} \lambda_j \sigma(t_j) \in \mathbb{R}.$$

Now, if this flux vanishes, then the directional flow for ρ cannot be uniquely ergodic. But, Masur's criterion now tells us that $g_t \text{SL}(X, \omega)$ leaves every compact set. This implies in turn that there are very short saddle connections on the corresponding translation surfaces $g_t \circ (X, \omega)$ for large t. Using the quadratic nature of K, elementary Diophantine approximation considerations (to wit: quadratic numbers cannot be well approximated by rationals) then allow McMullen to conclude that for t sufficiently large, such a short saddle connection must in fact lie in the direction of the foliation. Restricting to genus 2, he then can give a complete analysis of such loops, to conclude that either the foliation is periodic, or else surgery along a leaf presents (X, ω) as a connected sum of irrationally foliated tori. In particular, it turns out that if there is a Weierstrass point lying on a saddle connection in the direction of flow for ρ, then this a parabolic direction.

However (upon developing (X, ω) such that a singularity lies at the origin, every developed image of), each non-singular Weierstrass point has coordinates in K. Thus, any separatrix passing through a non-singular Weierstrass point lies in a direction whose slope is in $\mathbb{P}^1(K)$. From the above, this direction is hence a parabolic direction. But, for any given point of a translation surface, the directions of separatrices passing through this point are dense, see say Lemma 1 of [26]. The density of parabolic limit points then follows.

REMARK 5. A side-product of the above is that a Veech surface of genus two defined over $\mathbb{Q}(\sqrt{d})$ allows a normalization such that the set of slopes of its periodic directions equals $\mathbb{Q}(\sqrt{d}) \cup \{\infty\}$, see also [11]. This is specific to genus two, see [7].

McMullen [36] gives an infinite family of genus two translation surfaces of infinitely generated Veech group by explicit construction, see Figure 1 there. Indeed, given 3 squares, of side length 1, a and $a + 1$, respectively, one can place these squares so as to construct a genus two surface. If a is irrational of the form $b - 1 + \sqrt{b^2 - b + 1}$ for non-zero $b \in \mathbb{Q}$, then the Veech group of the translation surface is infinitely generated.

2.4.3. *Sketch: Proof of Theorem 15* On the other hand, the proof of Theorem 15 constructs examples by use of ramified covers of Riemann surfaces $f : Y \to X$: the pull-back $\alpha = f^*(\omega)$ can have an infinitely generated group even if $\text{SL}(X, \omega)$ is a lattice. (Some background for this can be found in [24].) Indeed, suppose that the ramification is at the singularities of ω and at a point p—called a *connection point*—such that every separatrix of (X, ω) passing through p extends to a saddle connection. Again by Lemma 1 of [26], this is a dense set of directions. Since $\text{SL}(X, \omega)$ is a lattice, the direction of any saddle connection is a parabolic direction; one easily shows that each of our dense set of parabolic directions for (X, ω) is also a parabolic direction for (Y, α). It follows that the parabolic limit points of $\text{SL}(Y, \alpha)$ are dense.

The main part of the proof of Theorem 15 consists of showing that there are (X, ω) with connection points p such that the corresponding $\mathrm{SL}(Y, \alpha)$ is not a lattice. For this, it suffices to show that one can find points that are at the same time connection points and have infinite orbit under the group of oriented affine diffeomorphisms. Amusingly enough, the genus two example of Figure 2 admits such points. After an innocuous normalization, these are the points of coordinates in $\mathbb{Q}(\sqrt{5})$ (other than the regular Weierstrass points, which are given by the middle of the sides). This results from the fact that the parabolic (limit) points of Γ_5 (recall that this is the Veech group here, up to a normalization) is $\mathbb{Q}(\sqrt{5})$ [33]. This latter fact can be recovered by direct use of Remark 5. By way of [24], one then finds that the translation surface to which the triangle angles $(3\pi/10, 3\pi/10, 2\pi/5)$ unfolds is a ramified cover of the genus two example, with ramification above singularities and connection points.

In [27], it is shown that the geometry of the projection to $\mathcal{M}_g$ of the $\mathrm{SL}(2, \mathbb{R})$-orbit of such (Y, α) is very complicated: $\mathrm{SL}(Y, \alpha)$ has infinitely many non-equivalent parabolic points and infinitely many "infinite ends".

2.5. *Classification*

The fundamental classification problem of determining when two given translation surfaces are in the same $\mathrm{SL}(2, \mathbb{R})$-orbit seems far from being resolved. Indeed, this remains open even for square-tiled surfaces, with the exception of the stratum $\mathcal{H}(2)$.

In the setting of square-tiled surfaces, it suffices to classify the *primitive* square-tiled surfaces: those such that the lattice generated by their relative periods is $\mathbb{Z}^2$. One easily shows that in this setting $\mathrm{SL}(X, \omega) \subset \mathrm{SL}(2, \mathbb{Z})$. There is an action of $\mathrm{SL}(2, \mathbb{Z})$ on the set of primitive square-tiled surfaces of fixed number of squares, n; two such surfaces are in the same $\mathrm{SL}(2, \mathbb{R})$-orbit if and only if they are in the same $\mathrm{SL}(2, \mathbb{Z})$-orbit.

In $\mathcal{H}(2)$, the position of the Weierstrass points give an invariant for the $\mathrm{SL}(2, \mathbb{Z})$-action. Informally: given a surface of our type, we develop in such a manner that a singularity lies at the origin, the six Weierstrass points then each has coordinates that are integers are half-integers. To be more precise, one explicitly parametrizes the square-tiled surfaces of $\mathcal{H}(2)$, as in [14,52].

PROPOSITION 5 [22]. *The number of integer coordinate Weierstrass points of a square-tiled surface of* $\mathcal{H}(2)$ *is invariant under the action of* $\mathrm{SL}(2, \mathbb{Z})$.

If the number n of square tiles is even, there are two such Weierstrass points; if n is odd, there are either three or one such point. The invariant completely classifies the orbits.

THEOREM 17 ([22], McMullen). *Given an integer* $n \geqslant 3$; *the square-tiled surfaces of* $\mathcal{H}(2)$ *form two* $\mathrm{SL}(2, \mathbb{Z})$*-orbits if* n *is odd and* $n \geqslant 5$; *they form a single orbit if either* n *is even or* $n = 3$.

The theorem was first proved in [22] for prime n. McMullen generalized this to not only square-tiled surfaces, but also so as to give an analogous result for all Veech surfaces of $\mathcal{H}(2)$.

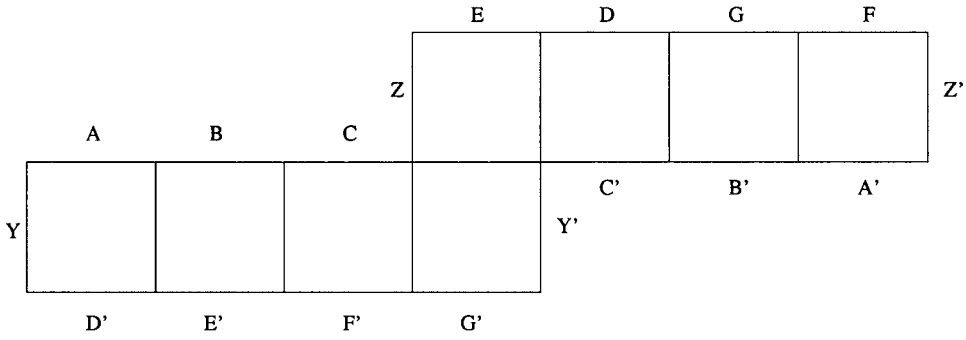


Fig. 5. A square-tiled surface with $\mathrm{SL}(X,\omega) = \mathrm{SL}(2,\mathbb{Z})$.

Combining Theorem 17 with a counting formula given by [14] shows that the genus of Teichmüller curves defined by primitive square-tiled surfaces tends to infinity with the number of tiles. This can be compared with the fact that there are no explicitly known Teichmüller curves of positive genus arising from non-arithmetic surfaces of $\mathcal{H}(2)$. (One expects that in fact almost all of these are of positive genus.)

One can also show the group $\mathrm{SL}(X,\omega)$ for a primitive square-tiled surface is a congruence subgroup of $\mathrm{SL}(2,\mathbb{Z})$ only in the case of surfaces of three square tiles. See [44] for an example of a non-congruence subgroup, and [23] for the general case. Nevertheless, there are non-trivial examples of square-tiled surfaces whose group is exactly the full group $\mathrm{SL}(2,\mathbb{Z})$, see [44]. There has been work on this phenomenon by Herrlich, Schmoll, as well as by Möller. We thank M. Möller for kindly providing Figure 5, which represents one such surface.

2.6. *Questions*

We conclude with some more open questions.

1. Is there a general converse to the Veech Dichotomy (as found by McMullen for genus $g = 2$)?
2. Which Fuchsian groups are realized as Veech groups?
3. Is there an algorithm for determining the Veech group of a general translation surface?
4. Do there exist non-trivial Veech groups without parabolic elements?

References

Surveys in volume 1A and this volume

[1] A. Eskin, *Counting problems in moduli space*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 581–595.

[2] H. Masur, *Ergodic theory of translation surfaces*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 527–547.

[3] H. Masur and S. Tabachnikov, *Rational billiards and flat structures*, Handbook on Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 1015–1090.

Other sources

[4] P. Arnoux, *Échanges d'intervalles et flots sur les surfaces*, Ergodic Theory (Sem., Les Plans-sur-Bex, 1980), Monograph. Enseign. Math., Vol. 29, Univ. Genève, Geneva (1981), 5–38.
[5] P. Arnoux, Thèse de 3^e cycle, Université de Reims (1981).
[6] P. Arnoux, *Un exemple de semi-conjugaison entre un échange d'intervalles et une translation sur le tore*, Bull. Soc. Math. France **116** (4) (1988), 489–500.
[7] P. Arnoux and T.A. Schmidt, *Fractions continues commensurables*, in preparation.
[8] P. Arnoux and J.-C. Yoccoz, *Construction de difféomorphismes pseudo-Anosov*, C. R. Acad. Sci. Paris Sér. I Math. **292** (1) (1981), 75–78.
[9] C. Birkenhake and H. Lange, *Complex Abelian Varieties*, 2nd edn, Grundl. der Math. Wissensch., Vol. 302, Springer-Verlag, Berlin (2004).
[10] M. Boshernitzan, *A condition for minimal interval exchange maps to be uniquely ergodic*, Duke Math. J. **52** (3) (1985), 723–752.
[11] K. Calta, *Veech surfaces and complete periodicity in genus two*, J. Amer. Math. Soc. **17** (4) (2004), 871–908.
[12] H. Clemens, *A Scrapbook of Complex Curve Theory*, Univ. Series in Math., Plenum, New York (1980).
[13] C.J. Earle and F.P. Gardiner, *Teichmüller disks and Veech's $\mathcal{F}$ structures*, *Extremal Riemann Surfaces*, Contemp. Math., Vol. 201, Amer. Math. Soc., Providence, RI (1997), 165–189.
[14] A. Eskin, H. Masur and M. Schmoll, *Billiards in rectangles with barriers*, Duke Math. J. **118** (3) (2003), 427–463.
[15] H.M. Farkas and I. Kra, *Riemann Surfaces*, 2nd edn, Grad. Text Math., Vol. 71, Springer-Verlag, Berlin (1992).
[16] A. Fathi, F. Laudenbach and V. Poénaru, *Travaux de Thurston sur les surfaces*, Séminaire Orsay, Astérisque, Soc. Math. France, Paris (1979), 66–67.
[17] P. Griffiths and J. Harris, *Principles of Algebraic Geometry*, Wiley (1978).
[18] E. Gutkin, P. Hubert and T.A. Schmidt, *Affine diffeomorphisms of translation surfaces: Periodic points, Fuchsian groups, and arithmeticity*, Ann. Sci. École Norm. Sup. **36** (2003), 847–866.
[19] E. Gutkin and C. Judge, *Affine mappings of translation surfaces: Geometry and arithmetic*, Duke Math. J. **103** (2000), 191–213.
[20] R. Hartshorne, *Algebraic Geometry*, Grad. Text Math., Vol. 52, Springer-Verlag, Berlin (1977).
[21] F. Hirzebruch, *Hilbert Modular Surfaces*, Enseign. Math. **19** (1973), 183–281.
[22] P. Hubert and S. Lelièvre, *Prime arithmetic Teichmüller disks in $\mathcal{H}(2)$*, Israel J. Math., to appear.
[23] P. Hubert and S. Lelièvre, *Noncongruence subgroups in $\mathcal{H}(2)$*, Internat. Math. Res. Notices **1** (2005), 47–64.
[24] P. Hubert and T.A. Schmidt, *Veech groups and polygonal coverings*, J. Geom. Phys. **35** (2000), 75–91.
[25] P. Hubert and T.A. Schmidt, *Invariants of translation surfaces*, Ann. Inst. Fourier **51** (2001), 461–495.
[26] P. Hubert and T.A. Schmidt, *Infinitely generated Veech groups*, Duke Math. J. **123** (2004), 49–69.
[27] P. Hubert and T.A. Schmidt, *Geometry of infinitely generated Veech groups*, Conform. Geom. Symb. Dynam., to appear.
[28] S. Katok, *Fuchsian Groups*, University of Chicago Press (1992).
[29] A. Katok and A. Zemlyakov, *Topological transitivity of billiards in polygons*, Math. Notes **18** (1975), 760–764.
[30] R. Kenyon and J. Smillie, *Billiards in rational-angled triangles*, Comment. Math. Helv. **75** (2000), 65–108.
[31] S. Kerckhoff, H. Masur and J. Smillie, *Ergodicity of billiard flows and quadratic differentials*, Ann. Math. **124** (1986), 293–311.
[32] I. Kra, *The Carathéodory metric on Abelian Teichmüller disks*, J. Anal. Math. **40** (1981), 129–143.
[33] A. Leutbecher, *Über die Heckeschen Gruppen $G(\lambda)$. II*, Math. Ann. **211** (1974), 63–86.
[34] C.T. McMullen, *Billiards and Teichmüller curves on Hilbert modular surfaces*, J. Amer. Math. Soc. **16** (4) (2003), 857–885 (electronic).
[35] C.T. McMullen, *Teichmüller geodesics of infinite complexity*, Acta Math. **191** (2003), 191–223.

[36] C.T. McMullen, *Dynamics of* $\mathrm{SL}(2,\mathbb{R})$ *over moduli space in genus two*, preprint. Available at: http://abel.math.harvard.edu/∼ctm/papers/index.html.
[37] C.T. McMullen, *Teichmüller curves in genus two: the decagon and beyond*, J. Reine Angew. Math. **582** (2005), 173–199.
[38] C.T. McMullen, *Teichmüller curves in genus two: torsion divisors and ratio of sines*, preprint. Available at: http://abel.math.harvard.edu/∼ctm/papers/index.html.
[39] M. Möller, *Variations of Hodge structure of a Teichmüller curve*, preprint. Available at: http://front.math.ucdavis.edu/math.AG/0401290.
[40] D. Mumford, *Abelian Varieties*, Oxford Univ. Press (1970).
[41] D. Mumford, *Curves and Their Jacobians*, The University of Michigan Press, Ann Arbor, MI (1975).
[42] J.-Ch. Puchta, *On triangular billiards*, Comment. Math. Helv. **76** (2001), 501–505.
[43] B. Runga, *Endomorphism rings of Abelian surfaces and projective models of their moduli spaces*, Tohoku Math. J. **51** (1999), 283–303.
[44] G. Schmithüsen, *An algorithm for finding the Veech group of an origami*, Experiment. Math. **13** (4) (2004), 459–472.
[45] S. Tabachnikov, *Billiards*, Panoramas et Synthèses 1, Soc. Math. France, Paris (1995).
[46] G. van der Geer, *Hilbert Modular Surfaces*, Springer (1980).
[47] W. Veech, *Boshernitzan's criterion for unique ergodicity of an interval exchange transformation*, Ergodic Theory Dynam. Systems **7** (1) (1987), 149–153.
[48] W. Veech, *Teichmüller curves in moduli space, Eisenstein series, and an application to triangular billiards*, Inv. Math. **97** (1989), 553–583.
[49] W. Veech, *Geometric realizations of hyperelliptic curves*, Algorithms, Fractals, and Dynamics (Okayama/Kyoto, 1992), Plenum, New York (1995), 217–226.
[50] C. Voisin, *Hodge Theory and Complex Algebraic Geometry, I*, Cambridge Univ. Press, Cambridge (2002).
[51] Ya. Vorobets, *Plane structures and billiards in rational polygons: the Veech alternative*, Russian Math. Surveys **51** (1996), 779–817.
[52] A. Zorich, *Square tiled surfaces and Teichmüller volumes of the moduli spaces of Abelian differentials*, Rigidity in Dynamics and Geometry (Cambridge, 2000), Springer, Berlin (2002), 459–471.

CHAPTER 7

Ergodic Theory of Translation Surfaces

Howard Masur*

Department of Mathematics, UIC, Chicago, IL 60607-7045, USA
E-mail: *masur@math.uic.edu*

Contents

*The author is partially supported by NSF DMS 2-496045.

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

This page intentionally left blank

1. Three definitions of translation surface or flat surface and examples

In this survey article we describe the ergodic theory of flows on translation surfaces. We relate this theory to the dynamics of the $\mathrm{SL}(2,\mathbb{R})$-action on the moduli space of translation surfaces. We describe recent results on the diagonal subgroup also known as the Teichmüller geodesic flow and results on the unipotent flow.

There is considerable overlap of material here with the survey article [4] as well as with the survey article of T. Schmidt and P. Hubert in this handbook [2].

We are going to give three (equivalent) definitions of translation surface. Equivalently, these will be called flat surface with trivial linear holonomy or just flat surfaces. The first definition is via charts. The second definition is the most geometric and is by glued polygons. The third definition is complex analytic. They arise as the flat structure associated to a holomorphic 1-form on a Riemann surface. We will indicate (but not provide a complete proof) their equivalence.

Let M be a closed topological surface, of genus $g \geqslant 1$.

DEFINITION 1. A translation surface consists of a finite set of points (the singularity set) $\Sigma = \{x_1, x_2, \ldots, x_m\}$ and an open cover of $M - \Sigma$ by sets $\{U_\alpha\}$ together with charts $\phi_\alpha : U_\alpha \to \mathbb{R}^2$ such that for all α, β, with $U_\alpha \cap U_\beta \neq \emptyset$,

$$\phi_\alpha \phi_\beta^{-1}(v) = v + c.$$

At each singular point the surface has a $2\pi c$ cone singularity.

Specifically, since the Euclidean metric on the plane is preserved by translations, the notion of direction and parallel lines makes sense on the complement of the singularity set. In fact we get a metric ds, by pulling back the Euclidean metric on the plane via these coordinate charts. In this metric geodesics that do not go through singularities are straight lines in a fixed direction, and such geodesics never intersect themselves, except possibly to close up.

DEFINITION 2. For each direction θ and each nonsingular point p define the flow $\phi_t(p)$ to be the point obtained after moving in the direction θ for time t, starting at p.

The flow $\phi_t : X \to X$ preserves the natural Euclidean measure (normalized to have total area one) on the surface. It is defined for all time only on the set of full measure of points that do not run into a singularity either in forwards or backwards time. A major part of these notes will be devoted to describing ergodic properties of this flow.

At each singular point we write $ds^2 = dr^2 + (cr\,d\theta)^2$, a conical singularity written in polar coordinates. We require c to be a positive integer. For $c = 1$, we simply recover the Euclidean metric. If $c > 1$, we have a $2\pi c$ cone angle. We can think of a point with a $2\pi c$ cone angle as $2c$ Euclidean half discs glued together along half lines—for the case $c = 2$ see Figure 1.

The total angle around each vertex is required to be $2\pi c$, c a positive integer.

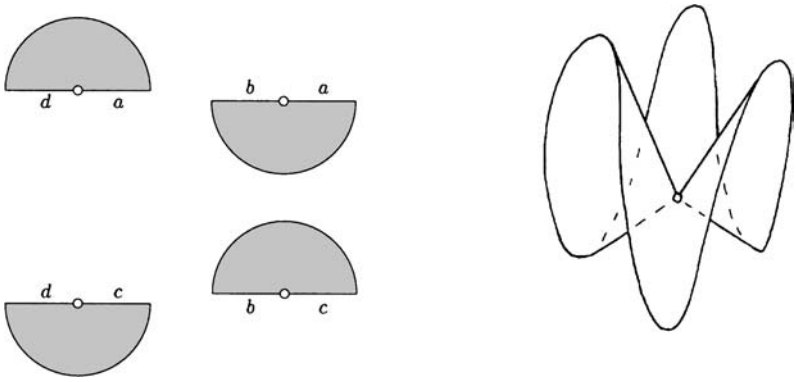


Fig. 1. Flat surface near a singularity.

Geodesics can change direction if they go through a singular point. A pair of straight lines through the singular point form a geodesic if the angle between them is at least 2π.

Metrically we can also describe these as flat surfaces with conical singularities of the above type and trivial linear holonomy. The latter means that parallel transport of a vector around a path missing the singularities comes back to the same vector. This explains why these surfaces are also called *flat surfaces with trivial linear holonomy*.

DEFINITION 3. A saddle connection is a geodesic joining two of the singularities with no singularities in its interior.

In each coordinate chart it is a straight line in the Euclidean metric. An oriented saddle connection determines a vector called the *holonomy vector* of the saddle connection.

It is a standard fact (see [26]) that between any two points there is a unique geodesic in any homotopy class. In particular, there is a unique geodesic joining any two singularities in each homotopy class. The geodesic is a union of saddle connections. We sketch an argument which says that the set of holonomy vectors of saddle connections is a discrete subset of $\mathbb{R}^2$. This fact is used in the proof of Veech dichotomy and is implicit in any discussion of counting problems. (See the articles of Hubert–Schmidt and Eskin in this handbook.) Another sketch is given in [2].

Lift the metric to the universal cover, to give a complete metric on the hyperbolic plane. Fix a Dirichlet fundamental domain $\mathcal{F}$ for the action of the covering group. Let D be its diameter. A ball of radius $R + D$ about a base point in $\mathcal{F}$ intersects only a finite number of translates of $\mathcal{F}$. Any saddle connection of length at most R must lift to a saddle connection joining a singularity in $\mathcal{F}$ to a singularity in a ball of radius $R + D$. There are only finitely many such points and hence only finitely many such saddle connections.

As mentioned in the article of P. Hubert and T. Schmidt [2], the $\mathrm{SL}(2, \mathbb{R})$ action on flat surfaces can be defined as postcomposition with charts—we discuss this action in more detail later.

Our next definition of a flat surface is the most geometric and often useful when we need to visualize these objects.

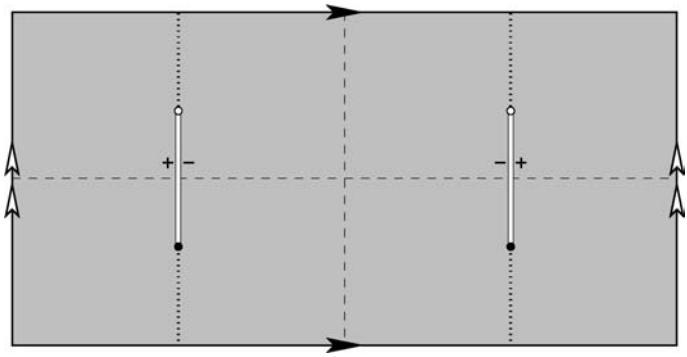


Fig. 2. Slit torus example.

DEFINITION 4. A *translation surface* is a finite union of Euclidean polygons $\{\Delta_1, \Delta_2, \ldots, \Delta_n\}$ such that

- the boundary of every polygon is oriented so that the polygon lies to the left;
- for every $1 \leqslant j \leqslant n$, for every oriented side s_j of Δ_j there is a $1 \leqslant k \leqslant n$ and an oriented side s_k of Δ_k so that s_j and s_k are parallel and of the same length. They are glued together in the opposite orientation by a parallel translation. (Note that this means that as one moves along a glued edge, one polygon appears to the left, the other to the right.)

It follows that the total angle around each vertex is $2\pi c$, c a positive integer. Note that when we speak of Euclidean polygons we fix their embedding into a standard Euclidean plane up to a parallel translation. In particular we distinguish two polygons obtained one from the other by a nontrivial rotation. Another way to say the same thing is that we equip a translation surface with a choice of vertical direction.

The rational billiard table examples (see [2]) yield surfaces of this form. However, note that in general we do not require the angles of the polygons to be rational, as is the case for the billiards. The best way to see this definition is by considering a few examples:

The first example is a regular octagon with opposite sides identified. This gives rise to a surface of genus two with one singularity of angle 6π (all the vertices collapse to one point, yielding an angle $8(3\pi/4)$). This is an example of a Veech surface which satisfies the Veech dichotomy (see [2]). Namely for any direction, either all the orbits in that direction are closed or equally distributed.

Another example also gives a surface in genus two but which turns out to have very different ergodic properties.

Consider a $1 \times 1/2$ rectangle with a barrier of length $\alpha/2$ hanging down from the top of the rectangle at its midpoint: that is, a vertical line segment from $(1/2, 1/2)$ to $(1/2, 1/2 - \alpha/2)$. billiards in this polygon gives rise [2] to a surface with opposite sides identified (of side length two), with two slits of length α inside it (see Figure 2). The left side of the left slit is identified with the right side of the right slit, and the right side of the left slit is identified with the left side of the right one. The rectangle with slits yields a torus with two holes, when opposite sides are identified, and when the slits are glued, the result is a

genus two surface. There are two singularities, each with a 4π cone angle coming from the endpoints of the glued slit.

In fact this example illustrates the definition by polygons. There are four generalized 7-gons, each of which has six vertex angles of $\pi/2$ and one angle 2π.

When α is rational, the surface is a Veech surface and the Veech group is a finite index subgroup of $\mathrm{SL}(2,\mathbb{Z})$. These are particular examples of arithmetic Veech surfaces.

Now in general, since the gluings of the polygons are realized by parallel translations, it is clear that a surface satisfying the definition by polygons satisfies the definition by charts. Conversely, one can show that a translation surface has a triangulation by geodesic triangles so a surface satisfying the first definition satisfies the second.

The third definition is complex-analytic.

DEFINITION 5. A *translation surface* is given by a pair (X,ω) where X is a Riemann surface and ω is a holomorphic 1-form (Abelian differential) on X.

Recall that this means that to each holomorphic chart z is assigned a holomorphic function $f(z)$ such that in an overlapping chart ζ with function $g(\zeta)$, the relation is

$$g(\zeta)\frac{d\zeta}{dz} = f(z).$$

In the article of P. Hubert and T. Schmidt [2] they show how to go from a pair (X,ω) to a collection of charts where the transition maps are translations, i.e., our first definition (zeroes of the 1-form correspond to singularities, etc.). Specifically in a neighborhood of a point p_0 which is not a zero of ω there are holomorphic coordinates z defined by

$$z(p) = \int_{p_0}^{p} \omega$$

which give $\omega = dz$. In an overlapping neighborhood similarly defined coordinates z' will satisfy

$$z' = z + c$$

so that the change of coordinates is a translation. At a zero of order k in appropriate coordinates

$$\omega = z^k dz = d\left(\frac{z^{k+1}}{k+1}\right)$$

and so the surface is locally a $k+1$ fold cover over the complex plane. This means that the zero of order k gives rise to a singularity with cone angle $2\pi(k+1)$.

In this language the (affine) holonomy of a saddle connection β coincides with $\int_\beta \omega = \int_\beta dz$, where we have identified the complex numbers with $\mathbb{R}^2$, and so we can consider the holonomy to be a complex number (with real and imaginary parts) or as a vector.

To get from the first definition to this one, simply pull back the natural 1-form dz via the charts. This defines a holomorphic 1-form ω on the surface. The cone singularity gives rise to a zero of ω.

2. Spaces of translations surfaces and Riemann surfaces

For the rest of this article we will use the notation (X, ω) to refer to a translation surface.

A translation surface has three pieces of topological data: the genus, the set of zeros, and the multiplicity of the singularities. We can represent the topological data by $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$, where α_i denotes the order of the ith zero. It is classical and in any case follows from either an Euler characteristic argument or from the Gauss–Bonnet theorem that

$$\sum_{i=1}^{k} \alpha_i = 2g - 2.$$

For example, given the data $\alpha = (2)$, the surface has genus two with one singularity with cone angle 6π. Given $\alpha = (1, 1)$, the genus is still two, but with two singularities, each of cone angle 4π.

We want to consider the space of all translation surfaces with fixed topological data. For this, we need to define an equivalence relation on such surfaces.

We say that two surfaces are equivalent, if there is an orientation preserving isometry from one to the other preserving the given preferred direction. This definition distinguishes between polygons that differ by rotations. In the complex analytic definition, it distinguishes between (X, ω) and $(X, e^{i\theta}\omega)$.

DEFINITION 6. Given topological data α, we define the moduli space $\mathcal{H}(\alpha)$ as the space of translation surfaces with topological data α together with a choice of direction under the above equivalence relation. If we add the condition that the surfaces have area 1 we denote the resulting space by $\mathcal{H}_1(\alpha)$. These moduli spaces are also called strata.

On the other hand for any genus g we may define the Riemann moduli space $\mathcal{M}_g$ as the space of Riemann surfaces of genus g up to conformal equivalence. Every closed Riemann surface of genus $g > 1$ carries a metric of constant curvature -1 in its conformal class, so $\mathcal{M}_g$ is also the space of hyperbolic metrics on a surface up to equivalence by isometries.

For each $\alpha = (\alpha_1, \ldots, \alpha_k)$, define g by $2g - 2 = \sum_{i=1}^{k} \alpha_i$. There is then a map

$$\pi : \mathcal{H}(\alpha) \to \mathcal{M}_g$$

which sends (X, ω) to X. The map only remembers the complex structure on the surface defined by the Abelian differential.

As a main motivating example, let us consider the space of tori with specified directions, i.e., $\mathcal{H}(\emptyset)$. Recall that while two tori differing by a rotation are identical as metric spaces,

the vertical direction on each torus is distinct, so we do not consider them the same point—as opposed to the moduli space of Riemann surfaces $\mathcal{M}_1$ where these are the same point.

The space of tori $\mathcal{H}(\emptyset)$ can also be viewed as the space of unit volume lattice in $\mathbb{R}^2$ (together with a specified direction), which is identified with the symmetric space $\mathrm{SL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z})$ (if one ignores the direction, we get instead $\mathbb{H}^2/\mathrm{SL}(2,\mathbb{Z})$, the moduli space of tori $\mathcal{M}_1$).

In general the moduli spaces $\mathcal{H}(\alpha)$ are not necessarily connected, though each has no more than three connected components. The components have been classified by Kontsevich and Zorich [18].

If we allow reflections as well as translations in gluings (or equivalently, allow transitions to be of the form $z \mapsto \pm z + c$), we get quadratic differential and the classification is different.

3. SL(2, $\mathbb{R}$)-action and invariant measures

Recall from the survey paper of Hubert and Schmidt [2] the $\mathrm{SL}(2,\mathbb{R})$-action. In the language of polygons, we can define the action as follows. Given a translation surface (X,ω) (i.e., a finite collection of polygons $\{\Delta_i\}$) and a matrix $A \in \mathrm{SL}(2,\mathbb{R})$, we define the translation surface $A\cdot(X,\omega)$ by the collection of polygons $\{A\Delta_i\}$. The gluing pattern is preserved since linear maps preserve parallel lines. One can check that the definition does not depend on how one represents the surface as a union of polygons.

In the language of complex analysis, the action of the rotation

$$r_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

is the same as multiplying the Abelian differential ω by $e^{i\theta}$.

The action of a matrix in $\mathrm{SL}(2,\mathbb{R})$ does not change the topological data of a flat surface. Thus, for each stratum $\mathcal{H}(\alpha)$, we have an $\mathrm{SL}(2,\mathbb{R})$-action. We are interested in defining a measure μ on $\mathcal{H}_1(\alpha)$ which is invariant under this action.

We do this by defining coordinates for this space, and then pulling back natural Lebesgue measure on the coordinate space. Our first coordinates will arise from our "visual" definition of the moduli space using polygons.

Suppose $\{\Delta_i\}$, a collection of polygons, represents a point in $\mathcal{H}(\alpha)$. It is obvious that there is some finite collection of sides $v_1, v_2, \ldots, v_N$ which determine the surface. For example, for a flat torus, the surface is determined by two sides v_1, v_2 of a parallelogram. For the surface to be in $\mathcal{H}_1(\emptyset)$ one has the further condition that the area determined by the polygon is *one*, which we denote by $v_1 \wedge v_2 = 1$. Another example is the octagon, for which we need four vectors (once a side is determined, so is its opposite).

These v_i yield local coordinates for $\mathcal{H}(\alpha)$ giving a map $\phi : \mathcal{H}(\alpha) \to (\mathbb{R}^2)^N$. We consider Lebesgue λ measure on $(\mathbb{R}^2)^N$, restricted to the hypersurface corresponding to the area 1 surfaces and define $\mu = \phi^*\lambda$. This measure is independent of the choice of coordinates and the way the surface is cut into polygons (in particular the number of polygons may change, but the number of sides necessary to determine the surface does not).

A more formal way to see this definition, is by starting with the surface (X, ω), and its set of singularities Σ. Consider the relative homology group $H_1(X, \Sigma; \mathbb{Z})$. This is an $N = 2g + n - 1$ dimensional space, where n is the number of singularities. Fix a basis $\{\beta_1, \beta_2, \ldots, \beta_N\}$. Define coordinates for (X, ω) by $\{\int_{\beta_i} \omega\} \in \mathbb{R}^{2N}$. Once again consider Lebesgue measure on the image of this map, and pull it back to get a measure on $\mathcal{H}(\alpha)$. This is more easily seen to be invariant of choices—in particular, any change of basis is a determinant one matrix. For the same reason it is invariant under the $\mathrm{SL}(2, \mathbb{R})$-action.

Returning to the torus, recall that the space of tori is

$$\mathcal{H}_1(\emptyset) = \mathrm{SL}(2, \mathbb{R})/\mathrm{SL}(2, \mathbb{Z}).$$

It has finite volume because $\mathrm{SL}(2, \mathbb{Z})$ is a lattice in $\mathrm{SL}(2, \mathbb{R})$. We can see this directly because the space of tori (without normalization) is simply the set of all pairs of noncolinear vectors $v_1, v_2 \in \mathbb{R}^2$. This space clearly has infinite Lebesgue measure. When restricting to the area one tori, the space is noncompact, since the vector v_1 can be arbitrarily short. However the space $\mathcal{H}_1(\emptyset)$ has finite volume because of the easily proven fact:

$$\mu\big\{(v_1, v_2) \in \mathbb{R}^2 \times \mathbb{R}^2 \colon |v_1 \wedge v_2| \leqslant 1\big\} < \infty.$$

A similar computation explains the finite measure in general. We will sketch this explanation. In A. Eskin's survey article [1] he explains how to actually compute the measures of these spaces.

On any flat surface (X, ω), consider a closed geodesic in some direction which does not hit any singularities. Then there is a cylinder, containing this curve, which is filled with closed curves, parallel of the same length. If we make it as large as possible, it is called a metric cylinder. If $g > 1$, the boundary of the metric cylinder is a union of saddle connections. It turns out that for each genus g, there is a universal constant $C(g)$ such that if $\mathrm{diam}(X, \omega) \geqslant C(g)$, there is a metric cylinder on the surface such that the distance h across the cylinder satisfies $h \sim \mathrm{diam}(X, \omega)$; that is, they are comparable up to a definite factor.

Since the measure is defined by the holonomy along saddle connections, the measure of the part of moduli space corresponding to surfaces of diameter at most $C(g)$ is finite. On the part of the moduli space consisting of surfaces with large diameter, the above consideration says that these (area 1) surfaces have cylinders with small circumference and large distance across them. We can take as part of the basis for the homology a curve parallel to the cylinder with holonomy v_1 and a curve across the cylinder with holonomy vector v_2. But recalling that

$$\mu\big\{(v_1, v_2) \in \mathbb{R}^2 \times \mathbb{R}^2 \colon |v_1 \wedge v_2| \leqslant 1\big\} < \infty,$$

we have that the measures of these "cusps" are finite, and thus we have the following theorem. Complete proofs can be found in [28] and [22].

THEOREM 1. *For each stratum $\mathcal{H}_1(\alpha)$, $\mu(\mathcal{H}_1(\alpha)) < \infty$.*

4. Ergodicity of flows defined by translation surfaces

In this section we begin the discussion of the properties of the flow ϕ_t defined for each direction θ.

To avoid the problem of measures which are concentrated on the singular set, we consider only measures supported on the *punctured surface* $X - \Sigma$.

The first notion is purely topological. We will say that the flow in direction θ is minimal if there are no closed curves in direction θ. Equivalently, other than a finite number of saddle connections, every orbit that does not run into a singularity is dense, and if an orbit runs into a singularity in forward (respectively backwards) time then it is dense in backward (respectively forward) time.

Recall that a flow is ergodic if any invariant set has measure zero or measure one. Let ν be surface area. In this case the Birkhoff ergodic theorem states that for $f \in L^1(X, \nu)$, and for almost all p,

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T f\big(\phi_t(p)\big)\, dt = \int_X f\, d\nu.$$

If this convergence holds for every point p, and every continuous function f, the flow is said to be uniquely ergodic. This is equivalent to saying that the measure ν is the unique normalized flow-invariant measure on $X \setminus \Sigma$.

For motivation, we once again turn to the case of the torus $\mathbb{R}^2/(\mathbb{Z} \oplus \mathbb{Z})$. If the direction θ has rational slope, then every orbit is closed. On the other hand, if θ has irrational slope, then the flow is minimal, and moreover by the classical theorem of Weyl, the flow is uniquely ergodic.

However, even in the case of the torus there is a flow constructed by Furstenberg [13], which is minimal, but not uniquely ergodic. For a general treatment of the subject of nonunique ergodicity, see Section 14.5 of the book [15] and Sections 12.3 and 12.4 of [14].

We now want to exhibit a minimal nonuniquely ergodic example on a translation surface of genus 2. Veech [27] considered the following dynamical system. Take a pair of unit circles and mark off a segment of length β on each circle in the counterclockwise direction with one endpoint at $(1, 0)$. Start on one circle and rotate counterclockwise by angle θ until the point lands in the segment. Then switch to the corresponding point on the other circle, rotate by θ until the orbit lands in the segment again, switch back to the first circle and so forth. Veech showed that for any irrational θ with unbounded partial quotients in its continued fraction expansion, there are irrational β so that the dynamical system is minimal, but not uniquely ergodic. What happens is that sets of orbits of positive measure spend asymptotically more than half their time on one circle and less than half the time on the other.

This dynamical system can be seen to be equivalent to the billiard flow on the billiard table with a slit described in Figure 2. Recall, it was given by a rectangular $1 \times 1/2$ table with a slit of length $\alpha/2 = (1 - \beta)/2$ hanging down from the midpoint of the top side. The surface (X, ω) associated to has genus two, with two singular points, each with angle 4π. It is formed from a 2×1 rectangle with a pair of slits and appropriate identifications. Now take two circles in the vertical direction. The first follows one side of the slit and then

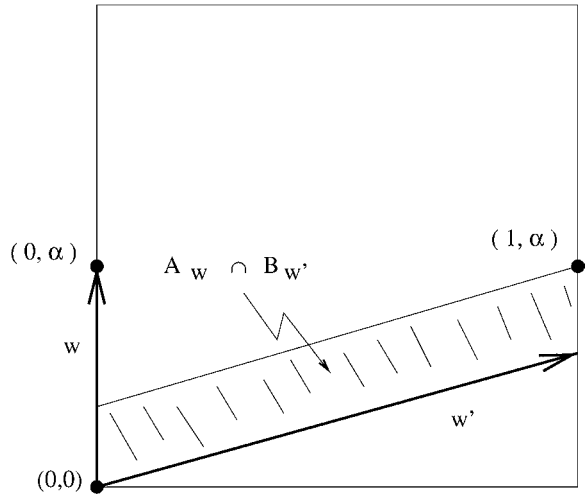


Fig. 3. Sheet interchange.

a vertical segment of length β joining the two singularities and which passes through the point $(1/2, 1) \sim (1/2, 0)$. The second follows the other slit and passes through $(3/2, 1) \sim (3/2, 0)$. The first return map to those circles of a flow in direction θ, gives the dynamical system described by Veech.

In this section we show how to build these minimal nonergodic examples geometrically. Additional details can be found in [4].

THEOREM 2. *When β is irrational there are uncountably many directions θ such that the flow in direction θ is minimal and not ergodic.*

In order to prove the theorem we will view the surface (X, ω) differently. Cut the surface along the pair of dotted vertical lines that go from P to Q in Figure 2. The result is a pair of tori each with a hole consisting of the pair of vertical lines. Each torus then can be thought of as a standard square tori T slit along a segment w_0 going from $p_1 = (0, 0)$ to $p_2 = (0, \alpha)$. The surface (X, ω) is reformed by gluing the tori together pairwise along w_0. The union of this pair of slits partitions the surface (X, ω) into two pieces A_{w_0} and B_{w_0} of equal area.

We will look for other slits w' defining (X, ω). That is, we want another pair of saddle connections w' joining p_1 to p_2 so that their union also splits (X, ω) into two pieces of equal area. The new slit w' will cut the original slit, and so the new partition $A_{w'} \cup B_{w'}$ of (X, ω) will differ from the original.

On the universal cover $\mathbb{R}^2$ of the torus T, the new slit w' is a line from $(0, 0)$ to $(m, \alpha + n)$ for some integers m, n. The condition that the pair of slits w' divide (X, ω) is equivalent to the condition that m and n are both even. Equivalently, on T, w' intersects w an odd number of times in its interior. It is also equivalent to saying that w and w' are homologous mod(2) on T. Then the change in partition on (X, ω) measured by $c = (A_w \cap B_{w'}) \cup (B_w \cap A_{w'})$ is a union of an even number of parallelograms with sides on w and w' (here thought of as vectors). Thus the area of c is bounded by $2|w \times w'|$ (see Figure 3).

The main step in the proof is to find uncountably many sequences $\{w_n\}$ of vectors determining partitions $\{A_n, B_n\}$ such that

$$\sum_{n=1}^{\infty} \nu(A_{n+1} \triangle A_n) < \infty.$$

Here ν is area on the surface. The directions of any sequence of these vectors will converge to a limiting direction θ. Assuming that such sequences can be found, we show first that the flow in direction θ is not ergodic.

Let

$$A_\infty = \liminf A_n = \{x\colon\ \exists N\colon\ x \in A_n,\ \forall n \geqslant N\}$$

and let B_∞ be defined similarly. The condition $\sum \nu(A_{n+1} \triangle A_n) < \infty$ and the Borel–Cantelli lemma imply

$$\nu\{x\colon\ x \in A_n \triangle A_{n+1} \text{ infinitely many } n\} = 0$$

so $\nu((X, \omega)\backslash(A_\infty \cup B_\infty)) = 0$. By symmetry, we get $\nu(A_\infty) = \nu(B_\infty) = 1$.

Now we claim that A_∞ is a.e. invariant under the flow $\{\phi_t\}$ in direction θ, i.e.,

$$\nu\big(\phi_t(A_\infty) \triangle A_\infty\big) = 0$$

for all times t. Assume that the claim is false so that there is some $\delta > 0$ and t_0 such that

$$\nu\big(\phi_{t_0}(A_\infty) \triangle A_\infty\big) \geqslant \delta > 0. \tag{1}$$

Without loss of generality we may assume that the limiting direction is vertical. It follows from the summability condition on the areas that

$$h_n \to 0,$$

where h_n is the horizontal component of the holonomy of w_n. (Recall the holonomy is a vector.) Pick n such that

$$\nu(A_n \triangle A_\infty) < \delta/8 \tag{2}$$

and

$$t_0 h_n < \delta/8. \tag{3}$$

The flow invariance of the measure, (2), (1) and the triangle inequality imply

$$\nu\big(\phi_{t_0}(A_n) \triangle A_n\big) > \delta - 2\delta/8 = 3\delta/4.$$

Thus at time t_0 at least $3\delta/8$ of the measure of A_n flows to its complement. However if a point crosses w_n, the boundary of A_n at time t_0 of the flow, its vertical distance to w_n must be at most t_0. The set of points whose vertical distance to w_n is at most t_0 lie in a parallelogram whose sides are w_n and a vertical segment of length t_0. The area of such a parallelogram is $h_n t_0 < \delta/8$ by (3). We have a contradiction, proving the claim.

From the claim there is an argument that says there is a set A' with $\nu(A' \triangle A_\infty) = 0$ such that A' is ϕ_t invariant. This implies that the flow in direction θ is not ergodic completing the first step.

Let us return to finding an uncountable number of sequences of w_n satisfying the condition $\sum_n \nu(A_{n+1} \triangle A_n) < \infty$. We wish to show that the limiting directions are distinct for then we will have constructed an uncountable number of nonergodic directions. This will guarantee an uncountable number of *minimal* nonergodic directions, since in a nonminimal direction there is a saddle connection, and there are only countably many saddle connections.

Fix any sequence ρ_n such that $\sum \rho_n < \infty$. We will build an infinite directed tree with each "parent" vertex w_j leading to a pair of "child" vertices w_{j+1}. At level j there will be 2^j vertices. Each vertex will correspond to a pair (p_j, q_j) which will yield a slit joining $(0,0)$ to $(p_j, q_j + \alpha)$.

Let $w_0 = (0, \alpha)$ and suppose inductively we have found 2^j vectors $w_j = (p_j, q_j + \alpha)$ at stage j. For any pair (p_j, q_j) form the ratio $\frac{q_j+\alpha}{p_j}$, the slope of the slit. Define δ_j to be the minimum distance between the slopes of any pair of distinct w_j at level j. For any (p_j, q_j) we will look for integer solutions r, s of

$$2\big|p_j s - (q_j + \alpha)r\big| < \rho_j.$$

Since α is irrational, so is each $\frac{p_j}{\alpha+q_j}$, and so there are infinitely many coprime solutions (r, s) of the above inequality. Choose any two sets of solutions (r_j, s_j) so that

$$\frac{\rho_j}{(q_j + \alpha)(q_j + \alpha + 2s_j)} \leqslant \delta_j/4$$

and for each, set $p_{j+1} = p_j + 2r_j$; $q_{j+1} = q_j + 2s_j$ and then

$$w_{j+1} = (p_{j+1}, q_{j+1} + \alpha).$$

A direct calculation also shows that

$$\nu(A_{j+1} \triangle A_j) < 2|w_{j+1} \times w_j| \leqslant 4\big|p_j s_j - (q_j + \alpha)r_j\big| < 2\rho_j,$$

giving the desired summability condition.

The proof will be complete when we show that the directions of a sequence of w_j converge and distinct sequences give distinct limiting directions. A calculation shows

$$\left|\frac{p_j}{q_j + \alpha} - \frac{p_{j+1}}{q_{j+1} + \alpha}\right| \leqslant \delta_j/4;$$

that is, the distance between the slopes of a parent and child is at most $\delta_j/4$. The triangle inequality says that the distance between slopes of children of the same parent is at most $\delta_j/2$ and so $\delta_{j+1} < \delta_j/2$.

Since the distance between the slopes of a parent and a child goes to 0, the slopes of the vertices w_j of any geodesic in the tree converges.

We finally show that limits of slopes of w_j along distinct geodesics are different. For suppose two geodesics l_1, l_2 are different for the first time at level j with vertices w_j^1, w_j^2 (thought of as parents). Let θ_1, θ_2 be the limiting slopes of the vertices along l_i. Since the slope of each child at level $m+1$ is within $\delta_m/4$ of the slope of the parent at level m, summing the geometric series says that the difference of the slope of θ_i and the slope of w_j^i is smaller than $\delta_j/2$. Since the slopes of w_j^1, w_j^2 are at least δ_j apart, we must have $\theta_1 \neq \theta_2$.

5. Further results on unique ergodicity

The above construction was generalized recently [11] to show that on any translation surface in genus 2 which is not a Veech surface there is some direction for which the flow is minimal and not ergodic. A natural question is which translation surfaces have minimal nonergodic directions. Veech surfaces do not, due to the Veech Dichotomy (which was proved in [2]).

The existence of minimal nonergodic directions led to work about their prevalence. For each (X, ω), define $\mathrm{NE}(X, \omega)$ to be the set of $\theta \in [0, 2\pi)$ such that the flow ϕ_t in the θ direction is not ergodic. Equivalently, $\mathrm{NE}(X, \omega)$ is the set of θ such that the flow ϕ_t in the vertical direction of $e^{i\theta}\omega$, is not ergodic. In [16] it was shown that the Lebesgue measure of $\mathrm{NE}(X, \omega)$ is 0. The idea of the proof is the following. Let

$$r_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

be the rotation group in $\mathrm{SL}(2, \mathbb{R})$ and let

$$g_t = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}$$

be the diagonal group.

The action of the diagonal subgroup is known as Teichmüller geodesic flow, since images of these orbits under the projection π to the Riemann moduli space $\mathcal{M}_g$ are geodesics in the Teichmüller metric on $\mathcal{M}_g$.

One shows that for large t most points on the circle $g_t r_\theta(X, \omega)$ are not near the cusp in moduli space. This is combined with the following theorem [21] (whose proof is sketched in the next section).

THEOREM 3. *Suppose (X, ω) is a translation surface. Suppose the flow in the vertical direction is not uniquely ergodic. Then $X_t = \pi g_t(X, \omega)$ eventually leaves every compact set in $\mathcal{M}_g$. That is, the Teichmüller geodesic associated to (X, ω) is divergent.*

We note that this theorem is also one of the ingredients in the proof of the Veech dichotomy [2].

We continue with some remarks about Theorem 3. It is a basic fact that the moduli space $\mathcal{M}_g$ is noncompact. The reason is that one may have a sequence of surfaces and curves on those surfaces whose lengths in the hyperbolic metric (assume $g > 1$) go to zero. Such surfaces cannot converge to a compact surface. On the other hand it is a basic fact [24] that this is the only way to leave compact sets in $\mathcal{M}_g$. Namely, if X_n is a sequence that eventually leaves every compact set, then there is a sequence of curves γ_n such that the length of γ_n (in the hyperbolic metric on X_n) goes to zero.

It is easy to see that if there is a closed leaf in the vertical direction (in particular, the flow is not minimal) of ω, then X_t eventually leaves every compact set of $\mathcal{M}_g$. Namely, since g_t shrinks lengths in the vertical direction by a factor of e^t, the length in the flat metric of $g_t(X, \omega)$ of any closed vertical leaf goes to 0. If there were a subsequence of X_t converging to a compact surface X_0, there would be a further subsequence of $g_t(X, \omega)$ converging to some (X_0, ω_0). This (X_0, ω_0) would assign 0 length to a closed curve, which is impossible.

In the minimal case there are no closed vertical leaves. This means that under the flow g_t the length of any *fixed* curve γ must go to infinity in the flat metric of $g_t(X, \omega)$ as $t \to \infty$. What the Theorem 3 says is that there is a sequence of distinct simple closed curves γ_n such that for any $\varepsilon > 0$, for sufficiently large t there is a curve $\gamma_n = \gamma_n(t)$ such that the length of γ_n in the flat metric of $g_t(X, \omega)$ is smaller than ε.

The measure 0 result was generalized by Veech [29] to Borel probability measures on $[0, 2\pi)$ that satisfy certain growth conditions. Normalized Cantor–Lebesgue measure on the Cantor middle third set is an example of such a measure.

Further work concerns the Hausdorff dimension of $\mathrm{NE}(X, \omega)$. In [22] it was shown that for each component of each stratum (other than several low dimensional exceptional cases covered by the Weyl theorem) there is a $\delta > 0$ such that for μ a.e. (X, ω) in the component, $\mathrm{NE}(X, \omega)$ has Hausdorff dimension δ. The construction of these nonergodic directions on a generic surface uses a method similar to that described in the Veech example.

In [21] it was shown that the Hausdorff dimension of $\mathrm{NE}(X, \omega)$ is always bounded by $1/2$. The proof of this result is also based on Theorem 3 and estimates on counting saddle connections. As discussed above, Theorem 3 says that for $\theta \in \mathrm{NE}(X, \omega)$, for all large times there is a short saddle connection. Typically there may be many such short intersecting saddle connections at any given time, and this collection of short saddle connections change with time. However one can make a choice of a short saddle connection in this collection at any time so that successive choices are *disjoint*. Thus the proof amounts to estimating the size of the set of angles θ such that along the orbit $g_t r_\theta(X, \omega)$ there is a sequence of saddle connections that become successively short, and such that each is disjoint from its predecessor. This problem can be reduced to counting problems for saddle connections. There is an estimate [20,12] which says that the number of saddle connections of length T grows at most quadratically in T, and another estimate which says for fixed saddle connection of length l, the number of *disjoint* saddle connections of length at most L grows roughly linearly in L/l. The comparison of linear growth and quadratic growth accounts for the dimension $1/2$.

Y. Cheung [8] has shown that this estimate is sharp. Specifically, suppose an irrational α satisfies a Diophantine condition that for some $s > 0$ there are no fractions p/q that satisfy

$$|\alpha - p/q| < \frac{1}{q^s}. \tag{4}$$

Then for the rectangular table with a slit of length $\alpha/2$ described in Section 4, $\dim \mathrm{NE}(X, \omega) = 1/2$.

6. Boshernitzan's Theorem and sketch of proof of Theorem 3

Before we turn to the proof of the Theorem 3, we state an alternative criterion for divergence, given by Boshernitzan [7], formulated in terms of interval exchange map.

If we consider the flow in the vertical direction, then by considering the first return to a piece of horizontal transversal I, we obtain an interval exchange map T, whose discontinuity points correspond to leaves that run into singular points before returning to I. Suppose T exchanges k intervals. Let $T^{(n)}$ be the nth iterate of T. It will be an interval exchange on approximately kn intervals. Let m_n denote the length of the shortest of these intervals.

THEOREM 4 [7]. *If T is not uniquely ergodic, then $nm_n \to 0$.*

This is slightly weaker than Theorem 3 as it only guarantees divergence of the geodesic in a stratum, whereas Theorem 3 guarantees divergence in moduli space.

To explain the difference, in a stratum $\mathcal{H}(\alpha_1, \ldots, \alpha_k)$ where $k > 1$, one may leave compact sets by a sequence of translation surfaces such that a pair of singularities come close together. If no closed curve becomes short then one stays in a compact set of $\mathcal{M}_g$.

Now the reason that the criterion $nm_n \to 0$ implies divergence in the stratum is as follows. The discontinuity points on I of the interval exchange $T^{(n)}$ are points of the form $T^{(-l)}(x)$ for $l \leqslant n$, where x is a discontinuity point of T. Hence each interval yields a saddle connection crossing it such that the vertical component of its holonomy has length $O(n)$. However, the short interval yields a saddle connection γ_n such that in addition, the horizontal component of its holonomy is $O(m_n)$. Since $nm_n \to 0$, for some interval of times t, the length of both the vertical and horizontal component of γ_n in the metric of $g_t(X, \omega)$ are small. One can show that for any time t there is such a γ_n.

We are now ready to sketch the proof of Theorem 3. Let $\{\phi_t\}$ denote the vertical flow of ω. As explained above, we can assume it is minimal but not uniquely ergodic. The set of invariant probability measures for ϕ_t is a finite dimensional convex set and the extreme points are mutually singular ergodic measures. For sake of argument assume there are exactly two ν_1, ν_2. (The general case is almost the same.) Since ν_i is invariant under the vertical flow, for any horizontal interval I, the measure ν_i decomposes into

$$\nu_i = \mu_i \times dy,$$

where μ_i is an ergodic measure on I invariant under the first return map ψ and y is the coordinate in the vertical direction. Since the ν_i are mutually singular, so are the μ_i and I can be chosen so that

$$\mu_1(I) \neq \mu_2(I).$$

Let χ_I be the indicator function of I. We say x is generic for μ_i if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \chi_I\big(\psi^n(x)\big) = \mu_i(I).$$

Thus if x_i is a generic point of μ_i then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \chi_I\big(\psi^n(x_1)\big) \neq \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \chi_I\big(\psi^n(x_2)\big). \tag{5}$$

It is a fact that μ_i almost all points of I are generic for μ_i.

We argue by contradiction. If the theorem is false, there is a sequence of times $t_n \to \infty$ and X_0 such that $X_{t_n} \to X_0 \in \mathcal{M}_g$. Since the part of $\mathcal{H}_1(\alpha)$ that lies over a compact set of $\mathcal{M}_g$ is also compact, by passing to further subsequences, we can assume there is ω_0 an Abelian differential on X_0 such that $g_{t_n}(X,\omega) \to (X_0,\omega_0)$.

Let $x_i \in I$ be generic for μ_i, $i = 1, 2$. We follow the image of x_i under the flow g_{t_n} and denote its image by $g_{t_n}(x_i)$. Note that each term in the sequence $g_{t_n}(x_i)$ is a point on a different Riemann surface X_i that evolves over time. Since the surfaces in question are compact, by passing to further subsequences we can assume that there exists $y_i \in X_0$ such that $g_{t_n}(x_i) \to y_i$. Since the surface X_0 is connected, and the set of generic points is of full measure for each μ_i and each of these sets is invariant under ϕ_t, it is not hard to show that we can pick x_i generic for μ_i such that y_1, y_2 lie on the same horizontal line h_1 of the limiting translation surface (X_0,ω_0). We will show that this contradicts (5).

Let l_1, l_2 short vertical lines of (X_0,ω_0) through y_1, y_2 and let R be the Euclidean rectangular box with vertical sides l_1 and l_2 and one horizontal side h_1. If l_1, l_2 are chosen small enough, R will have no singularities in its interior. Then the number of intersections of every connected horizontal line of (X_0,ω_0) with l_1 will differ with the number of its intersections with l_2 by at most 1.

For $i = 1, 2$, let $l_{i,n}$ denote bounded segments of the vertical leaf of $g_{t_n}(X,\omega)$ through $g_{t_n}(x_i)$ of equal length such that $l_{i,n} \to l_i$. Thus for n sufficiently large, with small error, every horizontal segment of $g_{t_n}(X,\omega)$ intersecting $l_{1,n}$ will intersect $l_{2,n}$. In particular, this is true for the long horizontal segment $g_{t_n}(I)$. Pulling back by g_{t_n}, we see that $g_{t_n}^{-1}(l_{i,n})$ are very long vertical leaves of the same length with respect to the original (X,ω) through x_1 and x_2, such that the ratio of the number of their intersections with I is approximately 1. But this is a contradiction to (5).

We describe how the Veech nonergodic example described earlier fits into the above theorem. The theorem says that for all large enough time t there is a curve $\gamma(t)$ on X_t with small hyperbolic length; the curve depends on the time. There is a sequence of dividing

curves formed from slits $w_j = (p_j, q_j + \alpha)$ such that each w_j becomes short in hyperbolic length for a *finite* interval of time before it becomes long. It is a standard fact in hyperbolic geometry [6] that intersecting curves are never simultaneously short, so the intervals of times that different slits are short in hyperbolic length are disjoint. The slit curves therefore cannot account for all the short curves in the family X_t. What happens is that each slit divides the surface into a pair of tori, and before that slit curve becomes long, a (r_j, s_j) curve on each torus becomes short, where recall from the construction, $p_{j+1} = p_j + 2r_j$, $q_{j+1} = q_j + 2s_j$. This also occurs for a finite interval of time before becoming long. It stays short until the next slit becomes short, defining a new pair of tori and the process repeats.

One way to think about why such examples are impossible in genus one, is that on a torus there are no disjoint nonhomotopic curves.

7. Further results on dynamics of actions of subgroups of SL(2, $\mathbb{R}$)

The first set of results have to do with the Teichmüller flow. The converse to Theorem 3 is not true. It is possible to construct examples of divergent geodesics such that the flow ϕ_t in the vertical direction is uniquely ergodic [10]. Another interesting line of work has concerned the rate of divergence of geodesics $g_t(X, \omega)$. Cheung [9] has recently shown that one can find geodesics with arbitrarily slow rates of divergence. Let $(X, \omega) \in \mathcal{H}(1, 1)$ be a surface which is a double cover over the torus and which is not a Veech surface. (An example is given by the slit torus considered in Section 4 with irrational α.) Then given any function $R(t) \to \infty$ there is a direction θ so that $\pi g_t(X, e^{i\theta}\omega)$ diverges in moduli space in $\mathcal{M}_g$, and such that

$$\tau\left(\pi g_t\left(X, e^{i\theta}\omega\right), \pi\left(X, e^{i\theta}\omega\right)\right) < R(t)$$

for all large t. Here $\tau(\cdot, \cdot)$ is the Teichmüller metric on $\mathcal{M}_g$.

He also showed that if α satisfies (4) and $e_0 > \max(2, s)$, then there is a Hausdorff dimension $1/2$ of directions $\theta \in \mathrm{NE}(X, \omega)$ such that the sublinear rate of divergence

$$r_+(\theta) = \limsup_{t\to\infty} \frac{\log \tau(\pi g_t(X, e^{i\theta}\omega), \pi(X, e^{i\theta}\omega))}{\log t} \leqslant 1 - \frac{1}{e_0}$$

holds. It would be interesting to know if slow rates of divergence in general implies unique ergodicity.

It is known [19,28] that the flow g_t is ergodic with respect to the natural "Lebesgue" measure μ on each component of each stratum. For the principal stratum (all simple zeroes) this implies in particular that the projection to the Riemann moduli space of almost every geodesic is dense. Thus the set of cobounded geodesics $g_t(X, \omega)$; those geodesics whose projection to the moduli space remain in some compact set (depending on the geodesic) has measure 0. One can then ask about the Hausdorff dimension of the set of cobounded geodesics.

The case of $g = 1$ is classical. Suppose X is the standard square torus and ω is the 1-form $e^{i(\pi/2-\alpha)}\, dz$. The lines in direction α are vertical with respect to ω. The behavior of

$g_t(X,\omega)$ in the moduli space $\mathrm{SL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z})$ is determined by the continued fraction expansion of α. In particular, the orbit is cobounded iff α has bounded partial quotients. The set of these irrational numbers has measure 0 and Hausdorff dimension 1.

The result for general (X,ω) is recent work of Kleinbock and Weiss [17]. They show that for any (X,ω), the set of $\theta \in [0,2\pi)$ such that $g_t(X, e^{i\theta}\omega)$ is cobounded has Hausdorff dimension 1.

Other interesting and important recent work in the dynamics in moduli space has been inspired by the dynamics of flows of subgroups of G acting on G/Γ, where G is a Lie group and Γ is a lattice subgroup. (See [3] for a survey.)

The most important analogy is with the horocycle flow

$$h_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}.$$

The Kleinbock–Weiss theorem is in turn based on work of Minsky and Weiss [23] on the horocycle flow. Let H denote this subgroup. It is a basic principle that g_t orbits can be quite wild. For example, the closure can be a Cantor set. On the other hand H orbits are constrained. In $\mathrm{SL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z})$ every horocycle orbit is either closed or dense. It is a basic question in the subject to find all H orbit closures of points (X,ω) and all measures invariant under the action of H. (See the article by Eskin in these proceedings for more on this problem, which one can call the Ratner problem in moduli space.)

Veech [29] showed that horocycle orbits do not diverge in the stratum. Minsky and Weiss gave a quantitative version of this result which shows that horocycle orbits spend most of their time in a compact set. To explain their result, introduce the terminology of $l(\gamma,(X,\omega))$ to represent the length of the saddle connection γ with respect to the metric of (X,ω), and K_ε the set of (X,ω) such that for every saddle connection γ, $l(\gamma,(X,\omega)) \geqslant \varepsilon$.

THEOREM 5. *There are positive constants C,α,ρ_0 depending only on the topology of the surface such that if (X,ω), an interval $I \subset \mathbb{R}$ and $0 \leqslant \rho \leqslant \rho_0$ satisfy the following condition*:

- *for any saddle connection γ there is $s \in I$ such that $l(\gamma, h_s(X,\omega)) \geqslant \rho$,*

then for any $\varepsilon > 0$,

$$\left|\left\{s \in I\colon\ h_s(X,\omega) \notin K_\varepsilon\right\}\right| \leqslant C\left(\frac{\varepsilon}{\rho}\right)^{\alpha} |I|.$$

Another recent result of Smillie and Weiss [25] classifies *minimal* sets for the horocycle flow. A set is minimal if it is invariant, closed, and there is no proper invariant closed subset. The authors first describe examples of minimal sets and then show that *every* minimal set is given by such an example. To describe the examples suppose in the horizontal direction all leaves of (X,ω) are closed so that (X,ω) decomposes into a union of cylinders each of which is swept out by closed horizontal leaves. Let $\mathcal{O} = \overline{H(X,\omega)}$. Then

- Every $(Y,\sigma) \in \mathcal{O}$ admits a cylinder decomposition $C_1 \cup \cdots \cup C_r$ where each C_i is swept out by closed horizontal leaves.

- There is an isomorphism between $\mathcal{O}$ and a d dimensional torus where d is the dimension of the $\mathbb{Q}$ linear subspace of $\mathbb{R}$ spanned by the moduli of $C_1, \ldots, C_r$. The isomorphism conjugates the H-action on $\mathcal{O}$ with a one parameter translational flow.
- The restriction of the H-action to $\mathcal{O}$ is minimal.

Acknowledgements

This survey article is based on lectures of the author at Luminy in June 2003. Thanks are given to Kristen Wickelgren, Corinna Ulcigrai, and Jayadev Athreya for taking detailed notes and to the latter for turning the notes into a manuscript.

References

Surveys in volume 1A and this volume

[1] A. Eskin, *Counting problems in moduli space*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 581–595.
[2] P. Hubert and T. Schmidt, *Affine diffeomorphisms and the Veech dichotomy*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 501–526
[3] D. Kleinbock, N. Shah and A. Starkov, *Dynamics of subgroup actions on homogeneous spaces of Lie groups and applications to number theory*, Handbook of Dynamical Systems, Vol. 1A, Elsevier, Amsterdam (2002), 813–932.
[4] H. Masur and S. Tabachnikov, *Rational billiards and flat structures*, Handbook of Dynamical Systems, Vol. 1A, Elsevier, Amsterdam (2002), 1015–1089.

Other sources

[5] L. Bers, *Quasiconformal mappings and Teichmuller's theorem*, Analytic Functions, Princeton Univ. Press (1960).
[6] L. Bers, *Finite dimensional Teichmuller spaces and generalizations*, Bull. Amer. Math. Soc. **5** (1981), 131–172.
[7] M. Boshernitzan, *A condition for minimal interval exchange maps to be uniquely ergodic*, Duke Math. J. **52** (3) (1985), 723–752.
[8] Y. Cheung, *Hausdorff dimension of the set of nonergodic directions*, Ann. of Math. **158** (2003), 661–678.
[9] Y. Cheung, *Slowly divergent geodesics in moduli space*, Conform. Geom. Dynam. **8** (2004), 167–189.
[10] Y. Cheung and H. Masur, *A divergent Teichmuller geodesic with uniquely ergodic vertical foliation*, Israel J. Math., to appear.
[11] Y. Cheung and H. Masur, *Minimal nonergodic directions on genus* 2 *translation surfaces*, Ergodic Theory Dynam. Systems, to appear.
[12] A. Eskin and H. Masur, *Asymptotic formulas on flat surfaces*, Ergodic Theory Dynam. Systems **21** (2001), 443–478.
[13] H. Furstenberg, *Strict ergodicity and transformations of the torus*, Amer. J. Math. **88** (1961), 573–601.
[14] A. Katok, *Combinatorial constructions in ergodic theory*, University Lecture Series, Vol. 30, Amer. Math. Soc. (2003).
[15] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge Univ. Press (1995).

[16] S. Kerckhoff, H. Masur and J. Smillie, *Ergodicity of billiard flows and quadratic differentials*, Ann. of Math. **124** (1986), 293–311.
[17] D. Kleinbock and B. Weiss, *Bounded geodesics in moduli space*, Internat. Math. Res. Notices **30** (2004), 1551–1560.
[18] M. Kontsevich and A. Zorich, *Connected components of the space of holomorphic differentials with prescribed singularities*, Invent. Math. **153** (2003), 631–678.
[19] H. Masur, *Interval exchange transformations and measured foliations*, Ann. of Math. **115** (1982), 169–200.
[20] H. Masur, *The growth rate of trajectories of a quadratic differential*, Ergodic Theory Dynam. Systems **10** (1990), 151–176.
[21] H. Masur, *Hausdorff dimension of the set of nonergodic foliations of a quadratic differential*, Duke Math. J. **66** (1992), 387–442.
[22] H. Masur and J. Smillie, *Hausdorff dimension of sets of nonergodic foliations*, Ann. of Math. **134** (1991), 455–543.
[23] Y. Minsky and B. Weiss, *Nondivergence of horocyclic flows on moduli space*, J. Reine Angew Math. **552** (2002), 131–177.
[24] D. Mumford, *A remark on Mahler's compactness theorem*, Proc. Amer. Math. Soc. **28** (1971), 289–294.
[25] J. Smillie and B. Weiss, *Minimal sets for flows on moduli space*, Israel J. Math. **142** (2004), 249–260.
[26] K. Strebel, *Quadratic Differentials*, Springer-Verlag (1984).
[27] W. Veech, *Strict ergodicity in zero dimensional dynamical systems and the Kronecker–Weyl theorem* mod 2, Trans. Amer. Math. Soc. **140** (1969), 1–34.
[28] W. Veech, *Teichmuller geodesic flow*, Ann. of Math. **124** (1986), 441–530.
[29] W. Veech, *Measures supported on the set of uniquely ergodic directions of an arbitrary holomorphic* 1 *form*, Ergodic Theory Dynam. Systems **19** (1999), 1093–1109.

This page intentionally left blank

CHAPTER 8

On the Lyapunov Exponents of the Kontsevich–Zorich Cocycle

Giovanni Forni

Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston, IL 60208-2730, USA
E-mail: *gforni@math.northwestern.edu*

Laboratoire de Topologie et Dynamique, Université de Paris-Sud, Bât. 425, 91405 Orsay cedex, France
E-mail: *giovanni.forni@math.u-psud.fr*

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

This page intentionally left blank

1. Introduction

The Kontsevich–Zorich cocycle, introduced in [25], is a cocycle over the Teichmüller flow on the moduli space of holomorphic (quadratic) differentials. The study of the dynamics of this cocycle, in particular of its Lyapunov structure, has important applications to the ergodic theory of interval exchange transformations (i.e.t.'s) and related systems such as measured foliations, flows on *translation surfaces* and rational polygonal billiards (see the article by H. Masur [5] in this handbook). The Kontsevich–Zorich cocycle is a continuous-time version of a cocycle introduced by G. Rauzy [35] as a "continued fractions algorithm" for i.e.t.'s and later studied by W. Veech, in his work on the unique ergodicity of the generic i.e.t. [38], and A. Zorich [45,46] among others.

1.1. *Deviation of ergodic averages and other applications*

Zorich (see [44,46,47]) made the key discovery that typical trajectories of generic (orientable) measured foliations on surfaces of higher genus (or equivalently of generic i.e.t.'s with at least 4 intervals) deviate from the mean according to a power law with exponents determined by the Lyapunov exponents of the cocycle.

In [45] he began a systematic study of the Lyapunov spectrum of the cocycle and conjectured, on the basis of careful numerical experiments, that all of its Lyapunov exponents are non-zero and simple. He also observed that, as a consequence of the close connection between the cocycle and the Teichmüller geodesic flow, the simplicity of the top exponent, sometimes called the *spectral gap* property, is equivalent to the (non-uniform) hyperbolicity of the Teichmüller flow, which had been proved earlier by W. Veech [40].

The applications of the Kontsevich–Zorich cocycle to the dynamics of i.e.t.'s and related systems are not limited to the deviation of ergodic averages. The spectral gap property of the cocycle also plays an important role in recent results of Marmi, Moussa and Yoccoz [27,28] on the *cohomological equation* for generic i.e.t.'s, which improve on previous work of the author [19].

In a different direction, A. Avila and the author [7] have recently shown that the positivity of the second exponent (for surfaces of higher genus) implies that almost every i.e.t. which is not a rotation is weakly mixing and that the generic directional flow on the generic translation surface of higher genus is weakly mixing as well. This result answers in the affirmative a longstanding conjecture on the dynamics of i.e.t.'s. Special cases of the conjecture were earlier settled by A. Katok and A. Stepin [24] (for i.e.t.'s on 3 intervals) and W. Veech [39] (for i.e.t.'s on any number of intervals, but with special combinatorics).

1.2. *Renormalization for parabolic systems*

The role of the Kontsevich–Zorich cocycle can be explained by the somewhat vague observation that it provides a *renormalization dynamics* for i.e.t.'s (and related systems). Such systems provide fundamental examples of *parabolic* dynamics, which by definition is characterized by sub-exponential (polynomial) divergence of nearby orbits.

All systems with behavior intermediate between *elliptic*, characterized by no or "very slow" divergence of nearby orbits, and *hyperbolic*, characterized by exponential divergence of nearby orbits, can be roughly classified as parabolic. A classical example of parabolic dynamics is the horocycle flow (on a surface of constant negative curvature). For i.e.t.'s (and related systems) there is no infinitesimal divergence of orbits, but parabolic orbit divergence is produced over time by the presence of singularities.

Key generic features of parabolic dynamics include unique ergodicity, polynomial deviation of ergodic averages from the mean and presence of invariant distributional obstructions, which are not measures, to the existence of smooth solutions of the cohomological equation. The elliptic, parabolic and hyperbolic paradigms are described in depth in the survey by B. Hasselblatt and A. Katok [3] in this handbook.

Parabolic (and elliptic) systems are often studied by means of appropriate renormalization schemes which enable to understand the dynamics of the generic system in a given family through the study of an auxiliary hyperbolic system. The hyperbolic system (renormalization) can in turn be studied by means of the well-developed tools of hyperbolic theory (Lyapunov exponents, invariant manifolds, Pesin theory, Lifschitz theory).

The Teichmüller flow and the Kontsevich–Zorich cocycle (and related systems such as the Rauzy–Zorich induction [35,45] or Veech "zippered rectangles" flow [38] and the corresponding cocycles) provide an effective renormalization scheme for i.e.t.'s and related systems.

Other well-known examples of renormalization include the classical Gauss map, which renormalizes rotations of the circle, and the geodesic flow (on a surface of constant negative curvature), which renormalizes the corresponding horocycle flow.

A tentative systematic approach to renormalization for a class of parabolic flows of algebraic nature, called "pseudo-homogeneous" flows, which includes conservative flows on surfaces, classical horocycle flows and nilflows in dimension 3, has been proposed by the author in [20].

1.3. *Contents*

In this article we outline the author's proof [21] of a substantial part of the *Zorich conjecture* on the Lyapunov spectrum of the Kontsevich–Zorich cocycle.

ZORICH CONJECTURE. *The Lyapunov exponents for the canonical absolutely continuous invariant measure on any connected component of any stratum of the moduli space are all non-zero and distinct.*

In [21] we have proved that the exponents are all non-zero, hence the cocycle is by definition *non-uniformly hyperbolic*. The full Zorich conjecture, which affirms that the Lyapunov spectrum is *simple*, that is, all Lyapunov exponents are distinct, was left open in [21] in genus higher than 3. A proof based on ideas different from ours has been recently announced by A. Avila and M. Viana [8]. In this outline, we have chosen to restrict ourselves to the proof of the positivity of the second exponent (Corollary 6.3) which is easier to explain and already contains all the main ideas of our method. As we have mentioned

above, this is the key property in applications to i.e.t.'s (deviations of ergodic averages, weak-mixing). In passing we give a new, rather elementary, *complete* proof of the spectral gap property (Theorem 2.2).

We then present a rather unexpected example of an $SL(2,\mathbb{R})$-invariant measure supported on a closed Teichmüller disk in genus 3 for which the second and the third Lyapunov exponents are zero (Corollary 7.4). This example shows that (in genus greater than 3) the Zorich conjecture does not hold for all $SL(2,\mathbb{R})$-invariant measures on the moduli space. The significance of this conclusion is best understood in the perspective of the ergodic theory of *rational polygonal billiards*. In fact, for the generic directional flow on a fixed rational polygonal billiard the questions on deviation of ergodic averages and weak mixing are wide open, except for special cases, as a consequence of the fact that holomorphic differentials arising from rational billiards form a zero Lebesgue measure subset of the moduli space (see the survey by H. Masur and S. Tabachnikov [6] in this handbook on the dynamics of rational polygonal billiards).

Finally, we present the bulk of our proof of a representation theorem for *Zorich cycles* (Theorem 8.2). The phase space of the Kontsevich–Zorich cocycle is a (orbifold) vector bundle over the moduli space of holomorphic (quadratic) differentials on Riemann surfaces with fiber at each holomorphic differential given by the real homology (or cohomology) of the underlying Riemann surface. This bundle is sometimes called the *real homology* (*or cohomology*) *bundle*. Zorich cycles (or cocycles) are the homology (or cohomology) classes forming the invariant stable/unstable space of the Kontsevich–Zorich cocycle. For a generic (holomorphic) quadratic differential, leaves of the horizontal/vertical measured foliation "wind around a surface" deviating from a straight line (spanned by the Schwartzman's asymptotic cycle) in the direction of Zorich cycles in the real homology of the surface (see [44,47] or [48, Appendix D]).

We prove that Zorich cycles can be represented in terms of special closed currents on the surface (in the sense of de Rham) related to the horizontal/vertical measured foliation, called *basic currents*. Basic currents for measured foliations are in turn closely related to *invariant distributions* appearing as obstructions to the existence of smooth solutions of the cohomological equation for directional flows on translation surfaces or for i.e.t.'s [19, 27,28].

1.4. *Organization*

In Section 2 we review some background on the dynamics of the Teichmüller flow on the moduli space of holomorphic (quadratic) differentials.

In Section 3 we give our definition of the Kontsevich–Zorich cocycle and state the main theorem on its Lyapunov spectrum (Theorem 3.1).

In Section 4 we derive the variational formulas which describe the evolution of cohomology classes and their norms under the action of the cocycle (Lemmas 4.2 and 4.3).

In Section 5 bounds (upper and lower) on the second Lyapunov exponent are derived from the variational formulas of Section 4. The upper bound is easily obtained and allows us to immediately prove the spectral gap property (Theorem 2.2). The proof of the lower bound is harder since there are subtle cancellations.

Following [25] we take a harmonic analysis point of view (boundary behavior of harmonic functions, Brownian motion) on the generic Teichmüller disk which happens to be an isometric copy of the Poincaré disk. In concrete terms, we compute formulas for the hyperbolic gradient and Laplacian of the norm of a (fixed) cohomology class along a Teichmüller disk (Lemma 5.2). These formulas allow us to prove a lower bound for the second exponent in terms of the lowest eigenvalue of a Hermitian form which represents a 'Hodge curvature' of the real cohomology bundle. However, we have yet to prove that such a bound is non-trivial, that is, strictly positive. In fact, the Hodge curvature is degenerate on a real analytic subvariety of codimension 2 of the moduli space of holomorphic differentials.

In Section 6 we describe such a subvariety that we have called the *determinant locus* since it coincides with the locus where the determinant of the Lie derivative of the classical period matrix along the Teichmüller flow vanishes (Lemma 6.1). The proof that the second exponent is positive on all connected components of all strata of the moduli space is reduced to the statement that no connected component of a stratum is contained in the determinant locus (Theorem 6.2). The proof of this theorem, based on asymptotic formulas for the period matrix and its Lie derivative near appropriate boundary points of the moduli space, is only sketched here. The complete argument can be found in [21, Section 4].

In Section 7 we answer in the affirmative a question asked by W. Veech on whether there exist Teichmüller disks entirely contained in the determinant locus. Our example consists of a closed Teichmüller disk in genus 3 (in the stratum of holomorphic differentials with 4 simple zeroes) generated by a non-primitive Veech surface obtained as a 2-sheeted branched cover over the square torus with 4 branching points of order 2. Such a Veech surface has appropriate symmetries, stable under the $SL(2,\mathbb{R})$-action, which imply that the Hodge curvature has the minimal rank 1 (Theorem 7.3). It follows that of the 3 non-negative exponents of the Kontsevich–Zorich cocycle only one (the trivial one) is non-zero on the corresponding closed $SL(2,\mathbb{R})$-orbit (Corollary 7.4).

Finally, in Section 8 we prove the representation theorem for Zorich cycles. The proof is based on the variational formulas of Section 4, on a Cheeger-type lower bound for the smallest eigenvalue of the flat Laplacian on a translation surface, equivalent to a Poincaré inequality for the appropriate Sobolev norms (Lemma 8.3), and on the logarithmic law for geodesic in the moduli space of holomorphic (quadratic) differentials [31].

2. Elements of Teichmüller theory

In this section we recall a few definitions and results of Teichmüller theory which are essential to understanding the material treated in later sections.

Let T_g, Q_g be the *Teichmüller spaces* of complex (conformal) structures and of holomorphic quadratic differentials on a surface of genus $g \geqslant 1$. The spaces T_g and Q_g can be roughly described as follows:

$$\begin{aligned} T_g &:= \{\text{complex (conformal) structures}\}/\operatorname{Diff}_0^+(M), \\ Q_g &:= \{\text{holomorphic quadratic differentials}\}/\operatorname{Diff}_0^+(M), \end{aligned} \tag{1}$$

where $\mathrm{Diff}_0^+(M)$ is the group of orientation preserving diffeomorphisms of the surface M which are isotopic to the identity (equivalently, it is the connected component of the identity in the Lie group of all orientation preserving diffeomorphisms of M).

If $g \geqslant 2$, the space T_g is topologically equivalent to an open ball of real dimension $6g-6$. In fact, a theorem of L. Ahlfors, L. Bers and S. Wolpert states that T_g has a complex structure holomorphically equivalent to that of a Stein (strongly pseudo-convex) domain in $\mathbb{C}^{3g-3}$ [9, §6], or [32, Chapters 3, 4 and Appendix §6]. The space Q_g of holomorphic quadratic differentials is a complex vector bundle over T_g which can be identified to the cotangent bundle of T_g. If $g=1$, the Teichmüller space T_1 of elliptic curves (complex structures on T^2) is isomorphic to the upper half plane $\mathbb{C}^+$ and the Teichmüller space Q_1 of holomorphic quadratic differentials on elliptic curves is a complex line bundle over T_1 [32, Example 2.1.8].

Let $R_g, \mathcal{M}_g$ be the *moduli spaces* of complex (conformal) structures and of holomorphic quadratic differentials on a surface of genus $g \geqslant 1$. The spaces R_g and $\mathcal{M}_g$ can be roughly described as the quotient spaces:

$$R_g := T_g/\Gamma_g, \qquad \mathcal{M}_g := Q_g/\Gamma_g, \tag{2}$$

where Γ_g denotes the *mapping class group* $\mathrm{Diff}^+(M)/\mathrm{Diff}_0^+(M)$. If $g=1$, the mapping class group can be identified with the lattice $SL(2,\mathbb{Z})$ which acts on the upper half plane $\mathbb{C}^+$ in the standard way. The moduli space $R_1 := \mathbb{C}^+/SL(2,\mathbb{Z})$ is a non-compact finite volume surface with constant negative curvature, called the *modular surface*. The moduli space $\mathcal{M}_1$ can be identified to the cotangent bundle of the modular surface.

The *Teichmüller* (*geodesic*) *flow* is a Hamiltonian flow on $\mathcal{M}_g$, defined as the geodesic flow with respect to a natural metric on R_g called the *Teichmüller metric*. Such a metric measures the amount of *quasi-conformal distortion* between two different (equivalent classes of) complex structures in R_g. In the higher genus case, the Teichmüller metric is not Riemannian, but only *Finsler* (that is, the norm on each tangent space does not come from an Euclidean product) and, as H. Masur proved, does not have negative curvature in any reasonable sense [9, §3 (E)]. If $g=1$, the Teichmüller metric coincides with the Poincaré metric on the modular surface R_1 [32, 2.6.5], in particular it is Riemannian with constant negative curvature.

In order to obtain a more geometric description of the Teichmüller flow, we introduce below a natural action of the Lie group $SL(2,\mathbb{R})$ on Q_g (see also [4, §1.4] or [5, §3], in this handbook). This action is equivariant with respect to the action of the mapping class group, hence it passes to the quotient $\mathcal{M}_g$.

A holomorphic quadratic differential q naturally defines two transverse *measured foliations* (in the Thurston's sense [37,17]), the *horizontal* foliation $\mathcal{F}_q$ and the *vertical* foliation $\mathcal{F}_{-q}$:

$$\begin{aligned} &\mathcal{F}_q := \left\{\mathrm{Im}\left(q^{1/2}\right)=0\right\}, \quad \text{with transverse measure } \left|\mathrm{Im}\left(q^{1/2}\right)\right|, \\ &\mathcal{F}_{-q} := \left\{\mathrm{Re}\left(q^{1/2}\right)=0\right\}, \quad \text{with transverse measure } \left|\mathrm{Re}\left(q^{1/2}\right)\right|. \end{aligned} \tag{3}$$

Vice versa, any pair $(\mathcal{F}, \mathcal{F}^\perp)$ of transverse measure foliations determines a complex structure and a holomorphic quadratic differential q such that $\mathcal{F}=\mathcal{F}_q$ and $\mathcal{F}^\perp = \mathcal{F}_{-q}$.

Transversality for measured foliations is taken in the sense that $\mathcal{F}$ and $\mathcal{F}^{\perp}$ have a common set Σ of (saddle) singularities, have the same index at each singularity and are transverse in the standard sense on $M \setminus \Sigma$. The set Σ of common singularities coincides with the set Σ_q of zeroes of the holomorphic quadratic differential $q \equiv (\mathcal{F}, \mathcal{F}^{\perp})$.

The $SL(2,\mathbb{R})$-action on Q_g is defined as follows. Every 2×2 matrix $A \in SL(2,\mathbb{R})$ acts naturally by left multiplication on the (locally defined) pair of real-valued 1-forms $(\mathrm{Im}(q^{1/2}), \mathrm{Re}(q^{1/2}))$. The resulting (locally defined) pair of 1-forms defines a new pair of transverse measured foliations, hence a new complex structure and a new holomorphic quadratic differential $A \cdot q$.

The Teichmüller flow G_t is given by the action of the diagonal subgroup $\mathrm{diag}(e^{-t}, e^{t})$ on Q_g (on $\mathcal{M}_g$). In other terms, if we identify holomorphic quadratic differentials with pairs of transverse measured foliations as explained above, we have:

$$G_t(\mathcal{F}_q, \mathcal{F}_{-q}) := \left(e^{-t}\mathcal{F}_q, e^{t}\mathcal{F}_{-q}\right). \tag{4}$$

In geometric terms, the action of the Teichmüller flow on quadratic differentials induces a one-parameter family of deformations of the conformal structure which consist in contracting along vertical leaves (with respect to the horizontal length) and expanding along horizontal leaves (with respect to the vertical length) by reciprocal (exponential) factors.

The reader can compare the definition in terms of the $SL(2,\mathbb{R})$-action with the analogous description of the geodesic flow on a surface of constant negative curvature (such as the modular surface). In fact, if $g = 1$ the above definition reduces to the standard Lie group presentation of the geodesic flow on the modular surface: the unit sub-bundle $\mathcal{M}_1^{(1)} \subset \mathcal{M}_1$ of all holomorphic quadratic differentials of unit total area on elliptic curves can be identified with the homogeneous space $SL(2,\mathbb{R})/SL(2,\mathbb{Z})$ and the geodesic flow on the modular surface is then identified with the action of the diagonal subgroup of $SL(2,\mathbb{R})$ on $SL(2,\mathbb{R})/SL(2,\mathbb{Z})$.

We list below, following [41,25], the main structures carried by the Teichmüller space Q_g and by the moduli space $\mathcal{M}_g$ of quadratic differentials (see also [5, §2] and [48, §4]):

(1) $\mathcal{M}_g$ is a (stratified) analytic space (orbifold); each stratum $\mathcal{M}_\kappa$ (corresponding to fixing the multiplicities $\kappa := (k_1, \ldots, k_\sigma)$ of the zeroes $\{p_1, \ldots, p_\sigma\}$ of the quadratic differentials) is $SL(2,\mathbb{R})$-invariant and, in particular, G_t-invariant.

(2) The total area function $A : \mathcal{M}_g \to \mathbb{R}^+$,

$$A(q) := \int_M |q|,$$

is $SL(2,\mathbb{R})$-invariant; hence the *unit bundle* $\mathcal{M}_g^{(1)} := A^{-1}(\{1\})$ and its strata $\mathcal{M}_\kappa^{(1)} := \mathcal{M}_\kappa \cap \mathcal{M}_g^{(1)}$ are $SL(2,\mathbb{R})$-invariant and, in particular, G_t-invariant.

Let $\mathcal{M}_\kappa$ be a stratum of *orientable* quadratic differentials, that is, quadratic differentials which are squares of holomorphic 1-forms. In this case, the natural numbers $(k_1, \ldots, k_\sigma)$ are all even.

(3) The stratum of squares $\mathcal{M}_\kappa$ has a locally affine structure modeled on the affine space $H^1(M, \Sigma_\kappa; \mathbb{C})$, with $\Sigma_\kappa := \{p_1, \ldots, p_\sigma\}$. Local charts are given by the period map $q \to [q^{1/2}] \in H^1(M, \Sigma_\kappa; \mathbb{C})$.

(4) The Lebesgue measure on the Euclidean space $H^1(M, \Sigma_\kappa; \mathbb{C})$, normalized so that the quotient torus

$$H^1(M, \Sigma_\kappa; \mathbb{C})/H^1(M, \Sigma_\kappa; \mathbb{Z} \oplus \imath\mathbb{Z})$$

has volume 1, induces an absolutely continuous $SL(2, \mathbb{R})$-invariant measure μ_κ on $\mathcal{M}_\kappa$. The conditional measure $\mu_\kappa^{(1)}$ induced on $\mathcal{M}_\kappa^{(1)}$ is $SL(2, \mathbb{R})$-invariant, hence G_t-invariant.

The ergodic theory of the Teichmüller flow begins with the natural questions whether the measure $\mu_\kappa^{(1)}$ has finite total mass and whether it is ergodic for the Teichmüller flow on $\mathcal{M}_\kappa^{(1)}$. However, it was discovered by W. Veech [41] that $\mathcal{M}_\kappa^{(1)}$ has in general several connected components. M. Kontsevich and A. Zorich [26] have been able to obtain a complete classification of the connected components of the strata. Taking this phenomenon into account, the following result holds:

THEOREM 2.1 [30,40]. *The total volume of the measure $\mu_\kappa^{(1)}$ on $\mathcal{M}_\kappa^{(1)}$ is finite and the Teichmüller geodesic flow G_t is ergodic on each connected component of $\mathcal{M}_\kappa^{(1)}$.*

Since the measure $\mu_\kappa^{(1)}$ has finite total mass, the Poincaré recurrence theorem applies. This is the core of Masur's proof [30] of the unique ergodicity for almost all i.e.t.'s and measured foliations, a statement known as the *Keane conjecture* (see the article by H. Masur [5] in this handbook on the ergodic theory of measured foliations, i.e.t.'s and translation surfaces).

Poincaré recurrence for a suitable "renormalization" flow (on the space of "zippered rectangles") is also the key idea of Veech's proof of the Keane conjecture [38]. In [40] Veech further investigated the ergodic theory of the Teichmüller flow and proved that the Teichmüller flow is *non-uniformly hyperbolic*, in the sense that all of its *Lyapunov exponents*, except one corresponding to the flow direction, are non-zero.

We recall that a Lyapunov exponent is the asymptotic exponential rate of expansion of a (tangent) vector along the orbit of a point in the phase space of a dynamical system. The Oseledec's *Multiplicative Ergodic Theorem* [34,23] establishes their existence as appropriately defined limits, for almost all points with respect to any ergodic invariant probability measure. The theory of Lyapunov exponents for *cocycles* over (smooth) dynamical systems is explained in [23, §S.1], and in the survey [1] in this handbook.

The Lyapunov spectrum (that is, the collection of Lyapunov exponents) of the Teichmüller flow with respect to any ergodic invariant probability measure μ on $\mathcal{M}_\kappa^{(1)}$is known to have symmetries. In fact, it can be written as follows [45, §5], [25, §7], [47, §2.3]:

$$\begin{aligned} 2 &\geqslant \big(1+\lambda_2^\mu\big) \geqslant \cdots \geqslant \big(1+\lambda_g^\mu\big) \geqslant \overbrace{1=\cdots=1}^{\sigma_\kappa - 1} \geqslant \big(1-\lambda_g^\mu\big) \\ &\geqslant \cdots \geqslant \big(1-\lambda_2^\mu\big) \geqslant 0 \geqslant -\big(1-\lambda_2^\mu\big) \geqslant \cdots \geqslant -\big(1-\lambda_g^\mu\big) \\ &\geqslant \underbrace{-1=\cdots=-1}_{\sigma_\kappa - 1} \geqslant -\big(1+\lambda_g^\mu\big) \geqslant \cdots \geqslant -\big(1+\lambda_2^\mu\big) \geqslant -2. \end{aligned} \tag{5}$$

By the ergodicity statement of Theorem 2.1, the non-uniform hyperbolicity of the Teichmüller flow, proved by W. Veech in [40], can be formulated as follows:

THEOREM 2.2 [40]. *Let μ denote the normalized absolutely continuous $SL(2,\mathbb{R})$-invariant ergodic measure on any connected component $\mathcal{C}_\kappa^{(1)}$ of a stratum $\mathcal{M}_\kappa^{(1)} \subset \mathcal{M}_g^{(1)}$ of the moduli space of orientable holomorphic quadratic differentials of unit total area. The non-negative number λ_2^μ satisfies the inequality*:

$$\lambda_2^\mu < \lambda_1^\mu = 1. \tag{6}$$

M. Kontsevich and A. Zorich have interpreted the non-negative numbers

$$\lambda_1^\mu = 1 \geqslant \lambda_2^\mu \geqslant \cdots \geqslant \lambda_g^\mu \tag{7}$$

as *Lyapunov exponents* of a cocycle over the Teichmüller flow that will be described below. This cocycle is obtained as the natural (fiber-wise linear) lift of the Teichmüller flow to an appropriate vector bundle over the moduli space. The non-negative Lyapunov exponents of the Kontsevich–Zorich cocycle turn out to be exactly the numbers in (7).

In this paper we discuss the Lyapunov spectrum and the Oseledec's splitting of this cocycle. In particular, we give a new elementary proof of the inequality (6) for *any* ergodic probability measure on a stratum of orientable holomorphic quadratic differentials (Theorem 5.1) and we outline the proof of the inequality $\lambda_2^\mu > 0$, when μ is the normalized absolutely continuous $SL(2,\mathbb{R})$-invariant ergodic measure on any connected component of a stratum of the moduli space of orientable holomorphic quadratic differentials (Corollary 6.3).

3. The Kontsevich–Zorich cocycle

M. Kontsevich (and A. Zorich) [25] have introduced a (multiplicative) 'renormalization' cocycle over the Teichmüller geodesic flow. This cocycle is a continuous-time version of a cocycle introduced by G. Rauzy [35] as a "continued fractions algorithm" for i.e.t.'s, and later studied by W. Veech, in his work on the Keane conjecture [38], and A. Zorich [45,46] among others. Zorich was motivated by the study of the asymptotic behavior in homology of (long) typical leaves of orientable measured foliations on closed surfaces of higher genus, which he initiated in [44].

Let Q_g be the Teichmüller space of holomorphic quadratic differentials on Riemann surfaces of genus $g \geqslant 2$. The *Kontsevich–Zorich cocycle* G_t^{KZ} can be defined as the quotient cocycle, with respect to the action of the mapping class group Γ_g, of the trivial cocycle

$$G_t \times \mathrm{id} : Q_g \times H^1(M,\mathbb{R}) \to Q_g \times H^1(M,\mathbb{R}). \tag{8}$$

The cocycle G_t^{KZ} acts on the orbifold vector bundle

$$\mathcal{H}_g^1(M,\mathbb{R}) := \big(Q_g \times H^1(M,\mathbb{R})\big)/\Gamma_g \tag{9}$$

over the moduli space $\mathcal{M}_g = Q_g/\Gamma_g$ of holomorphic quadratic differentials. The base dynamics of the Kontsevich–Zorich cocycle is the Teichmüller geodesic flow G_t on $\mathcal{M}_g$. Note that the mapping class group acts naturally on the cohomology $H^1(M, \mathbb{R})$ by pull-back. We recall that the real homology $H_1(M, \mathbb{R})$ and the real cohomology $H^1(M, \mathbb{R})$ of an orientable closed surface M are endowed with a natural symplectic form (the intersection form) and are (symplectically) isomorphic by Poincaré duality.

Since the vector bundle $\mathcal{H}^1_g(M, \mathbb{R})$ has a symplectic structure, the Lyapunov spectrum of the cocycle G^{KZ}_t (with respect to any G_t-invariant ergodic probability measure μ on $\mathcal{M}^{(1)}_g$) is *symmetric*:

$$\lambda^\mu_1 \geqslant \cdots \geqslant \lambda^\mu_g \geqslant 0 \geqslant -\lambda^\mu_g \geqslant \cdots \geqslant -\lambda^\mu_1. \tag{10}$$

The non-negative part of the Kontsevich–Zorich spectrum (10) coincides with the numbers (7) which appear in the Lyapunov spectrum (5) of the Teichmüller flow. This relation can be explained as follows. By Section 2 the tangent space $T\mathcal{M}_\kappa \equiv \mathcal{M}_\kappa \times H^1(M, \Sigma_\kappa; \mathbb{C})$ locally. There is a surjective map $H^1(M, \Sigma_\kappa; \mathbb{C}) \to H^1(M, \mathbb{C})$ which neglects cohomology classes dual to cycle joining two singularities. Such classes are responsible for the (trivial) part of the Lyapunov spectrum (5) consisting of $\sigma_\kappa - 1$ repeated 1's and -1's. Let then $\mathcal{H}^1_\kappa(M, \mathbb{C})$ be the bundle over the moduli space with fiber $H^1(M, \mathbb{C})$. There is the following natural isomorphism of vector bundles over $\mathcal{M}_\kappa$:

$$\mathcal{H}^1_\kappa(M, \mathbb{C}) \equiv \mathbb{C} \otimes \mathcal{H}^1(M, \mathbb{R}) \equiv \mathbb{R}^2 \otimes \mathcal{H}^1(M, \mathbb{R}), \tag{11}$$

induced by the corresponding isomorphism on the fibers. The tangent cocycle TG_t of the Teichmüller geodesic flow on $\mathcal{H}^1(M, \mathbb{C})$ can then be written in terms of the Kontsevich–Zorich cocycle:

$$TG_t = \operatorname{diag}\left(e^t, e^{-t}\right) \otimes G^{KZ}_t \quad \text{on } \mathbb{R}^2 \otimes \mathcal{H}^1(M, \mathbb{R}). \tag{12}$$

Formula (12) implies that the non-trivial Lyapunov spectrum of TG_t on $\mathcal{H}^1(M, \mathbb{C})$ can be obtained as a union of the translations of the Lyapunov spectrum of G^{KZ}_t by ± 1, hence (5) follows (see also [48, §5.7]).

We will discuss the main ideas of the proof of the following result originally conjectured by A. Zorich in [45] for the Rauzy–Veech–Zorich cocycle, a discrete-time version of the Kontsevich–Zorich cocycle, and by M. Kontsevich (and A. Zorich) in [25] for the Kontsevich–Zorich cocycle (see also [48, §5.6]):

THEOREM 3.1 [21, Theorem 8.5]. *Let μ denote the absolutely continuous $SL(2, \mathbb{R})$-invariant ergodic probability measure on any connected component $\mathcal{C}^{(1)}_\kappa$ of a stratum $\mathcal{M}^{(1)}_\kappa \subset \mathcal{M}^{(1)}_g$ of the moduli space of orientable holomorphic quadratic differentials of unit total area. The Lyapunov exponents of G^{KZ}_t over $\mathcal{C}^{(1)}_\kappa$ satisfy the inequalities*:

$$\lambda^\mu_1 = 1 > \lambda^\mu_2 \geqslant \cdots \geqslant \lambda^\mu_g > 0. \tag{13}$$

The inequality $\lambda_1^\mu = 1 > \lambda_2^\mu$ is the content of Veech's Theorem 2.2. We will give a complete new proof below. The other non-trivial inequality in (13) is $\lambda_g^\mu > 0$. We will describe the strategy of the proof that $\lambda_2^\mu > 0$. The full proof of the theorem for genus $g \geqslant 3$ is more complicated but it does not require substantial new ideas. The *Zorich conjecture* states that the exponents in (13) are all distinct, that is, the Lyapunov spectrum of the cocycle is *simple*. A proof of the conjecture, which yields as a corollary an independent proof of Theorem 3.1 based on completely different methods, has been recently given by A. Avila and M. Viana [8].

4. Variational formulas

The Kontsevich–Zorich cocycle can be written in the form of an O.D.E. in a fixed Hilbert space. This is accomplished as follows. Let R_q be (degenerate) Riemannian metric induced by a holomorphic quadratic differential q and let ω_q be the corresponding area form. With respect to a holomorphic local coordinate $z = x + iy$, the quadratic differential q has the form $q = \phi(z)\,dz^2$, where ϕ is a locally defined holomorphic function, and, consequently,

$$R_q = \big|\phi(z)\big|^{1/2}\big(dx^2 + dy^2\big)^{1/2}, \qquad \omega_q = \big|\phi(z)\big|\,dx \wedge dy. \tag{14}$$

The metric R_q is flat, it is degenerate at the finite set Σ_q of zeroes of q and, if q is orientable, it has trivial holonomy, hence q induces a structure of *translation surface* on M. It follows that, if q is orientable, there exists a (unique) frame $\{S, T\}$ of the tangent bundle of M over $M \setminus \Sigma_q$ with the following properties [19, §2]:

(1) The frame $\{S, T\}$ is orthonormal with respect to the Riemannian metric R_q on $M \setminus \Sigma_q$;
(2) The vector field $S[T]$ is tangent to the oriented horizontal [vertical] foliation $\mathcal{F}_q$ $[\mathcal{F}_{-q}]$ in the positive direction.

Let $L_q^2(M) := L^2(M, \omega_q)$ the space of complex-valued, square-integrable functions and $H_q^1(M)$ be the (Sobolev) subspace of functions $v \in L_q^2(M)$ such that $Sv \in L_q^2(M)$ and $Tv \in L_q^2(M)$. The flows generated by the vector fields S, T preserves the area form ω_q. In fact, the 1-forms

$$\imath_S \omega_q = \mathrm{Im}\big(q^{1/2}\big) \quad \text{and} \quad \imath_T \omega_q = -\,\mathrm{Re}\big(q^{1/2}\big) \tag{15}$$

are closed and the Lie derivatives

$$\begin{aligned} &\mathcal{L}_S \omega_q = d\imath_S \omega_q + \imath_S d\omega_q = 0, \\ &\mathcal{L}_T \omega_q = d\imath_T \omega_q + \imath_T d\omega_q = 0. \end{aligned} \tag{16}$$

Hence, the vector fields S, T yield densely defined anti-symmetric (in fact, essentially skew-adjoint) operators on the Hilbert space $L_q^2(M)$. In addition, these operators commute

in the following sense. Let $(\cdot,\cdot)_q$ denote the inner product in $L^2_q(M)$. For all functions $v_1, v_2 \in H^1_q(M)$,

$$(Sv_1, Tv_2)_q = (Tv_1, Sv_2)_q. \tag{17}$$

In conclusion, there is a well-defined action of the commutative Lie algebra $\mathbb{R}^2$ on $L^2_q(M)$ by essentially skew-adjoint operators [19].

The above properties are not surprising since, with respect to a local canonical (holomorphic) coordinate $z = x + \imath y$ at a point $p \in M \setminus \Sigma_q$, the holomorphic quadratic differential $q = dz^2$, the metric R_q is Euclidean, the area form $\omega_q = dx \wedge dy$ and the vector fields $S = \partial/\partial x$, $T = \partial/\partial y$. The formulas for S, T in a neighbourhood of a zero $p \in \Sigma_q$ of even order $k \geqslant 2$ are given in [19, (2.7)].

A key idea in [19,21] is to consider the *Cauchy–Riemann operators* determined by an orientable quadratic differential.

LEMMA 4.1 [19, Proposition 3.2]. *Let q be an orientable quadratic differential on M. The Cauchy–Riemann operators*

$$\partial^{\pm}_q := \frac{S \pm iT}{2} \tag{18}$$

with (dense) domain $H^1_q(M) \subset L^2_q(M)$ are closed and have closed range of finite codimension equal to the genus of M. Let $\mathcal{M}^{\pm}_q \subset L^2_q(M)$ be the subspaces of meromorphic, respectively, anti-meromorphic, functions. The following orthogonal splittings hold:

$$L^2_q(M) = \operatorname{Ran}\bigl(\partial^+_q\bigr) \oplus \mathcal{M}^-_q = \operatorname{Ran}\bigl(\partial^-_q\bigr) \oplus \mathcal{M}^+_q. \tag{19}$$

The spaces $\mathcal{M}^{\pm}_q$ consist of all meromorphic, respectively anti-meromorphic, functions with poles at Σ_q of orders bounded above in terms of the multiplicities of the points $p \in \Sigma_q$ as zeroes of the quadratic differential q. The complex dimension of $\mathcal{M}^{\pm}_q$ can therefore be computed by the Riemann–Roch theorem and it is equal to the genus of M. By (17) the adjoint operators $(\partial^{\pm}_q)^*$ are extensions of the operators $-\partial^{\mp}_q$. It follows that the kernels of $(\partial^{\pm}_q)^*$ are the subspaces $\mathcal{M}^{\mp}_q$, respectively, hence the splitting (19) follows immediately by Hilbert space theory.

(Absolute) real cohomology classes on M can be represented in terms of meromorphic (or anti-meromorphic) functions in $L^2_q(M)$. In fact, by the theory of Riemann surfaces [16, III.2], any $c \in H^1(M, \mathbb{R})$ can be represented as the real part of a holomorphic differential h on M. Let q be an orientable holomorphic quadratic differential on M and let $q^{1/2}$ a holomorphic square root of q. The quotient $h/q^{1/2}$ is a meromorphic function on M with poles at the set Σ_q of zeroes of q. A computation shows that $m^+ = h/q^{1/2} \in L^2_q(M)$, hence $m^+ \in \mathcal{M}^+_q$. The following *representation* of cohomology classes therefore holds:

$$c \in H^1(M, \mathbb{R}) \leftrightarrow c = \operatorname{Re}\bigl[m^+ \cdot q^{1/2}\bigr], \quad m^+ \in \mathcal{M}^+_q. \tag{20}$$

The map $\mathcal{M}_q^+ \to H^1(M,\mathbb{R})$ given by the representation (20) is bijective and it is in fact *isometric* if $\mathcal{M}_q^+$ is endowed with the Euclidean structure induced by $L_q^2(M)$ and $H^1(M,\mathbb{R})$ with the *Hodge product* relative to the complex structure of the Riemann surface M_q carrying the holomorphic quadratic differential $q \in \mathcal{M}_g$.

Let $q \in \mathcal{Q}_\kappa^{(1)}$ and $c \in H^1(M,\mathbb{R})$. Let $q_t := G_t(q)$ be the orbit of q under the Teichmüller flow and $c_t := G_t^{KZ}(c)$ the orbit of c under the Kontsevich–Zorich cocycle. Let M_t the Riemann surface carrying $q_t \in \mathcal{Q}_\kappa^{(1)}$. By (20),

$$c_t = \operatorname{Re}\big[m_t^+ \cdot q_t^{1/2}\big] \in H^1(M_t,\mathbb{R}), \tag{21}$$

where $m_t^+ \in \mathcal{M}_t^+$, the space of meromorphic function on M_t which are in $L_q^2(M)$. At this point, we have to make the following crucial remark. By the very definition of the Teichmüller flow G_t, the area form ω_t of the metric R_t induced by the quadratic differential q_t is *constant*. Hence the Hilbert space $L_q^2(M)$ is invariant under the action of the Teichmüller flow on $\mathcal{Q}_\kappa^{(1)}$. Let $\mathcal{M}_t^\pm \subset L_q^2(M)$ be the subspaces of meromorphic, respectively, anti-meromorphic, functions on the Riemann surface M_t. Such spaces are, respectively, the kernels of the adjoints of the Cauchy–Riemann operators $\partial_t^\mp$, related to the holomorphic quadratic differential q_t. By Lemma 4.1, the dimension of $\mathcal{M}_t^\pm$ is constant equal to the genus $g \geqslant 1$ of M. It can be proved that $\{\mathcal{M}_t^\pm \mid t \in \mathbb{R}\}$ are smooth families of g-dimensional subspaces of the fixed Hilbert space $L_q^2(M)$.

Let $\pi_q^\pm : L_q^2(M) \to \mathcal{M}_q^\pm$ denote the orthogonal projection onto the finite-dimensional subspace of meromorphic, respectively anti-meromorphic, functions. It follows immediately from (19) that, for every $u \in L_q^2(M)$, there exist functions $v^\pm \in H_q^1(M)$ such that

$$u = \partial_q^+ v^+ + \pi_q^-(u) = \partial_q^- v^- + \pi_q^+(u). \tag{22}$$

Let $\pi_t^\pm : L_q^2(M) \to \mathcal{M}_t^\pm$ denote the orthogonal projections in the (fixed) Hilbert space $L_q^2(M)$. By definition, the projections $\pi_t^\pm$ coincide with the projections $\pi_q^\pm$ for $q = q_t$, for any $t \in \mathbb{R}$.

LEMMA 4.2 [21, Lemma 2.1]. *The Kontsevich–Zorich cocycle is described by the following variational formulas*:

$$\begin{cases} m_t^+ = \partial_t^+ v_t + \pi_t^-(m_t^+), \\ \frac{d}{dt} m_t^+ = \partial_t^- v_t - \overline{\pi_t^-(m_t^+)}. \end{cases} \tag{23}$$

PROOF. By the definition (4) of the Teichmüller flow G_t, the quadratic differential $q_t := G_t(q)$ and the related Cauchy–Riemann operators $\partial_t^\pm$ can be explicitly written in terms of q and of corresponding frame $\{S, T\}$. In fact, we have $\operatorname{Re}(q_t^{1/2}) \equiv e^t \operatorname{Re}(q^{1/2})$, $\operatorname{Im}(q_t^{1/2}) \equiv e^{-t} \operatorname{Re}(q^{1/2})$ and

$$S_t \equiv e^{-t} S, \qquad T_t \equiv e^t T, \qquad \partial_t^\pm \equiv \frac{e^{-t} S \pm i e^t T}{2}; \tag{24}$$

hence, by straightforward computations,

$$\frac{d}{dt}q_t^{1/2} \equiv \overline{q_t}^{1/2}, \qquad \frac{d}{dt}\partial_t^{\pm} \equiv -\partial_t^{\mp}. \tag{25}$$

Equation (23) in the statement of the lemma follows from the formulas (25) by a computation based on the following two remarks. First, since the function m_t^+ is meromorphic on the Riemann surface M_t, it satisfies the equation $\partial_t^+ m_t^+ \equiv 0$ in the weak sense in $L_q^2(M)$. It follows that, by taking a time derivative,

$$-\partial_t^- m_t^+ + \partial_t^+\left(\frac{d}{dt}m_t^+\right) \equiv 0. \tag{26}$$

Second, by the definition (8) of the cocycle G_t^{KZ}, the one-parameter family of cohomology classes $c_t := G_t^{KZ}(c)$ is locally constant, that is, $c_t \equiv c \in H^1(M,\mathbb{R})$. It follows that the time derivative of the 1-form $\mathrm{Re}(m_t^+ q_t^{1/2})$ is equal to zero in $H^1(M,\mathbb{R})$, hence it is an exact form. There exists therefore a function $U_t \in H^1(M)$ such that

$$\mathrm{Re}\left[\left(\frac{d}{dt}m_t^+ + \overline{m_t^+}\right)q_t^{1/2}\right] = dU_t. \tag{27}$$

A straightforward computation based on formulas (26), (27) and on the splittings (22) for $q = q_t$, applied to the functions $m_t^+ \in \mathcal{M}_t^+ \subset L_q^2(M)$ and $dm_t^+/dt \in L_q^2(M)$, concludes the argument. In fact, the splitting in the first line of (23) is simply the first splitting in (22) for $q = q_t$, applied to the function $m_t^+ \in \mathcal{M}_t^+ \subset L_q^2(M)$. It is therefore an identity which determines the function $v_t \in H_q^1(M)$ up to an additive constant. The second line is a formula for the derivative dm_t^+/dt written in terms of the second splitting in (22) for $q = q_t$. □

An immediate consequence of Lemma 4.2 is the following result on the variation of the Hodge norm of cohomology classes under the action of the Kontsevich–Zorich cocycle. Let $B_q : L_q^2(M) \times L_q^2(M) \to \mathbb{C}$ be the complex bilinear form given by

$$B_q(u,v) := \int_M uv\omega_q, \quad \text{for all } u, v \in L_q^2(M). \tag{28}$$

LEMMA 4.3 [21, Lemma 2.1′]. *The variation of the Hodge norm $\|c_t\|$, which coincides with the L_q^2-norm $|m_t^+|_0$ under the identification* (21)*, is given by the following formulas*:

$$\text{(a)} \quad \frac{d}{dt}\left|m_t^+\right|_0^2 = -2\,\mathrm{Re}\,B_q\left(m_t^+\right) = -2\,\mathrm{Re}\int_M \left(m_t^+\right)^2\omega_q,$$

$$\text{(b)} \quad \frac{d^2}{dt^2}\left|m_t^+\right|_0^2 = 4\left\{\left|\pi_t^-\left(m_t^+\right)\right|_0^2 - \mathrm{Re}\int_M \left(\partial_t^+ v_t\right)\left(\partial_t^- v_t\right)\omega_q\right\}. \tag{29}$$

PROOF. The formulas (29) can be immediately deduced from (23) by taking into account the G_t-invariance of the inner product in $L^2_q(M)$ and the orthogonality of the splittings (19), (22) for $q = q_t$. □

5. Bounds on the exponents

Lemma 4.2 immediately implies Veech's Theorem 2.2. In fact, we have

THEOREM 5.1 [21, Corollary 2.2]. *Let μ denote any ergodic G_t-invariant probability measure on the moduli space $\mathcal{M}_g^{(1)}$ of orientable holomorphic quadratic differentials of unit total area. The Lyapunov exponents of the Kontsevich–Zorich cocycle G_t^{KZ} with respect to the ergodic measure μ satisfy the following inequality*:

$$\lambda_1^\mu = 1 > \lambda_2^\mu. \tag{30}$$

PROOF. By formula (a) in (29),

$$\frac{d}{dt}\log|m_t^+|_0^2 = -2\frac{\operatorname{Re} B_q(m_t^+)}{|m_t^+|_0^2}. \tag{31}$$

Since by the Schwarz inequality,

$$\left|B_q\left(m_t^+\right)\right| = \left|\left(m_t^+, \overline{m_t^+}\right)_q\right| \leqslant |m_t^+|_0^2, \tag{32}$$

Equation (31) implies that the upper Lyapunov exponent

$$\lambda_1^\mu := \limsup_{T\to\pm\infty}\frac{1}{T}\log\left|m_T^+\right|_0 \leqslant 1. \tag{33}$$

Moreover, the 1-dimensional subspace of complex constant functions is invariant under the flow of Equation (23), since for $m_t^+ \in \mathbb{C}$, the function $v_t \equiv 0$ and the orthogonal projection $\pi_t^-(m_t^+) \equiv m_t^+ \in \mathbb{C}$. By the definition of the isomorphism (20), this corresponds to the fact that the plane $E_q \subset H^1(M,\mathbb{R})$ generated by the cohomology classes $\{\operatorname{Re}(q^{1/2}), \operatorname{Im}(q^{1/2})\}$ is invariant under the cocycle G_t^{KZ}. The Lyapunov exponents of G_t^{KZ} restricted to this plane are ± 1, as it can be seen directly from the definition or by the formula (31) in the case of purely real or purely imaginary constant functions. Hence $\lambda_1^\mu = 1$. The exponent λ_2^μ is the top Lyapunov exponent of G_t^{KZ} on the bundle with fiber $H^1(M_q,\mathbb{R})/E_q$. Under the isomorphism (20), the vector space $H^1(M_q,\mathbb{R})/E_q$ is represented by meromorphic functions with *zero average* (orthogonal to constant functions). It can be seen that the subspace of zero average meromorphic functions is invariant under the flow of Equation (23). Let

$$\Lambda^+(q) := \max\left\{\frac{|B_q(m^+)|}{|m^+|_0^2}\,\Big|\, m^+ \in \mathcal{M}_q^+ \setminus \{0\},\ \int_M m^+\omega_q = 0\right\}. \tag{34}$$

By averaging (31) over the interval $[0, T]$, taking the upper limit and applying the Birkhoff ergodic theorem with respect to the G_t-invariant measure μ to the r.h.s., we have that, if $m_0^+ \in \mathcal{M}_q^+$ has zero average, for μ-almost all $q \in \mathcal{M}_\kappa$,

$$\limsup_{T\to+\infty} \frac{1}{T} \log \left|m_T^+\right|_0 \leqslant \int_{\mathcal{M}_\kappa} \Lambda^+(q)\, d\mu(q). \tag{35}$$

Since, by the Schwarz inequality (32), $\Lambda^+(q) \leqslant 1$ for all $q \in \mathcal{M}_\kappa$, it is sufficient to prove that $\Lambda^+(q) < 1$ on a positive measure set. In fact, $\Lambda^+(q) = 1$ if and on only if there exists a *non-zero* meromorphic function with zero average $m^+ \in \mathcal{M}_q^+$ such that $|(m^+, \overline{m^+})_q| = |m^+|_0^2$. A well-known property of the Schwarz inequality then implies that there exists $\lambda \in \mathbb{C}$ such that $m^+ = \lambda \overline{m^+}$. However, it cannot be so, since in that case m^+ would be meromorphic and anti-meromorphic, hence constant, and by the zero average condition it would be zero. We have therefore proved that $\Lambda^+(q) < 1$ for all $q \in \mathcal{M}_\kappa$. The argument is completed. □

The proof of *lower bounds* on the Lyapunov exponents of the Kontsevich–Zorich cocycle relies on the formula (29), (b), for the second derivative. Unfortunately, the r.h.s of the formula contains two terms and, while the first is at least clearly non-negative, the sign of second appears to be oscillating in a way difficult to control. In order to overcome this difficulty, we follow an idea of [25] which consists in averaging over the orbits of the circle group $SO(2, \mathbb{R})$ in the stratum $\mathcal{M}_\kappa$.

Let $SL(2, \mathbb{R})\, q$ be an orbit of $SL(2, \mathbb{R})$ in $\mathcal{M}_\kappa$. For almost all $q \in \mathcal{M}_\kappa$, the quotient $SL(2, \mathbb{R})\, q / SO(2, \mathbb{R})$ is a copy of the Poincaré disk, in the sense that it is an immersed two-dimensional disk on which the Teichmüller metric reduces to the standard Poincaré metric (with curvature -4). Such a disk is called a *Teichmüller disk* (see [32, 2.6.5]).

The *hyperbolic Laplacian* of the Hodge norm of a cohomology class on a Teichmüller disk can be computed as follows. We write formula (29), (b), for all quadratic differentials in a $SO(2, \mathbb{R})$-orbit, we then average with respect to the Haar measure on $SO(2, \mathbb{R})$. The averaging eliminates the 'bad' second term in the r.h.s. of formula (29), (b) (the oscillation is canceled!).

LEMMA 5.2 [21, Lemma 3.2]. *The following formulas hold for the hyperbolic gradient ∇_h and the hyperbolic Laplacian Δ_h of the norm of a cohomology class on a Teichmüller disk*:

$$\begin{aligned} &\text{(a)} \quad \nabla_h \left|m_z^+\right|_0^2 = -2\big(\operatorname{Re} B_q\big(m^+\big), \operatorname{Im} B_q\big(m^+\big)\big), \\ &\text{(b)} \quad \Delta_h \left|m_z^+\right|_0^2 = 8\left|\pi_q^-\big(m^+\big)\right|_0^2. \end{aligned} \tag{36}$$

Hence, by a straightforward calculation,

$$\Delta_h \log \left|m_z^+\right|_0 = 4\frac{|\pi_z^-(m_z^+)|_0^2}{|m_z^+|_0^2} - 2\frac{|B_q(m_z^+)|^2}{|m_z^+|_0^4} \geqslant 2\frac{|\pi_z^-(m_z^+)|_0^2}{|m_z^+|_0^2}. \tag{37}$$

An analysis of the solutions of the hyperbolic Poisson equation, combined with the Oseledec's theorem on the existence of Lyapunov exponents and Birkhoff ergodic theorem, leads to the following lower bound. Let

$$\Lambda^-(q) := \min\left\{ \frac{|\pi_q^-(m^+)|_0^2}{|m^+|_0^2} \,\middle|\, m^+ \in \mathcal{M}_q^+ \setminus \{0\} \right\}. \tag{38}$$

THEOREM 5.3 [21, Theorem 3.3]. *Let μ be any G_t-ergodic $SL(2,\mathbb{R})$-invariant probability measure on $\mathcal{M}_\kappa^{(1)}$. The second Lyapunov exponent λ_2^μ of the Kontsevich–Zorich cocycle with respect to the measure μ, satisfies the following lower bound*:

$$\lambda_2^\mu \geqslant \int_{\mathcal{M}_\kappa^{(1)}} \Lambda^-(q)\, d\mu(q). \tag{39}$$

Theorem 5.3 shows that to be able to prove that $\lambda_2^\mu > 0$ it is sufficient to prove that the non-negative continuous function $\Lambda^- : \mathcal{M}_g^{(1)} \to \mathbb{R}$ is strictly positive at some $q \in \mathrm{supp}(\mu) \subset \mathcal{M}_\kappa$. Hence we are led to consider the locus $\{\Lambda^- = 0\}$ in the moduli space $\mathcal{M}_g^{(1)}$.

6. The determinant locus

Let π_q^- be as above the orthogonal projection on the subspace $\mathcal{M}_q^- \subset L_q^2(M)$ of anti-meromorphic functions. Let H_q be the non-negative definite Hermitian form on the subspace $\mathcal{M}_q^+ \subset L_q^2(M)$ defined as follows. For all $(m_1^+, m_2^+) \in \mathcal{M}_q^+ \times \mathcal{M}_q^+$,

$$\mathrm{H}_q\big(m_1^+, m_2^+\big) := \big(\pi_q^-\big(m_1^+\big), \pi_q^-\big(m_2^+\big)\big)_q. \tag{40}$$

The non-negative number $\Lambda^-(q)$ is by definition the *smallest eigenvalue* of the Hermitian form H_q. The locus $\{\Lambda^- = 0\}$ coincides therefore with the set of quadratic differentials for which the Hermitian form H_q is degenerate, that is, represented by a $g \times g$ Hermitian matrix with zero determinant.

There is a close relation between the Hermitian form H_q and the derivative of the classical *period matrix* along the Teichmüller trajectory in the moduli space determined by the quadratic differential q on M.

Let us recall the definition of the period matrix. Let M be a marked Riemann surface of genus $g \geqslant 2$ and let $\{a_1, b_1, \ldots, a_g, b_g\} \subset H_1(M,\mathbb{Z})$ be a *canonical homology basis* (see [16, III.1]), characterized by the property that, for all $i,\ j \in \{1,\ldots,g\}$,

$$a_i \cap a_j = b_i \cap b_j = 0 \quad \text{and} \quad a_i \cap b_j = \delta_{ij}. \tag{41}$$

In other terms, a canonical homology basis is a symplectic basis with respect to the symplectic structure on the real homology $H_1(M,\mathbb{R})$ given by the (*algebraic*) *intersection*

form $\cap$. Let $\{\theta_1, \dots, \theta_g\}$ be the dual basis of the space of holomorphic (Abelian) differentials on M, characterized by the conditions $\theta_i(a_j) = \delta_{ij}$, for all $i, j \in \{1, \dots, g\}$. The $g \times g$ complex matrix Π given by

$$\Pi_{ij}(M) := \int_{b_j} \theta_i, \quad i, j \in \{1, \dots, g\}, \tag{42}$$

is the *period matrix* of the marked Riemann surface M. The period matrix yields a holomorphic mapping $\Pi : T_g \to \mathfrak{S}_g$ on the Teichmüller space of Riemann surfaces with values in the *Siegel space* $\mathfrak{S}_g$ of $g \times g$ complex symmetric matrices with positive definite imaginary part.

Let $q \in Q_g^{(1)}$ be a holomorphic quadratic differential on the Riemann surface M_q. Let $(M_t, q_t) := G_t(M_q, q)$, for $t \in \mathbb{R}$, be the Teichmüller orbit of (M_q, q) in the Teichmüller space $Q_g^{(1)}$. The equation

$$\det\left[\frac{d}{dt}\Pi(M_t)\Big|_{t=0}\right] = 0 \tag{43}$$

defines a real analytic hypersurface $D_g^{(1)} \subset Q_g^{(1)}$ of real codimension 2. In other words, the hypersurface $D_g^{(1)}$ is the locus where the derivative of the period matrix in the direction of the Teichmüller flow is degenerate.

It is immediate to see that Equation (43), hence the locus $D_g^{(1)}$, is invariant under change of marking on M, that is, invariant under the action of the mapping class group Γ_g. It follows that the projection $\mathcal{D}_g^{(1)} := D_g^{(1)}/\Gamma_g$ of $D_g^{(1)}$ into the moduli space $\mathcal{M}_g^{(1)}$ is well defined. The real analytic hypersurface $\mathcal{D}_g^{(1)} \subset \mathcal{M}_g^{(1)}$ of real codimension 2 was introduced in [21, §4], and called the *determinant locus*. The following lemma holds.

LEMMA 6.1 [21, Lemma 4.1]. *The locus* $\{\Lambda^- = 0\} \subset \mathcal{M}_g^{(1)}$ *coincides with the determinant locus* $\mathcal{D}_g^{(1)}$.

PROOF. Let $\{m_1^+, \dots, m_g^+\}$ be an orthonormal basis of $\mathcal{M}_q^+ \subset L_q^2(M)$. The (symmetric) matrix $B(q)$ of the projection operator $\pi_q^- : \mathcal{M}_q^+ \to \mathcal{M}_q^-$, with respect to the bases $\{m_1^+, \dots, m_g^+\} \subset \mathcal{M}_q^+$ and $\{\overline{m_1^+}, \dots, \overline{m_g^+}\} \subset \mathcal{M}_q^-$, and the Hermitian non-negative matrix $H(q)$ of the Hermitian form H_q, with respect to the basis $\{m_1^+, \dots, m_g^+\}$, are given by the following formulas:

$$\begin{aligned} B_{ij}(q) &= B_q\left(m_i^+, m_j^+\right) = \left(m_i^+, \overline{m_j^+}\right)_q, \\ H(q) &= B(q)^* B(q) = \overline{B(q)} B(q). \end{aligned} \tag{44}$$

The quotients $\phi_i^+ := \theta_i / q^{1/2}$ are meromorphic functions on M_q with poles at Σ_q, which belong to the space $L_q^2(M)$. The system $\{\phi_1^+, \dots, \phi_g^+\}$ is a basis of the space $\mathcal{M}_q^+$.

The infinitesimal deformation of the complex structure of the Riemann surface M_q induced by the Teichmüller flow in the direction of the quadratic differential $q \in Q_g^{(1)}$ can be represented by a canonical *Beltrami differential* $\mu_q := |q|/q$, hence by Rauch's formula [22, Proposition A.3]:

$$\frac{d}{dt}\Pi_{ij}(M_t)\bigg|_{t=0} = \int_M \theta_i \theta_j \mu_q = \int_M \phi_i^+ \phi_j^+ \omega_q = B_q\big(\phi_i^+, \phi_j^+\big). \tag{45}$$

Since $\{\phi_1^+, \dots, \phi_g^+\}$ is a basis of $\mathcal{M}_q^+$, there exists a non-singular $g \times g$ complex matrix $C(q)$ such that

$$\phi_i^+ = \sum_{j=1}^{g} C_{ij}(q) m_j^+ \quad \text{and} \quad C(q)C(q)^* = \operatorname{Im}(\Pi). \tag{46}$$

In fact, by [16, III.2.3],

$$\begin{aligned} \big(\phi_i^+, \phi_j^+\big)_q &= \frac{i}{2}\int_M \theta_i \wedge \overline{\theta_j} \\ &= \frac{i}{2}\sum_{k=1}^{g}\left\{\int_{a_k}\theta_i \int_{b_k}\overline{\theta_j} - \int_{b_k}\theta_i \int_{a_k}\overline{\theta_j}\right\} = \operatorname{Im}(\Pi_{ij}). \end{aligned} \tag{47}$$

By (45) and (46),

$$\begin{aligned} \left|\det\left(\frac{d}{dt}\Pi_{ij}(M_t)\bigg|_{t=0}\right)\right| &= \big|\det C(q)B(q)C(q)^t\big| = \big|\det C(q)\big|^2\,\big|\det B(q)\big| \\ &= \det \operatorname{Im}(\Pi)\big[\det H(q)\big]^{1/2}. \end{aligned} \tag{48}$$

Since $\operatorname{Im}(\Pi)$ is positive definite, the Hermitian form H_q is degenerate, hence $\Lambda^-(q) = 0$, if and only if $q \in \mathcal{D}_g^{(1)}$. □

The geometry of the determinant locus, in particular with respect to the foliation of the moduli space $\mathcal{M}_g^{(1)}$ by orbits of the $SL(2,\mathbb{R})$-action, plays an important role in the study of Lyapunov exponents of the Kontsevich–Zorich cocycle (and of the Teichmüller flow). We have proved the following non-trivial result:

THEOREM 6.2 [21, Theorem 4.5]. *Let $\mathcal{M}_\kappa^{(1)}$ be any stratum of the moduli space of orientable holomorphic quadratic differentials. No connected component of $\mathcal{M}_\kappa^{(1)}$ is contained in the determinant locus. In fact, the following stronger result holds. Let*

$$\Lambda_1(q) \equiv 1 \geqslant \Lambda_2(q) \geqslant \cdots \geqslant \Lambda_g(q) \geqslant 0 \tag{49}$$

be the eigenvalues of the Hermitian form H_q *in decreasing order. Let* $\mathcal{C}_\kappa^{(1)}$ *denote any connected component of* $\mathcal{M}_\kappa^{(1)}$. *We have*:

$$\sup_{q\in\mathcal{C}_\kappa^{(1)}} \Lambda_i(q) = 1, \quad \textit{for all } i \in \{1,\dots,g\}. \tag{50}$$

The proof of Theorem 6.2 shows that the supremum of the (continuous) functions Λ_i is achieved at a certain kind of *boundary points* of the moduli space which can be found in the closure of any connected component of any stratum. The argument is based on asymptotic expansions for the period matrix (and its derivatives) [18, Chap. III], [29,43], [21, §4].

The simplest and most intuitive choice of the appropriate boundary points is the disjoint sums of g tori with $2g-2$ paired punctures. At these points, the period matrix and its derivative along the Teichmüller flow are diagonal with all diagonal entries different from zero. It follows that the Hermitian form H_q is non-degenerate. In fact, it is immediate to see that $\Lambda_1 = \cdots = \Lambda_g = 1$. Riemann surfaces pinched along $g-1$ (separating) cycles homologous to zero converge to boundary points of that type.

Unfortunately, quadratic differentials on such pinched surfaces cannot in general belong to a stratum with a zero of high multiplicity as the pinching parameters converge to zero. In order to overcome this difficulty and treat all strata, we have considered a different type of boundary points. Such points are given by meromorphic quadratic differentials on Riemann spheres with $2g$ paired punctures, having poles of order 2 with strictly positive real residues at all punctures, equal at paired punctures (the residue of a quadratic differential at a pole $p \in M$ is the standard residue of the holomorphic 1-form $z\,\phi(z)\,dz$ with respect to a holomorphic coordinate $z : M \to \mathbb{C}$ such that $z(p) = 0$ and $q = \phi(z)\,dz^2$).

A basic step of the proof of Theorem 6.2 consists in constructing in every connected component of every stratum $\mathcal{M}_\kappa$ of the moduli space a family of quadratic differentials on Riemann surfaces pinched along a set of g distinct closed regular trajectories spanning a *Lagrangian subspace* in homology. The limit of any such family as the pinching parameters converge to zero is a meromorphic quadratic differential on a Riemann sphere of the type just described. The period matrix and its derivative converge to a diagonal matrix only in the *projective* sense, but this is enough for the proof.

As a corollary of Theorems 5.3 and 6.2, we obtain

COROLLARY 6.3 [21, Corollary 4.5′]. *Let* μ *be the normalized absolutely continuous invariant measure on any connected component* $\mathcal{C}_\kappa^{(1)}$ *of a stratum* $\mathcal{M}_\kappa^{(1)}$ *of the moduli space of orientable holomorphic quadratic differentials of unit total area. The second Lyapunov exponents of* G_t^{KZ} *over* $\mathcal{C}_\kappa^{(1)}$ *is strictly positive, in fact*

$$\lambda_2^\mu \geqslant \int_{\mathcal{M}_\kappa^{(1)}} \Lambda^-(q)\,d\mu(q) > 0. \tag{51}$$

The proof of Theorem 3.1 is complete only if $g = 2$. If $g \geqslant 3$, the complete proof of the theorem is based on formulas similar to (37) for the logarithm of the k-volume of k-dimensional isotropic subspaces of $H^1(M,\mathbb{R})$, for all $k \in \{1,\dots,g\}$. Unfortunately, only in the case $k = g$ these computations yield a closed formula for the Lyapunov exponents,

that is, independent of the Oseledec's splitting of the real cohomology bundle $\mathcal{H}^1_\kappa(M,\mathbb{R})$. As a consequence, the complete proof of Theorem 3.1 is rather convoluted and beyond the scope of this paper. In the case $k = g$ we find a somewhat different version of a formula discovered by M. Kontsevich and A. Zorich:

THEOREM 6.4 ([25] and [21, Corollary 5.3]). *Let μ be the normalized absolutely continuous invariant measure on any connected component $\mathcal{C}^{(1)}_\kappa$ of a stratum $\mathcal{M}^{(1)}_\kappa$ of the moduli space of orientable holomorphic quadratic differentials of unit total area. The Lyapunov exponents of G^{KZ}_t over $\mathcal{C}^{(1)}_\kappa$ satisfy the following formula*:

$$\lambda^\mu_1 + \cdots + \lambda^\mu_g = \int_{\mathcal{M}^{(1)}_g} \big(\Lambda_1(q) + \cdots + \Lambda_g(q)\big)\, d\mu(q). \tag{52}$$

We remark that, since $\Lambda_1(q) \equiv \lambda^\mu_1 = 1$, the above formula yields a closed formula for the sum $\lambda^\mu_2 + \cdots + \lambda^\mu_g$, hence for the second exponent λ^μ_2 if $g = 2$. We do not know of any other closed formulas for single exponents or partial sums of them if $g \geqslant 3$.

Kontsevich (and Zorich) [25] have conjectured that the sums of the Lyapunov exponents (52) are rational numbers for all connected components of all strata. These numbers are conjecturally related to the Siegel–Veech constants which arise in counting problems for embedded flat cylinders or saddle-connections on translation surfaces [42,13]. Siegel–Veech constants can in turn be computed (exactly!) by formulas expressing them in terms of the volumes of connected components of strata [14,15] (see the article by A. Eskin [2] in this handbook on counting problems, Siegel–Veech constants and volumes of strata).

7. An example

The problem of describing the intersections of $SL(2,\mathbb{R})$-orbits of quadratic differentials with the determinant locus $\mathcal{D}^{(1)}_g \subset \mathcal{M}^{(1)}_g$ is in general open. Since $\mathcal{D}^{(1)}_g$ is by its very definition invariant under the action of the circle subgroup $SO(2,\mathbb{R})$, this problem can be reduced to the one of describing the intersection of the projection $\mathcal{D}^{(1)}_g/SO(2,\mathbb{R})$ of the determinant locus with Teichmüller disks inside the quotient space $\mathcal{M}^{(1)}_g/SO(2,\mathbb{R})$.

The determinant locus has real codimension 2 while Teichmüller disks have dimension 2, hence it is natural to expect that the intersection with a *generic* disk be either empty or a discrete (possibly countable) set. In many cases, it is immediate to see that the intersection is non-empty. Examples of Teichmüller disks with non-empty intersection are provided by quadratic differentials with symmetries.

W. Veech asked whether there exists a Teichmüller disk (in the moduli space of *orientable* quadratic differentials) entirely contained in the projection of the determinant locus. We will show below that the answer to this question is affirmative by exhibiting an example in genus $g = 3$. It should be remarked that we can prove that the answer to Veech's question is negative in genus $g = 2$.

The idea behind our example is to consider (orientable) holomorphic quadratic differentials with appropriate symmetries which are stable under the $SL(2,\mathbb{R})$-action. We are able

to answer a refined version of Veech's question which has immediate consequences for the Lyapunov exponents of the Kontsevich–Zorich cocycle. We introduce a natural filtration

$$\mathcal{R}_g^{(1)}(1) \subset \mathcal{R}_g^{(1)}(2) \subset \cdots \subset \mathcal{R}_g^{(1)}(g-1) = \mathcal{D}_g^{(1)} \tag{53}$$

of the determinant locus $\mathcal{D}_g^{(1)}$ by the sets

$$\mathcal{R}_g^{(1)}(k) := \left\{ q \in \mathcal{M}_g^{(1)} \mid \Lambda_{k+1}(q) = \cdots = \Lambda_g(q) = 0 \right\}. \tag{54}$$

It is immediate to see that $\mathcal{R}_g^{(1)}(k)$ is a real analytic subvariety of the moduli space (described by the vanishing of all minors of order $k+1$ of the derivative of the period matrix along the Teichmüller flow), invariant under the action of the circle group, for all $k \in \{1, \ldots, g-1\}$.

We will describe below a closed $SL(2,\mathbb{R})$-orbit contained not only in the determinant locus $\mathcal{D}_3^{(1)}$ but in the smaller locus $\mathcal{R}_3^{(1)}(1)$. We do not know whether there are similar examples in any genus $g \geqslant 3$.

The relevance of the locus $\mathcal{R}_g^{(1)}(1)$ is given by the following vanishing result for the Lyapunov exponents of the Kontsevich–Zorich cocycle:

COROLLARY 7.1. *Let μ be an $SL(2,\mathbb{R})$-invariant ergodic probability measure on the moduli space $\mathcal{M}_g^{(1)}$. If* $\operatorname{supp}(\mu) \subset \mathcal{R}_g^{(1)}(1)$, *then*

$$\lambda_2^\mu = \cdots = \lambda_g^\mu = 0. \tag{55}$$

PROOF. It can be proved that the Kontsevich–Zorich formula (52) holds for any $SL(2,\mathbb{R})$-invariant ergodic probability measure on $\mathcal{M}_g^{(1)}$. Hence the result follows. □

We are unable to prove by our methods stronger vanishing results, based on conditions of type $\operatorname{supp}(\mu) \subset \mathcal{R}_g(k)$ for $k > 1$.

Let $q \in Q_g^{(1)}$ be a holomorphic (orientable) quadratic differential with a non-trivial group $\operatorname{Aut}(q)$ of symmetries. The group $\operatorname{Aut}(q) \subset \operatorname{Aut}(M_q)$ is defined as the subgroup formed by all automorphisms $a \in \operatorname{Aut}(M_q)$ such that $a^*(q) = q$. There is a natural unitary action (by pull-back) of $\operatorname{Aut}(q)$ on the finite-dimensional Euclidean space $\mathcal{M}_q^+ \subset L_q^2(M)$ of meromorphic functions.

For any $a \in \operatorname{Aut}(q)$, let $\{m_1^+(a), \ldots, m_g^+(a)\}$ be an orthonormal basis of eigenvectors and let $\{u_1(a), \ldots, u_g(a)\}$ the corresponding eigenvalues for the unitary operator induced by a on $\mathcal{M}_q^+$. Let $B^a(q)$ be the matrix of the projection operator $\pi_q^- : \mathcal{M}_q^+ \to \mathcal{M}_q^-$, with respect to the bases $\{m_1^+(a), \ldots, m_g^+(a)\} \subset \mathcal{M}_q^+$ and $\{\overline{m_1^+}(a), \ldots, \overline{m_g^+}(a)\} \subset \mathcal{M}_q^-$, that is

$$B_{ij}^a(q) = B_q\left(m_i^+(a), m_j^+(a)\right) = \int_M m_i^+(a) m_j^+(a) \omega_q. \tag{56}$$

For any $I, J \subset \{1, \ldots, g\}$ with $\#(I) = \#(J)$, let $B_{I,J}^a(q)$ be the minor of the matrix $B^a(q)$ with entries $B_{ij}^a(q)$ for $i \in I$ and $j \in J$.

LEMMA 7.2. *Let $q \in Q_g^{(1)}$ be a holomorphic quadratic differential with a non-trivial group* $\mathrm{Aut}(q)$ *of symmetries. For any* $a \in \mathrm{Aut}(q)$,

$$\prod_{i\in I}\prod_{j\in J} u_i(a)u_j(a) \neq 1 \quad \Longrightarrow \quad \det B^a_{I,J}(q) = 0. \tag{57}$$

PROOF. Since $a \in \mathrm{Aut}(q)$, by (56) and by change of variables, we have

$$B^a_{ij}(q) = \int_M a^* m_i^+(a) a^* m_j^+(a)\omega_q = u_i(a)u_j(a)B^a_{ij}(q). \tag{58}$$

The result follows. □

Let Q_0 be the stratum of *meromorphic* quadratic differentials with 4 simple poles on the (punctured) Riemann sphere $\mathbb{P}^1(\mathbb{C})$. The corresponding moduli space $\mathcal{M}_0^{(1)}$ of meromorphic quadratic differentials with unit total area consists of a single $SL(2,\mathbb{R})$-orbit.

Let $\kappa = (1,1,1,1)$ and let $\mathcal{M}_\kappa$ the stratum of holomorphic differentials (on Riemann surfaces of genus $g = 3$) with 4 simple zeroes. Let $V \subset \mathcal{M}_\kappa$ be subvariety of all orientable quadratic differentials obtained as the pull-back of a meromorphic quadratic differential $q_0 \in Q_0$ by a 4-sheeted branched covering, branched over the 4 poles of q_0 (with branching order equal to 4 at each pole).

The subvariety $V^{(1)} = V \cap \mathcal{M}_\kappa^{(1)}$ consists of a single closed $SL(2,\mathbb{R})$-orbit. In fact, it can be described as the (closed) $SL(2,\mathbb{R})$-orbit of the (*non-primitive*) *Veech surface* obtained as a 2-sheeted branched cover of the torus $\mathbb{C}/(\mathbb{Z}\oplus\imath\mathbb{Z})$, branched over the 4 half-integer points $(\mathbb{Z}/2\oplus\imath\mathbb{Z}/2)/(\mathbb{Z}\oplus\imath\mathbb{Z})$ (see the article by P. Hubert and T. Schmidt [4] in this handbook on the theory of Veech surfaces).

THEOREM 7.3. *The closed $SL(2,\mathbb{R})$-orbit $V^{(1)} \subset \mathcal{M}_\kappa^{(1)}$ is entirely contained in the locus* $\mathcal{R}_3(1)$.

PROOF. Let $q \in V^{(1)}$. By definition there exists a 4-sheeted branched covering $z : M_q \to \mathbb{P}^1(\mathbb{C})$, branched over 4 (distinct) points $x_1, \ldots, x_4 \in \mathbb{P}^1(\mathbb{C})$ and a meromorphic quadratic differential q_0 on $\mathbb{P}^1(\mathbb{C})$, with 4 simple poles at the points $x_1, \ldots, x_4$ such that $q = z^*(q_0)$. The Riemann surface M_q is a genus 3 surface determined by the algebraic equation:

$$w^4 = (z - x_1)(z - x_2)(z - x_3)(z - x_4). \tag{59}$$

The group $\mathrm{Aut}(M_q)$ of all automorphisms of the Riemann surface M_q is cyclic of order 4, generated by the automorphism $a : M_q \to M_q$ given by

$$a(z,w) = (z, \imath w). \tag{60}$$

The divisors of the meromorphic functions z, w and of the meromorphic differential dz are of the following form:

$$(z) = \frac{P_1P_2P_3P_4}{Q_1Q_2Q_3Q_4}, \qquad (w) = \frac{X_1X_2X_3X_4}{Q_1Q_2Q_3Q_4},$$
$$(dz) = \frac{X_1^3X_2^3X_3^3X_4^3}{Q_1^2Q_2^2Q_3^2Q_4^2}, \tag{61}$$

where $z^{-1}\{0\} = \{P_1, \dots, P_4\}$, $z^{-1}\{\infty\} = \{Q_1, \dots, Q_4\}$ and $X_1, \dots, X_4$ are the branching points of the covering $z : M_q \to \mathbb{P}^1(\mathbb{C})$. It follows that the differentials

$$\theta_1 := \frac{dz}{w^2}, \qquad \theta_2 := \frac{dz}{w^3}, \qquad \theta_3 := \frac{zdz}{w^3}, \tag{62}$$

form a basis of the space of holomorphic differentials on M_q which diagonalizes the action of the group $\mathrm{Aut}(M_q)$ on the vector space of holomorphic differentials on M_q. In fact, by (60) and (62), the action of the automorphism $a \in \mathrm{Aut}(M_q)$ on the basis (62) is diagonal with eigenvalues -1 (with multiplicity 1) and $\imath = \sqrt{-1}$ (with multiplicity 2):

$$a^*(\theta_1) = -\theta_1, \qquad a^*(\theta_2) = \imath\theta_2, \qquad a^*(\theta_3) = \imath\theta_3. \tag{63}$$

The orientable quadratic differentials $q \in V^{(1)}$ is therefore equal to θ_1^2 (up to multiplication by a non-zero complex number) and the spectrum of the action of $a \in \mathrm{Aut}(q)$ on the space $\mathcal{M}_q^+ \subset L_q^2(M)$ of meromorphic functions consists of the eigenvalues

$$u_1(a) = 1, \qquad u_2(a) = -\imath, \qquad u_3(a) = -\imath. \tag{64}$$

It follows that $q \in \mathcal{R}_3^{(1)}(1)$. In fact, by Lemma 7.2 all entries $B_{ij}^a(q) = 0$ for all $(i, j) \neq (1, 1)$, hence the matrix $B^a(q)$ and, consequently, the Hermitian form H_q have rank 1. The argument is concluded. □

By Corollary 7.1, we have

COROLLARY 7.4. *The normalized $SL(2, \mathbb{R})$-invariant measure μ supported on the closed $SL(2, \mathbb{R})$-orbit $V^{(1)}$ is an $SL(2, \mathbb{R})$-invariant ergodic probability measure on $\mathcal{M}_3^{(1)}$ such that*

$$\lambda_2^\mu = \lambda_3^\mu = 0. \tag{65}$$

8. Invariant sub-bundles

By Oseledec's theorem [34], [1, §5], for almost all holomorphic quadratic differentials $q \in \mathcal{M}_\kappa^{(1)}$, the fiber $H^1(M_q, \mathbb{R})$ of the cohomology bundle $\mathcal{H}_\kappa^1(M, \mathbb{R})$ has a direct splitting

$$H^1(M_q, \mathbb{R}) = E_q^+ \oplus E_q^- \oplus E_q^0, \tag{66}$$

where E_q^+, E_q^- and E_q^0 are the subspaces of cohomology classes with, respectively, strictly positive, strictly negative and zero Lyapunov exponent. Since the cohomology bundle has a symplectic structure E_q^+ and E_q^- are isotropic subspaces of the same dimension. In fact, according to Theorem 3.1, $E_q^0 = \{0\}$ and E_q^+, E_q^- are Lagrangian. We will not rely below on this result, hence the results of this section will be independent of the non-uniform hyperbolicity of the Kontsevich–Zorich cocycle.

The homology cycles in the Poincaré dual of E_q^+, E_q^- are called (following I. Nikolaev and E. Zhuzhoma [33, §7.9.3]) the *Zorich cycles* for the horizontal, respectively, vertical, measured foliation of the quadratic differential q. Zorich cycles for an orientable measured foliations $\mathcal{F}$ are a generalization of the *Schwartzman's asymptotic cycle* which coincides with the Poincaré dual of the cohomology class of the closed 1-form $\eta_{\mathcal{F}}$ such that $\mathcal{F} := \{\eta_{\mathcal{F}} = 0\}$.

In fact, by unique ergodicity, the Schwartzman's cycle yields the direction of the leading term in the asymptotic behavior in homology of a typical leaf of a generic orientable measured foliation on a surface of genus $g \geqslant 1$, while Zorich cycles yield the direction of the first g terms (under the hypothesis that the cocycle is non-uniformly hyperbolic) as the length of the leaf gets large. The remainder in this asymptotics, that is, the distance in homology of the typical leaf from the space of all Zorich's cycles, stays uniformly bounded (see [44,47] or [48, Appendix D]).

We will outline below the proof of a *representation theorem* which states that all Zorich cycles (or rather the corresponding dual cohomology classes) can be represented in terms of *currents* of order 1 satisfying certain properties with respect to the measured foliation $\mathcal{F}$.

A *basic current* (of dimension 1) for a measured foliation $\mathcal{F}$ (with singularities at a finite set $\Sigma_{\mathcal{F}} \subset M$) is a 1-dimensional current C (in the sense of G. de Rham [12], that is, a continuous functional on the vector space of smooth 1-forms with compact support) on $M \setminus \Sigma_{\mathcal{F}}$ which satisfies the vanishing conditions

$$\iota_X C = \mathcal{L}_X C = 0, \tag{67}$$

for all smooth vector fields X with compact support in $M \setminus \Sigma_{\mathcal{F}}$ tangent to the leaves of the foliation $\mathcal{F}$. (The operation of contraction ι_X and Lie derivative $\mathcal{L}_X$ are extended to currents in the standard distributional sense [36, Chapter IX, §3].)

Basic currents are a distributional generalization of basic forms, a well-known notion in the geometric theory of foliations. Since M has dimension 2, a current of dimension 1 satisfying (67) is closed, hence it represents, by the generalized de Rham theorem (see [12, Theorem 12], or [36, Chapter IX, §3, Theorem I]) a cohomology class in $H^1(M \setminus \Sigma_{\mathcal{F}}, \mathbb{R})$.

Let $q \in Q_\kappa^{(1)}$ be an orientable quadratic differential. Let $\mathcal{B}_{\pm q}(M)$ be, respectively, the space of basic currents for the measured foliations $\mathcal{F}_{\pm q}$ (we recall that $\mathcal{F}_q$ is the horizontal foliation and $\mathcal{F}_{-q}$ the vertical foliation). Let $\{S, T\}$ be the orthonormal frame of the tangent bundle described in Section 4 and $\{\eta_T, \eta_S\}$ be the dual frame of the cotangent bundle, which is defined by

$$\eta_T := -\iota_T \omega_q = \operatorname{Re}\left(q^{1/2}\right), \qquad \eta_S := \iota_S \omega_q = \operatorname{Im}\left(q^{1/2}\right). \tag{68}$$

For the statement of the representation theorem, the notion of order of a current, taken with respect to a scale of *Sobolev spaces*, is crucial. Let Σ_q be the set of the zeroes of q.

A current on $M \setminus \Sigma_q$ has *order* $r \in \mathbb{N}$ if it extends to a continuous functional on the Sobolev space $H^r_q(M)$ of all L^2_q forms with L^2_q derivatives (with respect to the vector fields S, T) up to order r. We remark that under this definition the order of a current is not uniquely defined. In fact, a current of order r has also order r' for all $r' \geqslant r$.

Let $\mathcal{B}^r_{\pm q}(M) \subset \mathcal{B}_{\pm q}(M)$ be the subsets of basic currents of order r. There is a close relation between basic currents (of order r) and *invariant distributions* (of order r). An S-invariant, respectively T-invariant, distribution (of order r) is a distributional solution $\mathcal{D}$ (of order r) of the equation

$$S\mathcal{D} = 0, \quad \text{respectively} \quad T\mathcal{D} = 0. \tag{69}$$

We have proved in [19] that invariant distributions of finite order for the vector field S, respectively T, yield a complete system of obstructions to the existence of smooth solutions u to the *cohomological equation*

$$Su = f, \quad \text{respectively} \quad Tu = f, \tag{70}$$

in the following sense. There exists $\gamma > 1$ such that for almost all quadratic differentials $q \in \mathcal{M}^{(1)}_\kappa$ and for any function $f \in H^r_q(M)$ which belongs to the kernel of all S-invariant, respectively T-invariant, distributions of order r, the cohomological equation $Su = f$, respectively $Tu = f$, has a solution $u \in H^s_q(M)$ for all $s < r - \gamma$ (finite loss of derivatives).

The following result describes the relation between basic currents and S-invariant, T-invariant distributions:

LEMMA 8.1 [21, Lemma 6.6]. *A current $C \in \mathcal{B}^r_q(M)$, respectively $C \in \mathcal{B}^r_{-q}(M)$, if and only if $C = \mathcal{D} \cdot \eta_S$, respectively $C = \mathcal{D} \cdot \eta_T$, where $\mathcal{D}$ is an S-invariant, respectively a T-invariant, distribution of order $r \in \mathbb{N}$.*

The main result of this section states that, for almost all $q \in \mathcal{M}^{(1)}_\kappa$, the Poincaré dual of every Zorich cycle is the cohomology class of a basic current of *order* 1. It can be proved that the natural cohomology maps

$$\mathcal{B}^1_{\pm q}(M) \to H^1(M \setminus \Sigma_q, \mathbb{R})$$

are injective and their images $H^{1,1}_{\pm q}(M, \mathbb{R})$ satisfy the inclusions

$$H^{1,1}_{\pm q}(M, \mathbb{R}) \subset H^1(M, \mathbb{R}) \subset H^1(M \setminus \Sigma_q, \mathbb{R}).$$

We can finally state the representation theorem for Zorich cycles:

THEOREM 8.2 [21, Theorem 8.3]. *For almost all $q \in \mathcal{M}^{(1)}_\kappa$, we have*

$$E^+_q = H^{1,1}_q(M, \mathbb{R}), \qquad E^-_q = H^{1,1}_{-q}(M, \mathbb{R}). \tag{71}$$

(*The Poincaré duals of Zorich cycles for a generic orientable measured foliation $\mathcal{F}$ are represented by basic currents for $\mathcal{F}$ of Sobolev order* 1).

We will outline below the proof of the main part of Theorem 8.2, that is, the inclusions $E^{\pm}_q \subset H^{1,1}_{\pm q}(M,\mathbb{R})$. The argument is based on the following *Cheeger-type estimate* for the constant in the *Poincaré inequality* (equivalently, for the first non-trivial eigenvalue of the Laplace–Beltrami operator of the flat metric R_q induced by the quadratic differential q on M).

The *Dirichlet form* of the metric R_q, introduced in [19, (2.6)], is defined as the Hermitian form on the Hilbert space $L^2_q(M)$ given by

$$\mathcal{Q}(u,v) := (Su, Sv)_q + (Tu, Tv)_q. \tag{72}$$

The domain of the Dirichlet form $\mathcal{Q}$ is the Sobolev space $H^1_q(M) \equiv H^1(M)$ of functions $u \in L^2_q(M)$ such that $Su, Tu \in L^2_q(M)$.

LEMMA 8.3 [21, Lemma 6.9]. *There is a constant $K_{g,\sigma} > 0$ such that the following holds. Let $q \in Q^{(1)}_g$ be a holomorphic* (*orientable*) *quadratic differential, let Σ_q be the set of its zeroes and let $\sigma := \#(\Sigma_q)$. Denote by $\|q\|$ the R_q-length of the shortest geodesic segment with endpoints in Σ_q. Then, for any $v \in H^1_q(M)$, the following inequality holds*:

$$\left| v - \int_M v\,\omega_q \right|_0 \leqslant \frac{K_{g,\sigma}}{\|q\|}\,\mathcal{Q}(v,v)^{1/2}. \tag{73}$$

The proof of Lemma 8.3 follows closely Cheeger's proof (see [11] or [10, Chapter III, D.4]) for the case of a smooth Riemann metric. The degenerate (or singular) character of the metric R_q at the finite set Σ_q does not affect Cheeger's argument. Moreover, we are able to give an explicit estimate of Cheeger's *isoperimetric constant* in terms of the quantity $\|q\|$.

PARTIAL PROOF OF THEOREM 8.2. We prove the inclusion $E^+_q \subset H^{1,1}_q(M,\mathbb{R})$. The inclusion $E^-_q \subset H^{1,1}_{-q}(M,\mathbb{R})$ can be proved by a similar argument.

Let $q \in Q^{(1)}_\kappa$ be any Oseledec regular point of the Kontsevich–Zorich cocycle and let $c_t := G^{KZ}_t(c)$, $t \in \mathbb{R}$, be the orbit under the cocycle of a cohomology class $c \in H^1(M,\mathbb{R})$.

Let $\mathcal{M}^+_t$ be the space of meromorphic functions, with respect to the complex structure induced by the quadratic differential $q_t := G_t(q) \in Q^{(1)}_\kappa$, which belong the space $L^2_{q_t}(M)$.

According to the representation formula (20), for each $t \in \mathbb{R}$ there exists a function $m^+_t \in \mathcal{M}^+_t$ such that

$$c_t = \mathrm{Re}\big[m^+_t q_t^{1/2}\big]. \tag{74}$$

Since the L^2_q norm is invariant under the action Teichmüller flow on the Teichmüller space, the space $\mathcal{M}^+_t \subset L^2_q(M)$ for all $t \in \mathbb{R}$, and it can be proved that the map $t \to m^+_t \in L^2_q(M)$ is smooth.

There exist a measurable function $K_1 > 0$ on $\mathcal{M}_\kappa^{(1)}$ and an exponent $0 < \lambda < 1$ such that, if $c \in E_q^+$, the Hodge norm

$$\|c\|_{q_t} = \big|m_t^+\big|_0 \leqslant K_1(q)\big|m_0^+\big|_0 \exp\big(-\lambda|t|\big), \quad t \leqslant 0. \tag{75}$$

Since $c_t \equiv c \in H^1(M, \mathbb{R})$ (by the definition (8) of G_t^{KZ}), there exists a unique zero average function $U_t \in L_q^2(M)$ such that

$$dU_t = \mathrm{Re}\big[m_t^+ q_t^{1/2}\big] - \mathrm{Re}\big[m_0^+ q^{1/2}\big]. \tag{76}$$

It follows that, by the variational formula (23), the function U_t satisfies the following Cauchy problem in $L_q^2(M)$:

$$\begin{cases} \frac{d}{dt}U_t = 2\,\mathrm{Re}(v_t), \\ U_0 = 0 \end{cases} \tag{77}$$

(if the function $v_t \in H^1(M)$ in (23) is chosen with zero average).

For any (orientable) quadratic differential $q \in Q_\kappa^{(1)}$, by the commutativity property (17) of the vector fields S, T, the Dirichlet form can be written as

$$\mathcal{Q}(v, v) = \big|\partial_q^\pm v\big|_0^2, \quad \text{for all } v \in H_q^1(M)$$

(where $\partial_q^\pm$ are the Cauchy–Riemann operators introduced in Section 4).

Since the function $v_t \in H_q^1(M) \equiv H^1(M)$ in (23) is chosen with zero average, by the Poincaré inequality Lemma 8.3 and by the orthogonality of the decomposition (19), (22) for $q = q_t$, we have

$$|v_t|_0 \leqslant K_{g,\sigma}\|q_t\|^{-1}\big|\partial_t^+ v_t\big|_0 \leqslant K_{g,\sigma}\|q_t\|^{-1}\big|m_t^+\big|_0, \tag{78}$$

where $\|q_t\|$ denotes as above the length of the shortest geodesic segment with endpoints in the set of zeroes of the quadratic differential q_t with respect to the induced metric.

It follows, by formulas (75), (77) and (78), that there exists a measurable function $K_2 > 0$ on $\mathcal{M}_\kappa^{(1)}$ such that, if $c \in E_q^+$,

$$\left|\frac{d}{dt}U_t\right|_0 \leqslant 2|v_t|_0 \leqslant K_2(q)\big|m_0^+\big|_0\|q_t\|^{-1}\exp\big(-\lambda|t|\big), \quad t \leqslant 0. \tag{79}$$

Since $U_0 = 0$, by Minkowski's integral inequality, formula (79) implies the following estimate:

$$|U_t|_0 \leqslant K_2(q)\big|m_0^+\big|_0 \int_0^{|t|} e^{-\lambda|s|}\|q_s\|^{-1}\,ds, \quad t \leqslant 0. \tag{80}$$

By the *logarithmic law* for the Teichmüller geodesic flow on the moduli space, proved by H. Masur in [31], the following estimate holds for almost all quadratic differentials $q \in \mathcal{M}_\kappa^{(1)}$ (see [31, Proposition 1.2]):

$$\limsup_{t\to\pm\infty} \frac{-\log \|q_t\|}{\log |t|} \leqslant \frac{1}{2}. \tag{81}$$

It follows that, for almost all $q \in \mathcal{M}_\kappa^{(1)}$, the integral in formula (80) converges as $t \to -\infty$, hence the family of functions $\{U_t \mid t \leqslant 0\}$ is uniformly bounded in the Hilbert space $L^2_q(M)$.

Let $U \in L^2_q(M)$ be any weak limit of U_t as $t \to -\infty$ (which exists since bounded subsets of separable Hilbert spaces are sequentially weakly compact). By contraction of the identity (76) with the vector field S and by taking the limit as $t \to -\infty$, we have

$$\begin{aligned} SU_t &= -\operatorname{Re}\left(m_0^+\right) + e^t \operatorname{Re}\left(m_t^+\right), \quad t \leqslant 0; \\ SU &= -\operatorname{Re}\left(m_0^+\right). \end{aligned} \tag{82}$$

The identities in (82) hold in the sense of distributions. It follows by a straightforward computation that there exists a distribution $\mathcal{D}$ such that

$$dU = -\operatorname{Re}\left[m_0^+ q^{1/2}\right] + \mathcal{D}\cdot\eta_S. \tag{83}$$

In fact, $dU = SU\,\eta_T + TU\eta_S$, hence by (68) the identity (83) holds with $\mathcal{D} := TU - \operatorname{Im}(m_0^+)$. Since $U \in L^2_q(M)$ the distribution $\mathcal{D}$ has Sobolev order 1 and the current $C := \mathcal{D}\cdot\eta_S$ is a basic current of order 1 for the horizontal foliation $\mathcal{F}_q$ representing the cohomology class $c \in E_q^+$.

In fact, it is immediate by (83) that C is closed and represents $c = \operatorname{Re}[m_0^+ q^{1/2}]$. Finally, C is basic for $\mathcal{F}_q = \{\eta_S = 0\}$ since, if X is any vector fields tangent to $\mathcal{F}_q$ on $M \setminus \Sigma_q$, we have in the distributional sense:

$$\begin{aligned} \imath_X C &= \mathcal{D}\cdot\imath_X\eta_S = \mathcal{D}\cdot 0 = 0, \\ \mathcal{L}_X C &= \imath_X dC + d\imath_X C = 0. \end{aligned} \tag{84}$$

Otherwise, since C is closed and by a standard formula $dC = S\mathcal{D}\cdot\omega_q$, the distribution $\mathcal{D}$ is S-invariant, hence C is basic for $\mathcal{F}_q$ by Lemma 8.1. □

References

Surveys in volume 1A and this volume

[1] L. Barreira and Y. Pesin, *Smooth ergodic theory and nonuniformly hyperbolic dynamics*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 57–263.

[2] A. Eskin, *Counting problems in moduli space*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 581–595.
[3] B. Hasselblatt and A.B. Katok, *Principal structures*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 1–203.
[4] P. Hubert and T. Schmidt, *An introduction to Veech surfaces*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 501–526.
[5] H. Masur, *Ergodic theory of translation surfaces*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 527–547.
[6] H. Masur and S. Tabachnikov, *Rational billiards and flat structures*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 1015–1089.

Other sources

[7] A. Avila and G. Forni, *Weak mixing for interval exchange transformations and translation flows*, Preprint (2004), arXiv:math.DS/0304469, Ann. of Math., to appear.
[8] A. Avila and M. Viana, *Simplicity of Lyapunov spectra: Proof of the Zorich–Kontsevich conjecture*, Preprint (2005).
[9] L. Bers, *Spaces of degenerating Riemann surfaces*, Discontinuous Groups and Riemann Surfaces, Proc. Conf. Univ. of Maryland, College Park, MD, 1973, Annals of Mathematical Studies, Vol. 79, Princeton Univ. Press, Princeton, NJ (1974), 43–55.
[10] M. Berger, P. Gauduchon and E. Mazet, *Le spectre d'une variété riemannienne*, Lecture Notes in Mathematics, Vol. 194, Springer-Verlag, Berlin (1971).
[11] J. Cheeger, *A lower bound for the smallest eigenvalue of the Laplacian*, Problems in Analysis, A Symposium in Honor of Salomon Bochner, Proceedings, Princeton University, April 1969, R.C. Gunning, ed., Princeton Univ. Press, Princeton, NJ (1970), 195–199.
[12] G. de Rham, *Varietés differentiables*, Hermann, Paris (1960).
[13] A. Eskin and H. Masur, *Asymptotic formulas on flat surfaces*, Ergodic Theory Dynam. Systems **21** (2001), 443–478.
[14] A. Eskin, H. Masur and A. Zorich, *Moduli spaces of Abelian differentials: The principal boundary, counting problems and the Siegel–Veech constants*, Publ. Math. Inst. Hautes Etudes Sci. **97** (2003), 61–179.
[15] A. Eskin and A. Okounkov, *Asymptotics of numbers of branched covers of a torus and volumes of moduli spaces of holomorphic differentials*, Invent. Math. **145** (1) (2001), 59–104.
[16] H.M. Farkas and I. Kra, *Riemann Surfaces*, 2nd edn, Springer-Verlag, New York, NY (1992).
[17] A. Fathi, F. Laudenbach and V. Poénaru, *Travaux de Thurston sur les surfaces*, Astérisque **66–67** (1979).
[18] J.D. Fay, *Theta functions on Riemann surfaces*, Lecture Notes in Mathematics, Vol. 352, Springer-Verlag, Heidelberg (1973).
[19] G. Forni, *Solutions of the cohomological equation for area-preserving flows on compact surfaces of higher genus*, Ann. of Math. (2) **146** (2) (1997), 295–344.
[20] G. Forni, *Asymptotic behaviour of ergodic integrals of renormalizable parabolic flows*, Proceedings of the ICM 2002, Vol. III, Higher Education Press, Beijing, China (2002), 317–326.
[21] G. Forni, *Deviation of ergodic averages for area-preserving flows on surfaces of higher genus*, Ann. of Math. (2) **155** (1) (2002), 1–103.
[22] Y. Imayoshi and M. Taniguchi, *An Introduction to Teichmüller Spaces*, Springer-Verlag, Tokyo (1992).
[23] A.B. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge Univ. Press, Cambridge (1995).
[24] A.B. Katok and A.M. Stepin, *Approximations in ergodic theory*, Uspekhi Mat. Nauk **22** (5) (1967), 81–106.
[25] M. Kontsevich, *Lyapunov exponents and Hodge theory*, The Mathematical Beauty of Physics, Saclay, 1996, Adv. Ser. Math. Phys., Vol. 24, World Scientific, River Edge, NJ (1997), 318–332.
[26] M. Kontsevich and A. Zorich, *Connected components of the moduli space of Abelian differentials with prescribed singularities*, Inv. Math. **153** (2003), 631–678.
[27] S. Marmi, P. Moussa and J.-C. Yoccoz, *On the cohomological equation for interval exchange maps*, C. R. Acad. Sci. Paris **336** (2003), 941–948.

[28] S. Marmi, P. Moussa and J.-C. Yoccoz, *The cohomological equation for Roth type interval exchange maps*, Preprint (2004), J. Amer. Math. Soc., to appear.
[29] H. Masur, *The extension of the Weyl–Petersson metric to the boundary of the Teichmüller space*, Duke Math. J. **43** (1976), 623–635.
[30] H. Masur, *Interval exchange transformations and measured foliations*, Ann. of Math. **115** (1982), 168–200.
[31] H. Masur, *Logarithmic law for geodesics in moduli space*, Mapping Class Groups and Moduli Spaces of Riemann surfaces, Proceedings, Göttingen June–August 1991, Contemporary Mathematics, Vol. 150, Amer. Math. Soc., Providence, RI (1993), 229–245.
[32] S. Nag, *The complex analytic theory of Teichmüller spaces*, Wiley, New York, NY (1988).
[33] I. Nikolaev and E. Zhuzhoma, *Flows on 2-dimensional manifolds—An overview*, Lecture Notes in Mathematics, Vol. 1705, Springer-Verlag, Berlin (1999).
[34] V.I. Oseledec, *A multiplicative ergodic theorem. Characteristic Lyapunov exponents of dynamical systems*, Trans. Moscow Math. Soc. **19** (1968), 179–210.
[35] G. Rauzy, *Échanges d'intervalles et transformations induites*, Acta Arith. **34** (1979), 315–328.
[36] L. Schwartz, *Théorie des distributions*, Herman, Paris (1966).
[37] W. Thurston, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. **19** (1988), 417–431.
[38] W. Veech, *Gauss measures for transformations on the space of interval exchange maps*, Ann. of Math. **115** (1982), 201–242.
[39] W. Veech, *The metric theory of interval exchange transformations. I. Generic spectral properties*, Amer. J. Math. **106** (6) (1984), 1331–1359.
[40] W. Veech, *The Teichmüller geodesic flow*, Ann. of Math. **124** (1986), 441–530.
[41] W. Veech, *Moduli spaces of quadratic differentials*, J. Anal. Math. **55** (1990), 117–171.
[42] W. Veech, *Siegel measures*, Ann. of Math. **148** (1998), 895–944.
[43] A. Yamada, *Precise variational formulas for Abelian differentials*, Kodai Math. J. **3** (1980), 114–143.
[44] A. Zorich, *Asymptotic flag of an orientable measured foliation on a surface*, World Scientific (1994), 479–498.
[45] A. Zorich, *Finite Gauss measure on the space of interval exchange transformations. Lyapunov exponents*, Ann. Inst. Fourier (Grenoble) **46** (1996), 325–370.
[46] A. Zorich, *Deviation for interval exchange transformations*, Ergodic Theory Dynam. Systems **17** (1997), 1477–1499.
[47] A. Zorich, *How do the leaves of a closed* 1*-form wind around a surface?* Pseudoperiodic Topology, M. Kontsevich, V.I. Arnol'd and A. Zorich, eds, Amer. Math. Soc. Transl. (2), Vol. 197, Amer. Math. Soc., Providence, RI (1999), 135–178.
[48] A. Zorich, *Rauzy induction, Veech zippered rectangles, Masur polygons, and Lyapunov exponents of Teichmüller geodesic flow*, Preprint (2005).

CHAPTER 9

Counting Problems in Moduli Space

Alex Eskin

Department of Mathematics, University of Chicago, Chicago, IL 60637, USA
E-mail: *eskin@math.uchicago.edu*

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

This page intentionally left blank

Abstract

In this series of lectures, we describe some counting problems in moduli space and outline their connection to the dynamics of the $SL(2,\mathbb{R})$ action on moduli space. Much of this is presented in analogy with the space of lattices $SL(n,\mathbb{R})/SL(n,\mathbb{Z})$.

1. LECTURE 1: Counting problems and volumes of strata

Recall that $\Omega_n = SL(n,\mathbb{R})/SL(n,\mathbb{Z})$ is the space of covolume 1 lattices in $\mathbb{R}^n$. This space is non-compact, since we can have arbitrarily short vectors in a lattice.

We will refer to moduli spaces of translation surfaces as defined in the lectures by Howard Masur in this handbook [3, Definition 6] as *strata*. Note that the case of $n = 2$ in the space of lattices and the case of the stratum $\mathcal{H}_1(\emptyset)$ boil down to the same thing, since we are considering the space of unit area holomorphic 1-forms on tori, which is given by $SL(2,\mathbb{R})/SL(2,\mathbb{Z})$.

Let $B(R)$ be the ball of radius R centered at 0 in $\mathbb{R}^n$. For a given lattice $\Delta \in \Omega_n$, we would like to find out how many lattice points, that is, how many points of Δ are contained in $B(R)$.

It is immediately clear that for a fixed lattice Δ, as $R \to \infty$,

$$\big|\Delta \cap B(R)\big| \sim \mathrm{Vol}\big(B(R)\big) = \mathrm{Vol}\big(B(1)\big)R^n \tag{1}$$

(i.e. the number of lattice points is asymptotic to the volume). However, this is not uniform in Δ. A uniform upper bound can be given as follows:

Let $\mathbb{R}^n$ be endowed with a Euclidean structure. Given a subspace L of $\mathbb{R}^n$, we say it is Δ-rational if $L \cap \Delta$ is a lattice in L. We define $d(L)$ to be the volume of $L/(\Delta \cap L)$. We then define the function α by

$$\alpha(\Delta) = \sup \frac{1}{d(L)},$$

where the supremum is taken over all Δ-rational subspaces L. We have the following result (see [25]): there is a constant C, depending only on the dimension n so that for all $\Delta \in \Omega_n$,

$$\big|\Delta \cap B(1)\big| < C\alpha(\Delta). \tag{2}$$

This estimate follows from what is called "the geometry of numbers".

The analogous problem in moduli space is as follows: let $\mathcal{H}(\beta)$ be a stratum, i.e. a moduli space of translation surfaces (defined in [3, Definition 6]), and let $S = (X,\omega) \in \mathcal{H}(\beta)$. Recall (see, e.g., [3, §1.1]) that the holonomy of a curve γ on S is given by

$$\mathrm{hol}(\gamma) = \int_\gamma \omega.$$

Let

$$V_{sc}(S) = \big\{ \mathrm{hol}(\gamma) \colon \gamma \text{ is a saddle connection on } S \big\},$$

so that $V_{sc}(S) \subset \mathbb{C} \simeq \mathbb{R}^2$ (saddle connections are defined in [3, Definition 3]). Note that $V_{sc}(S)$ is a discrete subset of $\mathbb{R}^2$, but it is not, in general, a subgroup. We are interested in $|V_{sc}(S) \cap B(1)|$, i.e. the number of saddle connections of length at most 1 on S.

The result is as follows: Fix $\epsilon > 0$. Then there is a constant $c = c(\beta, \epsilon)$ such that for all $S \in \mathcal{H}(\beta)$ of area 1,

$$\big|V_{sc}(S) \cap B(1)\big| \leqslant \frac{c}{\ell(S)^{1+\epsilon}}, \tag{3}$$

where $\ell(S)$ is the length of the shortest saddle connection on S.

The proof of this result (which can be found in [9]) is more difficult than that of (2). It uses techniques developed by Margulis for the quantitative version of the Oppenheim conjecture (see Lecture 3), as well as induction on the genus.

The following construction and its analogues play a key role. For any function of compact support $f \in C_c(\mathbb{R}^n)$, let $\hat{f}(\Delta) = \sum_{v \in \Delta \setminus 0} f(v)$. Note that if $f = \chi_{B(1)}$, we get $\hat{f}(\Delta) = |\Delta \cap B(1)|$. We have the *Siegel formula*: For any $f \in C_c(\mathbb{R}^n)$,

$$\frac{1}{\mu(\Omega_n)} \int_{\Omega_n} \hat{f}(\Delta)\, d\mu(\Delta) = \int_{\mathbb{R}^n} f\, d\lambda, \tag{4}$$

where μ is Haar measure on $\Omega_n = SL(n, \mathbb{R})/SL(n, \mathbb{Z})$, and λ is Lebesgue measure on $\mathbb{R}^n$.

The generalization of this formula to moduli space was developed, so the legend goes, by Veech while he listened to Margulis lecture on the Oppenheim conjecture. For $f \in C_c(\mathbb{R}^2)$ we define the Siegel–Veech transform $\hat{f}(S) = \sum_{v \in V_{sc}(S)} f(v)$. Just as above, if $f = \chi_{B(1)}$, $\hat{f}$ counts the number of saddle connections of length $\leqslant 1$.

Just as we had the Siegel formula for lattices, here we have the *Siegel–Veech formula*: There is a constant $b(\beta)$ such that for any $f \in C_c(\mathbb{R}^2)$, we have

$$\frac{1}{\mu(\mathcal{H}_1(\beta))} \int_{\mathcal{H}_1(\beta)} \hat{f}(S)\, d\mu(S) = b(\beta) \int_{\mathbb{R}^2} f, \tag{5}$$

where μ is the natural $SL(2, \mathbb{R})$ invariant measure on $\mathcal{H}_1(\beta)$, where $\mathcal{H}_1(\beta) \subset \mathcal{H}(\beta)$ is the hypersurface of translation surfaces of area 1 (this measure is defined in [3, §3], or in the next section).

Let us sketch the proof of this result (essentially from [28], also reproduced in [9]). The first step (which is by far the most technical) is to show that $\hat{f} \in L^1(\mathcal{H}_1(\beta))$, so that the left-hand side is finite. This can be deduced, e.g., from (3). Having done this, we denote the quantity on the left-hand side of (5) by $\varphi(f)$.

Thus we have a linear functional $\varphi : C_c(\mathbb{R}^2) \to \mathbb{R}$, i.e. a measure. But it also has to be $SL(2, \mathbb{R})$ invariant. Only Lebesgue measure and δ_0, the delta measure at 0 are $SL(2, \mathbb{R})$ invariant. Thus we have $\varphi(f) = af(0) + b \int_{\mathbb{R}^2} f$. It remains to show $a = 0$. Consider the

limit of indicator functions $f = \chi_{B(R)}$ as $R \to 0$. Both sides of the equation tend to 0, so we have that $a = 0$, and thus our result.

Returning to lattices, we can apply literally the same arguments to prove the Siegel formula (4). Note that nothing was special about dimension 2 in the above proof sketch. Thus, we have almost proved (4) as well. To be precise, we currently have:

$$\frac{1}{\mu(\Omega_n)} \int_{\Omega_n} \hat{f}(\Delta)\, d\mu(\Delta) = b \int_{\mathbb{R}^n} f\, d\lambda,$$

for some constant b. We need to show $b = 1$. Here, we once again use $f = \chi_{B(R)}$, but this time consider $R \to \infty$. Recall that $\hat{f}(\Delta) = |\Delta \cap B(R)| \sim \mathrm{Vol}(B(R))$, for $R \to \infty$ and Δ fixed. Thus, we get $b = 1$, and the Siegel formula.

We should remark that for the space of lattices the proof of the Siegel formula indicated above is not the easiest available. In fact, it is possible to avoid proving a priori that $\hat{f} \in L^1(\Omega_n)$. See [26] or [5] or [27] for the details.

We now show how to use the Siegel formula to calculate the volumes of the spaces Ω_n. We first prove a variant of the formula. Recall that $v \in \Delta$ is primitive if there is no integer n so that $v/n \in \Delta$. The analogue of (1) for counting primitive vectors is

$$\big|\Delta_{\mathrm{prim}} \cap B(R)\big| \sim \frac{1}{\zeta(n)} \mathrm{Vol}\big(B(1)\big) R^n, \tag{6}$$

where $\zeta(n)$ is Riemann's zeta function. Now for $f \in C_c(\mathbb{R}^n)$, let

$$\tilde{f}(\Delta) = \sum_{v \in \Delta}{}' f(v),$$

where the prime indicates that we are summing over primitive vectors only. Now the proof of the Siegel formula given above shows that

$$\frac{1}{\mu(\Omega_n)} \int_{\Omega_n} \tilde{f}(\Delta)\, d\mu(\Delta) = \frac{1}{\zeta(n)} \int_{\mathbb{R}^n} f\, d\lambda. \tag{7}$$

The rest of the argument is heuristic. Consider $f = \chi_{B(\epsilon)}$ for some small positive ϵ. We have that $\tilde{f}(\Delta) = 0$ unless Δ has a primitive vector of length less than ϵ. Note that if v is a primitive short vector, then so is $-v$. It turns out that we can, in the limit as $\epsilon \to 0$, ignore the contribution to the integral of the lattices which have more then two primitive short vectors; thus we may assume that $\tilde{f}(\Delta) = 2$. Now, let v any one of the two primitive short vectors in Δ, and consider a basis for Δ containing v. We may subtract multiples of v from the other elements of the basis, to make them as short as possible. After this "reduction" procedure is complete, we get a basis for Δ containing v where all the other elements are almost orthogonal to v. Then these other basis elements form an arbitrary lattice of dimension $n-1$, i.e. an element of Ω_{n-1}. Thus, the left-hand side of (7) is approximately

$$\frac{2}{\mu(\Omega_n)} \frac{1}{2} \mathrm{Vol}\big(B(\epsilon)\big) \mu(\Omega_{n-1}),$$

where the factor of 2 came from the value of $\tilde{f}$, the factor of $\frac{1}{2}\mathrm{Vol}(B(\epsilon))$ came from the integral over $v \in \mathbb{R}^n$, and the factor $\mu(\Omega_{n-1})$ came from the integral over the rest of the basis (and we assumed that $\tilde{f}$ is always either 0 or 1). The right-hand side of (7) is exactly equal to $\frac{1}{\zeta(n)}\mathrm{Vol}(B(\epsilon))$.

Doing this more carefully, and taking into account the normalizations of the measures (to be defined in the next lecture), we get, after sending $\epsilon \to 0$,

$$\frac{1}{\zeta(n)} = \frac{n-1}{n}\frac{\mu(\Omega_{n-1})}{\mu(\Omega_n)}. \tag{8}$$

Now after iterating the above formula, we get the desired formula for the volume:

$$\mu(\Omega_n) = \frac{1}{n}\zeta(2)\zeta(3)\ldots\zeta(n). \tag{9}$$

The above could be justified rigorously, but this is usually not done since (8) and (9) can be obtained from (7) in an easier way (see [26] or [5] or [27]). However, the analogue of the argument presented here is the only way we currently know how to proceed in the case of translation surfaces. This was done in [11] where we obtained the following result, which corresponds to (8). For any stratum (i.e. moduli space of translation surfaces) $\mathcal{H}_1(\beta)$, the coefficient $b(\beta)$ involved in (5) can be expressed in the following form:

$$b(\beta) = \sum_{\alpha<\beta} c(\alpha,\beta)\frac{\mu(\mathcal{H}_1(\alpha))}{\mu(\mathcal{H}_1(\beta))}, \tag{10}$$

where the sum is over lower dimensional strata α (which lie at the "boundary" of $\mathcal{H}_1(\beta)$), and $c(\alpha,\beta)$ are explicitly known rational numbers.

We note that (10) fails as a method for calculating the volumes, since (unlike the lattice case) we do not have an independent formula for $b(\beta)$. In the second lecture we will show that the volumes can be computed in a different way; then (10) can be used to evaluate $b(\beta)$. Also, we will see in the third lecture that $b(\beta)$ is the answer to a certain natural counting problem. The numbers $b(\beta)$, called the Siegel–Veech constants, appear in some other contexts as well, in particular in connection with the Lyapunov exponents of the geodesic flow (see, e.g., [1, end of §6]).

2. LECTURE 2: Lattice points and branched covers

In this lecture we describe briefly another strategy for calculating volumes of moduli spaces of translation surfaces, which also has a parallel for the space of lattices. Recall that we are considering the moduli spaces $\mathcal{H}(\beta)$ of translation surfaces with singularity structure $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$, where $\beta_i \in \mathbb{N}$, $\sum \beta_i = 2g-2$. Let the set of singularities be denoted by Σ. We have $|\Sigma| = n$, and we have the first relative homology group of S relative to Σ (with coefficients in $\mathbb{Z}$):

$$H_1(S, \Sigma; \mathbb{Z}) = \mathbb{Z}^{2g+n-1}.$$

We can pick a basis for the relative homology by selecting g a-cycles, g b-cycles (from absolute homology), and $n-1$ relative cycles, where a relative cycle is a path with starts at some point of Σ and ends at a different point of Σ.

Fix a $\mathbb{Z}$-basis $\gamma_1, \gamma_2, \ldots, \gamma_k$ of $H_1(S, \Sigma; \mathbb{Z})$, where $k = 2g + n - 1$. We recall the following fact (see [14]):

THEOREM 1. *The map* $(X, \omega) \to (\mathrm{hol}(\gamma_1), \ldots, \mathrm{hol}(\gamma_k))$ *from* $\mathcal{H}(\beta) \to (\mathbb{R}^2)^k$ *is a local coordinate system.*

By pulling back Lebesgue measure on $(\mathbb{R}^2)^k$, we obtain a normalized measure ν on $\mathcal{H}(\beta)$. (For more details on the above construction, see [3, §3].) Now, we would like to define a measure on the hypersurface $\mathcal{H}_1(\beta)$.

This is similar to the lattice setting, where if we pick a basis $v_1, v_2, \ldots, v_n$ for our lattice $\Delta \subset \mathbb{R}^n$, we get a matrix in $M_n(\mathbb{R})$ by letting v_i be the ith column. Note that since our lattice is unit volume, our matrix has determinant 1. We have a natural (Lebesgue) measure ν on $M_n(\mathbb{R})$. Consider the $\det = 1$ hypersurface Ω_1 (i.e. $SL(n, \mathbb{R})$). We define a measure μ on this space as follows: let $E \subset \Omega_1$, and let $C_1(E)$ be the cone over E (i.e. the union of all line segments which start at the origin and end at a point of E). We define $\mu(E) = \nu(C_1(E))$. This yields a finite measure since we are considering a fundamental domain under the $SL(n, \mathbb{Z})$-action. This is in fact the measure used in the previous section in the case of lattices.

Returning to the setting of translation surfaces, recall that the area of our surface $S = (X, \omega)$ is given by

$$\mathrm{area}(S) = \frac{1}{2i} \int_X \omega \wedge \bar{\omega} = \frac{1}{2i} \sum_{i=1}^{g} \int_{a_i} \bar{\omega} \int_{b_i} \omega - \int_{b_i} \bar{\omega} \int_{a_i} \omega,$$

where a_i and b_i are the a- and b-cycles on X, respectively.

This gives that the area is a quadratic form in the local coordinate system, i.e.

$$\mathrm{area}(X, \omega) = Q\big(\mathrm{hol}(\gamma_1), \ldots, \mathrm{hol}(\gamma_k)\big).$$

However, it is a degenerate form, since it only depends on the absolute cycles a_i and b_i. We can mimic the lattice picture now: we define $\mu(E) = \nu(C_1(E))$ for any subset $E \subset \mathcal{H}_1(\beta)$. This is the measure used in the previous section for the case of translation surfaces.

In what follows, we should really work inside each local coordinate chart as in Theorem 1 and then sum over the charts at the end (see [12, §3.2]). But to simplify the presentation, we pretend there is only one chart. Let $\mathcal{F} \subset \mathcal{H}_1(\beta)$ denote a fundamental domain (for the relation of equivalence of translation surfaces) with rectifiable boundary, so that each translation surface corresponds to a unique point in $\mathcal{F}$. Then,

$$\mu\big(\mathcal{H}_1(\beta)\big) = \mu(\mathcal{F}) = \nu\big(C_1(\mathcal{F})\big).$$

We now make a cosmetic step. Let $C_R(\mathcal{F})$ denote the cone of $\mathcal{F}$ extended to the hypersurface of area R-surfaces. Clearly,

$$\mu\big(\mathcal{H}_1(\beta)\big) = \nu\big(C_1(\mathcal{F})\big) = \frac{\nu(C_R(\mathcal{F}))}{R^k}.$$

We have the following fact:

$$\big|C_R(\mathcal{F}) \cap \big(\mathbb{Z}^2\big)^k\big| \sim \nu\big(C_R(\mathcal{F})\big)$$

as $R \to \infty$, i.e. the number of lattice points in a cone is asymptotic to the volume. Usually this is used to estimate the number of lattice points, but here we use this in reverse and estimate the volume by the number of lattice points. Thus, we get that

$$\mu\big(\mathcal{H}_1(\beta)\big) = \frac{\nu(C_R(\mathcal{F}))}{R^k} \sim \frac{|C_R(\mathcal{F}) \cap (\mathbb{Z}^2)^k|}{R^k},$$

or, equivalently,

$$\big|C_R(\mathcal{F}) \cap \big(\mathbb{Z}^2\big)^k\big| \sim \mu\big(\mathcal{H}_1(\beta)\big)R^k. \tag{11}$$

Equation (11) is not useful unless we can find an interpretation of the points of $C_R(\mathcal{F}) \cap (\mathbb{Z}^2)^k$. This is given by the following:

LEMMA 2. *$S = (X, \omega) \in C_R(\mathcal{F}) \cap (\mathbb{Z}^2)^k$ if and only if X is a holomorphic branched cover of the standard torus of degree $\leqslant R$, ω is the pullback of dz under the covering map, and all singularities branch over the same point.*

PROOF. Since $S \in C_R(\mathcal{F})$, $\text{area}(S) \leqslant R$. By definition, $S \in (\mathbb{Z}^2)^k$ is equivalent to $\text{hol}(\gamma_1), \ldots, \text{hol}(\gamma_k) \in \mathbb{Z}^2$. Fix a non-singular point z_0 on S, and define $\pi : S \to T$, where T is the standard torus, by $\pi(z) = \int_{z_0}^{z} \omega$. Since $\int_\gamma \omega \in \mathbb{Z} + i\mathbb{Z}$ for any closed curve or saddle connection γ, this is a well-defined covering map with all singularities branching over the same point. Since the torus is unit volume, the area of S is equal to the degree of the covering. □

Let $N_\beta(d)$ denote the number[1] of branched covers of T of degree d with branching type β. (Note that $N_\beta(d)$ is defined in purely combinatorial terms.)

Combining Lemma 2 with (11), we obtain the following: as $R \to \infty$,

$$\sum_{d=1}^{R} N_\beta(d) \sim \mu\big(\mathcal{H}_1(\beta)\big)R^k. \tag{12}$$

[1]In order for Theorem 3 below to hold, we should, when defining $N_\beta(d)$, weigh each cover by the inverse of its automorphism group. However, this does not affect the asymptotics and can be ignored here.

(This relation was discovered by Kontsevich and Zorich, and independently by Masur and the author.) Thus, we can compute $\mu(\mathcal{H}_1(\beta))$ if we can compute the asymptotics of the left-hand side of (12). This is a purely combinatorial problem.

Suppose we are considering a degree d cover of the torus. Consider the standard basis a and b of curves on the torus (when the torus is viewed as the unit square, the curves correspond to the sides of the square). They give rise to permutations of the sheets, that is, elements of the symmetric group S_d. We will abuse notation by denoting these permutations also by a and b. Singularity types of covers correspond to different conjugacy classes of the commutator $aba^{-1}b^{-1}$. A simple zero is a transposition, a double zero a three cycle, a two simple zeroes is a product of two transpositions, etc. (So, for example, if we are considering the stratum $\mathcal{H}(1,1)$, the commutator will be in the same conjugacy class as a product of two transpositions.) The number of pairs $(a,b) \in S_d \times S_d$ satisfying such a commutation relation can be expressed as a sum over the characters of the symmetric group S_d.

However, simply looking at the conjugacy class of the commutator permutation does not guarantee that the resulting surface is connected. We wish to count only the connected covers. However, the disconnected ones dominate the count. If one knows the number of disconnected covers exactly, one can compute the number of connected covers (by using inclusion/exclusion to subtract off all the possible ways a cover can disconnect). Unfortunately, as one does that, the first n terms in the asymptotic formula cancel. Still, it is possible, using the exact formula for the number of disconnected covers in [4], to carry out the computation (see [12]). The result, is a fairly messy but computable formula for the volume $\mu(\mathcal{H}_1(\beta))$.

There are two consequences of the above computations worth mentioning:

THEOREM 3. *The generating function $F_\beta(q) = \sum_{d=0}^{\infty} N_\beta(d) q^d$ is a quasi-modular form, that is, it is a polynomial in the Eisenstein series $G_k(q)$, $k = 2, 4, 6$.*

THEOREM 4. *$\pi^{-2g}\mu(\mathcal{H}_1(\beta)) \in \mathbb{Q}$, where g is the genus of any surface in $\mathcal{H}(\beta)$.*

Both of the above theorems were conjectured by Kontsevich. Further work showed that they hold also for the connected components of strata, and that similar results hold for spaces of quadratic differentials. We remark that Theorem 4 implies that the Siegel–Veech constants are rational.

For the space of lattices, one can carry out the same construction. The main difference is that one ends up counting *unbranched* covers of the standard torus T^n, or what is equivalent, sublattices of the standard lattice $\mathbb{Z}^n$. By computing the number of sublattices of $\mathbb{Z}^n$ of index at most R, and sending $R \to \infty$, it is not difficult to reproduce (9).

3. LECTURE 3: The Oppenheim conjecture and Ratner's theorem

3.1. *Counting cylinders and saddle connections*

Recall that $V_{sc}(S) = \{\text{hol}(\gamma)\colon\ \gamma \text{ is a saddle connection on } S\}$ where $S = (X, \omega)$ is a translation surface. We also define the analogous set:

$$V(S) = \{\mathrm{hol}(\gamma)\colon\ \gamma \text{is a closed geodesic on } S \text{ not passing through singularities}\}.$$

Note that any such closed geodesic is part of a cylinder (see [3, §3]), and all the closed geodesics in the cylinder have the same holonomy. Thus, $|V(S) \cap B(R)|$ is the number of cylinders on S of length at most R.

Masur proved the following:

THEOREM 5. *For all translation surfaces S in a compact set, there are constants c_1 and c_2 so that for $R \gg 1$,*

$$c_1 R^2 < \big|V(S) \cap B(R)\big| \leqslant \big|V_{sc}(S) \cap B(R)\big| < c_2 R^2.$$

The upper bound is proved in [17] and the lower bound is proved in [16]. The proof of the lower bound depends on the proof of the upper bound. Another proof of both the upper and lower bounds with explicit constants was given by Vorobets in [29] and [30]. Also see [9] for yet another proof of the upper bound, which is influenced by ideas of Margulis.

We also note that there is a dense set of directions with a closed trajectory and thus a cylinder.

The following theorem, gives asymptotic formulas for the number of saddle connections and cylinders of closed geodesics on a fixed surface. It was first proved in this form in [9], but many of the ideas came from [28], where a slightly weaker version was proved.

THEOREM 6. *For a.e. $S \in \mathcal{H}_1(\beta)$, we have*

$$\big|V_{sc}(S) \cap B(R)\big| \sim \pi b(\beta) R^2,$$

where $V_{sc}(S)$ is the collection of vectors in $\mathbb{R}^2$ given by holonomy of saddle connections on S, and $b(\beta)$ is the Siegel–Veech constant from Lecture 1 (*whose value is given by* (10)).

Similarly, for cylinders of closed geodesics, we have that there is a constant $b_1(\beta)$ so that

$$\big|V(S) \cap B(R)\big| \sim \pi b_1(\beta) R^2,$$

where $V(S)$ is the collection of vectors given by holonomy along (*imprimitive*) *closed geodesics not passing through singularities, and $b_1(\beta)$ is the associated Siegel–Veech constant.*

The following exposition will be along the lines of [9], which was heavily influenced by [28]. To simplify the notation, we only deal with the case of saddle connections. Define $g_t = \left(\begin{smallmatrix} e^t & 0 \\ 0 & e^{-t} \end{smallmatrix}\right)$ and $r_\theta = \left(\begin{smallmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{smallmatrix}\right)$. Let f be the indicator function of the trapezoid defined by the points

$$(1,1),\ (-1,1),\ (-1/2,1/2),\ (1/2,1/2).$$

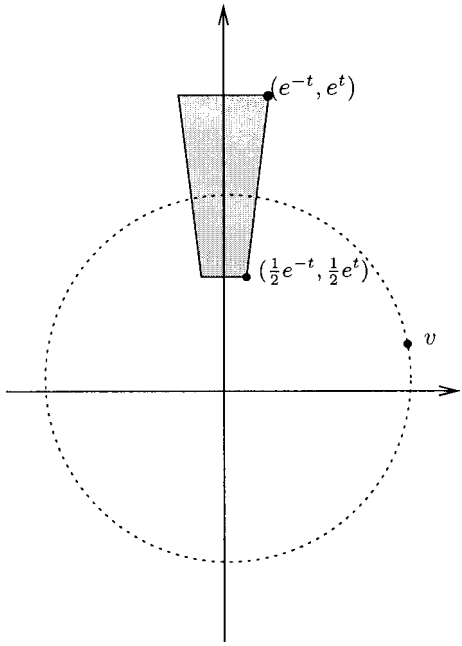


Fig. 1. Lemma 7.

LEMMA 7. *We have*

$$\int_0^{2\pi} f(g_t r_\theta v)\, d\theta \approx \begin{cases} 2e^{-2t} & \text{if } e^t/2 \leqslant \|v\| \leqslant e^t, \\ 0 & \text{otherwise.} \end{cases}$$

PROOF. Let U denote the trapezoid. Note that

$$f(g_t r_\theta v) \neq 0 \quad \Leftrightarrow \quad g_t r_\theta v \in U \quad \Leftrightarrow \quad r_\theta v \in g_t^{-1} U. \tag{13}$$

The set $g_t^{-1}U$ is the shaded region in Figure 1. From (13) it is clear that the integral in Lemma 7 is equal to (2π times) the fraction of the circle which lies inside the shaded region $g_t^{-1}U$. If v is too long or too short (not drawn), then the circle would completely miss the shaded region, and the integral would be zero. If it does not miss, then (2π times) the fraction of the circle in the shaded region is approximately $2e^{-2t}$, independent of $\|v\|$. □

We now prove Theorem 6. Summing our formula from Lemma 7 over all $v \in V_{sc}(S)$ and recalling the definition of the *Siegel–Veech transform* $\hat{f}(S) = \sum_{v \in V_{sc}(S)} f(v)$, we get

$$\tfrac{1}{2}e^{2t} \int_0^{2\pi} \hat{f}(g_t r_\theta S)\, d\theta \approx \left|V_{sc}(S) \cap B\left(e^t\right)\right| - \left|V_{sc}(S) \cap B\left(e^t/2\right)\right|.$$

Writing $R = e^t$, we can rewrite this as

$$\tfrac{1}{2}R^2 \int_0^{2\pi} \hat{f}(g_t r_\theta S)\, d\theta \approx \big|V_{sc}(S) \cap B(R)\big| - \big|V_{sc}(S) \cap B(R/2)\big|. \tag{14}$$

This equation is key to the counting problem, since the right-hand side counts saddle connections in an annulus, and the left-hand side is an integral over (part of) an $SL(2,\mathbb{R})$ orbit. (The fact that we only have approximate equality does not affect the leading order asymptotics.) Now we are supposed to use some sort of ergodic theory to analyze the behavior of integral on the left-hand side of (14) as $t \to \infty$ (or equivalently as $R \to \infty$).

There is an ergodic theorem of Nevo [19] which implies that[2] for almost all $S \in \mathcal{H}_1(\beta)$, and provided that $\hat{f} \in L^{1+\epsilon}(\mathcal{H}_1(\beta))$, the integral converges to $2\pi \int_{\mathcal{H}_1(\beta)} \hat{f}(S)\, dS = 2\pi b(\beta) \int_{\mathbb{R}^2} f$. The assertion that $\hat{f} \in L^{1+\epsilon}$ can be verified using (3). This immediately implies Theorem 6. □

However, this approach is a *failure* if one wants to prove things about billiards in rational polygons: our theorems hold for almost every point S, and the set of translation surfaces arising from rational polygons has measure zero.

3.2. *Oppenheim's conjecture*

We now describe a counting problem for lattices which has a solution very similar to the above approach. (In fact, the results in this subsection predated and heavily influenced the discussion in the previous subsection.) Let $Q = Q(x_1, x_2, \ldots, x_n)$ be a indefinite irrational quadratic form in n variables which is not a multiple of a rational form. In 1929 Oppenheim conjectured the following: for $n \geqslant 5$, $Q(\mathbb{Z}^n)$ *is dense in* $\mathbb{R}$. This was proved for $n \geqslant 3$ by Margulis in 1986 [15], using methods from dynamics and ergodic theory.

We will now assume that Q has signature (p, q), with $p \geqslant 3$ and $q \geqslant 1$. In [7], the following quantitative version of the conjecture is proved:

$$\big|\big\{x \in \mathbb{Z}^n\colon\ \|x\| \leqslant T,\ a \leqslant Q(x) \leqslant b\big\}\big| \sim c(Q)(b-a)T^{n-2}. \tag{15}$$

This is very similar to our above problem with saddle connections: we want to consider the lattice points in the ball of radius T intersected with the region in between the two hypersurfaces $Q(x) = a$ and $Q(x) = b$.

To solve this, one writes an integral very similar to the previous problem: this time, our compact group which we are integrating over is $H = SO(Q) \cap SO(n)$ and our diagonal subgroup, denoted by a_t, has 1's in every diagonal entry except the first and last, where they are e^t and e^{-t}, respectively. Our integral is as follows: $T^{n-2} \int_H \hat{f}(a_t h \Delta_Q)\, dh$, where Δ_Q is a certain lattice in $\mathbb{R}^n$ associated to Q.

[2] The theorem of Nevo used here is about a general $SL(2,\mathbb{R})$ action, and uses nothing about the geometry of the moduli space.

Hence, if one makes a formal analogy between the spaces of translation surfaces and the spaces of lattices, the problem of counting saddle connections corresponds to the quantitative Oppenheim conjecture. There is an important difference between the two problems: unlike the saddle connection case where the result is "almost everywhere", we can prove the asymptotic formula (15) for ALL quadratic forms Q not proportional to rational forms (and (15) fails for multiples of rational forms). This is due to the theorems we describe in the next part of the lecture, which are collectively known as Ratner's theorem. A major unsolved question is whether or not there is a version of Ratner's theorem for the action on the moduli space of translation surfaces. An affirmative answer would allow us to prove an asymptotic formula for billiards in every rational polygon (and every translation surface).

For more details on Oppenheim's conjecture and its solution, see [2, §3.3a, §5.1].

3.3. *Ratner's theorem*

Recall the *Birkhoff Ergodic Theorem* (see, e.g., [13, Theorem 4.1.2]):

THEOREM 8 (Birkhoff). *Let (X, μ) be a measure space with $\mu(X) = 1$, and let $T : X \to X$ be a ergodic measure preserving transformation. Let $f : X \to \mathbb{R}$ be in $L^1(X, \mu)$. Then, for almost every $x \in X$, we have that*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} f\left(T^i x\right) = \int_X f \, d\mu. \tag{16}$$

This is a great theorem, but the "almost every" is fatal for most applications to number theory. We would like to know what happens for those other points as well, and Ratner's theorem can describe the behavior in certain settings.

First, however, recall that $T : X \to X$ is said to be *uniquely ergodic* if there is a unique invariant probability measure μ on X.

We have the following consequence of unique ergodicity: if T is uniquely ergodic, and X is compact, then (assuming f in continuous) the convergence in Birkhoff's theorem holds *for all* $x \in X$. To see this, let $\nu_N(f) = \frac{1}{N} \sum_{i=0}^{N-1} f(T^i x)$. Since X is compact, the set of probability measures on X is weak-$*$ compact, so there is a subsequence ν_{n_j} and a probability measure ν_∞ so that $\nu_{n_j} \to \nu_\infty$. Its easy to see that ν_∞ is an invariant measure for T, so $\nu_\infty = \mu$. This is equivalent to (16).

Thus we can see that understanding the set of invariant measures is very important (or in particular, the set of ergodic invariant measures, since any invariant measure is a convex combination of ergodic measures). The other key issue in the topological setting is understanding the closure of orbits, and the two are related, since there will be invariant measures supported on orbit closures. This is the subject matter of Ratner's theorem (see [20–24]).

We now describe the setting. Let G be a semisimple Lie group with finite center (for example, $G = SL(n, \mathbb{R})$). Let Γ be a lattice in G (not necessarily cocompact, e.g., $\Gamma = SL(n, \mathbb{Z})$), and let U be a one parameter unipotent subgroup (for example, $u_t = \left(\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right)$). We let U act on G/Γ by left multiplication on cosets (for $n = 2$, this action is the horocycle flow).

The following theorem is stated somewhat informally, e.g., [24] or [2, §3.3c] for precise statements.

THEOREM 9 (Ratner).

(1) *The closure of every U-orbit is algebraic*: *that is, for all $x \in G/\Gamma$, there is a closed subgroup $L \subset G$ such that $\overline{Ux} = Lx$, and that $L \cap x\Gamma x^{-1}$ is a lattice in L* (*so that Lx is a closed subset of G/Γ*).
(2) *Every ergodic U-invariant measure ν is algebraic, that is there exists a subgroup L and $x \in G/\Gamma$, such that ν is the L-invariant measure on the closed subset Lx.*
(3) *Every orbit is uniformly distributed in its closure, that is, for every $x \in G$ there exists a* (*not necessarily proper*) *subgroup L of G such that $Lx = \overline{Ux}$ is closed, and $\frac{1}{T}\int_0^T f(u_t x)\,dt \to \int f(y)\,d\mu_L(y)$ as $t \to \infty$, where μ_L is the L-invariant probability measure on Lx.*

The second part of the theorem is the most difficult. The other two parts are essentially consequences of part 2. Also note that Birkhoff's theorem yields that for all $\epsilon > 0$ there is a set B of measure $< \epsilon$ so that outside of B, the convergence is uniform. Dani and Margulis obtained an explicit description of B using part 2 of Ratner's theorem (see [6]).

One eventual goal is to prove a version of Ratner's theorem for the $SL(2,\mathbb{R})$ action on $\mathcal{H}_1(\beta)$. That is, we would like to classify invariant measures, orbit closures, and prove uniform distribution, for both the full $SL(2,\mathbb{R})$ action, and for the horocycle flow (which is defined to be the action on $\mathcal{H}_1(\beta)$ of the subgroup $\left(\begin{smallmatrix} 1 & * \\ 0 & 1 \end{smallmatrix}\right)$ of $SL(2,\mathbb{R})$).

One partial result in this direction is due to McMullen [18]: he has classified the $SL(2,\mathbb{R})$ orbit closures and invariant measures for the moduli space of genus 2 surfaces (i.e. the strata $\mathcal{H}(1,1)$ and $\mathcal{H}(2)$). Note that the integral in (14) is over large circles in $SL(2,\mathbb{R})$, which can be approximated well by horocycles. Thus the horocycle flow is directly relevant to the counting problem. For other very partial results in this direction see [8] and [10], where this program (i.e. measure classification with respect to the horocycle flow and application to counting) has been carried out in the very special case of branched covers of Veech surfaces.

Acknowledgements

These notes are based on lectures of the author at Luminy, in June 2003. Thanks are given to Moon Duchin and Jayadev Athreya for taking detailed notes, and to the latter for typing them up.

References

Surveys in volume 1A and this volume

[1] G. Forni, *On the Lyapunov exponents of the Kontsevich–Zorich cocycle*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 549–580.

[2] D. Kleinbock, N. Shah and A. Starkov, *Homogeneous flows, applications to number theory, and related topics*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 1015–1089.
[3] H. Masur, *Ergodic theory of translation surfaces*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 527–547.

Other sources

[4] S. Bloch and A. Okounkov, *The character of the infinite wedge representation*, Adv. Math. **149** (1) (2000), 1–60.
[5] J.W.S. Cassels, *An Introduction to the Geometry of Neumbers*, Springer (1959).
[6] S.G. Dani and G.A. Margulis, *Limit distributions of orbits of unipotent flows and values of quadratic forms*, Adv. Soviet Math. **16** (1993), 91–137.
[7] A. Eskin, G. Margulis and S. Mozes, *Upper bounds and asymptotics in a quantitative version of the Oppenheim conjecture*, Ann. of Math. **147** (1998), 93–141.
[8] A. Eskin, J. Marklof and D. Witte Morris, *Unipotent flows and branched covers of Veech surfaces*, Eprint, arXiv:math.DS/0408090, Ergodic Theory Dynam. Systems, to appear.
[9] A. Eskin and H. Masur, *Asymptotic formulas on flat surfaces*, Ergodic Theory Dynam. Systems **21** (2001), 443–478.
[10] A. Eskin, H. Masur and M. Schmoll, *Billiards in rectangles with barriers*, Duke Math. J. **118** (3) (2003), 427–463.
[11] A. Eskin, H. Masur and A. Zorich, *Moduli spaces of Abelian differentials: The principal boundary, counting problems and the Siegel–Veech constants*, Publ. Math. Inst. Hautes Etudes Sci. **97** (2003), 61–179.
[12] A. Eskin and A. Okounkov, *Asymptotics of numbers of branched covers of a torus and volumes of moduli spaces of holomorphic differentials*, Invent. Math. **145** (1) (2001), 59–104.
[13] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems (paperback)*, Cambridge Univ. Press (1996), ISBN: 0-521-57557-5.
[14] M. Kontsevich, *Lyapunov Exponents and Hodge Theory*, The Mathematical Beauty of Physics (Saclay, 1996), Adv. Ser. Math. Phys., Vol. 24, World Scientific, River Edge, NJ (1997), 318–332.
[15] G.A. Margulis, *Indefinite quadratic forms and unipotent flows on homogeneous spaces*, Dynam. Systems Ergodic Theory, Vol. 23, Banach Center Publ., PWN, Warsaw (1989), 399–409.
[16] H. Masur, *Lower bounds for the number of saddle connections and closed trajectories of a quadratic differential*, Holomorphic Functions and Moduli, Vol. 1, D. Drasin, ed., Springer (1988), 215–228.
[17] H. Masur, *The growth rate of trajectories of a quadratic differential*, Ergodic Theory Dynam. Systems **10** (1990), 151–176.
[18] C. McMullen, *Dynamics of $SL_2(\mathbb{R})$ actions in genus* 2, Preprint.
[19] A. Nevo, *Equidistribution in measure preserving actions of semisimple groups*, Preprint.
[20] M. Ratner, *Strict measure rigidity for nilpotent subgroups of solvable groups*, Invent. Math. **101** (1990), 449–482.
[21] M. Ratner, *On measure rigidity of unipotent subgroups of semisimple groups*, Acta Math. **165** (1990), 229–309.
[22] M. Ratner, *On Raghunathan's measure conjecture*, Ann. of Math. **134** (1991), 545–607.
[23] M. Ratner, *Raghunathan's topological conjecture and distributions of unipotent flows*, Duke Math. J. **63** (1991), 235–290.
[24] M. Ratner *Interactions between Lie groups, ergodic theory and number theory*, Proc. of ICM, Zurich (1994).
[25] W. Schmidt, *Asymptotic formulae for point lattices of bounded determinant and subspaces of bounded height*, Duke Math. J. **35** (1968), 327–339.
[26] C.L. Siegel, *Lectures on the Geometry of Numbers*, Springer (1989).
[27] A. Terras, *Harmonic Analysis on Symmetric Spaces and Applications II*, Springer (1988).
[28] W. Veech, *Siegel measures*, Ann. of Math. **148** (1998), 895–944.
[29] Y. Vorobets, *Ergodicity of billiards in polygons*, Mat. Sb. **188** (3) (1997), 65–112.
[30] Y. Vorobets, *Periodic geodesics on translation surfaces*, Eprint, arXiv:math.DS/0307249.

This page intentionally left blank

CHAPTER 10

On the Interplay between Measurable and Topological Dynamics

E. Glasner

Department of Mathematics, Tel Aviv University, Tel Aviv, Israel
E-mail: *glasner@math.tau.ac.il*

B. Weiss

Institute of Mathematics, Hebrew University of Jerusalem, Jerusalem, Israel
E-mail: *weiss@math.huji.ac.il*

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

Introduction

Recurrent–wandering, conservative–dissipative, contracting–expanding, deterministic–chaotic, isometric–mixing, periodic–turbulent, distal–proximal, the list can go on and on. These (pairs of) words—all of which can be found in the dictionary—convey dynamical images and were therefore adopted by mathematicians to denote one or another mathematical aspect of a dynamical system.

The two sister branches of the theory of dynamical systems called *ergodic theory* (or *measurable dynamics*) and *topological dynamics* use these words to describe different but parallel notions in their respective theories and the surprising fact is that many of the corresponding results are rather similar. In the following chapter we have tried to demonstrate both the parallelism and the discord between ergodic theory and topological dynamics. We hope that the subjects we chose to deal with will successfully demonstrate this duality.

The table of contents gives a detailed listing of the topics covered. In the first part we have detailed the strong analogies between ergodic theory and topological dynamics as shown in the treatment of recurrence phenomena, equicontinuity and weak mixing, distality and entropy. In the case of distality the topological version came first and the theory of measurable distality was strongly influenced by the topological results. For entropy theory the influence clearly was in the opposite direction. The prototypical result of the second part is the statement that any abstract measure probability preserving system can be represented as a continuous transformation of a compact space, and thus in some sense ergodic theory embeds into topological dynamics.

We have not attempted in any way to be either systematic or comprehensive. Rather our choice of subjects was motivated by taste, interest and knowledge and to great extent is random. We did try to make the survey accessible to non-specialists, and for this reason we deal throughout with the simplest case of actions of $\mathbb{Z}$. Most of the discussion carries over to non-invertible mappings and to $\mathbb{R}$ actions. Indeed much of what we describe can be carried over to general amenable groups. Similarly, we have for the most part given rather complete definitions. Nonetheless, we did take advantage of the fact that this chapter is part of a handbook and for some of the definitions, basic notions and well known results we refer the reader to volume I of this handbook, mainly to Chapters 1, by B. Hasselblatt and A. Katok, and 2, by J.-P. Thouvenot. Finally, we should acknowledge the fact that we made use of parts of our previous expositions [87] and [36].

We made the writing of this survey more pleasurable for us by the introduction of a few original results. In particular the following results are entirely or partially new. Theorem 1.2 (the equivalence of the existence of a Borel cross-section with the coincidence of recurrence and periodicity), most of the material in Section 4 (on topological mild-mixing), all of Subsection 7.4 (the converse side of the local variational principle) and Subsection 7.6 (on topological determinism).

Part 1. Analogies

1. Poincaré recurrence vs. Birkhoff's recurrence

1.1. *Poincaré recurrence theorem and topological recurrence*

The simplest dynamical systems are the periodic ones. In the absence of periodicity the crudest approximation to this is approximate periodicity where instead of some iterate $T^n x$ returning exactly to x it returns to a neighborhood of x. We refer the reader to [1, Chapter 1, Section 2.3], in the first volume of this handbook, for a short review of recurrence in topological dynamics.

The first theorem in abstract measure dynamics is Poincaré's recurrence theorem which asserts that for a finite measure preserving system $(X, \mathcal{B}, \mu, T)$ and any measurable set A, μ-a.e. point of A returns to A (see [1, Chapter 1, Theorem 3.4.1]). The proof of this basic fact is rather simple and depends on identifying the set of points $W \subset A$ that never return to A. These are called the *wandering points* and their measurability follows from the formula

$$W = A \cap \left(\bigcap_{k=1}^{\infty} T^{-k}(X \setminus A) \right).$$

Now for $n \geqslant 0$, the sets $T^{-n}W$ are pairwise disjoint since $x \in T^{-n}W$ means that the forward orbit of x visits A for the last time at moment n. Since $\mu(T^{-n}W) = \mu(W)$ it follows that $\mu(W) = 0$ which is the assertion of Poincaré's theorem. Noting that $A \cap T^{-n}W$ describes the points of A which visit A for the last time at moment n, and that $\mu(\bigcup_{n=0}^{\infty} T^{-n}W) = 0$ we have established the following stronger formulation of Poincaré's theorem.

THEOREM 1.1. *For a finite measure preserving system $(X, \mathcal{B}, \mu, T)$ and any measurable set A, μ-a.e. point of A returns to A infinitely often.*

Note that only sets of the form $T^{-n}B$ appeared in the above discussion so that the invertibility of T is not needed for this result. In the situation of classical dynamics, which was Poincaré's main interest, X is also equipped with a separable metric topology. In such a situation we can apply the theorem to a refining sequence of partitions $\mathcal{P}_m$, where each $\mathcal{P}_m$ is a countable partition into sets of diameter at most $1/m$. Applying the theorem to a fixed $\mathcal{P}_m$ we see that μ-a.e. point comes to within $1/m$ of itself, and since the intersection of a sequence of sets of full measure has full measure, we deduce the corollary that μ-a.e. point of X is recurrent.

This is the measure theoretical path to the recurrence phenomenon which depends on the presence of a finite invariant measure. The necessity of such measure is clear from considering translation by one on the integers. The system is dissipative, in the sense that no recurrence takes place even though there is an infinite invariant measure.

There is also a topological path to recurrence which was developed in an abstract setting by G.D. Birkhoff. Here the above example is eliminated by requiring that the topological

space X, on which our continuous transformation T acts, be compact. It is possible to show that in this setting a finite T-invariant measure always exists, and so we can retrieve the measure theoretical picture, but a purely topological discussion will give us better insight.

A key notion here is that of minimality. A non-empty closed, T-invariant set $E \subset X$, is said to be *minimal* if $F \subset E$, closed and T-invariant implies $F = \emptyset$ or $F = E$. If X itself is a minimal set we say that the system (X, T) is a *minimal system.*

Fix now a point $x_0 \in X$ and consider

$$\omega(x_0) = \bigcap_{n=1}^{\infty} \overline{\{T^k x_0 \colon k \geqslant n\}}.$$

The points of $\omega(x_0)$ are called *ω-limit points of* x_0 (ω = last letter of the Greek alphabet) and in the separable case $y \in \omega(x_0)$ if and only if there is some sequence $k_i \to \infty$ such that $T^{k_i} x_0 \to y$. If $x_0 \in \omega(x_0)$ then x_0 is called a *positively recurrent point.*

Clearly $\omega(x_0)$ is a closed and T-invariant set. Therefore, in any non-empty minimal set E, any point $x_0 \in E$ satisfies $x_0 \in \omega(x_0)$ and thus we see that minimal sets have recurrent points.

In order to see that compact systems (X, T) have recurrent points it remains to show that minimal sets always exist. This is an immediate consequence of Zorn's lemma applied to the family of nonempty closed T-invariant subsets of X. A slightly more constructive proof can be given when X is a compact and separable metric space. One can then list a sequence of open sets $U_1, U_2, \ldots$ which generate the topology, and perform the following algorithm:

1. set $X_0 = X$,
2. for $i = 1, 2, \ldots$,
   if $\bigcup_{n=-\infty}^{\infty} T^{-n} U_i \supset X_{i-1}$ put $X_i = X_{i-1}$, else put $X_i = X_{i-1} \setminus \bigcup_{n=-\infty}^{\infty} T^{-n} U_i$.

Note that $X_i \neq \emptyset$ and closed and thus $X_\infty = \bigcap_{i=0}^{\infty} X_i$ is non-empty. It is clearly T-invariant and for any U_i, if $U_i \cap X_\infty \neq \emptyset$ then $\bigcup_{-\infty}^{\infty} T^{-n}(U_i \cap X_\infty) = X_\infty$, which shows that (X_∞, T) is minimal.

1.2. *The existence of Borel cross-sections*

There is a deep connection between recurrent points in the topological context and ergodic theory. To see this we must consider quasi-invariant measures. For these matters it is better to enlarge the scope and deal with continuous actions of $\mathbb{Z}$, generated by T, on a *complete separable metric space* X. A probability measure μ defined on the Borel subsets of X is said to be *quasi-invariant* if $T \cdot \mu \sim \mu$. Define such a system $(X, \mathcal{B}, \mu, T)$ to be *conservative* if for any measurable set A, $TA \subset A$ implies $\mu(A \setminus TA) = 0$.

It is not hard to see that the conclusion of Poincaré's recurrence theorem holds for such systems; i.e. if $\mu(A) > 0$, then μ-a.e. x returns to A infinitely often. Thus once again μ-a.e. point is topologically recurrent. It turns out now that the existence of a single topologically recurrent point implies the existence of a non-atomic conservative quasi-invariant measure. A simple proof of this fact can be found in [57] for the case when X is compact—but the

proof given there is equally valid for complete separable metric spaces. In this sense the phenomenon of Poincaré recurrence and topological recurrence are "equivalent" with each implying the other.

A Borel set $B \subset X$ such that each orbit intersects B in exactly one point is called a *Borel cross-section* for the system (X, T). If a Borel cross-section exists, then no non-atomic conservative quasi-invariant measure can exist. In [83] it is shown that the converse is also valid—namely if there are no conservative quasi-invariant measures then there is a Borel cross-section.

Note that the periodic points of (X, T) form a Borel subset for which a cross-section always exists, so that we can conclude from the above discussion the following statement in which no explicit mention is made of measures.

THEOREM 1.2. *For a system (X, T), with X a completely metrizable separable space, there exists a Borel cross-section if and only if the only recurrent points are the periodic ones.*

REMARK 1.3. Already in [44] as well as in [22] one finds many equivalent conditions for the existence of a Borel section for a system (X, T). However one doesn't find there explicit mention of conditions in terms of recurrence. Silvestrov and Tomiyama [77] established the theorem in this formulation for X compact (using C^*-algebra methods). We thank A. Lazar for drawing our attention to their paper.

1.3. *Recurrence sequences and Poincaré sequences*

We will conclude this section with a discussion of recurrence sequences and Poincaré sequences. First for some definitions. Let us say that D is a *recurrence set* if for any dynamical system (Y, T) with compatible metric ρ and any $\varepsilon > 0$ there is a point y_0 and a $d \in D$ with

$$\rho\big(T^d y_0,\ y_0\big) < \varepsilon.$$

Since any system contains minimal sets it suffices to restrict attention here to minimal systems. For minimal systems the set of such y's for a fixed ε is a dense open set.

To see this fact, let U be an open set. By the minimality there is some N such that for any $y \in Y$, and some $0 \leqslant n \leqslant N$, we have $T^n y \in U$. Using the uniform continuity of T^n, we find now a $\delta > 0$ such that if $\rho(u,\ v) < \delta$ then for all $0 \leqslant n \leqslant N$

$$\rho(T^n u,\ T^n v) < \varepsilon.$$

Now let z_0 be a point in Y and $d_0 \in D$ such that

$$p\big(T^{d_0} z_0,\ z_0\big) < \delta. \tag{1}$$

For some $0 \leqslant n_0 \leqslant N$ we have $T^{n_0} z_0 = y_0 \in U$ and from (1) we get $\rho(T^{d_0} y_0,\ y_0) < \varepsilon$. Thus points that ε return form an open dense set. Intersecting over $\varepsilon \to 0$ gives a dense G_δ in Y of points y for which

$$\inf_{d \in D} \rho\big(T^d y,\ y\big) = 0.$$

Thus there are points which actually recur along times drawn from the given recurrence set.

A nice example of a recurrence set is the set of squares. To see this it is easier to prove a stronger property which is the analogue in ergodic theory of recurrence sets.

DEFINITION 1.4. A sequence $\{s_j\}$ is said to be a *Poincaré sequence* if for any finite measure preserving system $(X,\ \mathcal{B},\ \mu,\ T)$ and any $B \in \mathcal{B}$ with positive measure we have

$$\mu\big(T^{s_j} B \cap B\big) > 0 \quad \text{for some } s_j \text{ in the sequence.}$$

Since any minimal topological system (Y, T) has finite invariant measures with global support, μ any Poincaré sequence is recurrence sequence. Indeed for any presumptive constant $b > 0$ which would witness the non-recurrence of $\{s_j\}$ for (Y, T), there would have to be an open set B with diameter less than b and having positive μ-measure such that $T^{s_j} B \cap B$ is empty for all $\{s_j\}$.

Here is a sufficient condition for a sequence to be a Poincaré sequence:

LEMMA 1.5. *If for every* $\alpha \in (0,\ 2\pi)$

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} e^{i\alpha s_k} = 0$$

then $\{s_k\}_1^\infty$ *is a Poincaré sequence.*

PROOF. Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and let U be the unitary operator defined on $L^2(X, \mathcal{B}, \mu)$ by the action of T, i.e.

$$(Uf)(x) = f(Tx).$$

Let H_0 denote the subspace of invariant functions and for a set of positive measure B, let f_0 be the projection of 1_B on the invariant functions. Since this can also be seen as a conditional expectation with respect to the σ-algebra of invariant sets $f_0 \geqslant 0$ and is not zero. Now since $\mathbf{1}_B - f_0$ is orthogonal to the space of invariant functions its spectral measure with respect to U doesn't have any atoms at $\{0\}$. Thus from the spectral representation we deduce that in L^2-norm

$$\left\| \frac{1}{n} \sum_{1}^{n} U^{s_k} (\mathbf{1}_B - f_0) \right\|_{L^2} \longrightarrow 0$$

or

$$\left\| \left(\frac{1}{n} \sum_1^n U^{s_k} \mathbf{1}_B \right) - f_0 \right\|_{L_2} \longrightarrow 0$$

and integrating against $\mathbf{1}_B$ and using the fact that f_0 is the projection of $\mathbf{1}_B$ we see that

$$\lim_{n\to\infty} \frac{1}{n} \sum_1^n \mu\big(B \cap T^{-s_k} B\big) = \| f_0 \|^2 > 0$$

which clearly implies that $\{s_k\}$ is a Poincaré sequence. □

The proof we have just given is in fact von Neumann's original proof for the mean ergodic theorem. He used the fact that $\mathbb{N}$ satisfies the assumptions of the proposition, which is Weyl's famous theorem on the equidistribution of $\{n\alpha\}$. Returning to the squares Weyl also showed that $\{n^2\alpha\}$ is equidistributed for all irrational α. For rational α the exponential sum in the lemma needn't vanish, however the recurrence along squares for the rational part of the spectrum is easily verified directly so that we can conclude that indeed the squares are a Poincaré sequence and hence a recurrence sequence.

The converse is not always true, i.e. there are recurrence sequences that are not Poincaré sequences. This was first shown by I. Kriz [61] in a beautiful example (see also [87, Chapter 5]). Finally here is a simple problem.

PROBLEM. If D is a recurrence sequence for all circle rotations is it a recurrence set?

A little bit of evidence for a positive answer to that problem comes from looking at a slightly different characterization of recurrence sets. Let $\mathcal{N}$ denote the collection of sets of the form

$$N(U,\ U) = \{n\colon\ T^{-n}U \cap U \neq \emptyset\} \quad (U \text{ open and non-empty}),$$

where T is a minimal transformation. Denote by $\mathcal{N}^*$ the subsets of $\mathbb{N}$ that have a non-empty intersection with every element of $\mathcal{N}$. Then $\mathcal{N}^*$ is exactly the class of recurrence sets. For minimal transformations, another description of $N(U,\ U)$ is obtained by fixing some y_0 and denoting

$$N(y_0,\ U) = \big\{n\colon\ T^n y_0 \in U\big\}.$$

Then $N(U,\ U) = N(y_0,\ U) - N(y_0,\ U)$. Notice that the minimality of T implies that $N(y_0,\ U)$ is a *syndetic* set (a set with bounded gaps) and so any $N(U,\ U)$ is the set of differences of a syndetic set. Thus $\mathcal{N}$ consists essentially of all sets of the form $S - S$ where S is a syndetic set.

Given a finite set of real numbers $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ and $\varepsilon > 0$ set

$$V(\lambda_1, \lambda_2, \ldots, \lambda_k; \varepsilon) = \Big\{n \in \mathbb{Z}\colon\ \max_j \big\{\|n\lambda_j\| < \varepsilon\big\}\Big\},$$

where $\|\cdot\|$ denotes the distance to the closest integer. The collection of such sets forms a basis of neighborhoods at zero for a topology on $\mathbb{Z}$ which makes it a topological group. This topology is called the *Bohr topology*. (The corresponding uniform structure is totally bounded and the completion of $\mathbb{Z}$ with respect to it is a compact topological group called the *Bohr compactification* of $\mathbb{Z}$.)

Veech proved in [79] that any set of the form $S - S$ with $S \subset \mathbb{Z}$ syndetic contains a neighborhood of zero in the Bohr topology *up to a set of zero density*. It is not known if in that statement the zero density set can be omitted. If it could then a positive answer to the above problem would follow (see also [33]).

2. The equivalence of weak mixing and continuous spectrum

In order to analyze the structure of a dynamical system $\mathbf{X}$ there are, a priori, two possible approaches. In the first approach one considers the collection of *subsystems* $Y \subset X$ (i.e. closed T-invariant subsets) and tries to understand how X is built up by these subsystems. In the other approach one is interested in the collection of *factors* $X \xrightarrow{\pi} Y$ of the system $\mathbf{X}$. In the measure theoretical case the first approach leads to the ergodic decomposition and thereby to the study of the "indecomposable" or ergodic components of the system. In the topological setup there is, unfortunately, no such convenient decomposition describing the system in terms of its indecomposable parts and one has to use some less satisfactory substitutes. Natural candidates for indecomposable components of a topological dynamical system are the "orbit closures" (i.e. the topologically transitive subsystems) or the "prolongation" cells (which often coincide with the orbit closures), see [5]. The minimal subsystems are of particular importance here. Although we can not say, in any reasonable sense, that the study of the general system can be reduced to that of its minimal components, the analysis of the minimal systems is nevertheless an important step towards a better understanding of the general system.

This reasoning leads us to the study of the collection of indecomposable systems (ergodic systems in the measure category and transitive or minimal systems in the topological case) and their factors. The simplest and best understood indecomposable dynamical systems are the ergodic translations of a compact monothetic group (a cyclic permutation on $\mathbb{Z}_p$ for a prime number p, the "adding machine" on $\prod_{n=0}^{\infty} \mathbb{Z}_2$, an irrational rotation $z \mapsto e^{2\pi i\alpha} z$ on $S^1 = \{z \in \mathbb{C}\colon |z| = 1\}$ etc.). It is not hard to show that this class of ergodic actions is characterized as those dynamical systems which admit a model $(X, \mathcal{X}, \mu, T)$ where X is a compact metric space, $T : X \to X$ a surjective isometry and μ is T-ergodic. We call these systems *Kronecker* or *isometric* systems. Thus our first question concerning the existence of factors should be: given an ergodic dynamical system $\mathbf{X}$ which are its Kronecker factors? Recall that a measure dynamical system $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ is called *weakly mixing* if the product system $(X \times X, \mathcal{X} \otimes \mathcal{X}, \mu \times \mu, T \times T)$ is ergodic. The following classical theorem is due to von Neumann. The short and elegant proof we give was suggested by Y. Katznelson.

THEOREM 2.1. *An ergodic system* $\mathbf{X}$ *is weakly mixing iff it admits no non-trivial Kronecker factor.*

PROOF. Suppose $\mathbf{X}$ is weakly mixing and admits an isometric factor. Now a factor of a weakly mixing system is also weakly mixing and the only system which is both isometric and weakly mixing is the trivial system (an easy exercise). Thus a weakly mixing system does not admit a non-trivial Kronecker factor.

For the other direction, if $\mathbf{X}$ is non-weakly mixing then in the product space $X \times X$ there exists a T-invariant measurable subset W such that $0 < (\mu \times \mu)(W) < 1$. For every $x \in X$ let $W(x) = \{x' \in X\colon\ (x, x') \in W\}$ and let $f_x = \mathbf{1}_{W(x)}$, a function in $L^\infty(\mu)$. It is easy to check that $U_T f_x = f_{T^{-1}x}$ so that the map $\pi : X \to L^2(\mu)$ defined by $\pi(x) = f_x$, $x \in X$, is a Borel factor map. Denoting

$$\pi(X) = Y \subset L^2(\mu) \quad \text{and} \quad \nu = \pi_*(\mu),$$

we now have a factor map $\pi : \mathbf{X} \to (Y, \nu)$. Now the function $\|\pi(x)\|$ is clearly measurable and invariant and by ergodicity it is a constant μ-a.e.; say $\|\pi(x)\| = 1$. The dynamical system (Y, ν) is thus a subsystem of the compact dynamical system (B, U_T), where B is the unit ball of the Hilbert space $L^2(\mu)$ and U_T is the Koopman unitary operator induced by T on $L^2(\mu)$. Now it is well known (see, e.g., [36]) that a compact topologically transitive subsystem which carries an invariant probability measure must be a Kronecker system and our proof is complete. □

Concerning the terminology we used in the proof of Theorem 2.1, B.O. Koopman, a student of G.D. Birkhoff and a co-author of both Birkhoff and von Neumann, introduced the crucial idea of associating with a measure dynamical system $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ the unitary operator U_T on the Hilbert space $L^2(\mu)$. It is now an easy matter to see that Theorem 2.1 can be re-formulated as saying that the system $\mathbf{X}$ is weakly mixing iff the point spectrum of the Koopman operator U_T comprises the single complex number 1 with multiplicity 1. Or, put otherwise, that the one-dimensional space of constant functions is the eigenspace corresponding to the eigenvalue 1 (this fact alone is equivalent to the ergodicity of the dynamical system) and that the restriction of U_T to the orthogonal complement of the space of constant functions has a continuous spectrum.

We now consider a topological analogue of this theorem. Recall that a topological system (X, T) is *topologically weakly mixing* when the product system $(X \times X, T \times T)$ is topologically transitive. It is *equicontinuous* when the family $\{T^n\colon\ n \in \mathbb{Z}\}$ is an equicontinuous family of maps. Again an equivalent condition is the existence of a compatible metric with respect to which T is an isometry. And, moreover, a minimal system is equicontinuous iff it is a minimal translation on a compact monothetic group. We will need the following lemma.

LEMMA 2.2. *Let (X, T) be a minimal system and $f : X \to \mathbb{R}$ a T-invariant function with at least one point of continuity* (*for example this is the case when f is lower or upper semi-continuous or more generally when it is the pointwise limit of a sequence of continuous functions*), *then f is a constant.*

PROOF. Let x_0 be a continuity point and x an arbitrary point in X. Since $\{T^n x\colon\ n \in \mathbb{Z}\}$ is dense and as the value $f(T^n x)$ does not depend on n it follows that $f(x) = f(x_0)$. □

THEOREM 2.3. *Let (X, T) be a minimal system then (X, T) is topologically weakly mixing iff it has no non-trivial equicontinuous factor.*

PROOF. Suppose (X, T) is minimal and topologically weakly mixing and let $\pi : (X, T) \to (Y, T)$ be an equicontinuous factor. If (x, x') is a point whose $T \times T$ orbit is dense in $X \times X$ then $(y, y') = (\pi(x), \pi(x'))$ has a dense orbit in $Y \times Y$. However, if (Y, T) is equicontinuous then Y admits a compatible metric with respect to which T is an isometry and the existence of a transitive point in $Y \times Y$ implies that Y is a trivial one-point space.

Conversely, assuming that $(X \times X, T \times T)$ is not transitive we will construct an equicontinuous factor (Z, T) of (X, T). As (X, T) is a minimal system, there exists a T-invariant probability measure μ on X with full support. By assumption there exists an open T-invariant subset U of $X \times X$, such that $\operatorname{cls} U := M \subsetneq X \times X$. By minimality the projections of M to both X coordinates are onto. For every $y \in X$ let $M(y) = \{x \in X: (x, y) \in M\}$, and let $f_y = \mathbf{1}_{M(y)}$ be the indicator function of the set $M(y)$, considered as an element of $L^1(X, \mu)$.

Denote by $\pi : X \to L^1(X, \mu)$ the map $y \mapsto f_y$. We will show that π is a continuous homomorphism, where we consider $L^1(X, \mu)$ as a dynamical system with the isometric action of the group $\{U_T^n: n \in \mathbb{Z}\}$ and $U_T f(x) = f(Tx)$. Fix $y_0 \in X$ and $\varepsilon > 0$. There exists an open neighborhood V of the closed set $M(y_0)$ with $\mu(V \setminus M(y_0)) < \varepsilon$. Since M is closed the set map $y \mapsto M(y)$, $X \to 2^X$ is upper semi-continuous and we can find a neighborhood W of y_0 such that $M(y) \subset V$ for every $y \in W$. Thus for every $y \in W$ we have $\mu(M(y) \setminus M(y_0)) < \varepsilon$. In particular, $\mu(M(y)) \leqslant \mu(M(y_0)) + \varepsilon$ and it follows that the map $y \mapsto \mu(M(y))$ is upper semi-continuous. A simple computation shows that it is T-invariant, hence, by Lemma 2.2, a constant.

With y_0, ε and V, W as above, for every $y \in W$, $\mu(M(y) \setminus M(y_0)) < \varepsilon$ and $\mu(M(y)) = \mu(M(y_0))$, thus $\mu(M(y) \Delta M(y_0)) < 2\varepsilon$, i.e. $\|f_y - f_{y_0}\|_1 < 2\varepsilon$. This proves the claim that π is continuous.

Let $Z = \pi(X)$ be the image of X in $L^1(\mu)$. Since π is continuous, Z is compact. It is easy to see that the T-invariance of M implies that for every $n \in \mathbb{Z}$ and $y \in X$, $f_{T^{-n}y} = f_y \circ T^n$ so that Z is U_T-invariant and $\pi : (Y, T) \to (Z, U_T)$ is a homomorphism. Clearly (Z, U_T) is minimal and equicontinuous (in fact isometric). □

Theorem 2.3 is due to Keynes and Robertson [58] who developed an idea of Furstenberg [23]; and independently to K. Petersen [71] who utilized a previous work of W.A. Veech [79]. The proof we presented is an elaboration of a work of McMahon [67] due to Blanchard, Host and Maass [13]. We take this opportunity to point out a curious phenomenon which recurs again and again. Some problems in topological dynamics—like the one we just discussed—whose formulation is purely topological, can be solved using the fact that a $\mathbb{Z}$ dynamical system always carries an invariant probability measure, and then employing a machinery provided by ergodic theory. In several cases this approach is the only one presently known for solving the problem. In the present case however purely topological proofs exist, e.g., the Petersen–Veech proof is one such.

3. Disjointness: measure vs. topological

In the ring of integers $\mathbb{Z}$ two integers m and n have no common factor if whenever $k|m$ and $k|n$ then $k = \pm 1$. They are disjoint if $m \cdot n$ is the least common multiple of m and n. Of course in $\mathbb{Z}$ these two notions coincide. In his seminal paper of 1967 [24], H. Furstenberg introduced the same notions in the context of dynamical systems, both measure-preserving transformations and homeomorphisms of compact spaces, and asked whether in these categories as well the two are equivalent. The notion of a factor in, say the measure category, is the natural one: the dynamical system $\mathbf{Y} = (Y, \mathcal{Y}, \nu, T)$ is a *factor* of the dynamical system $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ if there exists a measurable map $\pi : X \to Y$ with $\pi(\mu) = \nu$ that $T \circ \pi = \pi \circ T$. A common factor of two systems $\mathbf{X}$ and $\mathbf{Y}$ is thus a third system $\mathbf{Z}$ which is a factor of both. A *joining* of the two systems $\mathbf{X}$ and $\mathbf{Y}$ is any system $\mathbf{W}$ which admits both as factors and is in turn spanned by them. According to Furstenberg's definition the systems $\mathbf{X}$ and $\mathbf{Y}$ are *disjoint* if the product system $\mathbf{X} \times \mathbf{Y}$ is the only joining they admit. In the topological category, a joining of (X, T) and (Y, S) is any subsystem $W \subset X \times Y$ of the product system $(X \times Y, T \times S)$ whose projections on both coordinates are full; i.e. $\pi_X(W) = X$ and $\pi_Y(W) = Y$. (X, T) and (Y, S) are *disjoint* if $X \times Y$ is the unique joining of these two systems. It is easy to verify that if (X, T) and (Y, S) are disjoint then at least one of them is minimal. Also, if both systems are minimal then they are disjoint iff the product system $(X \times Y, T \times S)$ is minimal.

In 1979, D. Rudolph, using joining techniques, provided the first example of a pair of ergodic measure preserving transformations with no common factor which are not disjoint [73]. In this work Rudolph laid the foundation of joining theory. He introduced the class of dynamical systems having "minimal self-joinings" (MSJ), and constructed a rank one mixing dynamical system having minimal self-joinings of all orders.

Given a dynamical system $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ a probability measure λ on the product of k copies of X denoted $X_1, X_2, \ldots, X_k$, invariant under the product transformation and projecting onto μ in each coordinate is a *k-fold self-joining*. It is called an *off-diagonal* if it is a "graph" measure of the form $\lambda = \mathrm{gr}(\mu, T^{n_1}, \ldots, T^{n_k})$, i.e. λ is the image of μ under the map $x \mapsto (T^{n_1}x, T^{n_2}x, \ldots, T^{n_k}x)$ of X into $\prod_{i=1}^{k} X_i$. The joining λ is a *product of off-diagonals* if there exists a partition $(J_1, \ldots, J_m)$ of $\{1, \ldots, k\}$ such that (i) for each l, the projection of λ on $\prod_{i \in J_l} X_i$ is an off-diagonal, (ii) the systems $\prod_{i \in J_l} X_i$, $1 \leqslant l \leqslant m$, are independent. An ergodic system $\mathbf{X}$ has *minimal self-joinings of order k* if every k-fold ergodic self-joining of $\mathbf{X}$ is a product of off-diagonals.

In [73] Rudolph shows how any dynamical system with MSJ can be used to construct a counter example to Furstenberg's question as well as a wealth of other counter examples to various questions in ergodic theory. In [53] del Junco, Rahe and Swanson were able to show that the classical example of Chacón [17] has MSJ, answering a question of Rudolph whether a weakly but not strongly mixing system with MSJ exists. In [39] Glasner and Weiss provide a topological counterexample, which also serves as a natural counterexample in the measure category. The example consists of two horocycle flows which have no non-trivial common factor but are nevertheless not disjoint. It is based on deep results of Ratner [72] which provide a complete description of the self joinings of a horocycle flow. More recently an even more striking example was given in the topological category by

E. Lindenstrauss, where two minimal dynamical systems with no nontrivial factor share a common almost 1-1 extension [64].

Beginning with the pioneering works of Furstenberg and Rudolph, the notion of joinings was exploited by many authors; Furstenberg (1977) [25], Rudolph (1979) [73], Veech (1982) [82], Ratner (1983) [72], del Junco and Rudolph (1987) [54], Host (1991) [48], King (1992) [59], Glasner, Host and Rudolph (1992) [37], Thouvenot (1993) [78], Ryzhikov (1994) [74], Kammeyer and Rudolph (1995) (2002) [56], del Junco, Lemańczyk and Mentzen (1995) [52], and Lemańczyk, Parreau and Thouvenot (2000) [63], to mention a few. The negative answer to Furstenberg's question and the consequent works on joinings and disjointness show that in order to study the relationship between two dynamical systems it is necessary to know all the possible joinings of the two systems and to understand the nature of these joinings.

Some of the best known disjointness relations among families of dynamical systems are the following:

- id $\perp$ ergodic,
- distal $\perp$ weakly mixing ([24]),
- rigid $\perp$ mild mixing ([28]),
- zero entropy $\perp$ K-systems ([24]),

in the measure category and

- F-systems $\perp$ minimal ([24]),
- minimal distal $\perp$ weakly mixing,
- minimal zero entropy $\perp$ minimal UPE-systems ([10]),

in the topological category.

4. Mild mixing: measure vs. topological

DEFINITION 4.1. Let $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ be a measure dynamical system.

(1) The system $\mathbf{X}$ is *rigid* if there exists a sequence $n_k \nearrow \infty$ such that

$$\lim \mu\big(T^{n_k} A \cap A\big) = \mu(A)$$

for every measurable subset A of X. We say that $\mathbf{X}$ is $\{n_k\}$-*rigid*.

(2) An ergodic system is *mildly mixing* if it has no non-trivial rigid factor.

These notions were introduced in [28]. The authors show that the mild mixing property is equivalent to the following multiplier property.

THEOREM 4.2. *An ergodic system $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ is mildly mixing iff for every ergodic* (*finite or infinite*) *measure preserving system $(Y, \mathcal{Y}, \nu, T)$, the product system*

$$(X \times Y, \mu \times \nu, T \times T),$$

is ergodic.

Since every Kronecker system is rigid it follows from Theorem 2.1 that mild mixing implies weak mixing. Clearly strong mixing implies mild mixing. It is not hard to construct rigid weakly mixing systems, so that the class of mildly mixing systems is properly contained in the class of weakly mixing systems. Finally there are mildly but not strongly mixing systems; e.g., Chacón's system is an example (see Aaronson and Weiss [2]).

We also have the following analytic characterization of mild mixing.

PROPOSITION 4.3. *An ergodic system* $\mathbf{X}$ *is mildly mixing iff*

$$\limsup_{n\to\infty} \phi_f(n) < 1,$$

for every matrix coefficient ϕ_f, *where for* $f \in L^2(X,\mu)$, $\|f\| = 1$, $\phi_f(n) := \langle U_{T^n} f, f\rangle$.

PROOF. If $\mathbf{X} \to \mathbf{Y}$ is a rigid factor, then there exists a sequence $n_i \to \infty$ such that $U_{T^{n_i}} \to \mathrm{id}$ strongly on $L^2(Y,\nu)$. For any function $f \in L^2_0(Y,\nu)$ with $\|f\| = 1$, we have $\lim_{i\to\infty} \phi_f(n_i) = 1$. Conversely, if $\lim_{i\to\infty} \phi_f(n_i) = 1$ for some $n_i \nearrow \infty$ and $f \in L^2_0(X,\mu)$, $\|f\| = 1$, then $\lim_{i\to\infty} U_{T^{n_i}} f = f$. Clearly f can be replaced by a bounded function and we let A be the sub-algebra of $L^\infty(X,\mu)$ generated by $\{U_{T^n} f\colon\ n \in \mathbb{Z}\}$. The algebra A defines a non-trivial factor $\mathbf{X} \to \mathbf{Y}$ such that $U_{T^{n_i}} \to \mathrm{id}$ strongly on $L^2(Y,\nu)$. □

We say that a collection $\mathcal{F}$ of nonempty subsets of $\mathbb{Z}$ is a *family* if it is hereditary upward and *proper* (i.e. $A \subset B$ and $A \in \mathcal{F}$ implies $B \in \mathcal{F}$, and $\mathcal{F}$ is neither empty nor all of $2^{\mathbb{Z}}$).

With a family $\mathcal{F}$ of nonempty subsets of $\mathbb{Z}$ we associate the *dual family*

$$\mathcal{F}^* = \{E\colon\ E \cap F \neq \emptyset, \forall F \in \mathcal{F}\}.$$

It is easily verified that $\mathcal{F}^*$ is indeed a family. Also, for families, $\mathcal{F}_1 \subset \mathcal{F}_2 \Rightarrow \mathcal{F}_1^* \supset \mathcal{F}_2^*$, and $\mathcal{F}^{**} = \mathcal{F}$.

We say that a subset J of $\mathbb{Z}$ has *uniform density* 1 if for every $0 < \lambda < 1$ there exists an N such that for every interval $I \subset \mathbb{Z}$ of length $> N$ we have $|J \cap I| \geqslant \lambda |I|$. We denote by $\mathcal{D}$ the family of subsets of $\mathbb{Z}$ of uniform density 1. It is also easy to see that $\mathcal{D}$ has the finite intersection property.

Let $\mathcal{F}$ be a family of non-empty subsets of $\mathbb{Z}$ which is closed under finite intersections (i.e. $\mathcal{F}$ is a filter). Following [26] we say that a sequence $\{x_n\colon\ n \in \mathbb{Z}\}$ in a topological space X $\mathcal{F}$-*converges* to a point $x \in X$ if for every neighborhood V of x the set $\{n\colon\ x_n \in V\}$ is in $\mathcal{F}$. We denote this by

$$\mathcal{F}\text{-}\lim x_n = x.$$

We have the following characterization of weak mixing for measure preserving systems which explains more clearly its name.

THEOREM 4.4. *The dynamical system* $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ *is weakly mixing iff for every* $A, B \in \mathcal{X}$ *we have*

$$\mathcal{D}\text{-}\lim \mu(T^{-n} A \cap B) = \mu(A)\mu(B).$$

An analogous characterization of measure theoretical mild mixing is obtained by considering the families of *IP* and *IP** sets. An *IP-set* is any subset of $\mathbb{Z}$ containing a subset of the form $IP\{n_i\} = \{n_{i_1} + n_{i_2} + \cdots + n_{i_k}\colon\ i_1 < i_2 < \cdots < i_k\}$, for some infinite sequence $\{n_i\}_{i=1}^{\infty}$. We let $\mathcal{I}$ denote the family of *IP*-sets and call the elements of the dual family $\mathcal{I}^*$, *IP***-sets*. Again it is not hard to see that the family of *IP**-sets is closed under finite intersections. For a proof of the next theorem we refer to [26].

THEOREM 4.5. *The dynamical system* $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ *is mildly mixing iff for every* $A, B \in \mathcal{X}$ *we have*

$$\mathcal{I}^*\text{-}\lim \mu(T^{-n}A \cap B) = \mu(A)\mu(B).$$

We now turn to the topological category. Let (X, T) be a topological dynamical system. For two non-empty open sets $U, V \subset X$ and a point $x \in X$ set

$$N(U, V) = \{n \in \mathbb{Z}\colon\ T^n U \cap V \neq \emptyset\}, \qquad N_+(U, V) = N(U, V) \cap \mathbb{Z}_+$$
$$\text{and} \qquad N(x, V) = \{n \in \mathbb{Z}\colon\ T^n x \in V\}.$$

Notice that sets of the form $N(U, U)$ are symmetric.

We say that (X, T) is *topologically transitive* (or just *transitive*) if $N(U, V)$ is non-empty whenever $U, V \subset X$ are two non-empty open sets. Using Baire's category theorem it is easy to see that (for metrizable X) a system (X, T) is topologically transitive iff there exists a dense G_δ subset $X_0 \subset X$ such that $\overline{\mathcal{O}}_T(x) = X$ for every $x \in X_0$.

We define the family $\mathcal{F}_{\text{thick}}$ of *thick sets* to be the collection of sets which contain arbitrary long intervals. The dual family $\mathcal{F}_{\text{synd}} = \mathcal{F}^*_{\text{thick}}$ is the collection of *syndetic sets*—those sets $A \subset \mathbb{Z}$ such that for some positive integer N the intersection of A with every interval of length N is non-empty.

Given a family $\mathcal{F}$ we say that a topological dynamical system (X, T) is $\mathcal{F}$*-recurrent* if $N(A, A) \in \mathcal{F}$ for every non-empty open set $A \subset X$. We say that a dynamical system is $\mathcal{F}$*-transitive* if $N(A, B) \in \mathcal{F}$ for every non-empty open sets $A, B \subset X$. The class of $\mathcal{F}$-transitive systems is denoted by $\mathcal{E}_{\mathcal{F}}$. E.g., in this notation the class of *topologically mixing systems* is $\mathcal{E}_{\text{cofinite}}$, where we call a subset $A \subset \mathbb{Z}$ co-finite when $\mathbb{Z} \setminus A$ is a finite set. We write simply $\mathcal{E} = \mathcal{E}_{\text{infinite}}$ for the class of *recurrent transitive* dynamical systems. It is not hard to see that when X has no isolated points (X, T) is topologically transitive iff it is recurrent transitive. From this we then deduce that a weakly mixing system is necessarily recurrent transitive.

In a dynamical system (X, T) a point $x \in X$ is a *wandering point* if there exists an open neighborhood U of x such that the collection $\{T^n U\colon\ n \in \mathbb{Z}\}$ is pairwise disjoint.

PROPOSITION 4.6. *Let* (X, T) *be a topologically transitive dynamical system*; *then the following conditions are equivalent*:

(1) $(X, T) \in \mathcal{E}_{\text{infinite}}$.
(2) *The recurrent points are dense in* X.
(3) (X, T) *has no wandering points.*

(4) *The dynamical system* (X_∞, T)*, the one point compactification of the integers with translation and a fixed point at infinity, is not a factor of* (X, T).

PROOF. (1) $\Rightarrow$ (4) If $\pi : X \to X_\infty$ is a factor map then, clearly $N(\pi^{-1}(0), \pi^{-1}(0)) = \{0\}$.

(4) $\Rightarrow$ (3) If U is a non-empty open wandering subset of X then $\{T^j U\colon\ j \in \mathbb{Z}\} \cup (X \setminus \bigcup\{T^j U\colon\ j \in \mathbb{Z}\})$ is a partition of X. It is easy to see that this partition defines a factor map $\pi : X \to X_\infty$.

(3) $\Rightarrow$ (2) This implication is a consequence of the following:

LEMMA 4.7. *If the dynamical system* (X, T) *has no wandering points then the recurrent points are dense in* X.

PROOF. For every $\delta > 0$ put

$$A_\delta = \big\{x \in X\colon\ \exists j \neq 0,\ d(T^j x, x) < \delta\big\}.$$

Clearly A_δ is an open set and we claim that it is dense. In fact given $x \in X$ and $\varepsilon > 0$ there exists $j \neq 0$ with

$$T^j B_\varepsilon(x) \cap B_\varepsilon(x) \neq \emptyset.$$

If y is a point in this intersection then $d(T^{-j} y, y) < 2\varepsilon$. Thus for $\varepsilon < \delta/2$ we have $y \in A_\delta$ and $d(x, y) < \varepsilon$. Now by Baire's theorem

$$A = \bigcap_{k=1}^{\infty} A_{1/k}$$

is a dense G_δ subset of X and each point in A is recurrent. □

(2) $\Rightarrow$ (1) Given U, V non-empty open subsets of X and $k \in N(U, V)$ let U_0 be the non-empty open subset $U_0 = U \cap T^{-k} V$. Check that $N(U_0, U_0) + k \subset N(U, V)$. By assumption $N(U_0, U_0)$ is infinite and a fortiori so is $N(U, V)$. This completes the proof of Proposition 4.6. □

A well known characterization of the class **WM** of topologically weakly mixing systems is due to Furstenberg:

THEOREM 4.8. $\mathbf{WM} = \mathcal{E}_{\text{thick}}$.

Following [6] we call the systems in $\mathcal{E}_{\text{synd}}$ *topologically ergodic* and write **TE** for this class. This is a rich class as we can see from the following claim from [40]. Here **MIN** is the class of minimal systems and **E** the class of E-systems; i.e. those transitive dynamical systems (X, T) for which there exists a probability invariant measure with full support.

THEOREM 4.9. $\mathbf{MIN}, \mathbf{E} \subset \mathbf{TE}$.

PROOF. (1) The claim for **MIN** is immediate by the well known characterization of minimal systems: (X, T) is minimal iff $N(x, U)$ is syndetic for every $x \in X$ and non-empty open $U \subset X$.

(2) Given two non-empty open sets U, V in X, choose $k \in \mathbb{Z}$ with $T^k U \cap V \neq \emptyset$. Next set $U_0 = T^{-k} V \cap U$, and observe that $k + N(U_0, U_0) \subset N(U, V)$. Thus it is enough to show that $N(U, U)$ is syndetic for every non-empty open U. We have to show that $N(U, U)$ meets every thick subset $B \subset \mathbb{Z}$. By Poincaré's recurrence theorem, $N(U, U)$ meets every set of the form $A - A = \{n - m\colon\ n, m \in A\}$ with A infinite. It is an easy exercise to show that every thick set B contains some $D^+(A) = \{a_n - a_m\colon\ n > m\}$ for an infinite sequence $A = \{a_n\}$. Thus $\emptyset \neq N(U, U) \cap \pm D^+(A) \subset N(U, U) \cap \pm B$. Since $N(U, U)$ is symmetric, this completes the proof. □

We recall (see the previous section) that two dynamical systems (X, T) and (Y, T) are disjoint if every closed $T \times T$-invariant subset of $X \times Y$ whose projections on X and Y are full, is necessarily the entire space $X \times Y$. It follows easily that when (X, T) and (Y, T) are disjoint, at least one of them must be minimal. If both (X, T) and (Y, T) are minimal then they are disjoint iff the product system is minimal. We say that (X, T) and (Y, T) are *weakly disjoint* when the product system $(X \times Y, T \times T)$ is transitive. This is indeed a very weak sense of disjointness as there are systems which are weakly disjoint from themselves. In fact, by definition a dynamical system is topologically weakly mixing iff it is weakly disjoint from itself.

If **P** is a class of recurrent transitive dynamical systems we let $\mathbf{P}^{\perp}$ be the class of recurrent transitive dynamical systems which are weakly disjoint from every member of **P**

$$\mathbf{P}^{\perp} = \big\{(X, T)\colon\ X \times Y \in \mathcal{E} \text{ for every } (Y, T) \in \mathcal{P}\big\}.$$

We clearly have $\mathbf{P} \subset \mathbf{Q} \Rightarrow \mathbf{P}^{\perp} \supset \mathbf{Q}^{\perp}$ and $\mathbf{P}^{\perp\perp\perp} = \mathbf{P}^{\perp}$.

For the discussion of topologically mildly mixing systems it will be convenient to deal with families of subsets of $\mathbb{Z}_+$ rather than $\mathbb{Z}$. If $\mathcal{F}$ is such a family then

$$\mathcal{E}_{\mathcal{F}} = \big\{(X, T)\colon\ N_+(A, B) \in \mathcal{F} \text{ for every non-empty open } A, B \subset X\big\}.$$

Let us call a subset of $\mathbb{Z}_+$ a *SIP-set* (symmetric *IP*-set), if it contains a subset of the form

$$SIP\{n_i\} = \big\{n_\alpha - n_\beta > 0\colon\ n_\alpha, n_\beta \in IP\{n_i\} \cup \{0\}\big\},$$

for an *IP* sequence $IP\{n_i\} \subset \mathbb{Z}_+$. Denote by $\mathcal{S}$ the family of *SIP* sets. It is not hard to show that

$$\mathcal{F}_{\text{thick}} \subset \mathcal{S} \subset \mathcal{I}$$

(see [26]). Hence $\mathcal{F}_{\text{syndetic}} \supset \mathcal{S}^* \supset \mathcal{I}^*$, hence $\mathcal{E}_{\text{synd}} \supset \mathcal{E}_{\mathcal{S}^*} \supset \mathcal{E}_{\mathcal{I}^*}$, and finally

$$\mathcal{E}_{\text{synd}}^{\perp} \subset \mathcal{E}_{\mathcal{S}^*}^{\perp} \subset \mathcal{E}_{\mathcal{I}^*}^{\perp}.$$

DEFINITION 4.10. A topological dynamical system (X, T) is called *topologically mildly mixing* if it is in $\mathcal{E}_{\mathcal{S}^*}$ and we denote the collection of topologically mildly mixing systems by $\mathbf{MM} = \mathcal{E}_{\mathcal{S}^*}$.

THEOREM 4.11. *A dynamical system is in $\mathcal{E}$ iff it is weakly disjoint from every topologically mildly mixing system*:

$$\mathcal{E} = \mathbf{MM}^{\curlywedge}.$$

And conversely it is topologically mildly mixing iff it is weakly disjoint from every recurrent transitive system:

$$\mathbf{MM} = \mathcal{E}^{\curlywedge}.$$

PROOF. (1) Since $\mathcal{E}_{\mathcal{S}^*}$ is non-vacuous (for example, every topologically mixing system is in $\mathcal{E}_{\mathcal{S}^*}$), it follows that every system in $\mathcal{E}_{\mathcal{S}^*}^{\curlywedge}$ is in $\mathcal{E}$.

Conversely, assume that (X, T) is in $\mathcal{E}$ but $(X, T) \notin \mathcal{E}_{\mathcal{S}^*}^{\curlywedge}$, and we will arrive at a contradiction. By assumption there exists $(Y, T) \in \mathcal{E}_{\mathcal{S}^*}$ and a non-dense non-empty open invariant subset $W \subset X \times Y$. Then $\pi_X(W) = O$ is a non-empty open invariant subset of X. By assumption O is dense in X. Choose open non-empty sets $U_0 \subset X$ and $V_0 \subset Y$ with $U_0 \times V_0 \subset W$. By Proposition 4.6 there exists a recurrent point x_0 in $U_0 \subset O$. Then there is a sequence $n_i \to \infty$ such that for the *IP*-sequence $\{n_\alpha\} = IP\{n_i\}_{i=1}^{\infty}$, $IP\text{-}\lim T^{n_\alpha} x_0 = x_0$ (see [26]). Choose i_0 such that $T^{n_\alpha} x_0 \in U_0$ for $n_\alpha \in J = IP\{n_i\}_{i \geqslant i_0}$ and set $D = SIP(J)$. Given V a non-empty open subset of Y we have:

$$D \cap N(V_0, V) \neq \emptyset.$$

Thus for some α, β and $v_0 \in V_0$,

$$T^{n_\alpha - n_\beta}\left(T^{n_\beta} x_0, v_0\right) = \left(T^{n_\alpha} x_0, T^{n_\alpha - n_\beta} v_0\right) \in (U_0 \times V) \cap W.$$

We conclude that

$$\{x_0\} \times Y \subset \operatorname{cls} W.$$

The fact that in an $\mathcal{E}$ system the recurrent points are dense together with the observation that $\{x_0\} \times Y \subset \operatorname{cls} W$ for every recurrent point $x_0 \in O$, imply that W is dense in $X \times Y$, a contradiction.

(2) From part (1) of the proof we have $\mathcal{E} = \mathcal{E}_{\mathcal{S}^*}^{\curlywedge}$, hence $\mathcal{E}^{\curlywedge} = \mathcal{E}_{\mathcal{S}^*}^{\curlywedge\curlywedge} \supset \mathcal{E}_{\mathcal{S}^*}$.

Suppose $(X, T) \in \mathcal{E}$ but $(X, T) \notin \mathcal{E}_{\mathcal{S}^*}$, we will show that $(X, T) \notin \mathcal{E}^{\curlywedge}$. There exist $U, V \subset X$, non-empty open subsets and an *IP*-set $I = IP\{n_i\}$ for a monotone increasing sequence $\{n_1 < n_2 < \cdots\}$ with

$$N(U, V) \cap D = \emptyset,$$

where

$$D = \{n_\alpha - n_\beta \colon\ n_\alpha, n_\beta \in I,\ n_\alpha > n_\beta\}.$$

If (X, T) is not topologically weakly mixing then $X \times X \notin \mathcal{E}$ hence $(X, T) \notin \mathcal{E}^{\curlywedge}$. So we can assume that (X, T) is topologically weakly mixing. Now in $X \times X$

$$N(U \times V, V \times U) = N(U, V) \cap N(V, U) = N(U, V) \cap -N(U, V),$$

is disjoint from $D \cup -D$, and replacing X by $X \times X$ we can assume that $N(U, V) \cap (D \cup -D) = \emptyset$. In fact, if $X \in \mathcal{E}^{\curlywedge}$ then $X \times Y \in \mathcal{E}$ for every $Y \in \mathcal{E}$, therefore $X \times (X \times Y) \in \mathcal{E}$ and we see that also $X \times X \in \mathcal{E}^{\curlywedge}$.

By going to a subsequence, we can assume that

$$\lim_{k\to\infty} n_{k+1} - \sum_{i=1}^{k} n_i = \infty$$

in which case the representation of each $n \in I$ as $n = n_\alpha = n_{i_1} + n_{i_2} + \cdots + n_{i_k};\ \alpha = \{i_1 < i_2 < \cdots < i_k\}$ is unique.

Next let $y_0 \in \{0, 1\}^{\mathbb{Z}}$ be the sequence $y_0 = \mathbf{1}_I$. Let Y be the orbit closure of y_0 in $\{0, 1\}^{\mathbb{Z}}$ under the shift T, and let $[1] = \{y \in Y \colon\ y(0) = 1\}$. Observe that

$$N\big(y_0, [1]\big) = I.$$

It is easy to check that

$$IP\text{-}\lim T^{n_\alpha} y_0 = y_0.$$

Thus the system (Y, T) is topologically transitive with y_0 a recurrent point; i.e. $(Y, T) \in \mathcal{E}$.

We now observe that

$$N\big([1], [1]\big) = N\big(y_0, [1]\big) - N\big(y_0, [1]\big) = I - I = D \cup -D \cup \{0\}.$$

If $X \times Y$ is topologically transitive then in particular

$$\begin{aligned} N\big(U \times [1], V \times [1]\big) &= N(U, V) \cap N\big([1], [1]\big) \\ &= N(U, V) \cap \big(D \cup -D \cup \{0\}\big) = \text{ infinite set.} \end{aligned}$$

But this contradicts our assumption. Thus $X \times Y \notin \mathcal{E}$ and $(X, T) \notin \mathcal{E}^{\curlywedge}$. This completes the proof. □

We now have the following:

COROLLARY 4.12. *Every topologically mildly mixing system is weakly mixing and topologically ergodic*:

$$\mathbf{MM} \subset \mathbf{WM} \cap \mathbf{TE}.$$

PROOF. We have $\mathcal{E}_{\mathcal{S}^*} \subset \mathcal{E} = \mathcal{E}_{\mathcal{S}^*}^{\wedge}$, hence for every $(X,T) \in \mathcal{E}_{\mathcal{S}^*}$, $X \times X \in \mathcal{E}$, i.e. (X,T) is topologically weakly mixing. And, as we have already observed the inclusion $\mathcal{F}_{\text{syndetic}} \supset \mathcal{S}^*$, entails $\mathbf{TE} = \mathcal{E}_{\text{synd}} \supset \mathcal{E}_{\mathcal{S}^*} = \mathbf{MM}$. □

To complete the analogy with the measure theoretical setup we next define a topological analogue of rigidity. This is just one of several possible definitions of topological rigidity and we refer to [38] for a treatment of these notions.

DEFINITION 4.13. A dynamical system (X,T) is called *uniformly rigid* if there exists a sequence $n_k \nearrow \infty$ such that

$$\lim_{k\to\infty} \sup_{x\in X} d\big(T^{n_k}x, x\big) = 0,$$

i.e. $\lim_{k\to\infty} T^{n_k} = \mathrm{id}$ in the uniform topology on the group of homeomorphism of $H(X)$ of X. We denote by $\mathcal{R}$ the collection of topologically transitive uniformly rigid systems.

In [38] the existence of minimal weakly mixing but nonetheless uniformly rigid dynamical systems is demonstrated. However, we have the following:

LEMMA 4.14. *A system which is both topologically mildly mixing and uniformly rigid is trivial*.

PROOF. Let (X,T) be both topologically mildly mixing and uniformly rigid. Then

$$\Lambda = \mathrm{cls}\{T^n\colon n \in \mathbb{Z}\} \subset H(X)$$

is a Polish monothetic group.

Let T^{n_i} be a sequence converging uniformly to id, the identity element of Λ. For a subsequence we can ensure that $\{n_\alpha\} = IP\{n_i\}$ is an *IP*-sequence such that $IP\text{-}\lim T^{n_\alpha} = \mathrm{id}$ in Λ. If X is non-trivial we can now find an open ball $B = B_\delta(x_0) \subset X$ with $TB \cap B = \emptyset$. Put $U = B_{\delta/2}(x_0)$ and $V = TU$; then by assumption $N(U,V)$ is an SIP^*-set and in particular:

$$\forall \alpha_0\ \exists \alpha, \beta > \alpha_0, \quad n_\alpha - n_\beta \in N(U,V).$$

However, since $IP\text{-}\lim T^{n_\alpha} = \mathrm{id}$, we also have eventually, $T^{n_\alpha - n_\beta}U \subset B$; a contradiction. □

COROLLARY 4.15. *A topologically mildly mixing system has no non-trivial uniformly rigid factors*.

We conclude this section with the following result which shows how these topological and measure theoretical notions are related.

THEOREM 4.16. *Let (X, T) be a topological dynamical system with the property that there exists an invariant probability measure μ with full support such that the associated measure preserving dynamical system $(X, \mathcal{X}, \mu, T)$ is measure theoretically mildly mixing then (X, T) is topologically mildly mixing.*

PROOF. Let (Y, S) be any system in $\mathcal{E}$; by Theorem 4.11 it suffices to show that $(X \times Y, T \times S)$ is topologically transitive. Suppose $W \subset X \times Y$ is a closed $T \times S$-invariant set with $\operatorname{int} W \neq \emptyset$. Let $U \subset X$, $V \subset Y$ be two non-empty open subsets with $U \times V \subset W$. By transitivity of (Y, S) there exits a transitive recurrent point $y_0 \in V$. By theorems of Glimm and Effros [44,22], and Katznelson and Weiss [57] (see also Weiss [83]), there exists a (possibly infinite) invariant ergodic measure ν on Y with $\nu(V) > 0$.

Let μ be the probability invariant measure of full support on X with respect to which $(X, \mathcal{X}, \mu, T)$ is measure theoretically mildly mixing. Then by [28] the measure $\mu \times \nu$ is ergodic. Since $\mu \times \nu(W) \geqslant \mu \times \nu(U \times V) > 0$ we conclude that $\mu \times \nu(W^c) = 0$ which clearly implies $W = X \times Y$. □

We note that the definition of topological mild mixing and the results described above concerning this notion are new. However independently of our work Huang and Ye in a recent work also define a similar notion and give it a comprehensive and systematic treatment [50]. The first named author would like to thank E. Akin for instructive conversations on this subject.

Regarding the classes **WM** and **TE** let us mention the following result from [86].

THEOREM 4.17.

$$\mathbf{TE} = \mathbf{WM}^{\curlywedge}.$$

For more on these topics we refer to [26,4,86,6,49] and [50].

5. Distal systems: topological vs. measure

As noted above the Kronecker or minimal equicontinuous dynamical systems can be considered as the most elementary type of systems. What is then the next stage? The clue in the topological case, which chronologically came first, is to be found in the notion of distality. A topological system (X, T) is called *distal* if

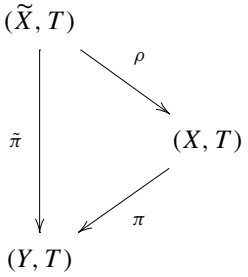
$$\inf_{n \in \mathbb{Z}} d(T^n x, T^n x') > 0$$

for every $x \neq x'$ in X. It is easy to see that this property does not depend on the choice of a metric. And, of course, every equicontinuous system is distal. Is the converse true? Are these notions one and the same? The dynamical system given on the unit disc

$D = \{z \in \mathbb{C}\colon |z| \leqslant 1\}$ by the formula $Tz = z\exp(2\pi i|z|)$ is a counter example, it is distal but not equicontinuous. However it is not minimal. H. Furstenberg in 1963 noted that skew products over an equicontinuous basis with compact group translations as fiber maps are always distal, often minimal, but rarely equicontinuous [23]. A typical example is the homeomorphism of the two torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ given by $T(x, y) = (x + \alpha, y + x)$ where $\alpha \in \mathbb{R}/Z$ is irrational. Independently and at about the same time, it was shown by L. Auslander, L. Green and F. Hahn that minimal nilflows are distal but not equicontinuous [7]. These examples led Furstenberg to his path breaking structure theorem [23].

Given a homomorphism $\pi : (X, T) \to (Y, T)$ let $R_\pi = \{(x, x')\colon \pi(x) = \pi(x')\}$. We say that the homomorphism π is an *isometric extension* if there exists a continuous function $d : R_\pi \to \mathbb{R}$ such that for each $y \in Y$ the restriction of d to $\pi^{-1}(y) \times \pi^{-1}(y)$ is a metric and for every $x, x' \in \pi^{-1}(y)$ we have $d(Tx, Tx') = d(x, x')$.

If K is a compact subgroup of $\mathrm{Aut}(X, T)$ (the group of homeomorphisms of X commuting with T, endowed with the topology of uniform convergence) then the map $x \mapsto Kx$ defines a factor map $(X, T) \xrightarrow{\pi} (Y, T)$ with $Y = X/K$ and $R_\pi = \{(x, kx)\colon x \in X,\ k \in K\}$. Such an extension is called a *group extension*. It turns out, although this is not so easy to see, that when (X, T) is minimal then $\pi : (X, T) \to (Y, T)$ is an isometric extension iff there exists a commutative diagram:



where $(\widetilde{X}, T)$ is minimal and $(\widetilde{X}, T) \xrightarrow{\tilde{\pi}} (X, T)$ is a group extension with some compact group $K \subset \mathrm{Aut}(\widetilde{X}, T)$ and the map ρ is the quotient map from $\widetilde{X}$ onto X defined by a closed subgroup H of K. Thus $Y = \widetilde{X}/K$ and $X = \widetilde{X}/H$ and we can think of π as a *homogeneous space extension* with fiber K/H.

We say that a (metrizable) minimal system (X, T) is an I *system* if there is a (countable) ordinal η and a family of systems $\{(X_\theta, x_\theta)\}_{\theta \leqslant \eta}$ such that (i) X_0 is the trivial system, (ii) for every $\theta < \eta$ there exists an isometric homomorphism $\phi_\theta : X_{\theta+1} \to X_\theta$, (iii) for a limit ordinal $\lambda \leqslant \eta$ the system X_λ is the inverse limit of the systems $\{X_\theta\}_{\theta<\lambda}$ (i.e. $X_\lambda = \bigvee_{\theta<\lambda}(X_\theta, x_\theta)$), and (iv) $X_\eta = X$.

THEOREM 5.1 (Furstenberg's structure theorem). *A minimal system is distal iff it is an I-system.*

W. Parry in his 1967 paper [70] suggested an intrinsic definition of measure distality. He defines in this paper a property of measure dynamical systems, called "admitting a separating sieve", which imitates the intrinsic definition of topological distality.

DEFINITION 5.2. Let $\mathbf{X}$ be an ergodic dynamical system. A sequence $A_1 \supset A_2 \supset \cdots$ of sets in $\mathcal{X}$ with $\mu(A_n) > 0$ and $\mu(A_n) \to 0$, is called a *separating sieve* if there exists a subset $X_0 \subset X$ with $\mu(X_0) = 1$ such that for every $x, x' \in X_0$ the condition "for every $n \in \mathbb{N}$ there exists $k \in \mathbb{Z}$ with $T^k x, T^k x' \in A_n$" implies $x = x'$, or in symbols:

$$\bigcap_{n=1}^{\infty} \left(\bigcup_{k \in \mathbb{Z}} T^k (A_n \times A_n) \right) \cap (X_0 \times X_0) \subset \Delta.$$

We say that the ergodic system $\mathbf{X}$ is *measure distal* if either $\mathbf{X}$ is finite or there exists a separating sieve.

Parry showed that every measure dynamical system admitting a separating sieve has zero entropy and that any T-invariant measure on a minimal topologically distal system gives rise to a measure dynamical system admitting a separating sieve.

If $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ is an ergodic dynamical system and $K \subset \mathrm{Aut}(\mathbf{X})$ is a compact subgroup (where $\mathrm{Aut}(\mathbf{X})$ is endowed with the weak topology) then the system $\mathbf{Y} = \mathbf{X}/K$ is well defined and we say that the extension $\pi : \mathbf{X} \to \mathbf{Y}$ is a *group extension*. Using (5) we can define the notion of isometric extension or homogeneous extension in the measure category. We will say that an ergodic system *admits a Furstenberg tower* if it is obtained as a (necessarily countable) transfinite tower of measure isometric extensions. In 1976 in two outstanding papers [88,89] R. Zimmer developed the theory of distal systems (for a general locally compact acting group). He showed that, as in the topologically distal case, systems admitting Parry's separating sieve are exactly those systems which admit Furstenberg towers.

THEOREM 5.3. *An ergodic dynamical system is measure distal iff it admits a Furstenberg tower.*

In [65] E. Lindenstrauss shows that every ergodic measure distal $\mathbb{Z}$-system can be represented as a minimal topologically distal system. For the exact result see Theorem 13.4 below.

6. Furstenberg–Zimmer structure theorem vs. its topological PI version

Zimmer's theorem for distal systems leads directly to a structure theorem for the general ergodic system. Independently, and at about the same time, Furstenberg proved the same theorem [25,26]. He used it as the main tool for his proof of Szemerédi's theorem on arithmetical progressions. Recall that an extension $\pi : (X, \mathcal{X}, \mu, T) \to (Y, \mathcal{Y}, \nu, T)$ is a *weakly mixing extension* if the relative product system $\mathbf{X} \underset{\mathbf{Y}}{\times} \mathbf{X}$ is ergodic. (The system $\mathbf{X} \underset{\mathbf{Y}}{\times} \mathbf{X}$ is defined by the $T \times T$ invariant measure

$$\mu \underset{\nu}{\times} \mu = \int_Y \mu_y \times \mu_y \, d\nu(y),$$

on $X \times X$, where $\mu = \int_Y \mu_y \, d\nu(y)$ is the disintegration of μ over ν.)

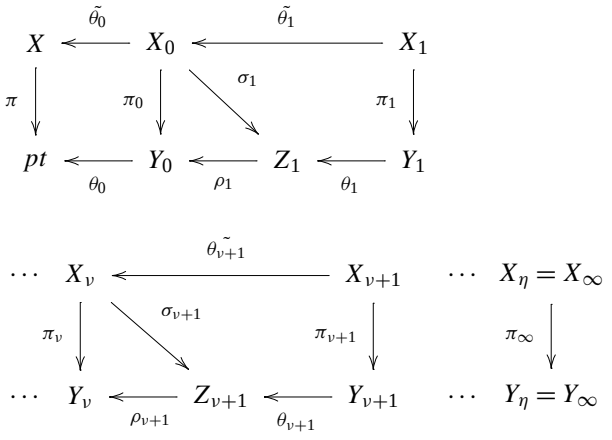
THEOREM 6.1 (The Furstenberg–Zimmer structure theorem). *Let* $\mathbf{X}$ *be an ergodic dynamical system.*

(1) *There exists a maximal distal factor* $\phi : \mathbf{X} \to \mathbf{Z}$ *with* ϕ *is a weakly mixing extension.*
(2) *This factorization is unique.*

Is there a general structure theorem for minimal topological systems? Here, for the first time, we see a strong divergence between the measure and the topological theories. The culpability for this divergence is to be found in the notions of proximality and proximal extension, which arise naturally in the topological theory but do not appear at all in the measure theoretical context. In building towers for minimal systems we have to use two building blocks of extremely different nature (isometric and proximal) rather than one (isometric) in the measure category. A pair of points $(x, x') \in X \times X$ is called *proximal* if it is not distal, i.e. if $\inf_{n\in\mathbb{Z}} d(T^n x, T^n x') = 0$. An extension $\pi : (X, T) \to (Y, T)$ is called *proximal* if every pair in R_π is proximal. The next theorem was developed gradually by several authors (Veech, Glasner, Ellis and Shapiro, and McMahon [80,30,66,81]). We need first to introduce some definitions. We say that a minimal dynamical system (X, T) is strictly *PI* (proximal isometric) if it admits a tower consisting of proximal and isometric extensions. It is called a *PI system* if there is a strictly PI minimal system $(\widetilde{X}, T)$ and a proximal extension $\theta : \widetilde{X} \to X$. An extension $\pi : X \to Y$ is a *RIC extension* (relatively incontractible) if for every $n \in \mathbb{N}$ and every $y \in Y$ the set of almost periodic points in $X_y^n = \pi^{-1}(y) \times \pi^{-1}(y) \times \cdots \times \pi^{-1}(y)$ (n times) is dense. (A point is called *almost periodic* if its orbit closure is minimal.) It can be shown that a every isometric (and more generally, distal) extension is RIC. Also every RIC extension is open. Finally a homomorphism $\pi : X \to Y$ is called *topologically weakly mixing* if the dynamical system $(R_\pi, T \times T)$ is topologically transitive.

The philosophy in the next theorem is to regard proximal extensions as 'negligible' and then the claim is, roughly (i.e. up to proximal extensions), that every minimal system is a weakly mixing extension of its maximal PI factor.

THEOREM 6.2 (Structure theorem for minimal systems). *Given a metric minimal system* (X, T)*, there exists a countable ordinal* η *and a canonically defined commutative diagram* (*the canonical PI-Tower*)



where for each $\nu \leqslant \eta$, π_ν is RIC, ρ_ν is isometric, $\theta_\nu, \tilde{\theta}_\nu$ are proximal extensions and π_∞ is RIC and topologically weakly mixing extension. For a limit ordinal ν, X_ν, Y_ν, π_ν, etc. are the inverse limits (or joins) of $X_\iota, Y_\iota, \pi_\iota$, etc. for $\iota < \nu$. Thus X_∞ is a proximal extension of X and a RIC topologically weakly mixing extension of the strictly PI-system Y_∞. The homomorphism π_∞ is an isomorphism (so that $X_\infty = Y_\infty$) iff X is a PI-system.

We refer to [34] for a review on structure theory in topological dynamics.

7. Entropy: measure and topological

7.1. *The classical variational principle*

For the definitions and the classical results concerning entropy theory we refer to Chapter 1 of [1]; Section 3.7 for measure theory entropy and Section 4.4 for metric and topological entropy. Chapter 2 of [1] has a short review of basic measure entropy theory. The variational principle asserts that for a topological $\mathbb{Z}$-dynamical system (X, T) the topological entropy equals the supremum of the measure entropies computed over all the invariant probability measures on X. It was already conjectured in the original paper of Adler, Konheim and McAndrew [3] where topological entropy was introduced; and then, after many stages (mainly by Goodwyn, Bowen and Dinaburg; see, for example, [18]) matured into a theorem in Goodman's paper [45].

THEOREM 7.1 (The variational principle). *Let (X, T) be a topological dynamical system, then*

$$h_{\text{top}}(X, T) = \sup\{h_\mu\colon\ \mu \in M_T(X)\} = \sup\{h_\mu\colon\ \mu \in M_T^{\text{erg}}(X)\}.$$

This classical theorem has had a tremendous influence on the theory of dynamical systems and a vast amount of literature ensued, which we will not try to trace here (see [1, Chapter 1, Theorem 4.4.4]). Instead we would like to present a more recent development.

7.2. *Entropy pairs and UPE systems*

As we have noted in the introduction, the theories of measurable dynamics (ergodic theory) and topological dynamics exhibit a remarkable parallelism. Usually one translates 'ergodicity' as 'topological transitivity','weak mixing' as 'topological weak mixing', 'mixing' as 'topological mixing' and 'measure distal' as 'topologically distal'. One often obtains this way parallel theorems in both theories, though the methods of proof may be very different.

What is then the topological analogue of being a K-system? In [9] and [10] F. Blanchard introduced a notion of 'topological *K*' for $\mathbb{Z}$-systems which he called UPE (uniformly positive entropy). This is defined as follows: a topological dynamical system (X, T) is called a UPE system if every open cover of X by two non-dense open sets U and V has

positive topological entropy. A local version of this definition led to the concept of an entropy pair. A pair $(x, x') \in X \times X,\ x \neq x'$, is an entropy pair if for every open cover $\mathcal{U} = \{U, V\}$ of X, with $x \in \text{int}(U^c)$ and $x' \in \text{int}(V^c)$, the topological entropy $h(\mathcal{U})$ is positive. The set of entropy pairs is denoted by $E_X = E_{(X,T)}$ and it follows that the system (X, T) is UPE iff $E_X = (X \times X) \setminus \Delta$. In general $E^* = E_X \cup \Delta$ is a $T \times T$-invariant closed symmetric and reflexive relation. Is it also transitive? When the answer to this question is affirmative then the quotient system X/E_X^* is the topological analogue of the Pinsker factor. Unfortunately this need not always be true even when (X, T) is a minimal system (see [42] for a counter example).

The following theorem was proved in Glasner and Weiss [41].

THEOREM 7.2. *If the compact system (X, T) supports an invariant measure μ for which the corresponding measure theoretical system $(X, \mathcal{X}, \mu, T)$ is a K-system, then (X, T) is UPE.*

Applying this theorem together with the Jewett–Krieger theorem it is now possible to obtain a great variety of strictly ergodic UPE systems.

Given a T-invariant probability measure μ on X, a pair $(x, x') \in X \times X,\ x \neq x'$ is called a μ-entropy pair if for every Borel partition $\alpha = \{Q, Q^c\}$ of X with $x \in \text{int}(Q)$ and $x' \in \text{int}(Q^c)$ the measure entropy $h_\mu(\alpha)$ is positive. This definition was introduced by Blanchard, Host, Maass, Martínez and Rudolph in [11] as a local generalization of Theorem 7.2. It was shown in [11] that for every invariant probability measure μ the set E_μ of μ-entropy pairs is contained in E_X.

THEOREM 7.3. *Every measure entropy pair is a topological entropy pair.*

As in [41] the main issue here is to understand the, sometimes intricate, relation between the combinatorial entropy $h_c(\mathcal{U})$ of a cover $\mathcal{U}$ and the measure theoretical entropy $h_\mu(\gamma)$ of a measurable partition γ subordinate to $\mathcal{U}$.

PROPOSITION 7.4. *Let $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ be a measure dynamical system. Suppose $\mathcal{U} = \{U, V\}$ is a measurable cover such that every measurable two-set partition $\gamma = \{H, H^c\}$ which (as a cover) is finer than $\mathcal{U}$ satisfies $h_\mu(\gamma) > 0$; then $h_c(\mathcal{U}) > 0$.*

Since for a K-measure μ clearly every pair of distinct points is in E_μ, Theorem 7.2 follows from Theorem 7.3. It was shown in [11] that when (X, T) is uniquely ergodic the converse of Theorem 7.3 is also true: $E_X = E_\mu$ for the unique invariant measure μ on X.

7.3. *A measure attaining the topological entropy of an open cover*

In order to gain a better understanding of the relationship between measure entropy pairs and topological entropy pairs one direction of a variational principle for open covers (Theorem 7.5 below) was proved in Blanchard, Glasner and Host [12]. Two applications of this principle were given in [12]; (i) the construction, for a general system (X, T), of

a measure $\mu \in M_T(X)$ with $E_X = E_\mu$, and (ii) the proof that under a homomorphism $\pi : (X, \mu, T) \to (Y, \nu, T)$ every entropy pair in E_ν is the image of an entropy pair in E_μ.

We now proceed with the statement and proof of this theorem which is of independent interest. The other direction of this variational principle will be proved in the following subsection.

THEOREM 7.5. *Let (X, T) be a topological dynamical system, and $\mathcal{U}$ an open cover of X, then there exists a measure $\mu \in M_T(X)$ such that $h_\mu(\alpha) \geqslant h_{\text{top}}(\mathcal{U})$ for all Borel partitions α finer than $\mathcal{U}$.*

A crucial element of the proof of the variational principle is a combinatorial lemma which we present next. We let $\phi : [0, 1] \to \mathbb{R}$ denote the function

$$\phi(x) = -t \log t \quad \text{for } 0 < t \leqslant 1;\ \phi(0) = 0.$$

Let $\mathfrak{L} = \{1, 2, \ldots, \ell\}$ be a finite set, called the *alphabet*; sequences $\omega = \omega_1 \ldots \omega_n \in \mathfrak{L}^n$, for $n \geqslant 1$, are called *words of length n on the alphabet $\mathfrak{L}$*. Let n and k be two integers with $1 \leqslant k \leqslant n$.

For every word ω of length n and every word θ of length k on the same alphabet, we denote by $p(\theta|\omega)$ the frequency of appearances of θ in ω, i.e.

$$p(\theta|\omega) = \frac{1}{n-k+1} \times \operatorname{card}\{i\colon\ 1 \leqslant i \leqslant n-k+1,\ \omega_i\omega_{i+1}\ldots\omega_{i+k-1} = \theta_1\theta_2\ldots\theta_k\}.$$

For every word ω of length n on the alphabet $\mathfrak{L}$, we let

$$H_k(\omega) = \sum_{\theta \in \mathfrak{L}^k} \phi\big(p(\theta|\omega)\big).$$

LEMMA 7.6. *For every $h > 0$, $\varepsilon > 0$, every integer $k \geqslant 1$ and every sufficiently large integer n,*

$$\operatorname{card}\big\{\omega \in \mathfrak{L}^n\colon\ H_k(\omega) \leqslant kh\big\} \leqslant \exp\big(n(h+\varepsilon)\big).$$

REMARK. It is equally true that, if $h \leqslant \log(\operatorname{card}\mathfrak{L})$, for sufficiently large n,

$$\operatorname{card}\big\{\omega \in \mathfrak{L}^n\colon\ H_k(\omega) \leqslant kh\big\} \geqslant \exp\big(n(h-\varepsilon)\big).$$

We do not prove this inequality here, since we have no use for it in the sequel.

PROOF. *The case $k = 1$.* We have

$$\operatorname{card}\big\{\omega \in \mathfrak{L}^n\colon\ H_1(\omega) \leqslant h\big\} = \sum_{q \in K} \frac{n!}{q_1! \cdots q_\ell!}, \tag{2}$$

where K is the set of $q = (q_1, \dots, q_\ell) \in \mathbb{N}^\ell$ such that

$$\sum_{i=1}^{\ell} q_i = n \quad \text{and} \quad \sum_{i=1}^{\ell} \phi\left(\frac{q_i}{n}\right) \leqslant h.$$

By Stirling's formula, there exist two universal constants c and c' such that

$$c\left(\frac{m}{e}\right)^m \sqrt{m} \leqslant m! \leqslant c'\left(\frac{m}{e}\right)^m \sqrt{m}$$

for every $m > 0$. From this we deduce the existence of a constant $C(\ell)$ such that for every $q \in K$,

$$\frac{n!}{q_1! \cdots q_\ell!} \leqslant C(\ell) \exp\left(n \sum_{i=1}^{\ell} \phi\left(\frac{q_i}{n}\right)\right) \leqslant C(\ell) \exp(nh).$$

Now the sum (2) contains at most $(n+1)^\ell$ terms; so that we have

$$\operatorname{card}\{\omega \in \mathfrak{L}^n\colon\ H_1(\omega) \leqslant h\} \leqslant (n+1)^\ell C(\ell) \exp(nh) \leqslant \exp\big(n(h+\varepsilon)\big)$$

for all sufficiently large n, as was to be proved.

The case $k > 1$. For every word ω of length $n \geqslant 2k$ on the alphabet $\mathfrak{L}$, and for $0 \leqslant j < k$, we let n_j be the integral part of $\frac{n-j}{k}$, and $\omega^{(j)}$ the word

$$(\omega_{j+1} \dots \omega_{j+k})\ (\omega_{j+k+1} \dots \omega_{j+2k})\ \dots\ (\omega_{j+(n_j-1)k+1} \dots \omega_{j+n_j k})$$

of length n_j on the alphabet $B = \mathfrak{L}^k$.

Let now θ be a word of length k on the alphabet $\mathfrak{L}$; we also consider θ as an element of B. One easily verifies that, for every word ω of length n on the alphabet $\mathfrak{L}$,

$$\left| p(\theta|\omega) - \frac{1}{k} \sum_{j=0}^{k-1} p\big(\theta|\omega^{(j)}\big) \right| \leqslant \frac{k}{n-2k+1}.$$

The function ϕ being uniformly continuous, we see that for sufficiently large n, and for every word ω of length n on $\mathfrak{L}$,

$$\sum_{\theta \in B} \left| \phi\big(p(\theta|\omega)\big) - \phi\left(\frac{1}{k} \sum_{j=0}^{k-1} p\big(\theta|\omega^{(j)}\big)\right) \right| < \frac{\varepsilon}{2}$$

and by convexity of ϕ,

$$\frac{1}{k} \sum_{j=0}^{k-1} H_1\big(\omega^{(j)}\big) = \frac{1}{k} \sum_{j=0}^{k-1} \sum_{\theta \in B} \phi\big(p\big(\theta|\omega^{(j)}\big)\big) \leqslant \frac{\varepsilon}{2} + \sum_{\theta \in \mathfrak{L}^k} \phi\big(p(\theta|\omega)\big) = \frac{\varepsilon}{2} + H_k(\omega).$$

Thus, if $H_k(\omega) \leqslant kh$, there exists a j such that $H_1(\omega^{(j)}) \leqslant \frac{\varepsilon}{2} + kh$.

Now, given j and a word u of length n_j on the alphabet B, there exist $\ell^{n-n_jk} \leqslant \ell^{2k-2}$ words ω of length n on $\mathfrak{L}$ such that $\omega^{(j)} = u$. Thus for sufficiently large n, by the first part of the proof,

$$\begin{aligned}\operatorname{card}\bigl\{\omega \in \mathfrak{L}^n\colon\ H_k(\omega) \leqslant kh\bigr\} &\leqslant \ell^{2k-2} \sum_{j=0}^{k-1} \operatorname{card}\left\{u \in B^{n_j}\colon\ H_1(u) \leqslant \frac{\varepsilon}{2} + kh\right\} \\ &\leqslant \ell^{2k-2} \sum_{j=0}^{k-1} \exp\bigl(n_j(\varepsilon + kh)\bigr) \\ &\leqslant \ell^{2k-2} k \exp\left(n\left(\frac{\varepsilon}{k} + h\right)\right) \leqslant \exp\bigl(n(h+\varepsilon)\bigr). \qquad \square\end{aligned}$$

Let (X, T) be a compact dynamical system. As usual we denote by $M_T(X)$ the set of T-invariant probability measures on X, and by $M_T^{\mathrm{erg}}(X)$ the subset of ergodic measures.

We say that a partition α is finer than a cover $\mathcal{U}$ when every atom of α is contained in an element of $\mathcal{U}$. If $\alpha = \{A_1, \ldots, A_\ell\}$ is a partition of X, $x \in X$ and $N \in \mathbb{N}$, we write $\omega(\alpha, N, x)$ for the word of length N on the alphabet $\mathfrak{L} = \{1, \ldots, \ell\}$ defined by

$$\omega(\alpha, N, x)_n = i \quad \text{if } T^{n-1}x \in A_i,\ 1 \leqslant n \leqslant N.$$

LEMMA 7.7. *Let $\mathcal{U}$ be a cover of X, $h = h_{\mathrm{top}}(\mathcal{U})$, $K \geqslant 1$ an integer, and $\{\alpha_l\colon\ 1 \leqslant l \leqslant K\}$ a finite sequence of partitions of X, all finer than $\mathcal{U}$. For every $\varepsilon > 0$ and sufficiently large N, there exists an $x \in X$ such that*

$$H_k\bigl(\omega(\alpha_l, N, x)\bigr) \geqslant k(h - \varepsilon) \quad \textit{for every } k, l \textit{ with } 1 \leqslant k, l \leqslant K.$$

PROOF. One can assume that all the partitions α_l have the same number of elements ℓ and we let $\mathfrak{L} = \{1, \ldots, \ell\}$. For $1 \leqslant k \leqslant K$ and $N \geqslant K$, denote

$$\Omega(N, k) = \bigl\{\omega \in \mathfrak{L}^N\colon\ H_k(\omega) < k(h - \varepsilon)\bigr\}.$$

By Lemma 7.3, for sufficiently large N

$$\operatorname{card}\bigl(\Omega(N, k)\bigr) \leqslant \exp\bigl(N(h - \varepsilon/2)\bigr) \quad \text{for all } k \leqslant K.$$

Let us choose such an N which moreover satisfies $K^2 < \exp(N\varepsilon/2)$. For $1 \leqslant k, l \leqslant K$, let

$$Z(k, l) = \bigl\{x \in X\colon\ \omega(\alpha_l, N, x) \in \Omega(N, k)\bigr\}.$$

The set $Z(k, l)$ is the union of $\operatorname{card}(\Omega(N, k))$ elements of $(\alpha_l)_0^{N-1}$. Now this partition is finer than the cover $\mathcal{U}_0^{N-1}$, hence $Z(k, l)$ is covered by

$$\operatorname{card}\bigl(\Omega(N, k)\bigr) \leqslant \exp\bigl(N(h - \varepsilon/2)\bigr)$$

elements of $\mathcal{U}_0^{N-1}$. Finally,

$$\bigcup_{1\leqslant k,l\leqslant K} Z(k,l)$$

is covered by $K^2\exp(N(h-\varepsilon/2)) < \exp(Nh)$ elements of $\mathcal{U}_0^{N-1}$. As every subcover of $\mathcal{U}_0^{N-1}$ has at least $\exp(Nh)$ elements,

$$\bigcup_{1\leqslant k,l\leqslant K} Z(k,l) \neq X.$$

This completes the proof of the lemma. □

PROOF OF THEOREM 7.5. Let $\mathcal{U} = \{U_1, \ldots, U_\ell\}$ be an open cover of X. It is clearly sufficient to consider Borel partitions α of X of the form

$$\alpha = \{A_1, \ldots, A_\ell\} \quad \text{with } A_i \subset U_i \text{ for every } i. \tag{3}$$

Step 1: Assume first that X is 0-dimensional.

The family of partitions finer than $\mathcal{U}$, consisting of clopen sets and satisfying (3) is countable; let $\{\alpha_l\colon\ l \geqslant 1\}$ be an enumeration of this family. According to the previous lemma, there exists a sequence of integers N_K tending to $+\infty$ and a sequence x_K of elements of X such that:

$$H_k\big(\omega(\alpha_l, N_K, x_K)\big) \geqslant k\left(h - \frac{1}{K}\right) \quad \text{for every } 1 \leqslant k, l \leqslant K. \tag{4}$$

Write

$$\mu_K = \frac{1}{N_K}\sum_{i=0}^{N_K-1} \delta_{T^i x_K}.$$

Replacing the sequence μ_K by a subsequence (this means replacing the sequence N_K by a subsequence, and the sequence x_K by the corresponding subsequence preserving the property (4)), one can assume that the sequence of measures μ_K converges weak* to a probability measure μ. This measure μ is clearly T-invariant. Fix $k, l \geqslant 1$, and let F be an atom of the partition $(\alpha_l)_0^{k-1}$, with name $\theta \in \{1, \ldots, \ell\}^k$. For every K one has

$$\big|\mu_K(F) - p\big(\theta|\omega(\alpha_l, N_K, x_K)\big)\big| \leqslant \frac{2k}{N_K}.$$

Now as F is clopen,

$$\mu(F) = \lim_{K\to\infty} \mu_K(F) = \lim_{K\to\infty} p\big(\theta|\omega(\alpha_l, N_K, x_K)\big) \quad \text{hence}$$
$$\phi\big(\mu(F)\big) = \lim_{K\to\infty} \phi\big(p\big(\theta|\omega(\alpha_l, N_K, x_K)\big)\big)$$

and, summing over $\theta \in \{1, \dots, \ell\}^k$, one gets

$$H_\mu\big((\alpha_l)_0^{k-1}\big) = \lim_{K\to\infty} H_k\big(\omega(\alpha_l, N_K, x_K)\big) \geqslant kh.$$

Finally, by sending k to infinity one obtains $h_\mu(\alpha_l) \geqslant h$.

Now, as X is 0-dimensional, the family of partitions $\{\alpha_l\}$ is dense in the collection of Borel partitions of X satisfying (3), with respect to the distance associated with $L^1(\mu)$. Thus, $h_\mu(\alpha) \geqslant h$ for every partition of this kind.

Step 2: The general case.

Let us recall a well known fact: there exists a topological system (Y, T), where Y is 0-dimensional, and a continuous surjective map $\pi : Y \to X$ with $\pi \circ T = T \circ \pi$.

(Proof: as X is a compact metric space, it is easy to construct a Cantor set K and a continuous surjective $f : K \to X$. Put

$$Y = \big\{y \in K^{\mathbb{Z}}\colon\ f(y_{n+1}) = Tf(y_n) \text{ for every } n \in \mathbb{Z}\big\}$$

and let $\pi : Y \to X$ be defined by $\pi(y) = f(y_0)$.

Y is a closed subset of $K^{\mathbb{Z}}$—where the latter is equipped with the product topology—and is invariant under the shift T on $K^{\mathbb{Z}}$. It is easy to check that π satisfies the required conditions.)

Let $\mathcal{V} = \pi^{-1}(\mathcal{U}) = \{\pi^{-1}(U_1), \dots, \pi^{-1}(U_d)\}$ be the preimage of $\mathcal{U}$ under π; one has $h_{\text{top}}(\mathcal{V}) = h_{\text{top}}(\mathcal{U}) = h$. By the above remark, there exists $\nu \in M(Y, T)$ such that $h_\nu(\mathcal{Q}) \geqslant h$ for every Borel partition $\mathcal{Q}$ of Y finer than $\mathcal{V}$. Let $\mu = \nu \circ \pi^{-1}$ the measure which is the image of ν under π. One has $\mu \in M_T(X)$ and, for every Borel partition α of X finer than $\mathcal{U}$, $\pi^{-1}(\alpha)$ is a Borel partition of Y which is finer than $\mathcal{V}$ with

$$h_\mu(\alpha) = h_\nu\big(\pi^{-1}(\alpha)\big) \geqslant h.$$

This completes the proof of the theorem. □

COROLLARY 7.8. *Let (X, T) be a topological system, $\mathcal{U}$ an open cover of X and α a Borel partition finer than $\mathcal{U}$, then, there exists a T-invariant ergodic measure μ on X such that $h_\mu(\alpha) \geqslant h_{\text{top}}(\mathcal{U})$.*

PROOF. By Theorem 7.5 there exists $\mu \in M_T(X)$ with $h_\mu(\alpha) \geqslant h_{\text{top}}(\mathcal{U})$; let $\mu = \int_\omega \mu_\omega \, dm(\omega)$ be its ergodic decomposition. The corollary follows from the formula

$$\int h_{\mu_\omega}(\alpha)\, dm(\omega) = h_\mu(\alpha).$$ □

7.4. *The variational principle for open covers*

Given an open cover $\mathcal{U}$ of the dynamical system (X, T), the results of the previous subsection imply the inequality

$$\sup_{\mu\in M_T(X)} \inf_{\alpha\succ\mathcal{U}} h_\mu(\alpha) \geqslant h_{\text{top}}(\mathcal{U}).$$

We will now present a new result which will provide the fact that

$$\sup_{\mu\in M_T(X)} \inf_{\alpha\succ\mathcal{U}} h_\mu(\alpha) = h_{\text{top}}(\mathcal{U})$$

thus completing the proof of a variational principle for $\mathcal{U}$. In fact we will obtain the explicit formula:

$$h_{\text{top}}(\mathcal{U}) = \max_{\mu\in M_T(X)} \inf_{\alpha\succ\mathcal{U}} h_\mu(\alpha).$$

To the best of our knowledge this is the first time that such an explicit formula is given for the topological entropy of a single open cover.

We first need a universal version of the Rohlin lemma.

PROPOSITION 7.9. *Let (X, T) be a (Polish) dynamical system and assume that there exists on X a T-invariant aperiodic probability measure. Given a positive integer n and a real number $\delta > 0$ there exists a Borel subset $B \subset X$ such that the sets $B, TB, \dots, T^{n-1}B$ are pairwise disjoint and for every aperiodic T-invariant probability measure $\mu \in M_T(X)$ we have $\mu(\bigcup_{j=0}^{n-1} T^j B) > 1 - \delta$.*

PROOF. Fix N (it should be larger than n/δ for the required height n and error δ). The set of points that are periodic with period $\leqslant N$ is closed. Any point in the complement (which by our assumption is non-empty) has, by continuity, a neighborhood U with N disjoint forward iterates. There is a countable subcover $\{U_m\}$ of such sets since the space is Polish. Take $A_1 = U_1$ as a base for a *Kakutani sky-scraper*

$$\begin{aligned}&\left\{T^j A_1^k\colon\ j = 0, \dots, k-1;\ k = 1, 2, \dots\right\},\\ &A_1^k = \left\{x \in A_1\colon\ r_{A_1}(x) = k\right\},\end{aligned}$$

where $r_{A_1}(x)$ is the first integer $j \geqslant 1$ with $T^j x \in A_1$. Next set

$$B_1 = \bigcup_{k\geqslant 1} \bigcup_{j=0}^{[(k-n-1)/n]} T^{jn} A_1^k,$$

so that the sets $B_1, TB_1, \dots, T^{n-1}B_1$ are pairwise disjoint.

Remove the full T orbit of U_1 from the space and repeat to find B_2 using as a base for the next Kakutani sky-scraper A_2 defined as U_2 intersected with the part of X not removed earlier. (Note that, by the proof of Poincaré's recurrence theorem, the set of points which are in the full orbit but not in the forward orbit of U_1 is a wandering set, whence of universal measure zero, and can therefore be ignored.)

Proceed by induction to define the sequence B_i, $i = 1, 2, \ldots$, and set $B = \bigcup_{i=1}^{\infty} B_i$. By Poincaré recurrence for any aperiodic invariant measure we exhaust the whole space except for n iterates of the union A of the bases of the Kakutani sky-scrapers. By construction $A = \bigcup_{m=1}^{\infty} A_m$ has N disjoint iterates so that $\mu(A) \leqslant 1/N$ for every $\mu \in M_T(X)$. Thus $B, TB, \ldots, T^{n-1}B$ fill all but $n/N < \delta$ of the space uniformly over the aperiodic measures $\mu \in M_T(X)$. □

Let (X, T) be a dynamical system and $\mathcal{U} = \{U_1, U_2, \ldots, U_\ell\}$ a finite open cover. We denote by $\mathcal{A}$ the collection of all finite Borel partitions α which refine $\mathcal{U}$, i.e. for every $A \in \alpha$ there is some $U \in \mathcal{U}$ with $A \subset U$. We set

$$\check{h}(\mathcal{U}) = \sup_{\mu \in M_T(X)} \inf_{\alpha \in \mathcal{A}} h_\mu(\alpha) \quad \text{and} \quad \hat{h}(\mathcal{U}) = \inf_{\alpha \in \mathcal{A}} \sup_{\mu \in M_T(X)} h_\mu(\alpha).$$

PROPOSITION 7.10. *Let (X, T) be a dynamical system, $\mathcal{U} = \{U_1, U_2, \ldots, U_\ell\}$ a finite open cover, then*

(1) $\check{h}(\mathcal{U}) \leqslant \hat{h}(\mathcal{U})$,

(2) $\hat{h}(\mathcal{U}) \leqslant h_{\text{top}}(\mathcal{U})$.

PROOF. (1) Given $\nu \in M_T(X)$ and $\alpha \in \mathcal{A}$ we obviously have $h_\nu(\alpha) \leqslant \sup_{\mu \in M_T(X)} h_\mu(\alpha)$. Thus

$$\inf_{\alpha \in \mathcal{A}} h_\nu(\alpha) \leqslant \inf_{\alpha \in \mathcal{A}} \sup_{\mu \in M_T(X)} h_\mu(\alpha) = \hat{h}(\mathcal{U}),$$

and therefore also $\check{h}(\mathcal{U}) \leqslant \hat{h}(\mathcal{U})$.

(2) Choose for $\varepsilon > 0$ an integer N large enough so that there is a subcover $\mathcal{D} \subset \mathcal{U}_0^{N-1} = \bigvee_{j=0}^{N-1} T^{-j}\mathcal{U}$ of cardinality $2^{N(h_{\text{top}}(\mathcal{U})+\varepsilon)}$. Apply Proposition 7.9 to find a set B such that the sets $B, TB, \ldots, T^{N-1}B$ are pairwise disjoint and for every T-invariant Borel probability measure $\mu \in M_T(X)$ we have $\mu(\bigcup_{j=0}^{N-1} T^j B) > 1 - \delta$. Consider $\mathcal{D}_B = \{D \cap B\colon\ D \in \mathcal{D}\}$, the restriction of the cover $\mathcal{D}$ to B, and find a partition β of B which refines $\mathcal{D}_B$. Thus each element $P \in \beta$ has the form

$$P = P_{i_0, i_1, \ldots, i_{N-1}} \subset \left(\bigcap_{j=0}^{N-1} T^{-j} U_{i_j} \right) \cap B,$$

where $\bigcap_{j=0}^{N-1} T^{-j} U_{i_j}$ represents a typical element of $\mathcal{D}$. Next use the partition β of B to define a partition $\alpha = \{A_i\colon\ i = 1, \ldots, \ell\}$ of $\bigcup_{j=0}^{N-1} T^j B$ by assigning to the set A_i all sets

of the form $T^j P_{i_0,i_1,\ldots,i_j,\ldots,i_{N-1}}$ where $i_j = i$ (j can be any number in $[0, N-1]$). On the remainder of the space α can be taken to be any partition refining $\mathcal{U}$.

Now if N is large and δ small enough then

$$h_\mu(\alpha) \leqslant h_{\text{top}}(\mathcal{U}) + 2\varepsilon. \tag{5}$$

Here is a sketch of how one establishes this inequality. For $n \gg N$ we will estimate $H_\mu(\alpha_0^{n-1})$ by counting how many (n,α)-names are needed to cover most of the space. We take $\delta > 0$ so that $\sqrt{\delta} \ll \varepsilon$. Denote $E = B \cup TB \cup \cdots \cup T^{N-1}B$ (so that $\mu(E) > 1-\delta$). Define

$$f(x) = \frac{1}{n}\sum_{i=0}^{n} \mathbf{1}_E(T^i x),$$

and observe that $0 \leqslant f \leqslant 1$ and

$$\int_X f(x)\,d\mu(x) > 1-\delta,$$

since T is measure preserving. Therefore $\int(1-f) < \delta$ and (Markov's inequality)

$$\mu\{x\colon (1-f) \geqslant \sqrt{\delta}\} \leqslant \frac{1}{\sqrt{\delta}}\int (1-f) \leqslant \sqrt{\delta}.$$

It follows that for points x in $G = \{f > 1-\sqrt{\delta}\}$, we have the property that $T^i x \in E$ for most i in $[0,n]$.

Partition G according to the values of i for which $T^i x \in B$. This partition has at most

$$\sum_{j \leqslant \frac{n}{N}} \binom{n}{j} \leqslant \frac{n}{N}\binom{n}{n/N}$$

sets, a number which is exponentially small in n (if N is sufficiently large).

For a fixed choice of these values the times when we are not in E take only $n\sqrt{\delta}$ values and there we have $< l^{n\sqrt{\delta}}$ choices.

Finally when $T^i x \in B$ we have at most $2^{(N(h_{\text{top}}(U)+\varepsilon))}$ names so that the total contribution is $< 2^{(N(h_{\text{top}}(U)+\varepsilon))\frac{n}{N}}$.

Collecting these estimations we find that

$$H\big(\alpha_0^{n-1}\big) < n\big(h_{\text{top}}(U) + 2\varepsilon\big),$$

whence (5). This completes the proof of the proposition. □

We finally obtain:

THEOREM 7.11 (The variational principle for open covers). *Let (X, T) be a dynamical system, $\mathcal{U} = \{U_1, U_2, \dots U_k\}$ a finite open cover and denote by $\mathcal{A}$ the collection of all finite Borel partitions α which refine $\mathcal{U}$, then*

(1) *for every $\mu \in M_T(X)$, $\inf_{\alpha\in\mathcal{A}} h_\mu(\alpha) \leqslant h_{\text{top}}(\mathcal{U})$, and*

(2) *there exists a measure $\mu_0 \in M_T(X)$ with $h_{\mu_0}(\alpha) \geqslant h_{\text{top}}(\mathcal{U})$ for every Borel partition $\alpha \in \mathcal{A}$.*

(3) $$h_{\text{top}}(\mathcal{U}) = \max_{\mu\in M_T(X)} \inf_{\alpha\succ\mathcal{U}} h_\mu(\alpha) = \inf_{\alpha\succ\mathcal{U}} h_{\mu_0}(\alpha).$$

(4) $$h_{\text{top}}(\mathcal{U}) = \check{h}(\mathcal{U}) = \hat{h}(\mathcal{U}).$$

PROOF. (1) This assertion can be formulated by the inequality $\check{h}(\mathcal{U}) \leqslant h_{\text{top}}(\mathcal{U})$ and it follows by combining the two parts of Lemma 7.10.

(2) This is the content of Theorem 7.5.

(3) Combine assertions (1) and (2).

(4) Clearly follows from (3). □

7.5. *Further results connecting topological and measure entropy*

Given a topological dynamical system (X, T) and a measure $\mu \in M_T(X)$, let $\pi: (X, \mathcal{X}, \mu, T) \to (Z, \mathcal{Z}, \eta, T)$ be the *measure-theoretical* Pinsker factor of $(X, \mathcal{X}, \mu, T)$, and let $\mu = \int_Z \mu_z \, d\eta(z)$ be the disintegration of μ over (Z, η). Set

$$\lambda = \int_Z (\mu_z \times \mu_z) \, d\eta(z),$$

the relatively independent joining of μ with itself over η. Finally let $\Lambda_\mu = \text{supp}(\lambda)$ be the topological support of λ in $X \times X$. Although the Pinsker factor is, in general, only defined measure theoretically, the measure λ is a well defined element of $M_{T\times T}(X \times X)$. It was shown in Glasner [32] that $E_\mu = \Lambda_\mu \setminus \Delta$.

THEOREM 7.12. *Let (X, T) be a topological dynamical system and let $\mu \in M_T(X)$.*

(1) *$E_\mu = \Lambda_\mu \setminus \Delta$ and $\Lambda_\mu = E_\mu \cup \{(x, x)\colon x \in \text{supp}(\mu)\}$.*

(2) *$\text{cls}\, E_\mu \subset \Lambda_\mu$.*

(3) *If μ is ergodic with positive entropy then $\text{cls}\, E_\mu = \Lambda_\mu$.*

One consequence of this characterization of the set of μ-entropy pairs is a description of the set of entropy pairs of a product system. Recall that an E-system is a system for which there exists a probability invariant measure with full support.

COROLLARY 7.13. *Let (X_1, T) and (X_2, T) be two topological E-systems then*:

(1) *$E_{X_1\times X_2} = (E_{X_1} \times E_{X_2}) \cup (E_{X_1} \times \Delta_{X_2}) \cup (\Delta_{X_1} \times E_{X_2})$.*

(2) *The product of two UPE systems is UPE.*

Another consequence is:

COROLLARY 7.14. *Let (X,T) be a topological dynamical system, P the proximal relation on X. Then*:

(1) *For every T-invariant ergodic measure μ of positive entropy the set $P \cap E_\mu$ is residual in the G_δ set E_μ of μ entropy pairs.*
(2) *When $E_X \neq \emptyset$ the set $P \cap E_X$ is residual in the G_δ set E_X of topological entropy pairs.*

Given a dynamical system (X,T), a pair $(x,x') \in X \times X$ is called a *Li–Yorke pair* if it is a proximal pair but not an asymptotic pair. A set $S \subseteq X$ is called *scrambled* if any pair of distinct points $\{x,y\} \subseteq S$ is a Li–Yorke pair. A dynamical system (X,T) is called *chaotic in the sense of Li and Yorke* if there is an uncountable scrambled set. In [14] Theorem 7.12 is applied to solve the question whether positive topological entropy implies Li–Yorke chaos as follows.

THEOREM 7.15. *Let (X,T) be a topological dynamical system.*

(1) *If (X,T) admits a T-invariant ergodic measure μ with respect to which the measure preserving system $(X,\mathcal{X},\mu,T)$ is not measure distal then (X,T) is Li–Yorke chaotic.*
(2) *If (X,T) has positive topological entropy then it is Li–Yorke chaotic.*

In [15] Blanchard, Host and Ruette show that in positive entropy systems there are also many asymptotic pairs.

THEOREM 7.16. *Let (X,T) be a topological dynamical system with positive topological entropy. Then*:

(1) *The set of points $x \in X$ for which there is some $x' \neq x$ with (x,x') an asymptotic pair, has measure 1 for every invariant probability measure on X with positive entropy.*
(2) *There exists a probability measure ν on $X \times X$ such that ν a.e. pair (x,x') is Li–Yorke and positively asymptotic; or more precisely for some $\delta > 0$*

$$\lim_{n\to+\infty} d(T^n x, T^n x') = 0, \quad \text{and}$$

$$\liminf_{n\to+\infty} d(T^{-n}x, T^{-n}x') = 0, \qquad \limsup_{n\to+\infty} d(T^{-n}x, T^{-n}x') \geqslant \delta.$$

7.6. *Topological determinism and zero entropy*

Following [55] call a dynamical system (X,T) *deterministic* if every T-factor is also a T^{-1}-factor. In other words every closed equivalence relation $R \subset X \times X$ which has the property $TR \subset R$ also satisfies $T^{-1}R \subset R$. It is not hard to see that an equivalent condition is as follows. For every continuous real valued function $f \in C(X)$ the function $f \circ T^{-1}$ is contained in the smallest closed subalgebra $\mathcal{A} \subset C(X)$ which contains the constant function **1** and the collection $\{f \circ T^n \colon n \geqslant 0\}$. The folklore question whether the

latter condition implies zero entropy was open for awhile. Here we note that the affirmative answer is a direct consequence of Theorem 7.16 (see also [55]).

PROPOSITION 7.17. *Let (X, T) be a topological dynamical system such that there exists a $\delta > 0$ and a pair $(x, x') \in X \times X$ as in Theorem* 7.16(2). *Then (X, T) is not deterministic.*

PROOF. Set

$$R = \{(T^n x, T^n x')\colon n \geqslant 0\} \cup \{(T^n x', T^n x)\colon n \geqslant 0\} \cup \Delta.$$

Clearly R is a closed equivalence relation which is T-invariant but not T^{-1}-invariant. □

COROLLARY 7.18. *A topologically deterministic dynamical system has zero entropy.*

PROOF. Let (X, T) be a topological dynamical system with positive topological entropy; by Theorem 7.16(2) and Proposition 7.17 it is not deterministic. □

Part 2. Meeting grounds

8. Unique ergodicity

The topological system (X, T) is called *uniquely ergodic* if $M_T(X)$ consists of a single element μ. If in addition μ is a full measure (i.e. $\operatorname{supp}\mu = X$) then the system is called *strictly ergodic* (see [1, Chapter 1, Section 4.3]). Since the ergodic measures are characterized as the extreme points of the Choquet simplex $M_T(X)$, it follows immediately that a uniquely ergodic measure is ergodic. For a while it was believed that strict ergodicity—which is known to imply some strong topological consequences (like in the case of $\mathbb{Z}$-systems, the fact that *every* point of X is a generic point and moreover that the convergence of the ergodic sums $\mathbb{A}_n(f)$ to the integral $\int f\,d\mu,\ f \in C(X)$ is *uniform*)—entails some severe restrictions on the measure-theoretical behavior of the system. For example, it was believed that unique ergodicity implies zero entropy. Then, at first some examples were produced to show that this need not be the case. Furstenberg in [24] and Hahn and Katznelson in [46] gave examples of uniquely ergodic systems with positive entropy. Later in 1970 R.I. Jewett surprised everyone with his outstanding result: every weakly mixing measure preserving $\mathbb{Z}$-system has a strictly ergodic model [51]. This was strengthened by Krieger [60] who showed that even the weak mixing assumption is redundant and that the result holds for every ergodic $\mathbb{Z}$-system.

We recall the following well known characterizations of unique ergodicity (see [36, Theorem 4.9]).

PROPOSITION 8.1. *Let (X, T) be a topological system. The following conditions are equivalent.*

(1) *(X, T) is uniquely ergodic.*

(2) *$C(X) = \mathbb{R} + \overline{B}$, where $B = \{g - g \circ T\colon g \in C(X)\}$.*

(3) *For every continuous function* $f \in C(X)$ *the sequence of functions*

$$\mathbb{A}_n f(x) = \frac{1}{n} \sum_{j=0}^{n-1} f(T^j x)$$

converges uniformly to a constant function.

(4) *For every continuous function* $f \in C(X)$ *the sequence of functions* $\mathbb{A}_n(f)$ *converges pointwise to a constant function.*

(5) *For every function* $f \in A$, *for a collection* $A \subset C(X)$ *which linearly spans a uniformly dense subspace of* $C(X)$, *the sequence of functions* $\mathbb{A}_n(f)$ *converges pointwise to a constant function.*

Given an ergodic dynamical system $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ we say that the system $\widehat{\mathbf{X}} = (\widehat{X}, \widehat{\mathcal{X}}, \hat{\mu}, T)$ is a *topological model* (or just a model) for $\mathbf{X}$ if $(\widehat{X}, T)$ is a topological system, $\hat{\mu} \in M_T(\widehat{X})$ and the systems $\mathbf{X}$ and $\widehat{\mathbf{X}}$ are measure theoretically isomorphic. Similarly we say that $\hat{\pi} : \widehat{\mathbf{X}} \to \widehat{\mathbf{Y}}$ is a *topological model* for $\pi : \mathbf{X} \to \mathbf{Y}$ when $\hat{\pi}$ is a topological factor map and there exist measure theoretical isomorphisms ϕ and ψ such that the diagram

$$\begin{array}{ccc} \mathbf{X} & \xrightarrow{\phi} & \widehat{\mathbf{X}} \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\hat{\pi}} \\ \mathbf{Y} & \xrightarrow[\psi]{} & \widehat{\mathbf{Y}} \end{array}$$

is commutative.

9. The relative Jewett–Krieger theorem

In this section we will prove the following generalization of the Jewett–Krieger theorem (see [1, Chapter 1, Theorem 4.3.10]).

THEOREM 9.1. *If* $\pi : \mathbf{X} = (X, \mathcal{X}, \mu, T) \to \mathbf{Y} = (Y, \mathcal{Y}, \nu, T)$ *is a factor map with* $\mathbf{X}$ *ergodic and* $\widehat{\mathbf{Y}}$ *is a uniquely ergodic model for* $\mathbf{Y}$ *then there is a uniquely ergodic model* $\widehat{\mathbf{X}}$ *for* $\mathbf{X}$ *and a factor map* $\hat{\pi} : \widehat{\mathbf{X}} \to \hat{\mathbf{Y}}$ *which is a model for* $\pi : \mathbf{X} \to \mathbf{Y}$.

In particular, taking $\mathbf{Y}$ to be the trivial one point system we get:

THEOREM 9.2. *Every ergodic system has a uniquely ergodic model.*

Several proofs have been given of this theorem, e.g., see [18] and [8]. We will sketch a proof which will serve the relative case as well.

PROOF OF THEOREM 9.1. A key notion for this proof is that of a *uniform* partition whose importance in this context was emphasized by G. Hansel and J.-P. Raoult [47].

DEFINITION 9.3. A set $B \in \mathcal{X}$ is uniform if

$$\lim_{N\to\infty} \text{ess-sup}_x \left| \frac{1}{N} \sum_{0}^{N-1} 1_B(T^i x) - \mu(B) \right| = 0.$$

A partition $\mathcal{P}$ is uniform if, for all N, every set in $\bigvee_{-N}^{N} T^{-i}\mathcal{P}$ is uniform.

The connection between uniform sets, partitions and unique ergodicity lies in Proposition 8.1. It follows easily from that proposition that if $\mathcal{P}$ is a uniform partition, say into the sets $\{P_1, P_2, \ldots, P_a\}$, and we denote by $\mathcal{P}$ also the mapping that assigns to $x \in X$, the index $1 \leqslant i \leqslant a$ such that $x \in P_i$, then we can map X to $\{1, 2, \ldots, a\}^{\mathbb{Z}} = A^{\mathbb{Z}}$ by:

$$\pi(x) = \big(\ldots, \mathcal{P}\big(T^{-1}x\big), \mathcal{P}(x), \mathcal{P}(Tx), \ldots, \mathcal{P}(T^n x), \ldots\big).$$

Pushing forward the measure μ by π, gives $\pi \circ \mu$ and the closed support of this measure will be a closed shift invariant subset, say $E \subset A^{\mathbb{Z}}$. Now the indicator functions of finite cylinder sets span the continuous functions on E, and the fact that $\mathcal{P}$ is a uniform partition and Proposition 8.1 combine to establish that $(E,$ shift) is uniquely ergodic. This will not be a model for $(X, \mathcal{X}, \mu, T)$ unless $\bigvee_{-\infty}^{\infty} T^{-i}\mathcal{P} = \mathcal{X}$ modulo null sets, but in any case this does give a model for a non-trivial factor of X.

Our strategy for proving Theorem 9.2 is to first construct a single non-trivial uniform partition. Then this partition will be refined more and more via uniform partitions until we generate the entire σ-algebra $\mathcal{X}$. Along the way we will be showing how one can prove a relative version of the basic Jewett–Krieger theorem. Our main tool is the use of Rohlin towers. These are sets $B \in \mathcal{X}$ such that for some N, $B, TB, \ldots, T^{N-1}B$ are disjoint while $\bigcup_0^{N-1} T^i B$ fill up most of the space. Actually we need Kakutani–Rohlin towers, which are like Rohlin towers but fill up the whole space. If the transformation does not have rational numbers in its point spectrum this is not possible with a single height, but two heights that are relatively prime, like N and $N+1$ are certainly possible. Here is one way of doing this. The ergodicity of $(X, \mathcal{X}, \mu, T)$ with μ non-atomic easily yields, for any n, the existence of a positive measure set B, such that

$$T^i B \cap B = \emptyset, \quad i = 1, 2, \ldots, n.$$

With N given, choose $n \geqslant 10 \cdot N^2$ and find B that satisfies the above. It follows that the return time

$$r_B(x) = \inf\{i > 0\colon T^i x \in B\}$$

is greater than $10 \cdot N^2$ on B. Let

$$B_\ell = \big\{x\colon r_B(x) = \ell\big\}.$$

Since ℓ is large (if B_ℓ is non-empty) one can write ℓ as a positive combination of N and $N \times 1$, say

$$\ell = Nu_\ell + (N+1)v_\ell.$$

Now divide the column of sets $\{T^i B_\ell\colon\ 0 \leqslant i < \ell\}$ into u_ℓ-blocks of size N and v_ℓ-blocks of size $N+1$ and mark the first layer of each of these blocks as belonging to C. Taking the union of these marked levels ($T^i B_\ell$ for suitably chosen i) over the various columns gives us a set C such that r_C takes only two values—either N or $N+1$ as required.

It will be important for us to have at our disposal K–R towers like this such that the columns of say the second K–R tower are composed of entire subcolumns of the earlier one. More precisely we want the base C_2 to be a subset of C_1—the base of the first tower. Although we are not sure that this can be done with just two column heights we can guarantee a bound on the number of columns that depends only on the maximum height of the first tower. Let us define formally:

DEFINITION 9.4. A set C will be called the base of a *bounded* K–R tower if for some N, $\bigcup_0^{N-1} T^i C = X$ up to a μ-null set. The least N that satisfies this will be called the *height* of C, and partitioning C into sets of constancy of r_C and viewing the whole space X as a tower over C will be called the K–R tower with columns the sets $\{T^i C_\ell\colon\ 0 \leqslant i < \ell\}$ for $C_\ell = \{x \in C\colon\ r_C(x) = \ell\}$.

Our basic lemma for nesting these K–R towers is:

LEMMA 9.5. *Given a bounded K–R tower with base C and height N, for any n sufficiently large there is a bounded K–R tower with base D contained in C whose column heights are all at least n and at most $n + 4N$.*

PROOF. We take an auxiliary set B such that $T^i\ B \cap B = \emptyset$ for all $0 < i < 10(n+2N)^2$ and look at the unbounded (in general) K–R tower over B. Using the ergodicity it is easy to arrange that $B \subset C$. Now let us look at a single column over B_m, with $m \geqslant 10(n+2N)^2$. We try to put down blocks of size $n+2N$ and $n+2N+1$, to fill up the tower. This can certainly be done but we want our levels to belong to C. We can refine the column over B_m into a finite number of columns so that each level is either entirely within C or in $X \setminus C$. This is done by partitioning the base C according to the finite partition:

$$\bigcap_{i=0}^{m-1} T^{-1}\{C,\ X \setminus C\}.$$

Then we move the edge of each block to the nearest level that belongs to C. The fact that the height of C is N means that we do not have to move any level more than $N-1$ steps, and so at most we lose $2N-2$ or gain that much thus our blocks, with bases now all in C, have size in the interval $[n,\ n+4N]$ as required. □

It is clear that this procedure can be iterated to give an infinite sequence of nested K–R towers with a good control on the variation in the heights of the columns. These can be used to construct uniform partitions in a pretty straightforward way, but we need one more lemma which strengthens slightly the ergodic theorem. We will want to know that when we look at a bounded K–R tower with base C and with minimum column height sufficiently large that for most of the fibers of the towers (that is for $x \in C,\ \{T^i x\colon\ 0 \leqslant i < r_C(x)\}$) the ergodic averages of some finite set of functions are close to the integrals of the functions. It would seem that there is a problem because the base of the tower is a set of very small measure (less than 1/min column height) and it may be that the ergodic theorem is not valid there. However, a simple averaging argument using an intermediate size gets around this problem. Here is the result which we formulate for simplicity for a single function f:

LEMMA 9.6. *Let f be a bounded function and $(X, \mathcal{X}, \mu, T)$ ergodic. Given $\varepsilon > 0$, there is an n_0, such that if a bounded K–R tower with base C has minimum column height at least n_0, then those fibers over $x \in C$: $\{T^i x\colon\ 0 \leqslant i < r_C(x)\}$ that satisfy*

$$\left| \frac{1}{r_C(x)} \sum_{i=0}^{r_C(x)-1} f(T^i x) - \int_X f\, d\mu \right| < \varepsilon$$

fill up at least $1 - \varepsilon$ of the space.

PROOF. Assume without loss of generality that $|f| \leqslant 1$. For a δ to be specified later find an N such that the set of $y \in X$ which satisfy

$$\left| \frac{1}{N} \sum_{0}^{N-1} f(T^i y) - \int f\, d\mu \right| < \delta \tag{6}$$

has measure at least $1 - \delta$. Let us denote the set of y that satisfy (6) by E. Suppose now that n_0 is large enough so that N/n_0 is negligible—say at most δ. Consider a bounded K–R tower with base C and with minimum column height greater than n_0. For each fiber of this tower, let us ask what is the fraction of its points that lie in E. Those fibers with at least a $\sqrt{\delta}$ fraction of its points not in E cannot fill up more than a $\sqrt{\delta}$ fraction of the space, because $\mu(E) > 1 - \delta$.

Fibers with more than $1 - \sqrt{\delta}$ of its points lying in E can be divided into disjoint blocks of size N that cover all the points that lie in E. This is done by starting at $x \in C$, and moving up the fiber, marking the first point in E, skipping N steps and continuing to the next point in E until we exhaust the fiber. On each of these N-blocks the average of f is within δ of its integral, and since $|f| \leqslant 1$ if $\sqrt{\delta} < \varepsilon/10$ this will guarantee that the average of f over the whole fiber is within ε of its integra. □

We are now prepared to construct uniform partitions. Start with some fixed non-trivial partition $\mathcal{P}_0$. By Lemma 9.6, for any tall enough bounded K–R tower at least $9/10$ of the columns will have the 1-block distribution of each $\mathcal{P}_0$-name within $\frac{1}{10}$ of the actual distribution. We build a bounded K–R tower with base $C_1(1)$ and heights $N_1,\ N_1 + 1$ with

N_1 large enough for this to be valid. It is clear that we can modify $\mathcal{P}_0$ to $\mathcal{P}_1$ on the bad fibers so that now all fibers have a distribution of 1-blocks within $\frac{1}{10}$ of a fixed distribution. We call this new partition $\mathcal{P}_1$. Our further changes in $\mathcal{P}_1$ will not change the $N_1,\ N_1+1$ blocks that we see on fibers of a tower over our ultimate C_1. Therefore, we will get a uniformity on all blocks of size $100N_1$. The 100 is to get rid of the edge effects since we only know the distribution across fibers over points in $C_1(1)$.

Next we apply Lemma 9.6 to the 2-blocks in $\mathcal{P}_1$ with $1/100$. We choose N_2 so large that N_1/N_2 is negligible and so that any uniform K–R tower with height at least N_2 has for at least $99/100$ of its fibers a distribution of 2-blocks within $1/100$ of the global $\mathcal{P}_1$ distribution. Apply Lemma 9.5 to find a uniform K–R tower with base $C_2(2) \subset C_1(1)$ such that its column heights are between N_2 and $N_2 + 4N_1$. For the fibers with good $\mathcal{P}_1$ distribution we make no change. For the others, we copy on most of the fiber (except for the top $10 \cdot N_1^2$ levels) the corresponding $\mathcal{P}_1$-name from one of the good columns. In this copying we also copy the $C_1(1)$-name so that we preserve the blocks. The final $10 \cdot N_1^2$ spaces are filled in with $N_1,\ N_1+1$ blocks. This gives us a new base for the first tower that we call $C_1(2)$, and a new partition $\mathcal{P}_2$. The features of $\mathcal{P}_2$ are that all its fibers over $C_1(2)$ have good (up to $1/10$) 1-block distribution, and all its fibers over $C_2(2)$ have good (up to $1/100$) 2-block distributions. These will not change in the subsequent steps of the construction.

Note too that the change from $C_1(1)$, to $C_1(2)$, could have been made arbitrarily small by choosing N_2 sufficiently large.

There is one problem in trying to carry out the next step and that is, the filling in of the top relatively small portion of the bad fibers after copying most of a good fiber. We cannot copy an exact good fiber because it is conceivable that no fiber with the precise height of the bad fiber is good. The filling in is possible if the column heights of the previous level are relatively prime. This was the case in step 2, because in step 1 we began with a K–R tower heights N_1, N_1+1. However, Lemma 9.5 does not guarantee relatively prime heights. This is automatically the case if there is no rational spectrum. If there are only a finite number of rational points in the spectrum then we could have made our original columns with heights $LN_1,\ L(N_1+1)$ with L being the highest power so that T^L is not ergodic and then worked with multiples of L all the time. If the rational spectrum is infinite then we get an infinite group rotation factor and this gives us the required uniform partition without any further work.

With this point understood it is now clear how one continues to build a sequence of partitions $\mathcal{P}_n$ that converge to $\mathcal{P}$ and $C_i(k) \to C_i$ such that the $\mathcal{P}$-names of all fibers over points in C_i have a good (up to $1/10^i$) distribution of i-blocks. This gives the uniformity of the partition $\mathcal{P}$ as required and establishes

PROPOSITION 9.7. *Given any $\mathcal{P}_0$ and any $\varepsilon > 0$ there is a uniform partition $\mathcal{P}$ such that $d(\mathcal{P}_0, \mathcal{P}) < \varepsilon$ in the ℓ_1-metric on partitions.*

As we have already remarked the uniform partition that we have constructed gives us a uniquely ergodic model for the factor system generated by this partition. We need now a relativized version of the construction we have just carried out. We formulate this as follows:

PROPOSITION 9.8. *Given a uniform partition $\mathcal{P}$ and an arbitrary partition $\mathcal{Q}_0$ that refines $\mathcal{P}$, for any $\varepsilon > 0$ there is a uniform partition $\mathcal{Q}$ that also refines $\mathcal{P}$ and satisfies*

$$\|\mathcal{Q}_0 - \mathcal{Q}\|_1 < \varepsilon.$$

Even though we write things for finite alphabets, everything makes good sense for countable partitions as well and the arguments need no adjusting. However, the metric used to compare partitions becomes important since not all metrics on ℓ_1 are equivalent. We use always:

$$\|\mathcal{Q} - \overline{\mathcal{Q}}\|_1 = \sum_j \int_X |1_{Q_j} - 1_{\overline{Q}_j}|\, d\mu,$$

where the partitions $\mathcal{Q}$ and $\overline{\mathcal{Q}}$ are ordered partitions into sets $\{Q_j\}$, $\{\overline{Q}_j\}$ respectively. We also assume that the σ-algebra generated by the partition $\mathcal{P}$ is non-atomic—otherwise there is no real difference between what we did before and what has to be done here.

We will try to follow the same proof as before. The problem is that when we redefine $\mathcal{Q}_0$ to $\mathcal{Q}$ we are not allowed to change the $\mathcal{P}$-part of the name of points. That greatly restricts us in the kind of names we are allowed to copy on columns of K–R towers and it is not clear how to proceed. The way to overcome the difficulty is to build the K–R towers inside the uniform algebra generated by $\mathcal{P}$. This being done we look, for example, at our first tower and the first change we wish to make in $\mathcal{Q}_0$. We divide the fibers into a finite number of columns according to the height and according to the $\mathcal{P}$-name.

Next each of these is divided into subcolumns, called $\mathcal{Q}_0$-columns, according to the $\mathcal{Q}_0$-names of points. If a $\mathcal{P}$-column has some good (i.e. good 1-block distribution of $\mathcal{Q}_0$-names) $\mathcal{Q}_0$-subcolumn it can be copied onto all the ones that are not good. Next notice that a $\mathcal{P}$-column that contains not even one good $\mathcal{Q}_0$-name is a set defined in the uniform algebra. Therefore if these sets have small measure then for some large enough N, uniformly over the whole space, we will not encounter these bad columns too many times.

In brief the solution is to change the nature of the uniformity. We do not make all of the columns of the K–R tower good—but we make sure that the bad ones are seen infrequently, uniformly over the whole space. With this remark the proof of the proposition is easily accomplished using the same nested K–R towers as before—*but inside the uniform algebra*.

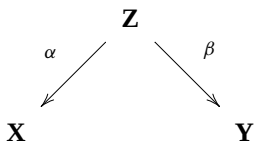
Finally the J–K theorem is established by constructing a refining sequence of uniform partitions and looking at the inverse limit of the corresponding topological spaces. Notice that if $\mathcal{Q}$ refines $\mathcal{P}$, and both are uniform, then there is a natural homeomorphism from $X_{\mathcal{Q}}$ onto $X_{\mathcal{P}}$. The way in which the theorem is established also yields a proof of the relative J–K theorem, Theorem 9.1. □

Using similar methods E. Lehrer [62] shows that in the Jewett–Krieger theorem one can find, for any ergodic system, a strictly ergodic model which is topologically mixing.

10. Models for other commutative diagrams

One can describe Theorem 9.1 as asserting that every diagram of ergodic systems of the form $\mathbf{X} \to \mathbf{Y}$ has a strictly ergodic model. What can we say about more complicated commutative diagrams? A moments reflection will show that a repeated application of Theorem 9.1 proves the first assertion of the following theorem.

THEOREM 10.1. *Any commutative diagram in the category of ergodic $\mathbb{Z}$ dynamical systems with the structure of an inverted tree, i.e. no portion of it looks like*



(7)

has a strictly ergodic model. On the other hand there exists a diagram of the form (7) *that does not admit a strictly ergodic model.*

For the proof of the second assertion we need the following theorem.

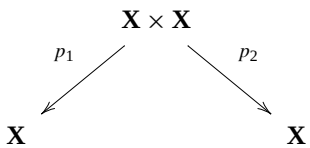
THEOREM 10.2. *If (Z, η, T) is a strictly ergodic system and $(Z, T) \xrightarrow{\alpha} (X, T)$ and $(Z, T) \xrightarrow{\beta} (Y, T)$ are topological factors such that $\alpha^{-1}(U) \cap \beta^{-1}(V) \neq \emptyset$ whenever $U \subset X$ and $V \subset Y$ are non-empty open sets, then the measure-preserving systems $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ and $\mathbf{Y} = (Y, \mathcal{Y}, \nu, T)$ are measure-theoretically disjoint. In particular this is the case if the systems (X, T) and (Y, T) are topologically disjoint.*

PROOF. It suffices to show that the map $\alpha \times \beta : Z \to X \times Y$ is onto since this will imply that the topological system $(X \times Y, T)$ is strictly ergodic. We establish this by showing that the measure $\lambda = (\alpha \times \beta)_*(\eta)$ (a joining of μ and ν) is full; i.e. that it assigns positive measure to every set of the form $U \times V$ with U and V as in the statement of the theorem. In fact, since by assumption η is full we have

$$\lambda(U \times V) = \eta\big((\alpha \times \beta)^{-1}(U \times V)\big) = \eta\big(\alpha^{-1}(U) \cap \beta^{-1}(V)\big) > 0.$$

This completes the proof of the first assertion. The second follows since topological disjointness of (X, T) and (Y, T) implies that $\alpha \times \beta : Z \to X \times Y$ is onto. □

PROOF OF THEOREM 10.1. We only need to prove the last assertion. Take $\mathbf{X} = \mathbf{Y}$ to be any non-trivial weakly mixing system, then $\mathbf{X} \times \mathbf{X}$ is ergodic and the diagram



(8)

is our counter example. In fact if (7) is a uniquely ergodic model in this situation then it is easy to establish that the condition in Theorem 10.2 is satisfied and we apply this theorem to conclude that **X** is disjoint from itself. Since in a non-trivial system $\mu \times \mu$ and $\mathrm{gr}(\mu, \mathrm{id})$ are different ergodic joinings, this contradiction proves our assertion. □

11. The Furstenberg–Weiss almost 1-1 extension theorem

It is well known that in a topological measure space one can have sets that are large topologically but small in the sense of the measure. In topological dynamics when (X, T) is a factor of (Y, T) and the projection $\pi : Y \to X$ is one to one on a topologically large set (i.e. the complement of a set of first category), one calls (Y, T) an *almost* 1-1 *extension* of (X, T) and considers the two systems to be very closely related. Nonetheless, in view of the opening sentence, it is possible that the measure theory of (Y, T) will be quite different from the measure theory of (X, T). The following theorem realizes this possibility in an extreme way (see [29]).

THEOREM 11.1. *Let (X, T) be a non-periodic minimal dynamical system, and let $\pi : Y \to X$ be an extension of (X, T) with (Y, T) topologically transitive and Y a compact metric space. Then there exists an almost* 1-1 *minimal extension, $\bar{\pi} : (\overline{Y}, T) \to (X, T)$ and a Borel subset $Y_0 \subset Y$ with a Borel measurable map $\theta : Y_0 \to \overline{Y}$ satisfying* (1) $\theta T = T\theta$, (2) $\bar{\pi}\theta = \pi$, (3) θ *is* 1-1 *on* Y_0, (4) $\mu(Y_0) = 1$ *for any T-invariant measure μ on Y.*

In words, one can find an almost 1-1 minimal extension of X such that the measure theoretic structure is as rich as that of an arbitrary topologically transitive extension of X.

An almost 1-1 extension of a minimal equicontinuous system is called an *almost automorphic* system. The next corollary demonstrates the usefulness of this modelling theorem. Other applications appeared, e.g., in [41] and [20].

COROLLARY 11.2. *Let $(X, \mathcal{X}, \mu, T)$ be an ergodic measure preserving transformation with infinite point spectrum defined by (G, ρ) where G is a compact monothetic group $G = \overline{\{\rho^n\}}_{n\in\mathbb{Z}}$. Then there is an almost* 1-1 *minimal extension of (G, ρ) (i.e. a minimal almost automorphic system), $(\widetilde{Z}, \sigma)$ and an invariant measure ν on Z such that (Z, σ, ν) is isomorphic to $(X, \mathcal{X}, \mu, T)$.*

12. Cantor minimal representations

A *Cantor minimal dynamical system* is a minimal topological system (X, T) where X is the Cantor set. Two Cantor minimal systems (X, T) and (Y, S) are called *orbit equivalent* (OE) if there exists a homeomorphism $F : X \to Y$ such that $F(\mathcal{O}_T(x)) = \mathcal{O}_S(Fx)$ for every $x \in X$. Equivalently: there are functions $n : X \to \mathbb{Z}$ and $m : X \to \mathbb{Z}$ such that for every $x \in X$ $F(Tx) = S^{n(x)}(Fx)$ and $F(T^{m(x)}) = S(Fx)$. An old result of M. Boyle implies that the requirement that, say, the function $n(x)$ be continuous already implies that the two systems are *flip conjugate*; i.e. (Y, S) is isomorphic either to (X, T) or to (X, T^{-1}).

However, if we require that both $n(x)$ and $m(x)$ have at most one point of discontinuity we get the new and, as it turns out, useful notion of *strong orbit equivalence* (SOE). A complete characterization of both OE and SOE of Cantor minimal systems was obtained by Giordano Putnam and Skau [31] in terms of an algebraic invariant of Cantor minimal systems called the *dimension group*. (See Glasner and Weiss [43] for further results, and Glasner [35] for a review of the subject of orbit equivalence in Cantor minimal dynamical systems.)

We conclude this section with the following remarkable theorems, due to N. Ormes [68], which simultaneously generalize the theorems of Jewett and Krieger and a theorem of Downarowicz [19] which, given any Choquet simplex Q, provides a Cantor minimal system (X, T) with $M_T(X)$ affinely homeomorphic with Q. (See also Downarowitcz and Serafin [21], and Boyle and Downarowicz [16].)

THEOREM 12.1.

(1) *Let $(\Omega, \mathcal{B}, \nu, S)$ be an ergodic, non-atomic, probability measure preserving, dynamical system. Let (X, T) be a Cantor minimal system such that whenever* $\exp(2\pi i/p)$ *is a* (*topological*) *eigenvalue of (X, T) for some $p \in \mathbb{N}$ it is also a* (*measurable*) *eigenvalue of $(\Omega, \mathcal{B}, \nu, S)$. Let μ be any element of the set of extreme points of $M_T(X)$. Then, there exists a homeomorphism $T' : X \to X$ such that* (i) *T and T' are strong orbit equivalent,* (ii) *$(\Omega, \mathcal{B}, \nu, S)$ and $(X, \mathcal{X}, \mu, T')$ are isomorphic as measure preserving dynamical systems.*

(2) *Let $(\Omega, \mathcal{B}, \nu, S)$ be an ergodic, non-atomic, probability measure preserving, dynamical system. Let (X, T) be a Cantor minimal system and μ any element of the set of extreme points of $M_T(X)$. Then, there exists a homeomorphism $T' : X \to X$ such that* (i) *T and T' are orbit equivalent,* (ii) *$(\Omega, \mathcal{B}, \nu, S)$ and $(X, \mathcal{X}, \mu, T')$ are isomorphic as measure preserving dynamical systems.*

(3) *Let $(\Omega, \mathcal{B}, \nu, S)$ be an ergodic, non-atomic, probability measure preserving dynamical system. Let Q be any Choquet simplex and q an extreme point of Q. Then there exists a Cantor minimal system (X, T) and an affine homeomorphism $\phi : Q \to M_T(X)$ such that, with $\mu = \phi(q)$, $(\Omega, \mathcal{B}, \nu, S)$ and $(X, \mathcal{X}, \mu, T)$ are isomorphic as measure preserving dynamical systems.*

13. Other related theorems

Let us mention a few more striking representation results.

For the first one recall that a topological dynamical system (X, T) is said to be *prime* if it has no non-trivial factors. A similar definition can be given for measure preserving systems. There it is easy to see that a prime system $(X, \mathcal{X}, \mu, T)$ must have zero entropy. It follows from a construction in [76] that the same holds for topological entropy, namely any system (X, T) with positive topological entropy has non-trivial factors. In [85] it is shown that any ergodic zero entropy dynamical system has a minimal model (X, T) with the property that any pair of points (u, v) not on the same orbit has a dense orbit in $X \times X$. Such minimal systems are necessarily prime, and thus we have the following result:

THEOREM 13.1. *An ergodic dynamical system has a topological, minimal, prime model iff it has zero entropy.*

The second theorem, due to Glasner and Weiss [41], treats the positive entropy systems.

THEOREM 13.2. *An ergodic dynamical system has a strictly ergodic, UPE model iff it has positive entropy.*

We also have the following surprising result which is due to Weiss [84].

THEOREM 13.3. *There exists a minimal metric dynamical system (X, T) with the property that for every ergodic probability measure preserving system $(\Omega, \mathcal{B}, \mu, S)$ there exists a T-invariant Borel probability measure ν on X such that the systems $(\Omega, \mathcal{B}, \mu, S)$ and $(X, \mathcal{X}, \nu, T)$ are isomorphic.*

In [65] E. Lindenstrauss proves the following:

THEOREM 13.4. *Every ergodic measure distal $\mathbb{Z}$-system $\mathbf{X} = (X, \mathcal{X}, \mu, T)$ can be represented as a minimal topologically distal system (X, T, μ) with $\mu \in M_T^{\mathrm{erg}}(X)$.*

This topological model need not, in general, be uniquely ergodic. In other words there are measure distal systems for which no uniquely ergodic topologically distal model exists.

PROPOSITION 13.5.

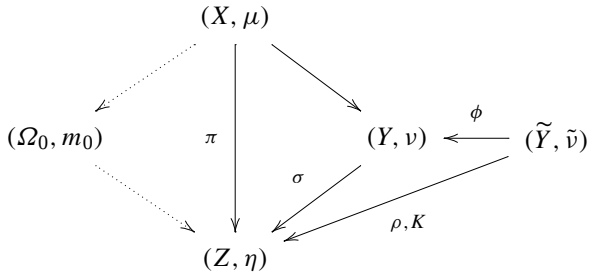
(1) *There exists an ergodic non-Kronecker measure distal system $(\Omega, \mathcal{F}, m, T)$ with non-trivial maximal Kronecker factor $(\Omega_0, \mathcal{F}_0, m_0, T)$ such that* (i) *the extension $(\Omega, \mathcal{F}, m, T) \to (\Omega_0, \mathcal{F}_0, m_0, T)$ is finite to one a.e. and* (ii) *every non-trivial factor map of $(\Omega_0, \mathcal{F}_0, m_0, T)$ is finite to one.*

(2) *A system $(\Omega, \mathcal{F}, m, T)$ as in part* (1) *does not admit a topologically distal strictly ergodic model.*

PROOF. (1) Irrational rotations of the circle as well as adding machines are examples of Kronecker systems satisfying condition (ii). There are several constructions in the literature of ergodic, non-Kronecker, measure distal, two point extensions of these Kronecker systems. A well known explicit example is the strictly ergodic Morse minimal system.

(2) Assume to the contrary that (X, μ, T) is a distal strictly ergodic model for $(\Omega, \mathcal{F}, m, T)$. Let (Z, T) be the maximal equicontinuous factor of (X, T) and let η be the unique invariant probability measure on Z. Since by assumption (X, μ, T) is not Kronecker it follows that $\pi : X \to Z$ is not one to one. By Furstenberg's structure theorem for minimal distal systems (Z, T) is non-trivial and moreover there exists an intermediate extension $X \to Y \xrightarrow{\sigma} Z$ such that σ is an isometric extension. A well known construction implies the existence of a minimal group extension $\rho : (\widetilde{Y}, T) \to (Z, T)$, with compact fiber group K, such that the following diagram is commutative (see Section 5 above). We denote by ν the unique invariant measure on Y (the image of μ) and let $\tilde{\nu}$ be an ergodic

measure on $\widetilde{Y}$ which projects onto ν. The dotted arrows denote measure theoretic factor maps.



Next form the measure $\theta = \int_K R_k \tilde{\nu}\, dm_K$, where m_K is Haar measure on K and for each $k \in K$, R_k denotes right translation by k on $\widetilde{Y}$ (an automorphism of the system $(\widetilde{Y}, T)$). We still have $\phi(\theta) = \nu$.

A well known theorem in topological dynamics (see [75]) implies that a minimal distal finite to one extension of a minimal equicontinuous system is again equicontinuous and since (Z, T) is the maximal equicontinuous factor of (X, T) we conclude that the extension $\sigma : Y \to Z$ is not finite to one. Now the fibers of the extension σ are homeomorphic to a homogeneous space K/H, where H is a closed subgroup of K. Considering the measure disintegration $\theta = \int_Z \theta_z\, d\eta(z)$ of θ over η and its projection $\nu = \int_Z \nu_z\, d\eta(z)$, the disintegration of ν over η, we see that a.e. $\theta_z \equiv m_K$ and $\nu_z \equiv m_{K/H}$. Since K/H is infinite we conclude that the *measure theoretical extension* $\sigma : (Y, \nu) \to (Z, \eta)$ is not finite to one. However considering the dotted part of the diagram we arrive at the opposite conclusion. This conflict concludes the proof of the proposition. □

In [69] Ornstein and Weiss introduced the notion of tightness for measure preserving systems and the analogous notion of mean distality for topological systems.

DEFINITION 13.6. Let (X, T) be a topological system.

(1) A pair (x, y) in $X \times X$ is *mean proximal* if for some (hence any) compatible metric d

$$\limsup_{n\to\infty} \frac{1}{2n+1} \sum_{i=-n}^{n} d(T^i x, T^i y) = 0.$$

If this lim sup is positive the pair is called *mean distal*.

(2) The system (X, T) is *mean distal* if every pair with $x \neq y$ is mean distal.

(3) Given a T-invariant probability measure μ on X, the triple (X, μ, T) is called *tight* if there is a μ-conull set $X_0 \subset X$ such that every pair of distinct points (x, y) in $X_0 \times X_0$ is mean distal.

Ornstein and Weiss show that tightness is in fact a property of the measure preserving system (X, μ, T) (i.e. if the measure system $(X, \mathcal{X}, \mu, T)$ admits one tight model then every topological model is tight). They obtain the following results.

THEOREM 13.7.
(1) *If the entropy of* (X, μ, T) *is positive and finite then* (X, μ, T) *is not tight.*
(2) *There exist strictly ergodic non-tight systems with zero entropy.*

Surprisingly the proof in [69] of the non-tightness of a positive entropy system does not work in the case when the entropy is infinite which is still open.

J. King gave an example of a tight system with a non-tight factor. Following this he and Weiss [69] established the following result. Note that this theorem implies that tightness and mean distality are not preserved by factors.

THEOREM 13.8. *If* $(X, \mathcal{X}, \mu, T)$ *is ergodic with zero entropy then there exists a mean-distal system* (Y, ν, S) *which admits* $(X, \mathcal{X}, \mu, T)$ *as a factor.*

References

Survey in volume 1A

[1] B. Hasselblatt and A. Katok, *Principle structures*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 1–203.

Other sources

[2] J. Aaronson and B. Weiss, *Remarks on tightness of cocycles*, Colloq. Math. **84/85** (Part 2) (2000), 363–376.
[3] R.L. Adler, A.G. Konheim and M.H. McAndrew, *Topological entropy*, Trans. Amer. Math. Soc. **114** (1965), 309–319.
[4] E. Akin, *Recurrence in Topological Dynamics, Furstenberg Families and Ellis Actions*, Plenum Press, New York (1997).
[5] E. Akin and E. Glasner, *Topological ergodic decomposition and homogeneous flows*, Contemp. Math. **215** (1998), 43–52.
[6] E. Akin and E. Glasner, *Residual properties and almost equicontinuity*, J. Anal. Math. **84** (2001), 243–286.
[7] L. Auslander, L. Green and F. Hahn, *Flows on Homogeneous Spaces*, Ann. of Math. Studies, Vol. 53, Princeton University Press, Princeton, NJ (1963).
[8] A. Bellow and H. Furstenberg, *An application of number theory to ergodic theory and the construction of uniquely ergodic models*, Israel J. Math. **33** (1979), 231–240.
[9] F. Blanchard, *Fully positive topological entropy and topological mixing*, Symbolic Dynamics and Applications, Contemp. Math., Vol. 135, Amer. Math. Soc., Providence (1992), 95–105.
[10] F. Blanchard, *A disjointness theorem involving topological entropy*, Bull. Soc. Math. France **121** (1993), 465–478.
[11] F. Blanchard, B. Host, A. Maass, S. Martínez and D. Rudolph, *Entropy pairs for a measure*, Ergodic Theory Dynamical Systems **15** (1995), 621–632.
[12] F. Blanchard, E. Glasner and B. Host, *A variation on the variational principle and applications to entropy pairs*, Ergodic Theory Dynamical Systems **17** (1997), 29–43.

[13] F. Blanchard, B. Host and A. Maass, *Topological complexity*, Ergodic Theory Dynamical Systems **20** (2000), 641–662.
[14] F. Blanchard, E. Glasner, S. Kolyada and A. Maass, *On Li–Yorke pairs*, J. Reine Angew. Math. **547** (2002), 51–68.
[15] F. Blanchard, B. Host and S. Ruette, *Asymptotic pairs in positive-entropy systems*, Ergodic Theory Dynamical Systems **22** (2002), 671–686.
[16] M. Boyle and T. Downarowicz, *The entropy theory of symbolic extensions*, Invent. Math. **156** (2004), 119–161.
[17] R.V. Chacón, *Weakly mixing transformations which are not strongly mixing*, Proc. Amer. Math. Soc. **22** (1969), 559–562.
[18] M. Denker, C. Grillenberger and K. Sigmund, *Ergodic Theory on Compact Spaces*, Lecture Notes in Math., Vol. 527, Springer (1976).
[19] T. Downarowicz, *The Choquet simplex of invariant measures for minimal flows*, Israel J. Math. **74** (1991), 241–256.
[20] T. Downarowicz and Y. Lacroix, *Almost* 1-1 *extensions of Furstenberg–Weiss type and applications to Toeplitz flows*, Studia Math. **130** (1998), 149–170.
[21] T. Downarowicz and J. Serafin, *Possible entropy functions*, Israel J. Math. **135** (2003), 221–250.
[22] E.G. Effros, *Transformation groups and C^*-algebras*, Ann. of Math. **81** (1976), 38–55.
[23] H. Furstenberg, *The structure of distal flows*, Amer. J. Math. **85** (1963), 477–515.
[24] H. Furstenberg, *Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation*, Math. System Theory **1** (1967), 1–49.
[25] H. Furstenberg, *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Anal. Math. **31** (1977), 204–256.
[26] H. Furstenberg, *Recurrence in Ergodic Theory and Combinatorial Number Theory*, Princeton University Press, Princeton, NJ (1981).
[27] H. Furstenberg and B. Weiss, *Topological dynamics and combinatorial number theory*, J. Anal. Math. **34** (1978), 61–85.
[28] H. Furstenberg and B. Weiss, *The finite multipliers of infinite transformation*, Lecture Notes in Math., Vol. 688, Springer (1978), 127–132.
[29] H. Furstenberg and B. Weiss, *On almost* 1-1 *extensions*, Israel J. Math. **65** (1989), 311–322.
[30] R. Ellis, E. Glasner and L. Shapiro, *Proximal-isometric flows*, Adv. Math. **17** (1975), 213–260.
[31] T. Giordano, I.F. Putnam and C.F. Skau, *Topological orbit equivalence and C^*-crossed products*, J. Reine Angew. Math. **469** (1995), 51–111.
[32] E. Glasner, *A simple characterization of the set of μ-entropy pairs and applications*, Israel J. Math. **102** (1997), 13–27.
[33] E. Glasner, *On minimal actions of Polish groups*, Topology Appl. **85** (1998), 119–125.
[34] E. Glasner, *Structure theory as a tool in topological dynamics*, Descriptive Set Theory and Dynamical Systems, LMS Lecture Note Series, Vol. 277, Cambridge University Press, Cambridge (2000), 173–209.
[35] E. Glasner, *Topics in topological dynamics, 1991 to 2001*, Recent Progress in General Topology, Vol. II, North-Holland, Amsterdam (2002), 153–175.
[36] E. Glasner, *Ergodic Theory via Joinings*, Surveys and Monographs, Vol. 101, Amer. Math. Soc. (2003).
[37] E. Glasner, B. Host and D. Rudolph, *Simple systems and their higher order self-joinings*, Israel J. Math. **78** (1992), 131–142.
[38] E. Glasner and D. Maon, *Rigidity in topological dynamics*, Ergodic Theory Dynamical Systems **9** (1989), 309–320.
[39] E. Glasner and B. Weiss, *Minimal transformations with no common factor need not be disjoint*, Israel J. Math. **45** (1983), 1–8.
[40] E. Glasner and B. Weiss, *Sensitive dependence on initial conditions*, Nonlinearity **6** (1993), 1067–1075.
[41] E. Glasner and B. Weiss, *Strictly ergodic, uniform positive entropy models*, Bull. Soc. Math. France **122** (1994), 399–412.
[42] E. Glasner and B. Weiss, *Topological entropy of extensions*, Proceedings of the 1993 Alexandria Conference, Ergodic Theory and its Connections with Harmonic Analysis, K.E. Petersen and I.A. Salama, eds, LMS Lecture Note Series, Vol. 205, Cambridge University Press, Cambridge (1995), 299–307.

[43] E. Glasner and B. Weiss, *Weak orbit equivalence of Cantor minimal systems*, Internat. J. Math. **6** (1995), 559–579.
[44] J. Glimm, *Locally compact transformation groups*, Trans. Amer. Math. Soc. **101** (1961), 124–138.
[45] T.N.T. Goodman, *Relating topological entropy with measure theoretic entropy*, Bull. London Math. Soc. **3** (1971), 176–180.
[46] F. Hahn and Y. Katznelson, *On the entropy of uniquely ergodic transformations*, Trans. Amer. Math. Soc. **126** (1967), 335–360.
[47] J. Hansel and J.-P. Raoult, *Ergodicity, uniformity and unique ergodicity*, Indiana Univ. Math. J. **23** (1974), 221–237.
[48] B. Host, *Mixing of all orders and independent joinings of systems with singular spectrum*, Israel J. Math. **76** (1991), 289–298.
[49] W. Huang and X. Ye, *An explicit scattering, non-weakly mixing example and weak disjointness*, Nonlinearity **15** (2002), 849–862.
[50] W. Huang and X. Ye, *Topological complexity, return times and weak disjointness*, Ergodic Theory Dynamical Systems, to appear.
[51] R.I. Jewett, *The prevalence of uniquely ergodic systems*, J. Math. Mech. **19** (1970), 717–729.
[52] A. del Junco, M. Lemańczyk and M.K. Mentzen, *Semisimplicity, joinings and group extensions*, Studia Math. **112** (1995), 141–164.
[53] A. del Junco, M. Rahe and L. Swanson, *Chacón's automorphism has minimal self-joinings*, J. Anal. Math. **37** (1980), 276–284.
[54] A. del Junco and D.J. Rudolph, *On ergodic actions whose self-joinings are graphs*, Ergodic Theory Dynamical Systems **7** (1987), 531–557.
[55] B. Kamiński, A. Siemaszko and J. Szymański, *The determinism and the Kolmogorov property in topological dynamics*, Preprint.
[56] J.W. Kammeyer and D. Rudolph, *Restricted Orbit Equivalence for Actions of Discrete Amenable Groups*, Cambridge Tracts in Mathematics, Vol. 146, Cambridge University Press, Cambridge (2002).
[57] Y. Katznelson and B. Weiss, *When all points are recurrent/generic*, Ergodic Theory and Dynamical Systems I, Proceedings, Special year, Maryland 1979–80, Birkhäuser, Boston (1981).
[58] H.B. Keynes and J.B. Robertson, *Eigenvalue theorems in topological transformation groups*, Trans. Amer. Math. Soc. **139** (1969), 359–369.
[59] J. King, *Ergodic properties where order 4 implies infinite order*, Israel J. Math. **80** (1992), 65–86.
[60] W. Krieger, *On unique ergodicity*, Proc. Sixth Berkeley Symposium Math. Statist. Probab., Univ. of California Press (1970), 327–346.
[61] I. Kriz, *Large independent sets in shift-invariant graphs. Solution of Bergelson's problem*, Graphs and Combinatorics **3** (1987), 145–158.
[62] E. Lehrer, *Topological mixing and uniquely ergodic systems*, Israel J. Math. **57** (1987), 239–255.
[63] M. Lemańczyk, F. Parreau and J.-P. Thouvenot, *Gaussian automorphisms whose ergodic self-joinings are Gaussian*, Fund. Math. **164** (2000), 253–293.
[64] E. Lindenstrauss, *Lowering topological entropy*, J. Anal. Math. **67** (1995), 231–267.
[65] E. Lindenstrauss, *Measurable distal and topological distal systems*, Ergodic Theory Dynamical Systems **19** (1999), 1063–1076.
[66] D.C. McMahon, *Weak mixing and a note on the structure theorem for minimal transformation groups*, Illinois J. Math. **20** (1976), 186–197.
[67] D.C. McMahon, *Relativized weak disjointness and relative invariant measures*, Trans. Amer. Math. Soc. **236** (1978), 225–137.
[68] N.S. Ormes, *Strong orbit realization for minimal homeomorphisms*, J. Anal. Math. **71** (1997), 103–133.
[69] D. Ornstein and B. Weiss, *Mean distality and tightness*, Proceedings of the Steklov Mathematical Institute, to appear.
[70] W. Parry, *Zero entropy of distal and related transformations*, Topological Dynamics, J. Auslander and W. Gottschalk, eds, Benjamin, New York (1967).
[71] K. Petersen, *Disjointness and weak mixing of minimal sets*, Proc. Amer. Math. Soc. **24** (1970), 278–280.
[72] M. Ratner, *Horocycle flows, joinings and rigidity of products*, Ann. of Math. **118** (1983), 277–313.
[73] D.J. Rudolph, *An example of a measure-preserving transformation with minimal self-joinings and applications*, J. Anal. Math. **35** (1979), 97–122.

[74] V.V. Ryzhikov, *Joinings, intertwining operators, factors and mixing properties of dynamical systems*, Russian Acad. Izv. Math. **42** (1994), 91–114.
[75] R. Sacker and G. Sell, *Finite extensions of minimal transformation groups*, Trans. Amer. Math. Soc. **190** (1974), 325–334.
[76] M. Shub and B. Weiss, *Can one always lower topological entropy?*, Ergodic Theory Dynamical Systems **11** (1991), 535–546.
[77] S.D. Silvestrov and J. Tomiyama, *Topological dynamical systems of type I*, Exposition. Math. **20** (2002), 117–142.
[78] J.-P. Thouvenot, *Some properties and applications of joinings in ergodic theory*, Proceedings of the 1993 Alexandria Conference, Ergodic Theory and its Connections with Harmonic Analysis, K.E. Petersen and I.A. Salama, eds, LMS Lecture Note Series, Vol. 205, Cambridge University Press, Cambridge (1995), 207–238.
[79] W.A. Veech, *The equicontinuous structure relation for minimal Abelian transformation groups*, Amer. J. Math. **90** (1968), 723–732.
[80] W.A. Veech, *Point distal flows*, Amer. J. Math. **92** (1970), 205–242.
[81] W.A. Veech, *Topological dynamics*, Bull. Amer. Math. Soc. **83** (1977), 775–830.
[82] W.A. Veech, *A criterion for a process to be prime*, Monatsh. Math. **94** (1982), 335–341.
[83] B. Weiss, *Measurable dynamics*, Conference in Modern Analysis and Probability, Contemp. Math., Vol. 26 (1984), 395–421.
[84] B. Weiss, *Countable generators in dynamics—Universal minimal models*, Contemp. Math. **94** (1989), 321–326.
[85] B. Weiss, *Multiple recurrence and doubly minimal systems*, Contemp. Math. **215** (1998), 189–196.
[86] B. Weiss, *A survey of generic dynamics*, *Descriptive Set Theory and Dynamical Systems*, LMS Lecture Note Series, Vol. 277, Cambridge University Press (2000), 273–291.
[87] B. Weiss, *Single orbit dynamics*, CBMS, Regional Conference Series in Math., Vol. 95, Amer. Math. Soc. Providence, RI (2000).
[88] R.J. Zimmer, *Extensions of ergodic group actions*, Illinois J. Math. **20** (1976), 373–409.
[89] R.J. Zimmer, *Ergodic actions with generalized discrete spectrum*, Illinois J. Math. **20** (1976), 555–588.

CHAPTER 11

Spectral Properties and Combinatorial Constructions in Ergodic Theory

Anatole Katok

Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, USA
E-mail: *katok_ a@math.psu.edu*

Jean-Paul Thouvenot

Laboratoire de Probabilites, Université de Paris VI, Jussieu, 75252 Paris cedex 05, France

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

This survey primarily deals with certain aspects of ergodic theory, i.e. the study of groups of measure preserving transformations of a probability (Lebesgue) space up to a metric isomorphism [8, Section 3.4a]. General introduction to ergodic theory is presented in [8, Section 3]. Most of that section may serve as a preview and background to the present work. Accordingly we will often refer to definitions, results and examples discussed there. For the sake of convenience we reproduce some of the basic material here as need arises.

Here we will deal exclusively with actions of Abelian groups; for a general introduction to ergodic theory of locally compact groups as well as in-depth discussion of phenomena peculiar to certain classes of non-Abelian groups see [4]. Furthermore, we mostly concentrate on the classical case of cyclic systems, i.e. actions of $\mathbb{Z}$ and $\mathbb{R}$. Differences between those cases and the higher-rank situations (basically $\mathbb{Z}^k$ and $\mathbb{R}^k$ for $k \geqslant 2$) appear already at the measurable level but are particularly pronounced when one takes into account additional structures (e.g., smoothness).

Expository work on the topics directly related to those of the present survey includes the books by Cornfeld, Fomin and Sinai [29], Parry [124], Nadkarni [114], Queffelec [128], and the first author [78] and surveys by Lemańczyk [104] and Goodson [64]. Our bibliography is far from comprehensive. Its primary aim is to provide convenient references where proofs of results stated or outlined in the text could be found and the topics we mention are developed to a greater depth. So we do not make much distinction between original and expository sources. Accordingly our references omit original sources in many instances. We make comments about historical development of the methods and ideas described only occasionally. These deficiencies may be partially redeemed by looking into expository sources mentioned above. We recommend Nadkarni's book and Goodson's survey in particular for many references which are not included to our bibliography. Goodson's article also contains many valuable historical remarks.

1. Spectral theory for Abelian groups of unitary operators

1.1. *Preliminaries*

1.1.1. *Spectral vs. metric isomorphism.* Any measure preserving action Φ of a group G on a measure space (X, μ) generates a unitary representation of G in the Hilbert space $L^2(X, \mu)$ by $U_g : \varphi \mapsto \varphi \circ \Phi^{g^{-1}}$. For an action of $\mathbb{Z}$ generated by $T : X \to X$ the notation U_T for the operator U_1 is commonly used; often this operator is called Koopman operator since this connection was first observed in [95]. If two actions are isomorphic then the corresponding unitary representations in L^2 are unitarily equivalent, hence any invariant of unitary equivalence of such operators defines an invariant of isomorphism. Such invariants are said to be *spectral invariants* or *spectral properties*. Actions for which the corresponding unitary representations are unitarily equivalent are usually called *spectrally isomorphic*. We will use terms "unitarily equivalent" and "unitarily isomorphic" interchangeably.

Let us quickly describe the difference between the spectral and metric isomorphism for groups of unitary operators generated by measure preserving actions. In addition to the structure of Hilbert space which is preserved by any unitary operator, the space $L^2(X, \mu)$

has an extra multiplicative structure. There is a certain subtlety in describing this structure in purely algebraic terms since the product of two functions from $L^2(X,\mu)$ may not be an L^2 function so the whole space is not a ring with respect to addition and multiplication. There are however various dense subsets (e.g., bounded functions) for which multiplication is always defined; a proper abstract description leads to the notion of *unitary ring* [136].

An easier way to capture the essential part of the multiplicative structure which avoids many technical complications is as follows. First there is preferred element, the constant function equal to one which is the multiplicative unity. Second, there are *the idempotents* characterized by the equation $f^2 = f$ which evidently correspond to characteristic functions. Products of characteristic functions correspond to intersection of the sets: $\chi_{A_1} \cdot \chi_{A_2} = \chi_{A_1 \cap A_2}$ and hence the union is also recovered: $\chi_{A_1 \cup A_2} = \chi_{A_1} + \chi_{A_2} - \chi_{A_1 \cap A_2} = \chi_{A_1} + \chi_{A_2} - \chi_{A_1} \cdot \chi_{A_2}$.

Now let us call a unitary operator $U : L^2(X,\mu) \to L^2(Y,\nu)$ *multiplicative* if it takes idempotents into idempotents and preserves the product of such elements. Assuming that (X,μ) and (Y,ν) are Lebesgue spaces [8, Section 3.2b], [141,86] such an operator is generated by an isomorphism of measure spaces, $h : (Y,\nu) \to (X,\mu)$, i.e. $U(f) = f \circ h$. Naturally, the Koopman operator generated by a measure preserving transformation of a Lebesgue space is multiplicative.

This can summarized as follows:

PROPOSITION 1.1. *Unitary representations generated by measure preserving actions of a group G are metrically isomorphic if and only is they are unitarily equivalent via a multiplicative operator.*

A closed subspace $H \subset L^2(X,\mu)$ is called a *unitary $*$-subalgebra* if H is invariant under complex conjugation, bounded functions are dense in H and product of any two bounded functions from H is again in H. In this case characteristic functions generate H and H defines a *measurable partition* ξ of the space X in the following way.

PROPOSITION 1.2. *Any unitary $*$-subalgebra consists of all functions in $L^2(X,\mu)$ which are constant mod* 0 *on elements of a measurable partition. If a unitary $*$-subalgebra is U_T invariant then the corresponding measurable partition is T invariant and defines a factor of the measure preserving transformation T.*

For a more detailed description see [21, Section 5], [141].

For a general discussion of spectral properties for groups of measure preserving transformations see [4]. In the remainder of this section we will discuss the case of locally compact Abelian groups. In the rest of the survey we will restrict our considerations to the classical cases of automorphisms and flows, i.e. actions of $\mathbb{Z}$ and $\mathbb{R}$ correspondingly (and primarily the former) with only occasional comments related to actions of other groups.

1.1.2. *Duality for locally compact Abelian groups* [126, Chapter 6], [115, Section 31]. Let G be a locally compact second countable topological Abelian group. A *character* of G is a continuous homomorphism $\chi : G \to S^1$. Characters form a group which is often called *the dual group* of G and is denoted by G^*. There is a natural locally compact topology

on G^*. It can be described as topology of uniform convergence on compact sets or, equivalently, as the weakest topology which makes any evaluation map $e_g : \chi \mapsto \chi(g)$ continuous. Obviously $e_g : G^* \to S^1$ thus defined is a continuous character of G^*. The Pontrjagin Duality Theorem asserts that any continuous character of G^* has the form e_g and that element $g \in G$ is uniquely defined [126, Section 40], [115, Section 31.6]. This is usually expressed in an attractive compact form

$$G^{**} = G.$$

A useful addition to the Pontrjagin Duality is the observation that G^* is compact if and only if G is discrete. In what follows the group G will be assumed not compact, but it may be discrete or continuous.

There are natural functorial properties of the duality, all easily derived from the fact that arrows in natural homomorphisms get reversed. For example, the dual to the direct sum of finitely many groups is the direct sum of their duals, the dual to the direct sum of countably many groups is the direct product of the duals, and there is a natural duality between subgroups and factors, and between direct and inverse limits.

EXAMPLE 1.3. $\mathbb{Z}^* = S^1 = \mathbb{R}/\mathbb{Z}$, $(\mathbb{Z}^k)^* = \mathbb{T}^k = \mathbb{R}^k/\mathbb{Z}^k$, $(\mathbb{R}^k)^* = \mathbb{R}^k$, Furthermore, $(\mathbb{Z}^\infty)^* = \mathbb{T}^\infty$, where $\mathbb{Z}^\infty$ is the discrete direct sum of countably many copies of $\mathbb{Z}$ and $\mathbb{T}^\infty$ is the compact direct product of countably many copies of S^1.

EXAMPLE 1.4. The multiplicative group of roots of unity of degrees 2^n, $n = 1, 2, \ldots$, with discrete topology is the direct limit of cyclic groups of order 2^n, $n = 1, 2, \ldots$. Its dual is the compact additive group $\mathbb{Z}_2$ of dyadic integers, which is the inverse limit of such cyclic groups. By replacing 2 with a natural number m one gets roots of unity of degrees m^n, $n = 1, 2, \ldots$, and the m-adic integers correspondingly.

Using the duality between direct sums and direct products one sees that the dual to the group of all roots of unity is the direct product of the p-adic integers $\prod \mathbb{Z}_p$ over all prime numbers p.

Here is another example of the duality between direct and inverse limits.

EXAMPLE 1.5. The dual to the group $\mathbb{Z}[1/2]$ of rational numbers whose denominators are powers of 2 (which is a direct limit of free cyclic groups) is the dyadic solenoid

$$\mathcal{S}_2 \stackrel{\text{def}}{=} \{(z_1, z_2, \ldots) \colon z_1 \in S^1, z_{n+1}^2 = z_n, n = 1, 2, \ldots\}.$$

1.2. *The spectral theorem*

1.2.1. *Formulation in the general case.* A character χ can be viewed as a one-dimensional unitary representation of the group, namely the element $g \in G$ acts on $\mathbb{C}$ by the multiplication by $\chi(g)$. Every irreducible unitary representation of an Abelian group is one-dimensional (see, e.g., [115, Section 31.7]). The spectral theorem states essentially

that every unitary representation of such a group in a separable Hilbert allows a canonical decomposition into a (in general, continuous) direct sum (i.e. direct integral) of characters. In this the spectral theorem represents a special case of the general theorem about the decomposition of a unitary group representation into irreducible representations [4, Theorem 3.1.3] (see [38] for a proof), but since in the Abelian case the structure of irreducible representations is simple and well understood it is considerably more specific than the general case.

Thus in the case of Abelian groups the spectral theorem gives a full collection of models for all unitary representations together with a necessary and sufficient condition for equivalence of such models.

Let G be a locally compact second countable Abelian group, ν be a σ-finite Borel measure on the dual group G^* and m be a ν-measurable function on G^* with values in $\mathbb{N} \cup \infty$. Let $H_{\nu,m}$ be the subspace of the ν-measurable square integrable functions $\varphi : G^* \to l^2$ such that at a point $\chi \in G^*$ all but the first $m(\chi)$ coordinates of $\varphi(\chi)$ vanish. The space $H_{\nu,m}$ is a separable Hilbert space with respect to the scalar product

$$\langle \varphi, \psi \rangle = \int_{G^*} \big(\varphi(\chi), \psi(\chi)\big)_{l^2} \, d\nu.$$

The group G acts unitarily on the space $H_{\nu,m}$ by the natural scalar multiplications:

$$U_g^{\nu,m} \varphi(\chi) = \chi(g)\varphi.$$

THEOREM 1.6 (The spectral theorem). *Any continuous in the strong operator topology unitary representation of G in a separable Hilbert space is unitarily equivalent to a representation $U^{\nu,m}$.*

Furthermore, representations U^{ν_1,m_1} and U^{ν_2,m_2} are unitarily equivalent if and only if measures ν_1 and ν_2 are equivalent (*i.e. have the same null-sets*) *and $m_1 = m_2$ almost everywhere.*

REMARK. Since every σ-finite measure is equivalent to a finite measure, one can assume without loss of generality that in the spectral theorem the measure ν is finite. If the group G is discrete (and hence G^* is compact) this is a customary assumption. However, in the case of a continuous group, such as $\mathbb{R}$, the most natural measure on the dual group, the Haar measure, is not finite. Accordingly, in the spectral theorem instead of finiteness of ν one assumes only local finiteness.

1.2.2. *Sketch of proof for single operator.* We outline a proof of the spectral theorem in the particular case of the action of a single operator U on a Hilbert space H.

DEFINITION 1.7. Consider a unitary operator U acting on a Hilbert space H. Let H_f be the norm closure of the linear span of the $U^n f$, $n \in \mathbb{Z}$. The space H_f is called *the cyclic subspace* generated by f.

Let us denote the scalar product in H by $\langle \cdot, \cdot \rangle$ and let θ be the natural cyclic coordinate on S^1.

THEOREM 1.8. *There exists a positive measure ν_f on $S^1 = \mathbb{R}/\mathbb{Z}$ with total mass $\|f\|^2$ such that for the unitary operator*

$$M : L^2\big(S^1, \nu_f\big) \to L^2\big(S^1, \nu_f\big), \quad g \mapsto e^{2\pi i\theta} g,$$

there exists an isometry V between H_f and $L^2(S^1, \nu_f)$ which conjugates the restriction of U to H_f and M (i.e. $VU = MV$), such that $Vf = 1$ (the constant function on S^1) and

$$\langle f, U^n f\rangle = \hat{\nu}_f(n), \quad n \in \mathbb{Z}. \tag{1.1}$$

PROOF. If (1.1) holds then the correspondence $U^n f \to e^{2\pi i n\theta}$, $n \in \mathbb{Z}$, extends to the isometry V with desirable properties. Thus it is sufficient to prove (1.1), i.e. to show that the *correlation coefficients* $\langle f, U^n f\rangle$ are Fourier coefficients of a measure. For that consider the following sequence of positive measures:

$$\nu_{N,f} = \frac{\big\|\sum_{n=1}^{N} e^{2\pi i n\theta} U^n f\big\|^2}{N}\, d\theta.$$

One can calculate the Fourier coefficients of these measures directly. In particular, if $|k| \leqslant N$, then

$$\begin{aligned}\hat{\nu}_{N,f}(k) &= \frac{1}{N}\int_{S^1} e^{-2\pi i k\theta} \sum_{1\leqslant m,n\leqslant N} \big\langle e^{2\pi i m\theta} U^m f, e^{2\pi i n\theta} U^n f\big\rangle d\theta \\ &= \frac{1}{N}\int_{S^1} \sum_{1\leqslant m,n\leqslant N} e^{2\pi i (m-n-k)\theta} \langle U^{m-n} f, f\rangle\, d\theta = \frac{N-|k|}{N}\langle f, U^k f\rangle.\end{aligned}$$

This equality for $k = 0$ means that the total mass of $\nu_{N,f}$ is constant, $\nu_{N,f}(S^1) = \|f\|^2$. Since for any $k \in \mathbb{Z}$, the Fourier coefficients $\hat{\nu}_{N,f}(k)$ converge to $\langle f, U^k f\rangle$, this implies that $\nu_{N,f}$ converge weakly to a measure ν_f on S^1 satisfying (1.1). □

The measure ν_f is called *the spectral measure* associated to f. If U_T is the Koopman operator acting on $L^2(X, \mu)$ and $f \in L^2(X, \mu)$ is a real-valued function, then the measure ν_f is symmetric with respect to the real axis.

We now state an important lemma, due to Wiener, which identifies all the invariant subspaces for the action of the operator $M : g \mapsto e^{2\pi i\theta} g$ in $L^2(S^1, \nu)$.

LEMMA 1.9. *If ν is a positive finite measure on S^1 and K is a closed M-invariant subspace of $L^2(S^1, \nu)$ then there exists a measurable set $E \subset S^1$ such that $K = \{f \in L^2(S^1, \nu)\colon\ f = 0 \text{ on } E^c\}$.*

PROOF. The projection of the constant function 1 on K, $\mathcal{P}_K 1$, is a characteristic function since if for every $n \in \mathbb{Z}$,

$$\int (1 - \mathcal{P}_K 1) e^{2\pi i n\theta} \overline{\mathcal{P}_K 1}\, d\nu = 0$$

then $\overline{\mathcal{P}_K 1}(1 - \mathcal{P}_K 1) = 0$, ν almost everywhere. This implies existence of a measurable set E such that $\mathcal{P}_K 1 = \chi_E$ and $1 - \mathcal{P}_K 1 = \chi_{E^c}$. □

As an immediate corollary one obtains

THEOREM 1.10. *Let U be a unitary operator acting on H. Let g_1 and g_2 be two elements in H such that the measures ν_{g_1} and ν_{g_2} are mutually singular. Then*

$$H_{g_1} \perp H_{g_2}.$$

Furthermore

$$H_{g_1+g_2} = H_{g_1} \oplus H_{g_2}.$$

Finally, if there exists $f \in H$ such that $H_{f_1} \subset H_f$, $H_{f_2} \subset H_f$ and $H_{f_1} \perp H_{f_2}$, then the measures ν_{g_1} and ν_{g_2} are mutually singular.

PROOF. This is an easy consequence in the circle model, constructed in Theorem 1.8, for the action of U on a cyclic subspace. For, since invariant subspaces are entirely characterized by subsets of the circle, we see that two such subspaces are orthogonal if and only if the corresponding sets are disjoint. In particular, a vector whose spectral measure has full support is cyclic. □

DEFINITION 1.11. Let U act on H as before. The *maximal spectral type* ν_U of the operator U is a positive measure on S^1 (which is defined up to equivalence) such that for every $f \in H$ the measure ν_f is absolutely continuous with respect to ν_U and no measure absolutely continuous with respect to ν_U but not equivalent to ν_U has the same property.

In the case of an action of $\mathbb{Z}$ the Spectral Theorem 1.6 which gives a complete set of invariants for a unitary operator, takes the following form.

THEOREM 1.12. *Let the unitary operator U act on H. There exists a family of positive measures on S^1, uniquely defined up to equivalence,*

$$\nu_1 \geqslant \nu_2 \geqslant \nu_3 \geqslant \cdots \geqslant \nu_n \geqslant \cdots,$$

where ν_1 is the maximal spectral type ν_U, such that the action of U on H is unitarily isomorphic to the action of M (the multiplication by $e^{2\pi i\theta}$) on the orthogonal sum

$$\bigoplus_{i \geqslant 1} L^2(S^1, \nu_i).$$

SKETCH OF PROOF. The theorem follows from the observation that if f and g in H have the property that $\nu_f \sim \nu_g$, then the two actions of U on $H_f^{\perp}$ and $H_g^{\perp}$ are unitarily equivalent. This can be seen as it suffices to check that the restrictions of U to the invariant

spaces $H_f^{\perp} \cap H_{f,g}$ and $H_g^{\perp} \cap H_{f,g}$ are equivalent. Here $H_{f,g}$ denotes the invariant subspace generated by f and g. These two spaces are cyclic and it is easily checked that they have equivalent spectral measures. □

REMARK. Alternatively, one can take a sequence of ν_1 measurable sets in S^1, A_i, $i \geqslant 1$, $A_{i+1} \subset A_i$ such that $\nu_i = \nu_1 \cdot \chi_{A_i}$.

1.3. *Spectral representation and principal constructions*

One of the advantages of the spectral representation is that it behaves nicely under the natural functorial constructions.

1.3.1. *Restrictions.* For the representation $U^{\nu,m}$ all closed invariant subspaces can be described. We will denote by l_2^m the subspace of l_2 which consists of all vectors for which all but first m coordinates vanish. The following statement generalizes the Wiener Lemma 1.9.

THEOREM 1.13. *Any $U^{\nu,m}$ invariant closed subspace of $H_{\nu,m}$ is determined by a ν-measurable field of closed subspaces $L_\chi \subset l_2^{m(\chi)}$, where by definition $l_2^\infty = l_2$, and consists of all φ such that $\varphi(\chi) \in L_\chi$.*

PROOF. First, consider the case of a cyclic subspace for $U^{\nu,m}$ generated by $f \in H_{\nu,m}$. Since $U^{\nu,m}$ acts by scalar multiplications, the subspace H_f of all functions proportional to f on the set

$$S_f \stackrel{\text{def}}{=} \{\chi \in G^* \colon f(\chi) \neq 0\}$$

and vanishing on $G^* \setminus S_f$, is $U^{\nu,m}$ invariant. The maximal spectral type on the subspace H_f is the restriction of ν to the set S_f. But then f generates this subspace since by the Wiener Lemma 1.9 any invariant subspace of H_f consists of functions vanishing on a certain subset of S_f of positive ν-measure and hence it cannot contain f.

Now consider an arbitrary invariant subspace H. It is generated by a finite or countable set of functions $f_1, \ldots$. Every cyclic subspace H_{f_n} determines a subset S_{f_n} and a field $L_{n,\chi}$ of one-dimensional subspaces on S_{f_n}. The sum of those subspaces at each $\chi \in G^*$ forms a ν-measurable field of subspaces L_χ and since every function g with values in L_χ is the limit of linear combinations of functions with values in $L_{n,\chi}$, we conclude that $g \in H$. □

1.3.2. *Direct products.* Similarly it is easy to represent the Cartesian product of representations of the form $U^{\nu,m}$ in a similar form.

THEOREM 1.14. *The Cartesian product of representations $U^{\nu,m}$ and $U^{\nu',m'}$ is unitarily equivalent to the representation $U^{\nu+\nu',m+m'}$.*

1.3.3. *Tensor products.* The tensor product of representations $U^{\nu,m}$ and $U^{\nu',m'}$ can be described as follows. Take the group $G^* \times G^* = (G \times G)^*$ with the measure $\nu \times \nu'$. Let $m(\chi_1, \chi_2) = m_1(\chi_1) \cdot m_2(\chi_2)$. Consider the space $H_{\nu\times\nu',m}$. The group G acts diagonally on that space:

$$(U_g\varphi)(\chi_1, \chi_2) = \chi_1(g)\chi_2(g)\varphi(\chi_1, \chi_2).$$

This is a representation of the tensor product of $U^{\nu,m}$ and $U^{\nu',m'}$. From this representation the spectral representation of the tensor product can be deduced. We do this explicitly for the case of two Koopman operators in Section 4.1.3.

1.4. *Spectral invariants*

DEFINITION 1.15. For a given unitary representation of G the equivalence class of the measure ν in a unitarily equivalent representation $U^{\nu,m}$ is called *the maximal spectral type of the representation*. The function m is called *the multiplicity function*.

The maximal spectral type and the multiplicity function form a complete set of invariants for a unitary representation of a locally compact Abelian group.

1.4.1. *Maximal spectral type.* The maximal spectral type is an equivalence class of measures on the locally compact group G^*. In the two classical cases $G = \mathbb{Z}$ and $G = \mathbb{R}$ the maximal spectral type is a class of measures on the circle and the real line correspondingly. The first of those cases has already been discussed in Section 1.2.2, the second is summarized in Section 1.4.4 below.

The crudest distinction among measures is between atomic and continuous. Any measure uniquely decomposes into an atomic (discrete) and continuous part and this decomposition is invariant under equivalence of measures. Atoms in the maximal spectral type correspond to the eigenvectors for the representation: the characteristic function of such an atom is an eigenfunction.

If the maximal spectral type is atomic the representation is said to have *pure point spectrum*. There is a difference with the notion of discrete spectrum common in many areas of analysis. The spectrum may be pure point but the eigenvalues may be dense in G^* or, more generally the eigenvalues may not be isolated; in other words, the spectrum as a set does not have to be discrete. Since genuine discrete spectrum appears in ergodic theory only in some trivial situations the term "discrete spectrum" is sometimes used instead of "pure point".

On the other hand, in our setting there is a special continuous measure on G^*, the Haar measure λ; any measure absolutely continuous with respect to Haar is called simply *absolutely continuous*. Any non-atomic measure singular with respect to the Haar measure is referred to as simply *singular*. Thus, an arbitrary measure on G^* allows a unique decomposition into atomic, absolutely continuous and singular parts invariant under equivalence. Representations whose maximal spectral type is atomic, absolutely continuous, or singular

are referred to correspondingly as representations with *pure point, absolutely continuous*, or *singular spectrum*.

Theorems 1.6 and 1.13 easily imply

COROLLARY 1.16. *The maximal spectral type of the restriction of a representation to any closed invariant subspace is absolutely continuous with respect to the maximal spectral type of the representation.*

1.4.2. *Correlation coefficients.* Similarly to the case of a single operator (Definition 1.7) for a unitary representation U of G in a Hilbert space H with the scalar product $\langle \cdot, \cdot \rangle$, every element $v \in H$ determines the *cyclic space* H_v, the minimal closed U-invariant subspace which contains v.

Theorem 1.13 implies that an invariant subspace is cyclic if and only if for almost every with respect to the maximal spectral type $\chi \in G^*$ the space L_χ has dimension at most one. The maximal spectral type in the space H_v is represented by the measure ν_v, called *the spectral measure of* v, where

$$\langle v, U(g)v \rangle = \int_{G^*} \chi(g)\, d\nu_v.$$

Similarly to the case of the single operator (Section 1.2.2) these scalar products are often called *the correlation coefficients* of the element v. Notice that the spectral measure is always finite, since $\|v\|^2 = \nu_v(G^*)$.

REMARK. Correlation coefficients can be viewed as the Fourier transform of the measure. It is useful to remember that the Fourier transform is linear and that the product of Fourier transforms corresponds to the convolution of measures. In ergodic theory convolutions appears in connection with multiplication of functions in the study of Cartesian products (Section 4.1.3) as well as in situation when multiplicative structure is well related with the spectral picture, such as the pure point spectrum (see Section 3.5), Gaussian systems (Section 6.4.1), and Gaussian Kronecker systems (Section 6.4.3).

For the representation $U^{\nu,m}$ the cyclic space determined by a function φ consists of all functions whose values are proportional to those of φ. This implies that $\nu_\varphi = |\varphi|^2 \nu$ and hence

COROLLARY 1.17. *For any finite measure μ absolutely continuous with respect to the maximal spectral type there exists $v \in H$ such that $\nu_v = \mu$.*

Recall that a set in a topological space is called *residual* if its complement is the union of countably many nowhere dense sets. In the space $H^{\nu,m}$ the set of elements which do not vanish is residual. Thus we obtain from Theorem 1.6.

COROLLARY 1.18. *The spectral measures for a residual set of elements belong to the maximal spectral type.*

1.4.3. *Multiplicity function*

DEFINITION 1.19. An *essential value* $n \in \mathbb{N} \cup \{\infty\}$ *of the spectral multiplicity* for a unitary representation of an Abelian group G is any number such that the multiplicity function m takes value n on a set of non-zero measure with respect to the maximal spectral type.

The *maximal multiplicity* is the supremum of essential values.

The representation is said to have a *homogeneous spectrum* if there is only one essential value of the spectral multiplicity. This value is then called *the multiplicity of the homogeneous spectrum*. If the only essential value is 1, the representation is said to have *simple spectrum*.

Simple spectrum is equivalent to cyclicity of the whole space: there exists a vector v such that the linear combinations of vectors $U_g(v)$, $g \in G$, are dense. Homogeneous spectrum of multiplicity m (finite or infinite) can be characterized as follows:

There exists a decomposition of the space H into an orthogonal sum of m cyclic subspaces such that the restrictions of the representation into all of those subspaces are unitarily equivalent.

The following closely related fact which follows immediately from Theorems 1.6 and 1.14, is often used in ergodic theory and is central in relating various symmetries with spectral properties (see Section 5.8).

COROLLARY 1.20. *Suppose U is a unitary representation of a locally compact Abelian group G in the Hilbert space H. Suppose that H decomposes into the orthogonal sum of $k \in \mathbb{N} \cup \{\infty\}$ invariant subspaces and the restrictions of U to all of those subspaces are unitarily equivalent. Then all essential values of the spectral multiplicity are multiples of k.*

Given a collection of elements $v_1, \dots, v_k \in H$, the subspace $H_{v_1,\dots,v_k}$ is defined as the minimal closed U-invariant subspace which contains $v_1, \dots, v_k$. It follows directly from Theorems 1.6 and 1.13 that the maximal multiplicity of a representation U is equal to the infimum of k such that $H = H_{v_1,\dots,v_k}$. In particular, if there is no such finite collection $v_1, \dots, v_k$ then the maximal multiplicity is infinite; this does not imply though that ∞ is an essential value.

In ergodic theory it often happens that one can construct sequences of cyclic subspaces, or, more generally, subspaces generated by a given number of vectors, which approximate every vector sufficiently well. Existence of such a sequence allows to estimate maximal spectral multiplicity from above thus improving the criterion above. For a $v \in H$ and a closed subspace $L \subset H$ let us denote as before by $\mathcal{P}_L v$ the orthogonal projection of the vector v onto L.

THEOREM 1.21. *Let U be a unitary operator on the Hilbert space H. If for every orthonormal m-tuple of vectors $v_1, \dots, v_m \in H$, there exists $w \in H$ such that the cyclic subspace generated by w, H_w satisfies*:

$$\sum_{i=1}^{i=m} \|\mathcal{P}_{H_w} v_i\|^2 > 1$$

then the maximal spectral multiplicity of U is $\leqslant m-1$.

In particular if $m=2$ the representation has simple spectrum.

PROOF. If the spectral multiplicity of U is $\geqslant m$, and if ν is the spectral measure of U we can find, by the spectral Theorem 1.12, a set $\mathcal{A}$ in S^1 such that there exists a U-invariant subspace K of H restricted to which U is isomorphic to the sum of m copies K_i, $i=1,\dots,m$, of the action of M on $L^2(S^1, \chi_{\mathcal{A}}\nu)$ (M as in Theorem 1.8). We choose for v_i, $i=1,2,\dots,m$, the functions $\chi_{\mathcal{A}}/\nu(\mathcal{A})^{1/2}$, $i=1,2,\dots,m$, of the above model. Then there exists $w_i \in K_i$, $i=1,2,\dots,m$, and w' orthogonal to K such that $w=\sum_{i=1}^{i=m} w_i + w'$. Let $\tilde{w}=\sum_{i=1}^{i=m} w_i$. Then $\mathcal{P}_{\mathcal{H}_{\tilde{w}}} v_i = \mathcal{P}_{\mathcal{H}_w} v_i$.

Now, if W is an invariant subspace of $\bigoplus_{i=1}^{i=m} L^2(S^1, \chi_{\mathcal{A}}\nu) = \mathcal{H}$, exactly the same proof as the one which was used for the Wiener lemma gives: $(v_i(x), \mathcal{P}_W v_i(x)) = \|\mathcal{P}_W v_i(x)\|^2$ (the scalar product is taken in $\mathcal{H}$). This says that the restriction of $\mathcal{P}_W$ to the fiber at x is a projection on a subspace W_x. Therefore, $\sum_{i=1}^{i=m} \|\mathcal{P}_W v(x)\|^2 = \dim W_x$. In our case, $\dim \mathcal{H}_w(x) = 1$ for ν almost all x, and we get, if B is the support of the spectral measure of w, $\sum_{i=1}^{i=m} \|\mathcal{P}_{\mathcal{H}_w} v_i\|^2 = \nu(B)/\nu(A) \leqslant 1$. □

1.4.4. *Spectral theorem and spectral invariants for one-parameter groups of unitary operators.* By the Stone Theorem any continuous one-parameter groups of unitary operators U_t: $t \in \mathbb{R}$, $U_{t+s} = U_t \cdot U_s$ in a Hilbert space H has the form $U_t = \exp itA$ where A is a Hermitian operator ($A^* = A$). Notice that A is not necessarily bounded. Nevertheless A is uniquely defined on a dense subset as $-i\frac{dU_t}{dt}|_{t=0}$. The operator A or sometimes the skew-Hermitian operator $-iA$ is called the (*infinitesimal*) *generator* of the group U_t. The spectral theorem for one-parameter groups of unitary operators takes the following form.

THEOREM 1.22. *Let $U_t = \exp itA$ be a one-parameter groups of unitary operators in a Hilbert space on H continuous in the strong operator topology. There exists a family of locally finite positive measures on $\mathbb{R}$ uniquely defined up to equivalence* (*and called the spectral types for the group*)

$$\nu_1 \geqslant \nu_2 \geqslant \nu_3 \geqslant \cdots \geqslant \nu_n \geqslant \cdots$$

such that the action of U_t on H is unitarily isomorphic to the multiplication by $e^{2\pi itx}$) on the orthogonal sum

$$\bigoplus_{i \geqslant 1} L^2(\mathbb{R}, \nu_i).$$

Accordingly the Hermitian infinitesimal generator A acts as multiplication by the independent variable x in each $L^2(\mathbb{R}, \nu_i)$.

Notice that in this case the maximal spectral type and multiplicity function for each individual operator U_t are defined on the circle: they are obtained from the spectral types of the group via the standard projection $\pi_t : \mathbb{R} \to S^1$, $\pi_t(s) = \exp its$. Thus spectral multiplicity of individual operators tend to be greater than for the group. A typical example is the case

of Lebesgue spectrum: every non-identity elements of a one-parameter group of unitary operators with simple Lebesgue spectrum has countable Lebesgue spectrum.

Here is a simple but useful criterion of Lebesgue spectrum for one-parameter groups:

PROPOSITION 1.23. *If the one-parameter group of unitary operators U_t is unitarily equivalent to the renormalized group U_{st} for any $s > 0$, then U_t has homogeneous Lebesgue spectrum.*

PROOF. It follows from the assumption that the infinitesimal generator A of U_t is unitarily equivalent to sA. But the spectral measures of sA are obtained from those of A by applying the multiplication by s on the real line. Hence the spectral measures are invariant under these multiplications and Lebesgue is the only type satisfying this property. □

2. Spectral properties and typical behavior in ergodic theory

Now we will consider a single unitary operator $U : H \to H$, or, equivalently, a unitary representation of the group $\mathbb{Z}$. The spectral measures in this case are measures on the circle S^1 (see Section 1.2.2). We will always assume that all measures we are considering are finite. Most of the discussion below can be extended straightforwardly to the case of discrete Abelian groups while in the continuous case certain subtle points appear. We will address some of these points for the case of one-parameter groups of operators, i.e. representations of $\mathbb{R}$.

A NOTE ON TERMINOLOGY. We will apply the spectral notions discussed below for unitary operators to measure preserving transformations if the Koopman operator in the orthogonal complement to the constants possesses the corresponding property. Thus we will speak about transformations with Lebesgue spectrum, mixing, mildly mixing, rigid, and so on. From now on, the scalar product will usually be denoted by $(\cdot, \cdot)$.

2.1. *Lebesgue spectrum*

2.1.1. *Correlation decay.* The maximal spectral type in a cyclic subspace $L \subset H$ is Lebesgue if and only if there exists $v \in L$ such that the iterates $U^n v$, $n \in \mathbb{Z}$, form an orthogonal basis in L. There are natural sufficient conditions for absolute continuity of the spectral measure, e.g., a certain decay rate for the correlation coefficients, such as l^2, but non of such conditions is necessary since an L^1 function on the circle may have very slowly decaying Fourier coefficients. The most general decay condition sufficient to guarantee that the spectral measure is actually equivalent to Lebesgue is an exponential decay

$$(v, U^n v) \leqslant c \exp\bigl(-\beta|n|\bigr)$$

for some positive numbers c and β. For, in this case the Fourier transform of the sequence $(v, U^n v)$ is a real-analytic function on the circle; it is nonnegative since it is a density of a measure and by analyticity it can only have finitely many zeroes.

The corresponding condition in the continuous time case is particularly useful because in that case there is no convenient direct counterpart of the orthogonality condition above.

2.1.2. *Countable Lebesgue spectrum in ergodic theory.* A particular type of spectrum which is ubiquitous in ergodic theory is *countable Lebesgue spectrum*, i.e. the Lebesgue maximal spectral type with the multiplicity function identically equal to ∞. The following criterion is evident from the definition.

A unitary operator $U : H \to H$ has countable Lebesgue spectrum if and only if there exists an infinite-dimensional closed subspace $H_0 \subset H$ such that

(i) *H_0 and $U^n H_0$ are orthogonal for $n > 0$ (or, equivalently for $n \neq 0$), and*

(ii) *$H = \bigoplus_{n\in\mathbb{Z}} U^n H_0$.*

As we already mentioned in the case of one-parameter group of operators if the infinitesimal generator A of the group has simple Lebesgue spectrum (i.e. the maximal spectral type is the type of Lebesgue measure on the line) then the unitary operators $\exp(-itA)$ have countable Lebesgue spectrum for every $t \neq 0$. Still the term "countable Lebesgue spectrum" is reserved for the case where the generator has Lebesgue spectrum of infinite multiplicity.

Here is a good illustration of how countable Lebesgue spectrum appears in ergodic theory.

EXAMPLE 2.1. Consider an automorphism A of a compact Abelian group G. It preserves Haar measure χ and the Koopman operator maps characters into characters. The characters form an orthonormal basis in $L^2(G, \chi)$. The cyclic subspace of each character is either finite-dimensional (and hence the spectral measure is atomic and the eigenvalues are roots of unity) or Lebesgue where the orbit of the character is infinite. Thus the spectrum of U_A in $L^2_0(G, \chi)$ is in general a combination of pure point and Lebesgue. If A is ergodic (see Section 3.3) the first case does not appear and the spectrum is Lebesgue. It is not difficult to show that the number of orbits in the dual group is always infinite so Lebesgue spectrum is always countable.

This conclusion extends with a slight modification to a more general class of *affine maps* on compact Abelian groups. Such a map is a product (composition) of an automorphism and a translation. In this case again the spectrum in general is a combination of pure point and countable Lebesgue, however it can be mixed even in the ergodic case, see Examples 3.17 and 3.18.

Other standard examples of transformations with countable Lebesgue spectrum are Bernoulli shifts introduced in Example 3.10 (see also [8, Section 3.3e]) and, more generally, transitive Markov shifts [8, Section 3.3f].

2.1.3. *Hyperbolic and parabolic paradigms*

Positive entropy, K-property, hyperbolic behavior. The main source of the presence of countable Lebesgue *part* in the spectrum is positivity of entropy [8, Theorem 3.7.13], [12]; in particular, the completely positive entropy (the K-property) implies that the spectrum in

the orthogonal complement L^2_0 to the constants is countable Lebesgue [8, Theorem 3.6.9], [12].

This kind of behavior appears in systems with *hyperbolic* and *partially hyperbolic* behavior [8, Section 6], [7,1,9]. Example 2.1 in the case when the group G is a torus $\mathbb{T}^n$ provides simple particular cases for both hyperbolic and partially hyperbolic situations. For, in this case the automorphism is determined by an integer $n \times n$ matrix with determinant ± 1. The hyperbolic case corresponds to the situation when the matrix has no eigenvalues of absolute value one; partially hyperbolic case appears when there are some such eigenvalues but no roots of unity among them. See [8, Sections 5.1h and 6.5a].

Zero entropy; parabolic behavior. Countable Lebesgue spectrum also appears in many zero entropy systems, sometimes accompanied by a pure-point part. This is typical for the *parabolic* paradigm [8, Section 8] which appears in particular in many systems of algebraic origin and their modifications. See Examples 3.17 and 3.18, Section 6.2.2 and [10] (especially Section 2.3a).

Horocycle flows. Now we will describe a particularly characteristic example of parabolic system which show how Lebesgue spectrum follow from a renormalization arguments.

Let X be the manifold $X = SL(2,\mathbb{R})/\Gamma$ where Γ is a discrete subgroup of finite covolume in $SL(2,\mathbb{R})$. Consider the following one-parameter subgroup of $SL(2,\mathbb{R})$: $H_t = \left(\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right)$, $t \in \mathbb{R}$.

The action of H_t by left translations on the right homogeneous space X preserves the measure m on X induced by the Haar volume in $SL(2,\mathbb{R})$. Let us denote this action h_t; it is called the *horocycle flow*.

PROPOSITION 2.2. *Every transformation h_t, $t \neq 0$ has countable Lebesgue spectrum.*

PROOF. Consider the one-parameter diagonal subgroup of $SL(2,\mathbb{R})$; $G_t = \left(\begin{smallmatrix} e^t & 0 \\ 0 & e^{-t} \end{smallmatrix}\right)$ and the corresponding left action of G_t on X by g_t; the latter is called the *geodesic flow*.[1]

Direct calculation shows that the commutation relation $G_t H_s G_{-t} = H_{se^t}$ holds and hence $g_t h_s g_{-t} = h_{se^t}$. Thus the flows h_t and h_{st} for any positive s are metrically and hence spectrally isomorphic. Hence by Proposition 1.23 the horocycle flow has homogeneous Lebesgue spectrum and each transformation h_t, $t \neq 0$ has countable Lebesgue spectrum. □

In fact, it is also true that the horocycle flow (i.e. its infinitesimal generator) has countable Lebesgue spectrum. For this it is enough to show that there are countably many mutually orthogonal subspaces in $L^2(X,m)$ simultaneously invariant under the geodesic and horocycle flows. Then the above argument can be applied separately to each of those subspaces producing Lebesgue spectrum there.

To find such subspaces one can use elements of theory of unitary representations for semisimple Lie groups; in this case $SL(2,\mathbb{R})$. Namely, notice that the whole group $SL(2,\mathbb{R})$

[1]The geometric terminology came from the interpretation of $SL(2,\mathbb{R})$ as the unit tangent bundle to the hyperbolic plane $\mathbb{H}^2$ which can be identified with the symmetric space $SO(2)\backslash SL(2,\mathbb{R})$, see, e.g., [79, Section 17.5].

acts by left translations on X and the corresponding Koopman operators produce a unitary representation of $SL(2,\mathbb{R})$ in $L^2(X,m)$. Consider the compact subgroup of rotations $SO(2) \subset SL(2,\mathbb{R})$. One sees easily that the action of that group decomposes into eigenspaces corresponding all the characters. Each such eigenspace is invariant under the whole group $SL(2,\mathbb{R})$.

2.2. *Mixing and recurrence*

2.2.1. *Mixing.* (See also [8, Section 3.6h].) A measure μ on the circle is called *mixing* (or sometimes *Rajchman*) if its Fourier coefficients (correlation coefficients) $\hat{\mu}_n = \int_{S^1} z^n \, d\mu(z)$ converge to 0 as $n \to \pm\infty$.

By the Riemann–Lebesgue lemma any absolutely continuous measure is mixing. However there are many mixing singular measures as well. To see this notice that the correspondence between taking convolutions and multiplication of Fourier coefficients implies the following

PROPOSITION 2.3. *Convolution of two mixing measures is mixing. If for a measure μ and for some m the m*th *convolution $\mu^{(m)} = \mu * \cdots * \mu$ of μ with itself is mixing, then μ is mixing.*

EXAMPLE 2.4. Let C be the projection of the standard (ternary) Cantor set on the unit interval to the circle. Construct the "uniform" measure μ on C by assigning the measures $1/2^n$ to the intersection of C with the intervals of nth order.[2] This measure is obviously singular. It is however mixing. This can be seen by looking at the convolution $\mu * \mu$ of μ with itself. The convolution is absolutely continuous, and hence mixing (its density is easy to calculate). Thus the Fourier coefficients of μ which are square roots of Fourier coefficients of $\mu * \mu$ also vanish at infinity.

PROPOSITION 2.5. *Any measure absolutely continuous with respect to a mixing measure is mixing.*

PROOF. Let μ be a mixing measure and $\rho \in L^1(S^1, \mu)$. We need to show that $\rho\mu$ is a mixing measure. We will prove decay of correlation coefficients without assuming non-negativity of ρ. First, notice that multiplication by the independent variable correspond to the shift in Fourier coefficients and hence preserves the decay of correlation coefficients at infinity. Second, this decay is a linear property. Thus for any trigonometric polynomial p the correlation coefficients of the complex measure $p\mu$ decay at infinity. Since trigonometric polynomials are dense in $L^1(S^1, \mu)$, the same property holds for $\rho\mu$. □

REMARK. The above argument naturally can be applied to the case of Lebesgue measure and thus it gives a proof of the Riemann–Lebesgue Lemma.

[2] This is the Hausdorff measure corresponding to the exponent $\frac{\log 2}{\log 3}$ which is equal to the Hausdorff dimension of C.

Mixing measures can be characterized in a geometric way as being asymptotically uniformly distributed. Let $E_n : S^1 \to S^1$ be the nth power map: $E_n(z) = z^n$. The pull-back $f^*\mu$ of the measure μ under a transformation f is defined by $f^*\mu(A) = \mu(f^{-1}A)$ for any μ-measurable set A.

PROPOSITION 2.6. *A measure μ on the circle is mixing if and only if the sequence $(E_n)^*\mu$ weakly converges to Lebesgue measure as $n \to \pm\infty$.*

PROOF. Since the mth Fourier coefficient of the measure $(E_n)^*\mu$ is equal to $\hat{\mu}_{mn}$, mixing implies that every non-zero Fourier coefficient of $(E_n)^*\mu$ converges to 0 as $n \to \pm\infty$ while the zero Fourier coefficients of all those measures are equal to one. Convergence of Fourier coefficients for probability measures on the circle is equivalent to weak convergence. This proves the "only if" part.

Conversely, weak convergence implies that the first Fourier coefficients of $(E_n)^*\mu$ which are equal to $\hat{\mu}_n$ converge to zero as $n \to \pm\infty$ implying mixing. □

By Proposition 2.5 mixing is a property of an equivalence class of measures. This justifies the following definition.

DEFINITION 2.7. A unitary operator is called *mixing* if some (and hence any) measure of maximal spectral type is mixing.

In fact, mixing can characterized directly:

PROPOSITION 2.8. *A unitary operator U is mixing if and only if U^n converges to* 0 *in the weak operator topology as $n \to \infty$.*

2.2.2. *Rigidity and pure point spectrum.* (See also [8, Section 3.6e].) Rigidity is a property of spectral measures which is opposite to mixing in a natural way.

DEFINITION 2.9. A measure μ on the circle is called *rigid* (or sometimes a *Dirichlet* measure) if $\hat{\mu}_{n_k} \to \mu(S^1)$ for some sequence $n_k \to \infty$.

The contrast between rigidity and mixing is seen from the following geometric characterization.

PROPOSITION 2.10. *The measure μ is rigid if and only if for certain sequence $n_k \to \infty$ the sequence of measures $(E_{n_k})^*\mu$ weakly converges to a δ-measure.*

LEMMA 2.11. *If for a certain sequence $n_k \to \infty$ $\hat{\mu}_{n_k} \to \alpha\mu(S^1)$ where $|\alpha| = 1$, then for any $m \in \mathbb{Z}$, $\hat{\mu}_{mn_k} \to \alpha^m\mu(S^1)$.*

PROOF. Fixing m, for every ε, there exists k_0 such that for $k > k_0$ if $A_k = \{\theta \in S^1 : |e^{2\pi i n_k \theta} - e^{2\pi i \alpha}| > \varepsilon\}$, then $m(A_k) < \varepsilon^2$. The conclusion now follows from the fact that, for complex numbers z_1 and z_2 such that $|z_1| = |z_2| = 1$, $|z_1^m - z_2^m| \leqslant m|z_1 - z_2|$. □

PROOF OF THE PROPOSITION. If $(E_{n_k})^*\mu \to \delta_\alpha$, then by Lemma 2.11 for any natural number p, $(E_{pn_k})^*\mu \to \delta_{\alpha^p}$ and we are able to produce a sequence $l_k = p_k n_k$ such that $(E_{l_k})^*\mu \to \delta_1$, which is just rigidity since the first Fourier coefficient of the measure $(E_{l_k})^*\mu)$ is equal to $\hat{\mu}_{l_k}$. This also gives the converse. □

COROLLARY 2.12. *Any measure absolutely continuous with respect to a rigid measure is rigid.*

Thus rigidity like mixing is also a property of an equivalence class of measures and hence one can speak about *rigid* unitary operators.

PROPOSITION 2.13. *Any atomic measure on S^1 is rigid.*

PROOF. For any atomic measure all but arbitrary small measure is concentrated on a finite set. But for any finite set $\lambda_1, \ldots, \lambda_n \in S^1$ one can find a sequence $n_k \to \infty$ such that $\lambda_i^{n_k} \to 1$, $i = 1, \ldots, n$. □

For a given unitary operator U the closure of powers U^n, $n \in \mathbb{Z}$ in the strong operator topology is a useful object whose structure is related to the spectral properties of U. First, all of its elements are unitary operators, and it forms a Polish Abelian group under composition. Let us denote this group by $\mathcal{G}(U)$.

It follows from the definition of rigidity that the operator U is rigid if and only if the group $\mathcal{G}(U)$ is perfect.

Notice that the group of Koopman operators is a closed subgroup of the group of all unitary operators in the strong operator topology (this is not true in weak topology). Thus we have the following useful corollary.

COROLLARY 2.14. *Any rigid measure preserving transformation T of Lebesgue space has an uncountable centralizer, i.e. there are uncountably many measure preserving transformations commuting with T.*

In fact, unitary operators with pure point spectrum (i.e. operators whose maximal spectral type is atomic) can be characterized by a property stronger than rigidity.

PROPOSITION 2.15. *A unitary operator U has pure point spectrum if and only if the group $\mathcal{G}(U)$ is compact.*

Thus, for any transformation with pure point spectrum a certain compact Abelian group can be associated. It is not surprising then that such transformations can be represented as shifts on compact Abelian groups. See Section 3.4.3 for a detailed discussion. At the moment we just notice that given a compact Abelian group G any translation on G preserves Haar measure χ and has pure point spectrum since characters are eigenfunctions for it and characters form an orthonormal basis in $L^2(G, \chi)$. Let us illustrate this by some concrete examples. Recall that a measure preserving transformation is *ergodic* if any invariant measurable set is either a null-set or has null-set complement.

EXAMPLE 2.16.

(1) *Circle rotation*. Let for $\alpha \in \mathbb{R}$

$$R_\alpha : S^1 \to S^1, \quad R_\alpha x = x + \alpha \pmod 1.$$

This rotation is ergodic if and only if α is irrational.

(2) *Translation on the torus*. For a vector $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ the translation T_α of the n-torus

$$T_\alpha x = x + \alpha \pmod 1$$

is ergodic with respect to Haar measure is and only if $\alpha_1, \ldots, \alpha_n$ and 1 are independent over rationals.

The translation vector α is sometimes called the vector of *frequencies* and the rational relations between its components are called *resonances*. Even if there are no resonances there may be near resonances which play important role in causing complications when the translation is modified in some way.

(3) *Adding machine* or *odometer*. An *adding machine* is an ergodic transformation with pure point spectrum all of whose eigenvalues are roots of unity. In other words, it is an ergodic shift on the dual to a subgroup of $\mathbb{Q}/\mathbb{Z}$, or by duality on a factor of the group of ideles (Example 1.4). It can also can be characterized as the inverse limit of cyclic permutations.

For example, a translation T_{x_0} on the group $\mathbb{Z}_p$ of p-adic integers (p prime) is ergodic is and only if x_0 is not divisible by p.

(4) *Shifts on solenoids*. A solenoid is the inverse limit of tori of the same dimension.

2.2.3. *Mild and weak mixing.* (See also [8, Sections 3.6f,g].)

DEFINITION 2.17. A measure μ on the circle is called *mildly mixing* if no measure absolutely continuous with respect to μ is rigid.

Notice that given a sequence $n_k \to \infty$, the space of all functions $f \in L^2(X, \mu)$ for which $U_T^{n_k} f \to f$ is a unitary $*$-subalgebra. Hence by Proposition 1.2 if U_T is not mildly mixing, T has a rigid factor. Thus

T is mildly mixing if and only if it has no non-trivial rigid factors.

Proposition 2.13 implies that any mildly mixing measure is continuous (non-atomic). The following characterization justifies calling non-atomic measures *weak* (*or weakly*) *mixing*.

Recall that a subset $S \subset \mathbb{Z}$ is called a *set of full density* if

$$\lim_{n\to\infty} \frac{S \cap [-n, n]}{2n+1} = 1.$$

PROPOSITION 2.18. *A measure μ on the circle is non-atomic if and only if for a set S of full density*

$$\lim_{n\in S, n\to\pm\infty} \hat{\mu}_n = 0.$$

SKETCH OF PROOF. Let Δ be the diagonal of $S^1 \times S^1$. Fubini's theorem implies that $(\mu\times\mu)(\Delta) = \sum |\hat{\mu}(\{\lambda\})|^2$, where the summation is taken over the atoms of μ. Now we have

$$\int \frac{1}{N}\sum_{n=1}^{N} \exp\bigl(2i\pi n(x-y)\bigr)\, d(\mu\times\mu) = \frac{1}{N}\sum_{n=1}^{N} |\hat{\mu}_n|^2.$$

By Lebesgue theorem the left-hand side of this last equality converges, when $N\to+\infty$, to $\mu\times\mu(\Delta)$. A simple calculation shows the equivalence between $\lim_{N\to\infty}\frac{1}{N}\sum_1^N |\hat{\mu}_n|^2 = 0$ and

$$\lim_{n\in S, n\to\pm\infty} \hat{\mu}_n = 0$$

for a set S of full density. □

While checking convergence along a sequence of full density may present problems there is an alternative criterion which is often convenient in the context of ergodic theory.

PROPOSITION 2.19. *An equivalence class of measures on S^1 is non-atomic if and only if there exists a sequence $n_k\to\infty$ such that for any measure μ from this class* (*or, equivalently, for an L^1 dense set of such measures*) $\hat{\mu}_{n_k}\to 0$.

2.2.4. *An elliptic paradigm*

Diophantine and Liouvillean behavior. Simple rigid spectrum, whether atomic, mixed or continuous, is the second type (after countable Lebesgue spectrum) ubiquitous in ergodic theory and other branches of dynamics. These spectral properties are associated with the *elliptic* behavior [8, Section 7] in its two manifestations, Diophantine and Liouvillean [47]. Simplicity of the spectrum relies on criteria like Theorem 1.21, rigidity on Proposition 2.10.

Diophantine paradigm is associated with rather simple and fully understood type of behavior: pure point spectrum with frequency vector which avoids too close near resonances, see Example 2.16(2); it is of great importance in classical mechanics due to KAM theory [30].

Liouvillean behavior is associated with simple singular (and usually continuous) rigid spectrum and with a very fast periodic approximation, see Proposition 5.39, and for more details, [81,78]; it is typical in the weak topology in the space of measure preserving transformations and various other spaces of dynamical systems, see Theorems 5.47 and 5.49, [47,78]. Although more exceptional from the point of view of classical and Hamiltonian

mechanics, it is still unavoidable in typical perturbations of completely integrable systems, twist maps and so on [70].

Time change in linear flows on $\mathbb{T}^2$. We will present now the most classical and very simple situation where non-trivial Liouvillean behavior appears.

We begin with an irrational linear flow on the two-dimensional torus. We will denote cyclic coordinates on $\mathbb{T}^2$ by x and y. Given a vector $\gamma = (\gamma_1, \gamma_2)$ the flow $\{T_\gamma^t\}$ is generated by the constant vector field with coordinates γ_1, γ_2 and has the form

$$T_\gamma^t(x, y) = (x + t\gamma_1, y + t\gamma_2) \pmod 1.$$

We assume that the slope γ_1/γ_2 is irrational which is equivalent to minimality or ergodicity of the flow.

Now consider a time change of the flow. Namely take a positive C^∞ function ρ and consider the flow generated by the vector field $(\rho\gamma_1, \rho\gamma_2)$. Denote the new flow by $\{S^t\}$. This flow preserves the smooth measure $\rho^{-1}\lambda$. It is also rigid, and has simple spectrum. If the number γ_1/γ_2 is *Diophantine*, i.e. there exist positive numbers N and C such that for any integers p and q,

$$|\gamma_1/\gamma_2 - p/q| > C/q^N$$

then there exists a C^∞ diffeomorphism preserving the orbits which conjugates the flow $\{S^t\}$ to a linear flow and hence has pure point spectrum. This goes back to Kolmogorov [94], see [8, Section 7.3], [78, Section 11.2] for proofs and discussions. Thus in the Diophantine situation the orbit structure of time changes is rigid.

On the other hand, if the slope is not a Diophantine number then generically in the C^∞ topology for ρ the flow $\{S^t\}$ is weakly mixing [42] (see also [78, Section 13.3] for related results and historical discussion). Furthermore, for some special values of the slope one can find a real-analytic ρ for which the flow has mixed spectrum [48]. We will continue discussion of this and similar situations in Section 5.6.3.

2.3. *Homogeneous systems*

We briefly mention now a very important class of dynamical systems which is discussed in much greater detail in [10]; see also [4], especially Section 3.

Let G be a Lie group, $\Gamma \subset G$ a lattice, i.e. a discrete group with the factor of finite volume (compact or not). A *homogeneous dynamical system* is the action of a subgroup $H \subset G$ on the homogeneous space G/Γ by left translations. Both horocycle and geodesic flows are examples of homogeneous dynamical systems where H is a one-parameter subgroup. Even more basic examples are translations on the torus or one-parameter groups of such translations (linear flows).

Homogeneous systems possess large symmetry since any such system is a part of a transitive action of G by left translations. Due to this symmetry spectral analysis of homogeneous dynamical systems can be carried out with the help of the theory of unitary

representations of Lie groups (see [10, Section 2.3]). While ergodic properties of homogeneous flows may be complicated and surprising (see Section 6.2.2) their spectral properties are rather simple.

If H is a one-parameter subgroup of G the left action by H is called a *homogeneous flow*.

THEOREM 2.20 [25]. *The spectrum of any homogeneous flow is the sum of pure point and countable Lebesgue.*

A similar conclusion holds for *homogeneous maps*, i.e. the homogeneous actions of $\mathbb{Z}$; this of course follows immediately from Theorem 2.20 for most homogeneous maps since such maps are parts of homogeneous flows.

This is similar to the case of automorphisms and affine maps on compact Abelian groups (Section 2.1.2).

3. General properties of spectra for measure preserving transformations and group actions

3.1. *The realization problem and the spectral isomorphism problem*

3.1.1. *Formulation of the problems.* Unitary operators which appear as Koopman operators associated with measure preserving transformations and, more generally, group actions, possess some special properties. The interface between the unitary operator theory (and the theory of unitary group representations), and ergodic theory centers on two general problems:

SPECTRAL REALIZATION PROBLEM. What are possible spectral properties for a Koopman operator or a group of such operators?

SPECTRAL ISOMORPHISM PROBLEM. Given two Koopman operators U_T and U_S (or groups) which are unitarily equivalent (i.e. have the same spectral invariants) what extra information is needed to conclude that the measure preserving transformations T and S (or the corresponding group actions) are isomorphic? More specifically, one is interested in the cases when such extra non-spectral invariants can be reasonably described, and, in particular, when they are not needed at all.

Both of those problems go back to the founding text of the modern ergodic theory, the 1932 article by John von Neumann [155]. Concerning the Spectral Realization Problem there are very few known restrictions, all of them quite general. The proofs are not difficult and all results in this direction will be presented in the rest of this section.

Still, many simply sounding questions are unanswered. Here is a famous example.

PROBLEM 3.1. Does there exist a measure preserving transformation whose Koopman operator has simple Lebesgue spectrum (or even Lebesgue spectrum of bounded multiplicity) in $L^2_0(X, \mu)$.

On the other hand, there is a large number of "positive" results asserting existence of measure preserving transformations with specific properties. This is achieved via a variety of specific constructions. Essentially the whole later part of this survey is dedicated to the development of these constructions and presenting specific examples, sometimes natural, sometimes "exotic".

3.1.2. *Elementary restrictions*

Invariance of constants. The most basic restriction on the spectral realization is presence of eigenvalue 1 in the spectrum, since constants are invariant functions. Thus the maximal spectral type of any Koopman operator always has an atom at 1. Due to this simple observation by spectral properties of a measure preserving transformation one usually means the corresponding properties of the operator U_T in the orthogonal complement to the space of constants, i.e. the space $L^2_0(X,\mu)$ of square-integrable functions with zero average. Sometimes, however, it is useful to remember presence of the atom at one which we will naturally denote by δ_1. Let μ_0 be a measure of maximal spectral type in the space $L^2_0(X,\mu)$. Then μ, the maximal spectral type in $L^2(X,\mu)$ can be represented by $\mu_0+\delta_1$. Consider the convolution

$$\mu*\mu=(\mu_0+\delta_1)*(\mu_0+\delta_1)=\mu_0*\mu_0+\mu_0+\delta_1=\mu_0*\mu_0+\mu>\mu$$

(equality and inequality signs refer to measure types).

This simple fact can be expressed in a way useful for the discussion of the convolution problem (Section 3.5).

PROPOSITION 3.2. *The maximal spectral type μ of the Koopman operator in the whole space $L^2(X,\mu)$ is dominated by its convolution $\mu*\mu$.*

Symmetry of the spectrum. Another, almost equally basic, is the symmetry of both the maximal spectral type and the multiplicity function with respect to the involution $\chi\to\chi^{-1}$ of the dual group G^*. This immediately follows from the fact the Koopman operator preserves the complex conjugation in $L^2(X,\mu)$.

In particular this implies the following restriction of the spectral realization.

PROPOSITION 3.3. *Any Koopman operator is unitarily isomorphic to its inverse.*

3.2. *Rokhlin lemma and its consequences*

3.2.1. *The Rokhlin lemma.* Recall that an action of a group is called *free* if the stationary subgroup of almost every point in the space is trivial; for $\mathbb{Z}$ this means aperiodicity.

The Rokhlin lemma gives a way to produce an approximate section for a free action for certain kinds of discrete groups, and therefore to control large pieces of orbits on a large part of the space.

In general, this is related to the existence of sets in the group which tile it. A set A in a group G is said to *tile* G if there exists a family of elements of G, g_i, $i \in I$, such that $G = \bigcup_{i\in I} g_i A$ and the sets $g_i A$ are mutually disjoint. All countable Abelian groups can be endowed with Fölner sequences such that every set in the sequence tiles the group, and therefore a version of the Rokhlin lemma can be stated in this framework. The proper setting for the most general version of the Rokhlin lemma is in fact for actions of countable amenable group. In this case there need not be any set in a Fölner sequence which tiles the group. However the group can always be almost tiled by a finite number of elements in the Fölner sequence which furthermore can be chosen as invariant as one wishes. This gives rise to the Ornstein–Weiss version of the Rokhlin lemma for amenable groups which has many important applications and in particular is the first key step in extending the Ornstein isomorphism theory to actions of arbitrary amenable groups [121,122].

Rokhlin [137] considered only $\mathbb{Z}$ actions, see [12, Section 5] for a proof in that case. Here we consider a free measure preserving action of $\mathbb{Z}^d$ on a measure space (X, μ), see [28,84].

THEOREM 3.4. *Consider a free action of $\mathbb{Z}^d$ on (X, μ) generated by d commuting automorphisms $T_1, T_2, \ldots, T_d$. For every $\varepsilon > 0$, and an integer N, there exists a set $F \subset X$ such that the sets $T_1^{n_1} T_2^{n_2}, \ldots, T_d^{n_d} F$, $0 \leqslant n_1, n_2, \ldots, n_d \leqslant N-1$, are mutually disjoint and their union has measure greater than $1-\varepsilon$.*

REMARK. Notice that the assumption of freeness which is natural in the ergodic theory setting is very restrictive in other branches of dynamics such as topological dynamics [8, Section 2] or theory of smooth dynamical systems [8, Section 5] since periodic orbits form an important ingredient of the orbit structure in many cases. For example, for hyperbolic systems [8, Section 6] periodic orbits are dense.

PROOF. To simplify notations we consider the case $d = 2$, i.e. we consider a $\mathbb{Z}^2$ action on (X, μ) generated by two commuting measure preserving transformations S and T. Fix an integer $L > N^2/\varepsilon^2$. Since the action is free one can find, using the ergodic decomposition (Section 3.3), a measurable set A such that:

(1) the sets $S^i T^j(A)$, $-L < i, j < +L$ are pairwise disjoint

(2) $\mu_y(A) > 0$ for almost every y in the ergodic decomposition.

Thus $\bigcup_{m\geqslant 0, n\geqslant 0} S^m T^n A = X$ and there exists M' such that for all $M > M'$, $\mu(\bigcup_{0\leqslant m,n\leqslant M} S^m T^n A) > 1-\varepsilon^2$. For an element $x \in X$ its itinerary is an element ω in $\{0,1\}^{\mathbb{Z}\times\mathbb{Z}}$ where $\omega_{i,j} = 1$ if $S^i T^j x$ is in A, $\omega_{i,j} = 0$ otherwise. We call M-itinerary the restriction of the previous itinerary to the values of (i, j) which lie inside the square $C_M = \{(i, j)\colon\ 0 \leqslant i, j \leqslant M-1\}$. An itinerary ω being given, we consider $Y_\omega \subset \mathbb{Z}^2$ the union of these indices $(i, j) \in \mathbb{Z}^2$ such that $\omega_{i,j} = 1$. We call $(C_y, y \in Y_\omega)$ the tiling of $\mathbb{R}^2$ determined by the Voronoi cells

$$C_y = \{x\colon\ |x-y| < |x-y'| \text{ for all } y' \neq y \text{ in } Y_\omega\}.$$

For every such cell C_y,we call y its center. We consider the partition P_M of A whose atoms are made of points which have the same M-itinerary. A cell C_y being given, we consider T_y

the union of the squares of the tiling of $\mathbb{Z}^2$ by squares of size N with base point y which lie entirely inside C_y. Each square c in T_y is of the form $(i_c, j_c \leqslant i, j \leqslant i_c + N - 1, j_c + N - 1)$; (i_c, j_c) is called the base of the square c. Assume that

(3) M is large enough (and in particular greater than M') so that the union of those y in A such that C_y is not contained in a square of size less than $\varepsilon^2 M$ occupies a fraction less that ε^2 of A.

For $p \in P_M$ with itinerary ω_p, we consider F_p the union of $S^{i_c} T^{j_c}(p)$ for all i_c, j_c which are the bases of squares in T_y for all $y \in Y_{\omega_p}$ such that C_y is entirely in C_M. Let $F = \bigcup_{p \in P_M} F_p$. Clearly the sets $S^i T^j F$, $(i, j) \in C_N$ are pairwise disjoint. (1), (2) and (3) imply that the measure of their union is greater than $1 - \varepsilon$. □

3.2.2. *Density of the maximal spectral type.* An important corollary of the Rokhlin lemma is the following restriction on the spectral realization.

THEOREM 3.5. *The support of the spectral measure of the Koopman operator for an aperiodic transformation is the whole circle S^1.*

PROOF. If λ is not in the support of the spectral measure of $U_{\mathbb{T}}$ then $U_T - \lambda \times \mathrm{Id}$ is invertible in $L^2(X, \mathcal{A}, m)$. However, for every ε, for every λ there exists $f \in L^2(X)$ such that $\|f\| = 1$ and $\|U_T f - \lambda f\| < \varepsilon$. This is sufficient to imply what we asserted. Given ε and λ, we construct f in the following way: take $n \ll 1/\varepsilon$, and take a set F, given by the Rokhlin lemma, such that the family of sets $T^i F$, $0 \leqslant i \leqslant n - 1$, is a disjoint family and such that the measure of their union is $\geqslant 1 - \varepsilon$. Define now f as taking the constant value λ^i on $T^i F$, $0 \leqslant i \leqslant n - 1$, 1 on the complement of the union of the $T^i F$. □

3.2.3. *Combinatorial constructions.* Rokhlin lemma has an interesting "negative" aspect. It implies that all asymptotic behavior of a measure preserving transformation depends on sets of arbitrary small measure and hence can be altered in an arbitrary way by changing the action on such a set. In the case of a single transformation this can be rephrased by saying that if one defines the *uniform topology* by the metric

$$d_u(T, S) = \mu\{x\colon Tx \neq Sx\}$$

then

PROPOSITION 3.6. *Conjugates of any aperiodic transformation are dense in the uniform topology in the set of all aperiodic measure preserving transformations.*

PROOF. Fixing n and ε construct Rokhlin towers with given n and ε for two aperiodic transformations T and S. Thus the towers have the form $T^i F$ and $S^i F'$, $i = 0, 1, \ldots, n-1$, correspondingly. Without loss of generality we may assume that the *bases* F and F' of two towers have the same measure. Pick some measure preserving transformation $h : F \to F'$

and define H on $T^i F$ as $S^i \circ h \circ T^{-i}$ for every $i = 0, 1, \ldots, n-1$. Complete H in an arbitrary way to a measure preserving transformation of X. Obviously

$$H \circ T \circ H^{-1} = S \quad \text{on} \bigcup_{i=0}^{n-1} S^i F',$$

hence

$$d_u\big(H \circ T \circ H^{-1}, S < \varepsilon\big). \qquad \square$$

This is somewhat deceptive however. Small sets determining asymptotic behavior become more and more complicated as their measure decreases.

A related fact is that the base F of a Rokhlin tower and its images although of small measure normally become "diffused" all over the space. The idea of looking at transformations for which the level sets of Rokhlin towers stay sufficiently "compact" leads to the notion of *rank* (Section 5.2.2) and the concept of *periodic approximation* (Section 5.4) as well as to the class of constructions known as *cutting and stacking* discussed in Section 5.2.

3.3. *Ergodicity and ergodic decomposition*

3.3.1. *Definitions*

DEFINITION 3.7. A measure preserving transformation $T : (X, \mu) \to (X, \mu)$ is *ergodic* if 1 is a simple eigenvalue of the Koopman operator U_T.

Equivalently, T is ergodic is any T-invariant measurable set A is either null ($\mu(A) = 0$) or co-null ($\mu(X \setminus A) = 0$).

For an arbitrary measure preserving transformation T consider the space I_T of invariant functions for the Koopman operator U_T. This space is generated by characteristic functions of invariant sets and by multiplicativity the product of U_T-invariant functions is also U_T invariant. Thus I_T is a unitary subalgebra of $L^2(X, \mu)$ (Proposition 1.2) and hence it defines a factor of T on which T obviously acts as the identity. Denote the measurable partition corresponding to that factor by η_T. The transformation T acts on elements of this partition preserving the system of conditional measures. Ergodic Decomposition Theorem [8, Theorem 3.4.3] states that for almost every $c \in \eta_T$ T acts ergodically with respect to the conditional measure μ_c. See [8, Sections 4.2d, 4.2e] for a more detailed discussion and references to detailed proofs.

DEFINITION 3.8. A measure preserving transformation is called *totally ergodic* if any of its non-zero powers is ergodic.

Total ergodicity is equivalent to the absence of roots of unity (other than 1 itself) among the eigenvalues. The inverse limit of totally ergodic transformations is totally ergodic.

Adding machines from Example 2.16(3) are interesting examples of ergodic but not totally ergodic transformations. This simple property is important in various non-spectral aspects of ergodic theory. A typical situation where total ergodicity plays a role is the following: when one considers the ergodic averages of an L^2 function taken at iterates which are perfect squares, there is convergence in L^2 and also almost everywhere (this is a deep theorem of Bourgain [23]); however the limit is the integral of the function only in the case when the transformation is totally ergodic.

3.3.2. *Ergodicity and spectrum.* Thus, the study of spectral properties of general measure preserving transformations can be separated into two questions: (i) finding ergodic decomposition, in particular establishing ergodicity, and (ii) studying spectral properties of the operators which appear on the ergodic components. Establishing ergodicity for a particular transformation or a class of transformation may be highly non-trivial. However in this survey we will primary (although not exclusively) discuss spectral and other closely related properties for ergodic measure preserving transformations. The argument for separating the study of ergodic decomposition from spectral analysis in the ergodic case may be illustrated by the following example which demonstrate that some properties of ergodic decomposition are non-spectral.

EXAMPLE 3.9. Let T and S be two ergodic measure preserving transformations on the measure spaces (X, μ) and (Y, ν) respectively. For any $0 < t < 1$ consider the space $X_t \stackrel{\text{def}}{=} X \cup (Y \times [0, t])$ with the probability measure $\mu_t \stackrel{\text{def}}{=} (1-t)\mu + \nu \times \lambda$, where λ is Lebesgue measure. Let T_t be defined on X_t as T on the X part and as $S \times \mathrm{Id}$ on the $Y \times [0, t]$. Obviously the spaces of ergodic components for X_t for different t are not isomorphic because this space contains exactly one atom of measure t. Hence T_t for different t are not isomorphic. However, they are spectrally isomorphic since they all have countable multiplicity for the eigenvalue one and the spectrum in the orthogonal complement to invariant functions is the union of the spectrum of U_T and the spectrum with the maximal spectral type of U_S and countable multiplicity.

3.3.3. *Difference between spectral and metric isomorphism in the ergodic case*

Entropy as an extra invariant. The following classical example shows that the Spectral Isomorphism Problem is non-trivial even in the ergodic situation.

EXAMPLE 3.10. Consider the Bernoulli shift σ_N on the space Ω_N of bi-infinite sequences of an alphabet N symbols provided with the product measure μ_p where $p = (p_0, \ldots, p_{N-1})$ is a probability distribution on the alphabet.

The spectrum of this transformation is always countable Lebesgue. This can be readily seen as follows. Let for $n \in \mathbb{Z}$, H_n be the subspace of $L^2_0(\Omega_N, \mu_p)$ of all functions which depend only on coordinates ω_k of the sequence $\omega \in \Omega_N$ with $k \leqslant n$. By definition of the shift one has $U_{\sigma_N} H_n = H_{n+1}$. The spaces H_n generate $L^2_0(\Omega_N, \mu_p)$ since every function can be approximated by a function which depends only on finitely many coordinates. Similarly $\bigcap_{n\in\mathbb{Z}} H_n = \{0\}$. Now let G_n be the orthogonal complement to H_n in H_{n+1}. Obvi-

ously these spaces are infinite-dimensional; they are orthogonal to each other by definition, $U_{\sigma_N} G_n = G_{n+1}$ and $\bigoplus_{n\in\mathbb{Z}} G_n = L^2_0(\Omega_N, \mu_p)$.

However the *entropy* $-\sum_{i=0}^{N-1} p_i \log p_i$ is an invariant of metric isomorphism [8, Section 3.7] so there are uncountably many non-isomorphic measure preserving transformation with countable Lebesgue spectrum.

This example directly extends to the case of $\mathbb{Z}^k$ actions and less directly to the continuous-time case [12].

In the case of zero entropy extra invariants including Kushnirenko's sequence entropy [97] and *slow entropy* [8, Section 3.7l], [83] sometimes distinguish spectrally isomorphic systems; see [97] for a classical example of non-isomorphic flows with countable Lebesgue spectrum and zero entropy which are distinguished by sequence entropy.

Asymmetry of metric isomorphism. Entropy shares with the spectral invariants the property of being symmetric with respect to the reversal of time [8, Section 3.7i(4)] and thus never distinguishes a transformation from its inverse. However there are instances where T and T^{-1} are not metrically isomorphic. The earliest examples of that phenomena were found in 1968 by S. Malkin [110] and are not particularly exotic: the spectrum is simple and the transformation itself is a two-point extension of an irrational rotation R_α with only four discontinuity points. These transformations have zero entropy. An interesting criterion which helps to decide whether a transformation T is conjugate to its inverse is in [66]. It implies for example that is the square of the conjugating map S is ergodic then all essential values of the multiplicity function for T are even.

3.4. *Pure point spectrum and extensions*

3.4.1. *Multiplicative structure of eigenfunctions.* As we pointed out, ergodicity is a spectral invariant: it is equivalent to 1 being a simple eigenvalue.

The complex conjugate of an eigenfunction is also an eigenfunction with the complex conjugate eigenvalue.

Ergodicity implies that eigenfunctions have constant absolute value: if $U_T f = \lambda f$ then

$$U_T(f \cdot \bar{f}) = U_T(f) \cdot U_T(\bar{f}) = \lambda\bar{\lambda} f \bar{f} = f\bar{f},$$

hence $f\bar{f} \equiv \text{const}$. Furthermore, both the eigenfunctions and the eigenvalues for an ergodic transformation form a group invariant under complex conjugation. Consequently linear combinations of eigenfunctions form an $*$-algebra and hence their L^2 closure is an invariant unitary $*$-subalgebra of $L^2(X, \mu)$ which we will denote by $\mathcal{K}(T)$. Thus by Proposition 1.2 $\mathcal{K}(T)$ determines a factor of T called the *Kronecker factor* of T. We will denote this factor transformation by $T_{\mathcal{K}}$; it is the maximal factor with pure point spectrum [156, 29]. The measurable partition corresponding to the Kronecker factor will also be denoted by $\mathcal{K}(T)$.

3.4.2. *The isomorphism theorem.* In the case of pure point spectrum the Spectral Isomorphism Problem has a complete and optimal solution.

THEOREM 3.11 (von Neumann Discrete Spectrum Theorem). *Any two ergodic measure preserving transformations with pure point spectrum that are spectrally isomorphic* (*i.e. have the same groups of eigenvalues*) *are metrically isomorphic. A complete system of invariants is given by the countable subgroup $\Gamma \subset S^1$ of eigenvalues.*

SKETCH OF PROOF. Let $T : (X, \mu) \to (X, \mu)$ be an ergodic measure preserving transformation with pure point spectrum and let Γ be the group of eigenvalues for U_T. Let x_0 be a common Lebesgue point for all eigenfunctions of U_T. Denote for each eigenvalue $\gamma \in \Gamma$ by f_γ the unique eigenfunction for which the Lebesgue value at x_0 is 1. Then

$$f_{\gamma_1\gamma_2} = f_{\gamma_1} f_{\gamma_2}. \tag{3.1}$$

Now identify Γ with the group of characters of the compact dual group Γ^* and denote the character on Γ^* corresponding to the evaluation at γ by χ_γ. Thus, we have orthonormal bases $\{f_\gamma\}_{\gamma\in\Gamma}$ and $\{\chi_\gamma\}_{\gamma\in\Gamma}$ in the Hilbert spaces $L^2(X, \mu)$ and $L^2(\Gamma^*, \lambda)$ correspondingly, where λ is the normalized Haar measure.

Now extend the correspondence $f_\gamma \to \chi_\gamma$ by linearity to a unitary operator $V : L^2(X, \mu) \to L^2(\Gamma^*, \lambda)$, which is multiplicative on the eigenfunctions by (3.1) and preserves complex conjugation. Their finite linear combinations are dense in $L^2(X, \mu)$, so V is generated by a measure preserving invertible transformation $H : (X, \mu) \to (\Gamma^*, \lambda)$. One immediately sees that $VU_TV^{-1}\chi_\gamma(s) = \gamma\chi_\gamma(s) = \chi_\gamma(s_0 s)$ for any $s \in \Gamma^*$, hence $H \circ T \circ H^{-1} = L_{s_0}$. □

For another proof see Section 4.1.2(6). See also [29, Section 12.2] for yet another proof and detailed discussion.

3.4.3. *Representation by compact Abelian groups translations*

THEOREM 3.12. *An ergodic transformation with pure point spectrum whose group of eigenvalues is Γ is metrically isomorphic to the translation on the compact group Γ^* of characters of Γ, considered as a discrete group, by the character s_0 that defines the inclusion $\Gamma \hookrightarrow S^1$. The invariant measure is Haar measure.*

Furthermore, every countable subgroup of the unit circle appears as the group of eigenvalues for an ergodic measure preserving transformations of a Lebesgue space with pure point spectrum.

Thus, translations on compact Abelian groups provide universal models for ergodic transformation with pure point spectrum. This justifies looking for criteria of ergodicity for such translations as well as considering characteristic examples.

PROPOSITION 3.13. *Translation T_{h_0} on a compact Abelian group H, $T_{h_0}(h) = hh_0$ is ergodic with respect to Haar* (*Lebesgue*) *measure if and only if for any character $\chi \in H^*$ $\chi(h_0) \neq 1$.*

Furthermore, ergodicity with respect to Haar measure is equivalent to topological transitivity, minimality and unique ergodicity.

Recall that the *weak topology* on the group of all measure preserving transformations of a Lebesgue space coincides with the strong operator topology for the Koopman operators.

PROPOSITION 3.14. *The centralizer of an ergodic translation T_{h_0} on a compact Abelian group H in the weak topology on the group of all Haar measure preserving transformations of H consists of all translations of H.*

This implies that ergodic transformations with pure point spectrum possess a certain kind of rigidity: Isomorphism and factor maps between such systems are rather limited.

Notice that the centralizer described in Proposition 3.14 coincides with the closure $\mathcal{G}(U_T)$ of the powers of U_T. By Proposition 2.15 if T has pure point spectrum then $\mathcal{G}(U_T)$ is a compact Abelian group. The multiplication by U_T is a translation on that group which preserves Haar measure χ. It follows from Theorems 1.6 and 3.11 that

PROPOSITION 3.15. *If a measure preserving transformation T has pure point spectrum than the multiplication by U_T on $(\mathcal{G}(U_T), \chi)$ is metrically isomorphic to T.*

3.4.4. *Invariance of the spectrum with respect to the discrete part.* By comparing the correlation coefficients for an arbitrary function $g \in L_0^2$ with those of the function $f \cdot g$ where f is an eigenfunction of absolute value one with the eigenvalue $\exp 2\pi i\alpha$ one sees that the spectral measure λ_{gf} is obtained from λ_g by rotation by α. The same argument applies to orthogonal functions with the same spectral measure. Hence we obtain the following general spectral property of measure preserving transformations.

THEOREM 3.16. *The maximal spectral type and the multiplicity function of the operator U_T induced by an ergodic measure preserving transformation T is invariant under multiplication by any eigenvalue.*

3.4.5. *The Kronecker factor.* By Theorem 3.12 the Kronecker factor defined in 3.4.1 is isomorphic to a particular translation on the dual to the group of eigenvalues. The Kronecker factor is the simplest example of a *characteristic factor* for an ergodic measure preserving transformation. Other examples include the maximal distal factor defined in the next subsection whose characteristic property appears in Proposition 4.5.

As was explained in Section 1.1.1 T itself is isomorphic to a skew product transformation over its Kronecker factor.

EXAMPLE 3.17 (*Affine twist on the torus*). An *affine* map of an Abelian group is a composition of an automorphism and a translation. Fix an irrational number α and consider the following affine map of $\mathbb{T}^2$:

$$A_\alpha(x, y) = (x + \alpha, x + y) \pmod 1.$$

This map has mixed spectrum. The Kronecker factor is the circle rotation R_α, the spectrum in the orthogonal complement to this factor is countable Lebesgue. This is the simplest example of a transformation with a *quasi-discrete spectrum* [13].

Transformations with quasi-discrete spectrum provide easiest examples of ergodic spectrally isomorphic transformations with zero entropy which are not metrically isomorphic. This possibility was mentioned in a different context in Section 3.3.3. Here is a simple example in the present context.

EXAMPLE 3.18. Consider the following affine map on $\mathbb{T}^3$,

$$B_\alpha(x, y, z) = (x + \alpha, x + y, y + z) \pmod 1.$$

The maps A_α and B_α are spectrally isomorphic. In both cases there is the same pure point part (the Kronecker factor is the rotation R_α) plus countable Lebesgue spectrum in the orthogonal complement. However A_α is a factor of B_α and a simple argument shows that any multiplicative correspondence must preserve this factor [13].

Proposition 3.14 provides for certain restrictions on isomorphisms between transformations with a pure point component in the spectrum. Such a transformation is an extension of its Kronecker factor. A particularly interesting case is those of a *finite extensions* when the measurable partition $\mathcal{K}(T)$ has finite elements. By ergodicity it follows that the number of elements is almost everywhere constant, say, equal to n, and hence such a transformation is metrically isomorphic to a skew product transformation on $H \times \{0, 1, \dots, n-1\}$ of the form

$$T(x, m) = (T_h x, \sigma_x m),$$

where $\sigma_x \in S_n$, the permutation group. We will briefly return to this subject in Section 3.6.3 and in more detail is Section 5.8.

3.4.6. *Distal systems.* Transformations with quasi-discrete spectrum and finite extensions are specimens of a more general class of systems which appears in many cases, in particular in the Furstenberg ergodic theoretical proof of the Szemeredi's theorem) [60,2].

DEFINITION 3.19. Consider an ergodic transformation $(Y, \mathcal{B}, \mu, S)$, a compact group G with a closed subgroup H and a measurable mapping $\phi : Y \to G$. Call the quotient G/H Z and equip Z with the Borel algebra $\mathcal{C}$ and the Haar measure ν. The transformation S_ϕ acting on $X = Y \times Z$ by $T_\phi(y, z) = (S(y), \phi(y)z)$ leaves the product measure $\mu \times \nu$ invariant. S_ϕ is called an *isometric extension* of S.

DEFINITION 3.20. A transformation T is said to be *distal* if there exists a countable family of T-invariant factor algebras indexed by ordinals $\mathcal{A}_\eta$, $\eta \leqslant \eta_0$, such that $\mathcal{A}_1 = \nu$ (the trivial algebra), $\mathcal{A}_{\eta_0} = \mathcal{A}$, for every $\xi < \eta$, $\mathcal{A}_\xi \subset \mathcal{A}_\eta$, T restricted to $\mathcal{A}_{\eta+1}$ is an isometric extension of its restriction to $\mathcal{A}_\eta$ and if ξ is a limit ordinal, $\mathcal{A}_\xi = \lim \uparrow \mathcal{A}_\eta$, $(\eta \uparrow \xi)$.

PROPOSITION 3.21. *Every ergodic measure preserving transformation has a unique maximal distal factor, i.e. a distal factor such that any other distal factor is contained in it.*

The distal factor contains Kronecker factor and is another example of a characteristic factor. It is trivial if and only if the transformation is weakly mixing. On the other hand, it may contain functions whose spectral type is mixing or even Lebesgue as in Examples 3.17 and 3.18.

Thus it is not defined in spectral terms.

3.5. *The convolution problem*

3.5.1. *Discrete and mixed spectrum.* In this section we will mean by the maximal spectral type of a transformation the maximal spectral type in the whole space L^2 including the atom δ_1 as was discussed in Section 3.1.2. Notice that the group property of the eigenvalues can be expressed equivalently as equivalence of the maximal spectral type of an ergodic transformation with pure point spectrum and its convolution. Thus we obtain the following statement which strengthens Proposition 3.2 in this case.

COROLLARY 3.22. *An atomic measure μ on the unit circle belongs to the maximal spectral type of the Koopman operator for an ergodic measure preserving transformation if and only if μ is equivalent to $\mu * \mu$.*

Furthermore, Theorem 3.16 is equivalent to the following statement.

COROLLARY 3.23. *If μ is a measure of the maximal spectral type for an ergodic measure preserving transformation and μ_d its atomic part then the convolution $\mu * \mu_d$ is equivalent to μ.*

3.5.2. *Continuous spectrum.* Observations above lead to a following question related to the general Spectral Realization Problem.

PROBLEM 3.24. What are connections between the maximal spectral type of an ergodic measure preserving transformation and its convolution with itself?

We will see below (Theorem 5.15, Propositions 5.43 and 5.44, and Theorem 5.49) that in general those measures are not directly connected. On the other hand, let us notice that for a weakly mixing transformation T the Cartesian powers $T \times T$, $T \times T \times T$, etc. including the infinite Cartesian power $T^{(\infty)}$ can be easily analyzed spectrally. In particular, if μ is the maximal spectral type of U_T in L_0^2 then for any $n \in \mathbb{N} \cup \infty$ the maximal spectral type of the n Cartesian power of T is equal to

$$\sum_{i=1}^{n} \mu^{(i)},$$

where $\mu^{(n)}$ is the convolution of n copies of μ. In particular, the measure $\sum_{n=1}^{\infty} \mu^{(n)}$, the maximal spectral type of $T^{(\infty)}$, is equivalent to its convolution. (See also Section 4.1.3.) This implies the following partial result related to the Spectral Realization Problem.

PROPOSITION 3.25. *If the class of a non-atomic measure μ appears as the maximal spectral type of an ergodic measure preserving transformation then for any $n \in \mathbb{N} \cup \{\infty\}$ the class of the measure $\sum_{i=1}^{n} \mu^{(i)}$ also appears a maximal spectral type of an ergodic measure preserving transformation.*

An effective method for realizing maximal spectral types is given by the construction of Gaussian dynamical systems (Section 6.4). It implies one of the few general results in the direction of realization of spectral types.

THEOREM 3.26. *Any non-atomic measure μ on the unit circle symmetric under the reflection in the real axis and equivalent to $\mu * \mu$ appears as a measure of maximal spectral type of an ergodic measure preserving transformation.*

This theorem follows directly from Proposition 6.12 by taking the Gaussian transformation T_μ.

3.6. *Summary*

3.6.1. *General restrictions.* In this section we have described all known general restrictions on the spectral properties of ergodic measure preserving transformations which then has to be taken into account in the discussion of the Spectral Realization Problem. For the sake of convenience let us summarize these restrictions:

Let T be an ergodic measure preserving transformation of a Lebesgue space. Then the Koopman operator U_T has the following properties:

(1) *One is always a simple eigenvalue of U_T.*
(2) *All eigenvalues are simple and form a finite or countable subgroup of the unit circle $S^1 \subset \mathbb{C}$.*
(3) *The maximal spectral type and the multiplicity function are symmetric under the reflection in the real axis.*
(4) *The maximal spectral type and the multiplicity function are invariant under multiplication by the eigenvalues.*
(5) *The support of the maximal spectral type is the whole circle.*

3.6.2. *Realization results.* Possibility of particular spectral properties for Koopman operators is proven by demonstrating pertinent examples which may either appear in the course of study of specific classes of systems or are constructed on demand. The state of our knowledge for the cases of the full spectral invariants or even just the maximal spectral type is much less advanced then for the case of the possible sets of values for the multiplicity function.

For the former problem there are very few results asserting that a given specific set of spectral data or even a given maximal spectral type can be realized. Theorem 3.26 is almost an exception in that respect. On the other hand, there are many examples showing possibility of realization of certain properties of the spectral type. An outstanding example is the possibility (see Theorem 5.15) and in fact genericity of the mutual singularity of the maximal spectral type in $L_0^2(X, \mu)$ and all its convolutions discussed in Section 5.4 and [78, Section 3.3], which demonstrates an extreme "negative" situation for the Problem 3.24. Another example is extreme "thinness" of the maximal spectral type for a generic measure preserving transformation which follows from very fast periodic approximation, cyclic (see Section 5.4), or, more generally, homogeneous [78, Section 5] which is a spectral property [78, Corollary 5.3].

In one considers the spectral multiplicity by itself, in other words, asks about what subsets of $\mathbb{N} \cup \infty$ appear as the sets of essential values of the multiplicity function for the Koopman operator in L_0^2, the constructive approach goes much further toward a definitive answer. No restrictions on the set of essential values are known and there is an impressive list of sets which do appear as well as certain technology which allows to add many new examples once some key cases have been constructed. Here is an incomplete list of cases when realization is possible:

(1) *If s subset* $S \subset \mathbb{N}$ *is realized then* $S \cup \{\infty\}$ *is realized.*
(2) *Any finite or infinite subset of* $\mathbb{N}$ *containing* 1 [65,100].
(3) *Any finite or infinite subset of even numbers containing* 2.
(4) $\{2, 3\}$, $\{3, 6\}$ [78].
(5) $\{n\}$ *for any* $n \in \mathbb{N}$ [17].

So one may venture to conjecture that no restriction on the set of essential values of spectral multiplicity in L_0^2 exist.

Let us mention that the notion of multiplicity makes sense also for the action of the Koopman operator in L^p. An open question is the following: Does every transformation have simple spectrum L^1? An equivalent way to formulate the question is to ask whether for every ergodic transformation T, there exists an L^1 function ϕ such that the L^1 closure of the linear span of the $T^n\phi$ is the whole of L^1. More generally, does there exist, for every $p < q$ a transformation whose Koopman operator has a cyclic vector in L^p but has no cyclic vector in L^q?

3.6.3. *Extra-spectral information.* Theorem 3.4.2 proved by von Neumann in [155] originally arose some hope that spectrum may serve as a basis of classification for measure preserving transformations up to metric isomorphism.

It later became apparent that for certain classes systems with non-trivial Kronecker factors such as finite or compact group extensions metric isomorphisms exhibit certain rigidity properties. The simplest of those is of course is Proposition 3.14, namely the fact that for an ergodic translation on a compact Abelian group measurable centralizer coincides with algebraic one (other translations) and hence every measurable isomorphism between two such translations is algebraic. Since the Kronecker factors of isomorphic transformations should match this restricts isomorphisms between extensions [19,13,110]. In some cases this allows a complete metric classification of extensions. Abramov's classification of transformations with quasi-discrete is a prime example [13]. In other situations classifi-

cation depends on cohomology classes of certain cocycles which may or may not behave regularly. Rigidity phenomena also appear in certain weakly mixing transformations, for example for those where measurable centralizer is sufficiently small. The notion of self-joining discussed in Section 4.3 is a useful tool of studying rigidity properties beyond pure point spectrum and simple extensions.

There are some exceptional cases when continuous spectrum provides the complete metric invariant in analogy with the pure point spectrum case. The Kronecker Gaussian systems provide the prime example, see Section 6.4.3 [54]. In is not quite clear to what extent very thin continuous spectral measures with strong arithmetic properties (concentration around roots of unity of particular orders) may carry substantial information about metric isomorphism; this information is certainly not complete as [110] and similar examples with continuous spectrum show.

In general, natural non-spectral invariants do not match well with the spectrum. One example where classification of systems with a fixed spectral type looks hopeless is the case of countable Lebesgue spectrum. Recall that every K-system has countable Lebesgue spectrum. On the other hand, every ergodic transformation with positive entropy induces a K-automorphism on some subset, see Theorem 5.65 [120]. Thus any positive entropy class of Kakutani equivalent transformations contains a transformation with countable Lebesgue spectrum. But complete classification up to Kakutani equivalence does not seem more feasible than classification up to metric isomorphism. For basic information on Kakutani (monotone) equivalence see [75,118] and for a summary [12, Section 13].

4. Some aspects of theory of joinings

4.1. *Basic properties*

See [12, Section 3.1, 3.2]. Unlike the other parts of this survey in this section we will often indicate the σ-algebra of measurable sets in our description of dynamical systems. The reason is that we will consider several different invariant measures for the same transformation.

4.1.1. *Definitions*

DEFINITION 4.1. Given two dynamical systems (measure preserving transformations) T acting on $(X, \mathcal{A}, m)$ and S acting on $(Y, \mathcal{B}, \mu)$, a *joining* is a probability measure λ on the Cartesian product $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ which is $T \times S$ invariant and such that $\lambda(A \times Y) = m(A)$ for all A in $\mathcal{A}$ and $\lambda(X \times B) = \mu(B)$ for all B in $\mathcal{B}$.

Joining of several transformations are defined similarly. Joinings were introduced by H. Furstenberg [59]. It is a powerful tool in a great variety of questions in ergodic theory, both spectral and non-spectral. The survey [151] presents a compact treatment of the subject. The book [62] contains extensive information about joinings and in fact represents an attempt to develop the core part of ergodic theory around that notion. See also [6, Section 1.3] for interesting insights and especially for comparison of relevant measure-theoretic and topological concepts and results.

Note that the set of joinings is never empty since there is always the *independent joining* $\lambda = m \times \mu$ but this may be the only one, see Section 4.2.

4.1.2. *Principal constructions.* We list now several basic constructions related to joinings. We will restrict ourselves to the case of two transformations since multiple joinings usually are treated similarly.

(1) *Ergodic decomposition of a joining.* When two systems are ergodic, there always exists an ergodic joining between them. For, ergodic components of a joining measure between ergodic systems are joinings too.

(2) *Factors as joinings.* Considering two systems given as in the definition we call $\mathcal{V}$ and $\mathcal{H}$ the algebras $\mathcal{A} \times Y$and $X \times \mathcal{B}$ respectively. If $\mathcal{H} \subset \mathcal{V}$ (λ) (by which we mean that for every set A in $\mathcal{H}$ there exist a set B in $\mathcal{V}$ such that $\lambda(A \Delta B) = 0$) then $(Y, \mathcal{B}, \mu, S)$ is a factor of $(X, \mathcal{A}, m, T)$. (For this we need that both measure spaces are Lebesgue.) Conversely, if ϕ is the factor map from X to Y, and $A \times B$ is a rectangle in $\mathcal{A} \otimes \mathcal{B}$, the joining defined by $\lambda(A \times B) = m(A \cap \phi^{-1}(B))$ satisfies the inclusion $\mathcal{H} \subset \mathcal{V}$.

(3) *Isomorphisms as joinings.* In the same way a joining λ such that $\mathcal{V} = \mathcal{H}$ (λ) defines an isomorphism between the two transformations, with the same converse as before: An isomorphism gives rise to a joining for which $\mathcal{V} = \mathcal{H}$.

Weak isomorphism means that there exists two joinings λ_1 and λ_2 such that $\mathcal{H} \subset \mathcal{V}$ (λ_1) and $\mathcal{V} \subset \mathcal{H}$ (λ_2).

(4) *Relatively independent joining over a common factor.* If two transformations have isomorphic factors, a useful construction is the *relatively independent joining* above this common factor. If $\mathcal{A}_1$ and $\mathcal{B}_1$ are the two invariant subalgebras of $\mathcal{A}$ and $\mathcal{B}$ respectively such that T restricted to $\mathcal{A}_1$ is isomorphic to S restricted to $\mathcal{B}_1$, we extend the joining λ_1 between these two algebras given by the isomorphism as in (3) which identifies them (we call the global algebra of this object $\mathcal{C}$) to a joining $\lambda_{\mathcal{C}}$ of the whole product in such a way that $\mathcal{A}$ and $\mathcal{B}$ are relatively independent given $\mathcal{C}$. This is done by defining, for a product set $A \times B$ its $\lambda_{\mathcal{C}}$ measure by taking the integral for the measure λ_1 of the product of $E^{\mathcal{A}_1} 1_A \times E^{\mathcal{B}_1} 1_B$. This makes sense as λ_1 is a measure on $\mathcal{A}_1 \otimes \mathcal{B}_1$.

(5) *Topology in the set of joinings.* We introduce a topology in the set of joinings of $(X, \mathcal{A}, m, T)$ and $(Y, \mathcal{B}, \mu, S)$ in the following way: Take A_n, $n \geqslant 1$, and B_n, $n \geqslant 1$, two sequences of sets dense in $\mathcal{A}$ and $\mathcal{B}$ respectively (the density is for the topology associated to the distance between sets which is the measure of the symmetric difference). Given two joinings λ_1 and λ_2 define

$$\delta(\lambda_1, \lambda_2) = \sum_{m,n \geqslant 1} \frac{1}{2^{m+n}} \times \left| \lambda_1(A_m \times B_n) - \lambda_2(A_m \times B_n) \right|.$$

δ is obviously a distance. The set of joinings is compact in the topology generated by this distance.

(6) *Proof of Theorem 3.4.2 via joinings.* There is a nice proof using joinings, due to Lemańczyk and Mentzen [106] of the von Neumann Isomorphism Theorem 3.4.2. We are going to show that any ergodic joining between two such transformations

(which always exists by (1)) is in fact an isomorphism. The L^2 space of both transformations is generated by the eigenfunctions (because they have discrete spectrum), and since the joining is ergodic, every eigenvalue in this joining must be simple. But by spectral isomorphism both transformations have the same eigenvalues and therefore the corresponding eigenfunctions for the ergodic joining are necessarily $L^2(\mathcal{V})$ and $L^2(\mathcal{H})$ measurable. This forces $\mathcal{V} = \mathcal{H}$ for our joining whence the announced isomorphism.

4.1.3. *Spectral analysis of Cartesian products.* Since there is always the independent joining between any two transformations it is appropriate now to describe the spectral properties of the Cartesian product of two measure preserving transformations with respect to the product measure. The Koopman operator of the Cartesian (direct) product $T \times S$ is isomorphic to the tensor product of U_T and U_S. This is a particular case of the tensor products of representations described in Section 1.3.3. Assume that the maximal spectral types of the two Koopman operators are represented by the measures μ and ν (including the δ measure at 1) with multiplicity functions m and m'.

PROPOSITION 4.2. *The maximal spectral type of $T \times S$ is represented by the convolution $\mu * \nu$.*

The multiplicity function m at the point $\lambda \in S^1$ is calculated as follows: *take the product $\mu \times \nu$ on the two-dimensional torus $\mathbb{T}^2$ and consider the system of conditional measures with respect to the partition of $\mathbb{T}^2$ into the "diagonal" circles $\lambda_1 + \lambda_2 = c$.*

(1) *If the conditional measure at $c = \lambda$ is not supported in the finite number of points then $m(\lambda) = \infty$.*
(2) *Otherwise, let the support of the conditional measure be the points $(\lambda_1^1, \lambda_2^1), \ldots, (\lambda_1^n, \lambda_2^n)$. Then*

$$m(\lambda) = \sum_{i=1}^{n} m_1\left(\lambda_1^i\right) \times m_2\left(\lambda_2^i\right).$$

A similar albeit more complicated description can be given in the case of Cartesian product of several transformations.

4.2. *Disjointness*

Joinings provide a good way to compare transformations; more precisely, how far is the isomorphism class of a transformation from that of another. We saw that when two transformations are isomorphic, there is a joining which identifies $\mathcal{V}$ and $\mathcal{H}$. At the opposite end, two transformations are said to be *disjoint* when the product joining is the only joining between them. (That is for every joining λ, $\mathcal{V} \perp \mathcal{H}(\lambda)$.) This notion was introduced by H. Furstenberg in his seminal paper [59]. One may say that disjoint transformations have as little is common as possible, e.g., no common factors since if there is one there is also the relatively independent joining over it.

Disjointness is also a tool, if we know that the restrictions of a transformation T to two invariant algebras are disjoint, to show independence of these algebras. The simple observation that the identity is disjoint from any ergodic transformation has shown surprising efficiency in various contexts.

PROPOSITION 4.3. *Two transformations whose spectral types are mutually singular are disjoint. In particular, rigid transformations are disjoint from mildly mixing transformations.*

PROOF. Assume that T_1 and T_2 are two transformations with spectral types ν_1 and ν_2 on the orthogonal complement of constants which are mutually singular. Consider a joining λ between them and let f and g be in $L^2(\mathcal{V})$ and $L^2(\mathcal{H})$ respectively, both with 0 integral. The projection of f in H_g (the cyclic subspace generated by g) for the joining λ has a spectral measure which is absolutely continuous both with respect to ν_1 and ν_2 and must therefore be 0. This says that $\int fg\,d\lambda = 0$, and λ is the product measure. □

PROPOSITION 4.4. *Distal transformations are disjoint from weakly mixing transformations.*

More generally, joinings allow to give a characterization of the maximal distal factor defined in 3.4.3–6.

Call a transformation *weakly mixing relative to a factor* if its relatively independent joining with itself above the given factor is ergodic.

PROPOSITION 4.5. *The maximal distal factor is the smallest factor algebra relative to which the transformation is weakly mixing.*

Since transformations with positive entropy have Bernoulli factors we see that

PROPOSITION 4.6. *Any two transformations with positive entropy are not disjoint.*

PROPOSITION 4.7. *K-automorphisms are disjoint from 0-entropy transformations.*

Here is a nice application of this last fact which goes back to the original paper of Furstenberg [59]. It is sometimes called the possibility of perfect filtering.

THEOREM 4.8. *Assume that are given two independent stationary processes (X_n) and (Y_n) such that (X_n) generates a K-automorphism and such that (Y_n) generates a zero entropy transformation. Assume furthermore that X_0 and Y_0 are both in L^2. Then (X_n) is measurable with respect to the $(X_n + Y_n)$ process. That is to say the (X_n) process can be recovered from the system perturbed by a random noise $(Z_n) = (X_n + Y_n)$.*

PROOF. Consider the relatively independent joining of (X_n, Z_n) with itself above (Z_n). This is a triple (X_n, Z_n, X'_n) such that (X'_n, Z_n) is a copy of (X_n, Z_n) and (X_n) and (X'_n) are relatively independent over (Z_n). (In the previous constructions, we identify a

process to the measure preserving transformation to which it gives rise.) $(Y'_n) = (Z_n - X'_n)$ is obviously isomorphic to (Y_n) and therefore K. $(X'_n + Y'_n) = (X_n + Y_n)$. We compute $E(X_n - X'_n)^2 = E(X_n - X'_n)(Y'_n - Y_n)$. As (X_n) and (Y_n) are independent (as well as (X'_n) and (Y_n) as a consequence of the disjointness of 0-entropy transformations with K-automorphisms), we get that the preceding expectation is 0 and therefore that $X_n = X'_n$ a.e. which is saying that (X_n) is measurable with respect to the (Z_n) process. □

Note that this can be considered an extension of the same statement obtained under the spectral hypothesis that the spectral measures of the two processes (X_n) and (Y_n) are mutually singular.

4.3. *Self-joinings*

4.3.1. *Basic properties.* Every transformation is isomorphic to itself which is reflected by the presence of the trivial diagonal joining: the measure Δ on $X \times X$ defined by

$$\Delta(A \times B) = m(A \cap B)$$

is a self-joining. Studying the collection of all joinings of a transformation with itself and the structure of such joinings provides deep insights into the orbit structure of the system. In particular presence of few joinings indicates a certain rigidity of the orbit structure while abundance of joinings indicates its richness and "plasticity". Thus, the family of self-joinings Δ_n, $n \geqslant 1$, defined by $\Delta_n = (\mathrm{Id} \times T^n)_* \Delta$ is quite interesting.

(1) T is mixing if and only if $\Delta_n \to m \times m$. Self-joinings of higher order are closely related to mixing of higher order.
(2) T is rigid if there exists a sequence n_i such that $\Delta_{n_i} \to \mathrm{Id}$.
(3) If S is an automorphism which commutes with T, then there is a joining $\Delta_S = (\mathrm{Id} \times S)_* \Delta$. As a consequence of Section 4.1.2(3) it is equivalent for a self-joining λ to be of this form, or to satisfy $\mathcal{V} = \mathcal{H}(\lambda)$.

Something analogous to Proposition 4.4 for weakly mixing but not mixing transformations follows from a recent work of F. Parreau (unpublished) who proved that if a transformation T is weakly mixing and not mixing, it possesses a non-trivial factor which is disjoint from all mixing transformations. A starting point for the construction of this factor is the consideration of a non-trivial limit for Δ_{n_i}.

It can be useful to consider joinings from a more functional analytic viewpoint [143]. Assume that we are given a linear operator $\phi : L^2(X, \mathcal{A}, m) \to L^2(Y, \mathcal{B}, \mu)$ satisfying the following properties: $U_T \phi = \phi U_S$, $\phi 1 = 1$, $\phi^* 1 = 1$, $\phi(f) \geqslant 0$ if $f \geqslant 0$.

Then the measure λ defined by $\lambda(A \times B) = \int_B \phi(1_A)$ gives a joining.

The converse is obvious: given a joining λ take for ϕ the conditional expectation with respect to $\mathcal{H}$ restricted to $L^2(\mathcal{V})$. As an application, we see that if λ is a self-joining of $(X, \mathcal{A}, m, T)$ with itself, and if T has simple spectrum then λ is $S \times S$ invariant for every automorphism S which commutes with T. Therefore λ is a self-joining for the S-action.

4.3.2. *Joinings and group extensions.* There is an important theorem, due to Veech, which contains many of the compactness arguments which appear in ergodic theory.

THEOREM 4.9. *Consider an ergodic transformation $(X, \mathcal{A}, m, T)$ together with a factor (a T invariant subalgebra) $\mathcal{B}$. The following statements are equivalent*:

(1) *Almost all ergodic components of the relatively independent joining of $(X, \mathcal{A}, m, T)$ with itself above $\mathcal{B}$ identify $\mathcal{V}$ and $\mathcal{H}$.*

(2) *There exists a compact group G and a measurable mapping $\phi : (X, \mathcal{B}) \to G$ such that $(X, \mathcal{A}, m, T)$ is isomorphic to the skew product on $(X_{\mathcal{B}}, \mathcal{B}, m) \times (G, \mathcal{G}, \mu_G)$ (μ_G the Haar measure on G) given by $T(x, g) = (Tx, \phi(x)g)$ by an isomorphism which is the identity restricted to $\mathcal{B}$. (This is to compare with isometric extensions which have been defined in* 3.4.9.)

4.3.3. *Minimal self-joinings.* D. Rudolph [140] introduced, for a transformation, the notion of minimal self-joinings, which basically says that a transformation has no other joinings with itself than the obvious ones, and proved existence of mixing transformations with that property.

DEFINITION 4.10. A weakly mixing transformation $(X, \mathcal{A}, m, T)$ has *minimal self-joinings* (*MSJ*) if the following is true: for all $n \geqslant 2$ any ergodic joining λ of n copies of $(X, \mathcal{A}, m, T)$ (λ is a probability measure on $\prod_1^n (X_i, \mathcal{A}_i)$ invariant under $\prod_{i=1}^n T_i$, which satisfies

$$\lambda\left(A_i \times \prod_{j \neq i} X_j\right) = m_i(A_i)$$

for all $1 \leqslant i \leqslant n$ and all $A_i \in \mathcal{A}_i$. $(X_i, \mathcal{A}_i, m_i, T_i)$, $1 \leqslant i \leqslant n$, is a copy of $(X, \mathcal{A}, m, T)$) satisfies the following: the set $[1, n]$ can be decomposed into a disjoint union of subsets E_k, $1 \leqslant k \leqslant r$, such that:

(1) The algebras

$$\mathcal{B}_k = \bigotimes_{i \in E_k} \mathcal{A}_i \times \prod_{j \in E_k^c} X_j, \quad 1 \leqslant k \leqslant r,$$

are λ-independent.

(2) For every $1 \leqslant k \leqslant r$ there exists integers

$$n_{i_1}, n_{i_2}, \ldots, n_{i_{s-1}} \quad (s = |E_k|)$$

such that λ restricted to $\mathcal{B}_k$ is exactly

$$\left(\mathrm{Id} \times T^{n_{i_1}} \times T^{n_{i_2}} \times \cdots \times T^{n_{i_{s-1}}}\right)_* \Delta.$$

Δ is the diagonal measure.

Since factors and commuting transformations other than powers produce joinings of types other than those described in the definition of MSJ as an immediate corollary of the definition we obtain

PROPOSITION 4.11. *Any MSJ transformation has no factors and its centralizer consists only of its powers.*

Weak mixing is not a restriction here since a pure point spectrum transformation has many self-joinings coming from the centralizer and presence of a non-trivial Kronecker factor provides for the independent joining over it. In fact one can show more.

PROPOSITION 4.12. *Any MSJ transformation is mildly mixing.*

PROOF. If T is MSJ and has a rigid factor the factor must be the whole T. But then by Proposition 2.14 T has an uncountable centralizer. □

We will see later that non-mixing MSJ transformations exist (Theorems 5.12 and 5.13).

4.3.4. *Minimal self-joinings for flows and simple transformations.* A transformation from a flow cannot have minimal self-joinings since it commutes not only with its powers but also with other transformations from the flow. This is taken into account in the definition of minimal self-joinings for flows. More generally, it turned out to be useful to have a notion which is somewhat weaker than minimal self-joining and which roughly speaking allows for joinings coming from non-trivial commuting transformations. This was done by Veech [152]. The class of simple transformations which he defined includes in particular transformations from flows with minimal self-joinings as well as certain rigid transformations.

DEFINITION 4.13. A weakly mixing transformation is *simple* if the following property is true:

For all $n \geqslant 2$ any ergodic joining λ of n copies of $(X, \mathcal{A}, m, T)$ satisfies the following: the set $[1, n]$ can be decomposed into a disjoint union of subsets E_k, $1 \leqslant k \leqslant r$, such that:

(1) The algebras

$$\mathcal{B}_k = \bigotimes_{i \in E_k} \mathcal{A}_i \times \prod_{j \in E_k^c} X_j, \quad 1 \leqslant k \leqslant r,$$

are λ-independent.

(2) For every $1 \leqslant k \leqslant r$ the $|E_k|$ algebras $\mathcal{A}_i \times \prod_{j \neq i}$, $i \in E_k$, are λ-identical (which is the same as saying that there exists $|E_k|$ automorphisms commuting with T, S_j such that λ restricted to $\mathcal{B}_k$ is exactly $(\prod_{j \in E_k} S_j)_* \Delta$).

We note that we could have labeled these two definitions according to the number of copies which were used. But in fact a theorem of Glasner, Host and Rudolph [63] asserts that as soon as the definition is satisfied for a joining of three copies, it is satisfied for any number of copies. It is not known whether the definition for two copies only would suffice to imply that it is satisfied for three copies (and therefore for any number of copies).

It follows from Theorem 4.9 that if T is simple, it is a compact group extension of any of its non-trivial factors.

By Section 4.1(2) two ergodic transformations with isomorphic common factors are never disjoint. In general the converse is not true. However the following holds [34]:

THEOREM 4.14. *Two simple transformations with no isomorphic common factors are disjoint.*

4.3.5. *Mixing properties and joinings.* We saw that $\Delta_n \to m \times m$ is equivalent to mixing. The study of self-joinings of higher order is closely related to higher order mixing properties. The next definition is due to del Junco and Rudolph [34].

DEFINITION 4.15. An ergodic transformation $(X, \mathcal{A}, m, T)$ is said to be *pairwise independently determined* if the following is true: for every integer k a joining λ of k copies of $(X, \mathcal{A}, m, T)$ which is such that any two of the k factors of the product of the k copies are pairwise independent (λ) must be the product joining (which is the one for which the k factors are globally independent).

One immediate fact is the following: if a transformation is mixing and pairwise independently determined, it is mixing of all orders. B. Host [72] has proved the following important theorem:

THEOREM 4.16. *An ergodic transformation with singular spectral measure is pairwise independently determined.*

COROLLARY 4.17. *A mixing transformation with singular spectral measure is mixing of all orders.*

The last corollary is one of the few deep structural results in ergodic theory. It sheds light on a long-standing unsolved problem (Does mixing apply mixing of all orders?) by giving an affirmative answer in one of the "most suspicious" cases. See also Theorems 5.18 and 5.19.

Let us remark that a simple transformation is one which is 2-simple (that is every ergodic joining of two copies of the transformation is either product measure or identifies $\mathcal{V}$ and $\mathcal{H}$) and pairwise independently determined. In case of an $\mathbb{R}$ action, V. Ryzhikov [143] has proved that it is always true that 2-simplicity implies pairwise independently determined (and therefore simplicity).

There are no examples known of transformations which are weakly mixing, have 0 entropy, and which are not pairwise independently determined.

5. Combinatorial constructions and applications

5.1. *From Rokhlin lemma to approximation*

5.1.1. *Genericity in the weak and uniform topologies.* Let us recall definitions of the two principal topologies in the group of all measure preserving transformations of a Lebesgue space (X, μ) [68].

The uniform topology first mentioned in Section 3.2.3 is quite strong: it is defined by the metric

$$d_u(T, S) = \mu\{x\colon Tx \neq Sx\} \tag{5.1}$$

invariant by both left and right multiplications.

Notice however that it is weaker than the topology induced from the uniform operator topology on the Koopman operators which is simply discrete.

The weak topology which appeared in Section 3.4.3 is metrizable but no canonical two-side invariant metric similar to (5.1) is available to define it. One way to define a metric is to pick a countable dense collection of measurable sets $A_1, \ldots$ and define the distance as

$$d_w(T, S) = \sum_{n=1}^{\infty} \mu(TA_n \Delta SA_n). \tag{5.2}$$

This topology coincides with the topology induced from the strong operator topology on Koopman operators. Weak topology is weaker than uniform and aperiodic transformations are dense in weak topology. Hence the density of conjugates of any aperiodic transformation in all aperiodic transformations in uniform topology (Proposition 3.6) implies

PROPOSITION 5.1. *Conjugates of any aperiodic transformation are dense in the group of all measure preserving transformations in weak topology.*

Here is an immediate corollary which due to the Baire Category Theorem plays a great role in proving existence and abundance of measure preserving transformations with many interesting properties including spectral ones.

COROLLARY 5.2. *Any conjugacy invariant G_δ in weak topology set which does not contain transformations with sets of periodic points of positive measure is dense and hence residual.*

This fact is widely used in existence proofs.

Another related method is to establish a property via checking its approximate versions which can be shown to be satisfied on open dense sets. This works with properties which can be expressed by the behavior along an unspecified subsequence of iterates (e.g., ergodicity, rigidity, weak mixing) but not along the whole sequence (mixing, Lebesgue spectrum).

5.1.2. *Towers and cityscapes*

DEFINITION 5.3. An *n-tower* in a Lebesgue space (X, μ) is a collection of disjoint subsets $F_1, \ldots, F_n$ of equal measure together with measure preserving transformations $T_i : F_i \to F_{i+1}, i = 1, \ldots, n-1$.

The sets F_i are called the *levels* of the tower; in particular, the set F_1 is called the *base* of the tower and the set F_n the *roof* of the tower.

The union of all levels is called the *support* of the tower.

The number n is sometimes called the *height* of the tower. The quantity $n\mu(F_1)$, i.e. the measure of tower's support, is called the *measure* or the *size* of the tower.

We will say that the tower $\mathcal{T}$ *agrees* with a measure preserving transformation T if $T = T_i$ on the ith level of the tower.

The Rokhlin Theorem 3.4 says that for every aperiodic measure preserving transformation T, there exists arbitrarily high (or long) towers of measure arbitrary close to one which agrees with T. If the measure of a Rokhlin tower is greater than $\frac{n}{n+1}$ then the image of its roof must overlap with the base. And if the size is very close to one then most of the roof is mapped into the base. However the Rokhlin Theorem says nothing about how most of the roof is mapped into the base. Thus an approximation of a measure preserving transformation by a single tower does not say much about the asymptotic properties of the transformation apart from the crudest one, the aperiodicity.

DEFINITION 5.4. A *cityscape* is a union of disjoint towers, in general, of varying heights. The *measure* of a cityscape is defined as the sum of measures of towers comprising the cityscape.

A cityscape *agrees* with a measure preserving transformation T if every tower comprising it agrees with T.

5.1.3. *Uniform approximation.* In order to make certain conclusions from approximation of a measure preserving transformation by towers, or more generally, cityscapes the latter should in some sense be representative of the σ-algebra of all measurable sets. This of course makes sense only if one considers not a single approximation but a sequence of such approximations. A useful model to visualize the requirement of being representative is to think of X as a metric space and of the levels of the towers (or of towers comprising the cityscape) as sets of small diameter. In this situation every fixed measurable set can be approximated up to a set of small measure by a union of levels and combinatorics of transformations in towers approximates the dynamics of the map at sufficiently long time ranges.

This model is suggestive but restrictive in two ways: first, the appropriate topological structure is not always available, and second, even if it is, the levels need not really be sets of small diameter: only after throwing away a set of small measure the intersections of the levels with the remainder would have this property. Anyway, there is a purely measure theoretic way to formulate the property we have in mind as well as its variations.

Every measurable partition ξ of the space X generates the σ-algebra $\mathfrak{B}(\xi)$ of sets *measurable with respect to the partition*. For every set $A \in \mathfrak{B}(\xi)$ one can find another set A' which is the union of elements of ξ such that the symmetric difference of A and A' is a null-set.

To each cityscape $\mathcal{C}$ we associate partition $\xi(\mathcal{C})$ on the space whose elements are level of towers comprising the cityscape and the complement to the union of all such levels.

Recall that the sequence of measurable partitions $\eta_n \to \varepsilon$ as $n \to \infty$, if for every measurable set $A \subset X$ there exists a sequence of sets

$A_n \in \mathfrak{B}(\eta_n)$ such that $\mu(A \Delta A_n) \to 0$ as $n \to \infty$.

This notion can be reformulated as follows. We will say that partition ξ δ-*refines* partition η if for every $A \in \mathfrak{B}(\eta)$ there exists $A' \in \mathfrak{B}(\xi)$ such that $\mu(A \Delta A') < \delta$. Then $\xi_n \to \varepsilon$ if for any finite partition η and every $\delta > 0$ there exists $N = N(\eta, \delta)$ such that for $n \geqslant N$, ξ_n ε-refines η.

DEFINITION 5.5. A sequence $\mathcal{C}_n$ of cityscapes is called *exhaustive* if $\xi(\mathcal{C}_n) \to \varepsilon$ as $n \to \infty$.

DEFINITION 5.6. An exhaustive sequence of cityscapes which agrees with a measure preserving transformation T is called a *uniform approximation* of T.

It follows from the Rokhlin Theorem 3.4 that every measure preserving transformation allows a uniform approximation. To see that one needs to take a Rokhlin tower and split its base in such a way that the partition into levels of resulting towers would be a refinement of a given partition. However if one restricts the type of cityscapes (e.g., consider cityscapes consisting of a single tower or a fixed number of towers) existence of a uniform approximation becomes a restrictive property and implies interesting properties of T, see Section 5.2.2.

Uniform approximation and its variations are used to produce measure preserving transformations with interesting properties. We will consider three ways to produce such approximations: cutting and stacking, coding with respect to a given generating partition, and periodic approximation.

5.2. *Cutting and stacking and applications*

5.2.1. *The method of cutting and stacking.* (See also [138].) The cutting and stacking method is a particular way to produce inductively an exhaustive sequence of cityscapes which form a uniform approximation of a measure preserving transformation.

At nth step a cityscape $\mathcal{C}_n$ is defined. The transformation is thus defined everywhere except for the roofs of the towers from $\mathcal{C}_n$ and a certain set A_n which is the complement to the union of supports of the towers in the cityscape. Then each tower of $\mathcal{C}_n$ is divided into towers and new levels are added from A_n to some of the towers. Then the roofs of most of new towers are mapped into bases of other towers. This produces the cityscape $\mathcal{C}_{n+1}$ and the set $A_{n+1} \subset A_n$. Specifically, those parts of the bases of old towers which do not belong to the images of the roofs of extended old towers serve as bases of new towers. Each new tower is defined by an itinerary, namely a sequence of old towers which are visited in succession. This is why the construction is called cutting and stacking: bases of old towers are cut according to the itineraries and this new thin towers are stacked on top of each other.

The list of important examples constructed with the cutting and stacking method is quite large. Let us mention the "Chacon transformation" described below in Section 5.2.3, the rank one mixing transformations (Section 5.2.4), the first examples of Ornstein of K-automorphisms which are not Bernoulli later developed in [119] (as well as his counterexamples to the Pinsker conjecture), the Feldman examples of non-standard transformations with zero entropy [51]. To illustrate the usefulness of the method for other groups let

us mention [83] where the cutting and stacking method is used to construct examples of $\mathbb{Z}^k$ and $\mathbb{R}^k$ actions with $k \geqslant 2$ where individual elements have zero entropy which cannot be realized by diffeomorphisms of compact manifolds with respect to any Borel measure. We will mention other specific constructions in due course.

5.2.2. *Approximations with towers or large size; rank*

DEFINITION 5.7 [117]. A measure preserving transformation T has *rank one* if it admits uniform approximation by single towers.

Equivalently, T is rank one if for every finite partition η and every $\delta > 0$ there is a tower $\mathcal{T}$ which agrees with T and such that the partition $\xi(\mathcal{T})$ into the levels of the tower and the complement to its support δ-refines the partition η.

Importance of the rank one property for the spectral theory of measure preserving transformations is based on the following fact.

PROPOSITION 5.8. *Any rank one transformation has simple spectrum and is hence ergodic*.

PROOF. Consider a tower $\mathcal{T}$ of height n approximating T with base F. The images of the characteristic function ξ_F under U_T^i, $i = 0, 1, \ldots, n-1$, are characteristic functions of the disjoint levels of the tower. Thus there is a cyclic subspace which contains all characteristic functions of the levels of the tower and their linear combinations. Consider these cyclic subspaces for an exhaustive sequence of towers. From the approximation property it follows that for any given $f \in L^2(X, \mu)$ projections to these cyclic subspaces converge to f. Hence by Theorem 1.21, U_T has simple spectrum. □

The spectral multiplicity estimates based on uniform approximation can be obtained under more general conditions than rank one.

DEFINITION 5.9. An ergodic transformation T is *locally rank one* if there exists $a > 0$ such that for every finite partition

$$\eta = (p_0, p_1, \ldots, p_l)$$

and for every $\varepsilon > 0$, there exists a tower $\mathcal{T}$ of size $\geqslant a$ and a partition

$$\bar{\eta} = (\bar{p}_0, \bar{p}_1, \ldots, \bar{p}_l)$$

of $\mathcal{T}$ whose elements are unions of levels such that

$$\sum_{s=0}^{l} m(\bar{p}_s \setminus p_s) < \varepsilon.$$

We call any number a satisfying the above definition an *order* of T.

REMARK. Property of local rank one of order a is equivalent to existence of a uniform approximation by cityscapes where one tower has measure at least a.

If the transformation T allows uniform approximation by cityscapes with k towers, the transformation is said to have *rank no greater than* k. Since in each cityscape at least one tower has measure at least $1/k$ any transformation of rank no greater than k is locally rank one of order at least $1/k$. The following theorem generalizes Proposition 5.8.

THEOREM 5.10. *If an ergodic transformation is locally rank one of order a, its spectral multiplicity is bounded by* $[1/a]$.

SKETCH OF PROOF. As before in the proof of Proposition 5.8 this easily follows form the definition and Theorem 1.21. Given $k = [1/a] + 1$ orthonormal functions $f_1, f_2, \ldots, f_k$, we first approximate them in L^2 by finite valued functions. We call η the partition which makes all these finite valued functions measurable. We consider a tower $\mathcal{T}$ with base F which locally approximates this partition (as in the definition) and which is sufficiently long to have the ergodic theorem giving that the frequency of appearances of each set in η in the tower is close to its measure. If we take for H the cyclic space generated by ξ_F, we see that the conditions of Theorem 1.21 are satisfied. □

Simplicity of the spectrum does not force anything on the rank of the transformation, see [52,36,108] for examples of transformations with simple spectrum which are not locally rank one. The relations between rank and spectral multiplicity have been thoroughly studied by J. Kwiatkowski and Y. Lacroix [99].

Another interesting property of local rank one transformations is related to Kakutani equivalence theory.

PROPOSITION 5.11 [80]. *Any locally rank one transformation is standard* (*zero entropy loosely Bernoulli, sometimes also called loosely Kronecker*), *i.e. it is induced by any odometer and any irrational circle rotation and induces any of those transformations*.

Ferenczi [52] and De la Rue [36] (see Theorem 6.23) constructed transformations with simple spectrum which are not standard and therefore also not locally rank one.

5.2.3. *Chacon transformation* [26]. The Chacon Transformation which is a particular rank one transformation is one of the jewels of ergodic theory. As we shall see, it can be used as a source of examples with interesting, often exotic, properties. Its particular interest is that while it exhibits very moderate and rather regular pattern of orbit growth properties it does not fit into either of the three main paradigms of smooth ergodic theory: elliptic (Section 2.2.4), hyperbolic and parabolic (Section 2.1.3). Smooth realization of this map is unknown and seems to be beyond the reach of available methods.

The transformation is defined inductively on the unit interval equipped with Lebesgue measure I. At stage n, there are $h(n)$ intervals of equal length $I_1, I_2, \ldots, I_{h(n)}$ and T maps I_k onto I_{k+1}, $1 \leqslant k \leqslant h(n) - 1$, by translations. T is not defined on $I_{h(n)}$. To go from stage

n to stage $n+1$, we divide the interval I_1 into three intervals of equal length, I_1^1, I_1^2, I_1^3, and therefore divide the tower

$$\tau_n = \bigcup_{i=0}^{h(n)-1} T^i I_1$$

into three columns

$$\tau_n^j = \bigcup_{i=0}^{h(n)-1} T^i I_1^j,$$

$1 \leqslant j \leqslant 3$. We now pick an interval J_n disjoint from τ_n with length equal to the length of I_1^1 and define τ_{n+1} mapping by translations $T^{h(n)-1} I_1^1$ onto I_1^2 then $T^{h(n)-1} I_1^2$ onto J_n and finally J_n onto I_1^3 (as all these intervals have the same width). The interval I_1^1 is thus the basis of a new tower τ_{n+1} of height $3h(n)+1$. It is easy to adjust the length of the interval at stage 0 ($h(0)=1$) in such a way that the limit transformation T will be defined on I. This transformation is rank one since the sequence of towers defines a refining sequence of partitions into intervals of length going to 0 which will generate the Lebesgue algebra.

THEOREM 5.12. *The Chacon transformation is weakly mixing but not mixing.*

PROOF. Absence of mixing is a consequence of the fact that any set A which is the union of intervals in τ_n satisfies

$$m\big(A \cap T^{h(n)} A\big) \geqslant \frac{1}{3} m(A).$$

Weak mixing comes from the fact that if f is an eigenfunction corresponding to the eigenvalue λ, $\lambda \neq 1$, then given $\varepsilon > 0$, there will be an n and a level J in τ_n on which f will not vary by more than ε on a fraction $9/10$ of J. Call a the value to which f is close on J. But $T^{h(n)} f$ will be close to $\lambda^n a$ on a third of J, and $T^{h(n)+1} f$ will be close to $\lambda^{n+1} a$ on another third of J, forcing

$$\big|\lambda^n a - \lambda^{n+1} a\big| < \varepsilon,$$

$|\lambda - 1| < \varepsilon$. As ε was arbitrary, we obtain a contradiction. □

The following theorem is due to del Junco, Rahe and Swanson [33].

THEOREM 5.13. *The Chacon transformation has minimal self-joinings.*

Note that on immediate consequence of the definition implies that a transformation with MSJ commutes only which its powers and has no non-trivial factors. Thus the Chacon

transformation is not rigid since the centralizer of a rigid transformation contains its orbit closure which is perfect and hence uncountable and has no rigid factors. Hence

COROLLARY 5.14. *The Chacon transformation is mildly mixing but not mixing.*

The Chacon transformation can be used to give an answer to the convolution problem. In fact M. Lemańczyk first proved that if σ is the spectral measure of the Chacon transformation, then $\sigma * \sigma \perp \sigma$. This was extended by A. Prikhod'ko and V. Ryzhikov [127] to the following.

THEOREM 5.15. *Let σ be the spectral measure of the Chacon transformation. Then for every $n \neq m$, $\sigma^{*n} \perp \sigma^{*m}$.*

For other methods of proving singularity of convolutions see Propositions 5.43 and 5.44 and Theorem 5.49.

A transformation with minimal self-joinings can be used as a "building block" for a great variety of examples. D. Rudolph in [140] developed a useful unifying concept of "counterexample machine". Very roughly, the counterexample machine can be thought of as a functor from the category of permutations of the set of integers $\mathbb{N}$ to measure preserving transformations. The arrows in the first category are injections $\mathbb{N} \to \mathbb{N}$ which are such that together with the corresponding permutations, they make the diagram commutative. In this last category, for example, it is easy to see that weak isomorphism does not imply isomorphism.

An interesting open question is related with Kakutani equivalence. The Chacon transformation itself is standard by Proposition 5.11, but it is not known whether its Cartesian square is standard.

5.2.4. *Rank one mixing transformations.* There is a method, due to D. Ornstein [117], to construct "random" rank one transformations which almost surely show very interesting properties.

We are given two sequences of integers $p(n)$ and $t(n)$ and a family of integers

$$a_{n,i}, \quad 1 \leqslant i \leqslant p(n), \quad a_{n,i} \leqslant t(n).$$

The construction is as in the Chacon example, with a tower τ_n which is made of $h(n)$ intervals $I_1, I_2, \ldots, I_{h(n)}$ of equal length such that

$$T I_k = I_{k+1}, \quad 1 \leqslant k \leqslant h(n) - 1,$$

and the map acts by translations. To go to τ_{n+1}, I_1 is divided this time in $p(n)$ intervals of equal length, producing $p(n)$ columns $\tau_n^i, 1 \leqslant i \leqslant p(n)$, and τ_{n+1} is constructed by stacking τ_n^{i+1} onto τ_n^i, $1 \leqslant i \leqslant p(n) - 1$, after the insertion, between the last level of τ_n^i and the basis of τ_n^{i+1} of $a_{n,i}$ intervals (which all have the same length); t_n is chosen so that $t_n \leqslant h_{n-1}$ and $t_n \to \infty$. These added intervals are called spacers. The randomness is on the $a_{n,i}$ which are chosen independently, such that for given n, all the $a_{n,i}$ take values on

$[1, t_n]$ which uniform probability $1/t(n)$. In this probability space, a point ω is the sequence of $a_{n,i}$, $1 \leqslant i \leqslant p(n)$, $n \geqslant 1$, and to every such ω corresponds a rank one transformation T_ω. D. Ornstein has proved:

THEOREM 5.16. *In the previous model, almost surely T_ω is mixing.*

We have seen (Proposition 5.8) that rank one transformations have simple spectrum, the fact that they could be mixing made them interesting candidates for examples with simple Lebesgue spectrum. However J. Bourgain [24] has proved

THEOREM 5.17. *In the previous model, almost surely the spectral measure of T_ω is singular with respect to Lebesgue measure.*

In fact it looks quite plausible that every rank one transformation has purely singular spectral measure. This is justified by the previous theorem which implies by Host's theorem that these transformations are almost surely mixing of all orders, and by the following result of S. Kalikow [73].

THEOREM 5.18. *A mixing rank one transformation is mixing of all orders*.

V. Ryzhikov [142] has extended the previous theorem to the following:

THEOREM 5.19. *A mixing finite rank transformation is mixing of all orders*.

It is not known whether the same holds for mixing locally rank one transformations.
As as consequence of the theorem of Kalikow, J. King proved

THEOREM 5.20. *A mixing rank one transformation has minimal self-joinings*.

This implies in particular that a mixing rank one transformation commutes only with its powers (this was proved in the original paper of Ornstein) and has no factors.

For a long time existence of mixing rank one transformations was only known through the construction of Ornstein. Much later T. Adams [14] gave an explicit construction of mixing rank one transformations (the staircase Smorodinsky's rank one where the spacers are added in such a way that they follow the shape of a staircase). And recently B. Fayad has constructed C^1 flows which are mixing and rank one (as flows) [45].

5.2.5. *Riesz products and spectra of rank one transformations.* Riesz products appear naturally as spectral measures in several natural examples in ergodic theory. For detailed definitions and extensive discussion of the subject see [114, Chapter 16]. Riesz products in the context of ergodic theory first appeared in the paper by Ledrappier [102], where a certain finite extension of a system with pure point spectrum is shown to have a component in its spectral measure which is a Riesz product. It is important and interesting because there exist in many cases explicit criteria which can determine whether the corresponding

measures are singular or absolutely continuous [125]. Riesz products occur also explicitly as components of the spectral measure of many substitutions [128].

A revival of the use of Riesz product techniques arose from already mentioned result of Bourgain (Theorem 5.17) where he first gives an explicit formula for the spectral measure of the general rank one transformation. Such measures can be viewed as generalized Riesz products. In his proof, Bourgain produces, using the fact that for any ergodic transformation $(X, \mathcal{A}, m, T)$, for any function f in L^2, almost surely the sequence of measures

$$\frac{1}{N}\left|\sum_{k=1}^{k=N} f(T^k x)e^{2i\pi k\theta}\right|^2 d\theta$$

converges weakly to ν_f, the spectral measure ν_T of the general rank one transformation as a generalized Riesz product

$$\prod_{n=1}^{n=\infty} |P_n|^2.$$

This can be seen exactly in the same way as in the proof of Theorem 1.8. By the weak convergence we mean that the measures $\prod_{n=1}^{n=N} |P_n|^2\, d\theta$ converge weakly to the spectral measure ν_T. The polynomials $P_n(\theta)$ are equal to

$$\big(p(n)\big)^{-1/2} \sum_{k=0}^{p(n)-1} e^{2\pi i(kh(n)+\sum_{j=1}^{j=k} a(j,n)\theta)}.$$

This is obtained by applying the previous formula to the characteristic functions of the base of the tower τ_1. Then Bourgain shows that it is sufficient to prove singularity for a product of a subsequence of the previous polynomials which are dissociated and to which classical Riesz product techniques can be applied. Note that the mixing property can not be verified by the use of this formula.

The same ideas are present in the paper of Klemes [89] where he shows that the spectral measure of the Adams example [14] is singular. It is also with a proof in the same spirit that El Abdalaoui [40] has shown that if we endow the Cartesian product of the parameter space of the Ornstein example with the product measure, for almost every pair ω, ω', T_ω and $T_{\omega'}$ have mutually singular spectral measures (and are therefore disjoint by Proposition 4.3).

5.2.6. *Cutting and stacking and orbit growth.* In Sections 5.2.3 and 5.2.4 we described constructions of rank one transformations where interesting behavior is achieved by time delays in the return from a part of the roof of the single tower to its base. For the Chacon transformation this delay was by time one on one third of the tower and for rank one mixing transformations the delays were uniformly distributed in an appropriate sense. Thus non-trivial combinatorics was achieved by the distribution of the delay times.

These examples represent instances of intermediate orbit growth; not slow elliptic and not exponential hyperbolic or uniform polynomial parabolic like horocycle flows (Sec-

tion 6.2.2). They probably are best described are being outside of three principal paradigms. This fits well with the fact that no smooth realization for Chacon transformation is known and for rank one mixing map realization has only been achieved in C^1 which is considered somewhat pathological in the smooth setting.

We should point out that interesting behavior (but not mixing or even mild mixing) may be achieved also in the context of fast cyclic approximation (elliptic behavior) (see Section 5.4.2) which can be interpreted as uniform approximation with single towers and direct return of most of the roof to the base. In this case there are spacers too but their effect becomes noticeable only after running the cycle for many times.

5.2.7. *Constructions with many towers.* At the other end of the spectrum of possibilities for cutting and stacking lie situations where the number of towers grows and the roofs of towers at an inductive step are mapped to the base in a complicated way. All positive entropy examples constructed by cutting and stacking necessarily have such structure as well as examples with subexponential but still substantial orbit growth such as transformations from the actions in [83]. The most straightforward way to carry out such constructions is to match the roofs to the bases more or less independently. This method allows to produce any desirable speed and regularity of orbit growth by controlling the number of towers in the approximating cityscapes.

In such constructions if spacers are not used at all (in other words, if at every step the cityscape C_n fills the whole space) the resulting transformation has an odometer (Example 2.16(3)) as a factor. In order to achieve weak mixing, not speaking of mixing or K-property, spacers are needed in addition to the distribution of roofs. Non-isomorphic K-automorphisms with the same entropy from [119] as well as non-loosely Bernoulli K-automorphisms from [51] are examples produced by cutting and stacking constructions of that type. The original Feldman example has been extended [118] to provide uncountably many zero entropy transformations which are pairwise not equivalent.

There are various types of cutting and stacking constructions: the ones we mentioned are based on the idea that a fixed pattern is repeated at every stage. Some others alternate two very different patterns. A typical one in that class is the Rothstein's construction of non-loosely Bernoulli transformation [138]: there is an alternation of stages where independent cutting and stacking is performed thereby creating so many names that most of them are far apart in the $\bar{f}$ distance (based on the Kakutani distance between string of symbols) and is next followed by a stage where names are just repeated twice, which has an effect of dropping the entropy without altering too much the separation of names previously created in $\bar{f}$ metric.

Very beautiful cutting and stacking constructions have been found by C. Hoffman [71]: he has developed a version of Rudoplh's counterexample machine (see Section 5.2.3) for K-automorphisms and in particular, produced two weakly isomorphic but not isomorphic K-automorphisms with finite entropy.

5.3. *Coding*

The coding constructions are very close to symbolic dynamics, see [8, Section 2.6] for an overview of that subject. For a comprehensive introduction and many interesting examples

see [109]. In some of those constructions invariant measure is given as is the case for interval exchange transformations (Section 5.3.1) in others it is not fixed from the beginning but is constructed as the asymptotic distribution on the chosen names as for substitution dynamical systems discussed in Section 5.3.2.

To put the coding-based constructions into the general framework of the inductive combinatorial constructions we consider a space with a partition into "symbols" of an "alphabet" and define certain rules by which allowable words are produced. Similarly to the cutting and stacking (uniform approximation) constructions the coding method is very general since any ergodic finite entropy transformation allows a finite generator and hence a symbolic representation [96]. However when we speak of combinatorial constructions of coding type we mean certain recursive procedures which allow inductively to produce distributions of longer words from those of shorter ones.

Now we will consider several specific classes of such constructions.

5.3.1. *Interval exchange transformations*

Definition and parametrization. Consider $n \geqslant 2$ and π an irreducible permutation of $\{1, \dots, n\}$. A permutation π is called *irreducible* if $\pi\{1, \dots, d\} \neq \{1, \dots, d\}$, $1 \leqslant d < n$. Let Δ be the simplex in R^n,

$$\lambda = (\lambda_i), \quad 1 \leqslant i \leqslant n, \lambda_i \geqslant 0, \sum_{i=1}^{i=n} \lambda_i = 1.$$

The unit interval $I = [0, 1)$ is divided into semi-open intervals $I_d = [\sum_{i<d} \lambda_i, \sum_{i \leqslant d} \lambda_i)$, $1 \leqslant d \leqslant n$.

The *interval exchange transformation* $T_{\pi,\lambda}$ acts on every I_d by a translation in such a way that the intervals are rearranged according to the permutation π. That is, on I_d, $T_{\pi,\lambda}$ is the translation by $\sum_{\pi(i)<\pi(d)} \lambda_i - \sum_{i<d} \lambda_i$.

Interval exchange transformations preserve Lebesgue measure. Sometimes more general transformations which change orientation on some of the intervals are also considered.

Interval exchange transformations are briefly mentioned in [8, Sections 4.3g and 8.4] and more thoroughly discussed in [11, Section 6]. For an elementary self-contained introduction to the subject see [79, Section 14.5]. The area has developed into a major subject of research with some of the deepest and most beautiful results and constructions in the whole of ergodic theory. Some of the recent work in the area is described in [5].

The parameter space for the set of exchanges of n intervals is the simplex of the lengths of the intervals multiplied by the finite set of irreducible permutations. Notice that dynamics obviously depend on the choice of parameters and is fairly simple in some cases. For example, if all λ's are rational all points are periodic albeit with different periods. This is of course similar to the case of translations on the torus. Another similarity with toral translations is prevalence of minimality.

THEOREM 5.21 ([79, Corollary 14.5.12], originally appeared in [85]). *If one excludes from the simplex of lengths intersections with countably many hyperplanes then every orbit*

of the interval exchange transformation corresponding to the remaining set of parameters is dense.

The proof is based on the observation that unless there is a "saddle connection", i.e. an orbit segment beginning and ending in discontinuity points then all orbits are dense. For an irreducible interval exchange any type of saddle connection generates a rational relation between the lengths of the intervals.

However, ergodic properties of interval exchange transformations with respect to Lebesgue measure exhibit more complicated dependence of the parameters than is the case with toral translations. The same apply to the question of unique ergodicity.

Finiteness properties. The following three theorems summarize the basic distinctive properties of interval exchange transformations which do not depend on the choice of parameters and which can be described as something like "finiteness of dynamical complexity". The key observation here is that the transformation induced by an exchange on n intervals on any interval, however small, is again an exchange of at most $n+1$ intervals.

THEOREM 5.22 [76]. *An aperiodic interval exchange transformation on k intervals is of finite rank at most k. Furthermore, it is rank k by intervals*: *that is all the levels of the towers which appear in the definition of finite rank are intervals. Furthermore these towers fill the whole space.*

Unlike the general finite rank (or even rank one) property this kind of uniform approximation implies absence of mixing.

THEOREM 5.23 ([11, Theorem 6.10], originally proved in [76]). *An interval exchange transformation is never mixing.*

Another consequence of Theorem 5.22 is an estimate on the number of ergodic measures and the spectral multiplicity of any such measure.

THEOREM 5.24. *An aperiodic interval exchange transformation of n intervals has at most $n-1$ ergodic Borel probability invariant measures. Spectral multiplicity of the transformation with respect to any invariant measure* (*ergodic or not*) *does not exceed n.*

The estimate on the number of ergodic measures can be improved. The best estimate which depends only on the number of intervals is $n/2$ for n even (this includes unique ergodicity of irrational rotation for $n=2$) and $n-1/2$ for n odd. This estimate is sharp for the reverse permutation $\pi(i)=n-i$. On the other hand, there is a sharp estimate for any permutation which depends not only on the number of intervals but on the permutation π. For a "generic" combinatorics, the resulting estimate is slightly above $n/4$. This may sound mysterious but becomes transparent when one constructs for any interval exchange transformation an oriented surface with a flow for which the original transformation serves as a section map on a certain arc connecting two (not necessarily different) saddles. The sharp estimate for the number of ergodic invariant measures is the genus of the surface which

depends only on the permutation π. See [79, Theorem 14.7.6] for the inequality and [145] for constructions of minimal examples with any number of ergodic measures between one and the genus. Satayev's method in [145] makes use of symmetry in a way somewhat similar to that used in the construction of transformations with given values of multiplicity in the spectrum which we discuss in Section 5.8.4.

Typical behavior in the parameter space: direct methods. Finiteness of the number of ergodic invariant measures implies in particular that Lebesgue measure has finite number of ergodic components. Hence one may ask when an interval exchange transformation is ergodic with respect to Lebesgue measure, or, which is even more natural in the present context when Lebesgue measure is the only invariant measure for an interval exchange transformation. This is one of the few places in this survey when we do not have ergodicity given *a priori* or following from a construction but discuss conditions for ergodicity instead.

The answer is easy and explicit for $n = 2$ and 3 because in those cases the surface discussed above is a torus.

In both cases the only irreducible permutations are reverse permutations. Exchange of two intervals of lengths λ and $1 - \lambda$ becomes the circle rotation R_λ once the interval $[0, 1)$ is identified with the circle. Thus irrationality of λ is equivalent to unique ergodicity.

The situation is only slightly more complicated for the exchange of three intervals of lengths λ_1, λ_2 and $1 - \lambda_1 - \lambda_2$ in reverse order. Direct inspection shows that this transformation is identified with the transformation induced by the rotation $R_{\frac{1-\lambda_1}{1+\lambda_2}}$ on any interval of length $\frac{\lambda_2}{1+\lambda_2}$. Hence the interval exchange is ergodic with respect to Lebesgue measure if and only if it is uniquely ergodic and this happens exactly when the number $\frac{1-\lambda_1}{1+\lambda_2}$ is irrational.

For $n \geqslant 4$ the picture becomes considerably more complicated. First, there is no more dichotomy between periodicity and unique ergodicity. Necessary and sufficient conditions for ergodicity with respect to Lebesgue measure or unique ergodicity (which are also not equivalent anymore) are not available. However, the following fundamental result holds.

THEOREM 5.25. *Almost every with respect to Lebesgue measure on the simplex of length interval exchanges transformations is uniquely ergodic*.

This theorem was originally proved independently by Veech [153] and Masur [111] using advanced indirect methods which are discussed below. Shortly afterwards Boshernitzan [22] found a direct (albeit fairly complicated) proof based on the following sufficient criterion for unique ergodicity.

For a given interval exchange transformations let ξ be the partition into its intervals of continuity and $\xi_n = \bigvee_{k=0}^{n-1} T_{\pi,\lambda}^k \xi$ be the iterated partition. Notice that the number of elements in ξ_n grows linearly with n. Aperiodicity of the transformation is equivalent to fact that the maximal length of elements in ξ_n goes to zero as $n \to \infty$. Let m_n be the *minimal* length of an element in ξ_n. Given $\varepsilon > 0$ we will call positive integer n *ε-regular* if $m_n \geqslant \frac{\varepsilon}{n}$.

An interval exchange transformation satisfies *property P* if for any $l \geqslant 2$ there exists $\lambda(l)$ such that there are infinitely many sequences of ε-regular numbers of length l, $n_1, \ldots, n_l$ with $n_{i+1} > 2n_i$, $i = 1, \ldots, l-1$, and $n_l < \lambda n_1$. Any set of natural numbers which contains such sequences will be called *essential*.

THEOREM 5.26. *Any interval exchange transformation which satisfies property P is uniquely ergodic*.

Now consider a family of interval exchange transformations parametrized by a space Ω with a probability measure μ. For a given ε let $u(n, \varepsilon)$ be the measure of the set of parameters for which the number n is ε-regular for the corresponding interval exchange transformation.

A family of interval exchange transformations parametrized by Ω satisfies *collective property P* if for any $\varepsilon > 0$ one can find $\delta > 0$ and a single essential set $A(\varepsilon)$ such that $u(n, \delta) > 1 - \varepsilon$ for all $n \in A(\varepsilon)$.

PROPOSITION 5.27. *If a family satisfies collective property P then almost every element in the family satisfies property P.*

One can show that for any admissible permutation the whole simplex of interval exchange transformations with this permutation and with Lebesgue measure satisfies collective property P. Theorem 5.25 then follows from Proposition 5.27 and Theorem 5.26.

An earlier and more elementary example of use of direct methods for showing prevalence of certain properties concerns spectral properties of exchanges of three intervals.

Many interesting phenomena in ergodic theory can be realized within the class of interval exchange transformations. In particular, this is connected with a possibility to realize certain kinds of symmetry within this class. Notice in particular that any piecewise constant finite extension of a rotation (or, more generally of an interval exchange transformation) can be represented as an interval exchange transformation. See Section 5.8.4.

THEOREM 5.28 [81]. *Almost every exchange of three intervals has simple singular continuous spectrum*.

This result follows form existence of both good cyclic approximation and good approximation of type $(n, n+1)$ (Section 5.4.2), which are constructed using properties of approximation of parameters by rationals, and Propositions 5.39 and 5.40.

Renormalization dynamics and advanced results. A powerful indirect approach to the study of interval exchange transformations is based on renormalization type dynamics introduced by Rauzy [133]. It was first developed by Veech [153] for his proof of Theorem 5.25. Let us mention a couple of relevant results.

Veech [154] has proved:

THEOREM 5.29. *Almost every interval exchange transformation is of rank one.*

Recently Avila and Forni [20] solved the long-standing open problem.

THEOREM 5.30. *Almost every interval exchange transformation is weakly mixing.*

The following related question by Veech remains open.

PROBLEM 5.31. Is it true that almost every interval exchange of $m \geqslant 3$ intervals is simple?

5.3.2. *Substitution dynamical systems.* Another interesting class of examples related to symbolic dynamics comes from what is called substitutions. Literature on substitution dynamical systems is quite extensive, maybe a bit out of proportion of the place of the subject within the general context of ergodic theory and symbolic dynamics. In particular, a detailed albeit not fully up-to-date account of the spectral properties for this class of systems exists in book form [128]. We restrict ourselves to the definition and a couple of interesting examples.

We consider a finite set $A = \{0, 1, 2, \ldots, n-1\}$. We let $A^* = \bigcup_{k \geqslant 1} A^k$ be the set of all finite words in the alphabet of A.

DEFINITION 5.32. A *substitution* ζ on A is a map from A to A^*. It defines a map from A^* to A^* in the following way: if $x = x_0 x_1 \ldots x_n \in A^*$, then

$$\zeta(x) = \zeta(x_0)\zeta(x_1)\ldots\zeta(x_n).$$

This obviously extends to a map from A^N to A^N.

We consider substitutions such that
(a) the length of $\zeta^n(i)$ goes to infinity when $n \to \infty$ for every $i \in A$,
(b) there exists a symbol 0 in A such that $\zeta(0)$ starts with (0),
(c) there exists an integer k such that for every two $i, j \in A$, $\zeta^k(i)$ contains j.

A substitution satisfying (a), (b) and (c) is called *primitive*. The most famous transformation which can be described by a substitution is the *Morse sequence*, which is defined on the alphabet 0, 1 by

$$\zeta(0) = 01, \qquad \zeta(1) = 10.$$

THEOREM 5.33. *Given a primitive substitution ζ any fixed point $x = \zeta(x)$, $x \in A^N$ (which is easily shown to exist) has on orbit closure X on which the shift T is a uniquely ergodic transformation (independent of the fixed point).*

THEOREM 5.34. *All the transformations (X, T) as described in the previous theorem are finite rank transformations.*

Another important example is the *Rudin–Shapiro sequence* which is generated by the following primitive substitution on 4 symbols:

$$\zeta(0) = 02, \qquad \zeta(1) = 32, \qquad \zeta(2) = 01, \qquad \zeta(3) = 31.$$

The remarkable spectral property of the transformation T associated to the Rudin–Shapiro sequence is the following, first proved by T. Kamae [74]

THEOREM 5.35. *T has a Lebesgue component in its spectral measure.*

Since T is of finite rank, this implies U_T is a unitary operator with finite spectral multiplicity and a Lebesgue component in its spectrum. We note however that U_T also has a discrete component in its spectrum. This is one of very few known examples with Lebesgue component of finite multiplicity in the spectrum. Other examples are discussed in Section 5.8.5; these constructions are somewhat more flexible than Rudin–Shapiro; in particular, they can be made weakly mixing (Theorem 5.75).

Notice that no examples are known with a *simple* Lebesgue component in the spectrum as well as with Lebesgue (or absolutely continuous) spectrum of finite multiplicity.

5.4. *Periodic approximation*

The method of periodic approximations is in a number of respects parallel and complementary to the cutting and stacking method. It is based on the ideas of fast approximation of a measure preserving transformations in weak as opposed to uniform topology. This allows to define approximating transformations everywhere if need arises. The method has been introduced in [81]; see also [29]. For the most up-to-date albeit not comprehensive presentation of the methods and some of its applications see [78]. We mostly follow the last source in this section.

5.4.1. *Periodic processes.* Let (X, μ) be a Lebesgue space. A *periodic tower* t is an ordered sequence of disjoint subsets $t = \{c_1, \dots, c_h\}$ of X having equal measure which we will usually denote $m(t)$. The number $h = h(t)$ will be called *the height* of the tower t. Associated with a tower, there will be a cyclic measure-preserving permutation σ sending c_1 to c_2, c_2 to c_3, etc., and c_n to c_1. The set c_1 will be called *the base* of the tower.

DEFINITION 5.36. A *periodic process* is a collection of disjoint towers covering X, together with an equivalence relation among these towers which identifies their bases. A periodic process which consists of a single tower is called a *cyclic process.*

The notion of periodic tower is a counterpart of the notion of tower in the construction of uniform approximation while the notion of periodic process corresponds that of cityscape from Section 5.1.2.

The partition into all elements of all towers will normally be denoted by ξ, sometimes with indices. The permutation σ sends every element of ξ into the next element of its tower in cyclic order. Another partition naturally associated with a periodic process consists of the unions of bases of towers in each equivalence class and their images under the iterates of σ, where when we go beyond the height of a certain tower in the class we drop this tower and continue until the highest tower in the equivalence class has been exhausted. We will denote this partition by η, with appropriate indices. Obviously $\eta \leqslant \xi$.

DEFINITION 5.37. The sequence $(\xi_n, \eta_n, \sigma_n)$ of periodic processes is called *exhaustive* if $\eta_n \to \varepsilon$ as $n \to \infty$, i.e. for every measurable set $A \subset X$ there exists a sequence of sets $A_n \in \mathfrak{B}(\eta_n)$ such that $\mu(A \Delta A_n) \to 0$ as $n \to \infty$. An exhaustive sequence of periodic processes $(\xi_n, \eta_n, \sigma_n)$ is called *consistent* if for every measurable set $A \subseteq X$, the sequence $\sigma_n A$ converges to a set B, i.e. $\mu(\sigma_n A \Delta B) \to 0$ as $n \to \infty$.

Since $\xi_n \geqslant \eta_n$, then for an exhaustive sequence of periodic processes, $\xi_n \to \varepsilon$ as $n \to \infty$. For a consistent exhaustive sequence of periodic processes, independently of particular realizations of σ_n as measure-preserving transformations, the sequence $\{\sigma_n\}$ converges in the weak topology. For a given transformation T and an exhaustive sequence of periodic processes $(\xi_n, \eta_n, \sigma_n)$, a sufficient condition for the weak convergence of $\sigma_n \to T$ is $d(\xi_n, T, \sigma_n) = \sum_{c \in \xi_n} \mu(Tc \Delta \sigma_n c) \to 0$ as $n \to \infty$.

DEFINITION 5.38. If the last condition is satisfied we will say that the exhaustive sequence of periodic processes $(\xi_n, \eta_n, \sigma_n)$ forms a *periodic approximation* of T. In particular, if the periodic processes are cyclic the periodic approximation is called *cyclic*.

5.4.2. *Speed of approximation.* The *type* of approximation is defined in [78, Definition 1.9]. It involves a somewhat technical equivalence relation between sequences of periodic processes. However there is going to be no ambiguity for natural types of approximation discussed below, such as cyclic, type $(n, n+1)$ and so on. Given a type $\mathcal{T} = \{\tau_n\}$ in that sense defined above and a sequence $g(n)$ of positive numbers, we will say that a measure preserving transformation T *admits a periodic approximation of type* $\{\tau_n\}$ *with speed* $g(n)$ if for a certain subsequence $\{n_k\}$ there exists an exhaustive sequence of periodic processes $(\xi_k, \eta_k, \sigma_k)$ of type τ_{n_k} such that

$$d(\xi_k, T, \sigma_k) < g(n_k).$$

The speed of approximation will usually be measured against a certain *characteristic parameter* q depending on the type. There is a natural notion of a good speed of approximation, which generally means that a typical orbit of the limit transformation reproduces the behavior of one of the orbits of the approximation for sufficiently many periods. Usually the characteristic parameter q is chosen in such a way that *good approximation means approximation with any speed of the form* $g(q) = o(1/q)$. In the particular case of cyclic approximation the only parameter for a cyclic process is the height q of its single tower, which naturally serves as the characteristic parameter. Cyclic approximation with speed $o(1/q)$ is usually called *good cyclic approximation*. Good cyclic approximation is characteristic for the elliptic paradigm in smooth ergodic theory (see Section 2.2.4). Principal properties of transformations allowing good cyclic approximation which are thus typical for the elliptic paradigm are summarized in the following proposition. When it is possible we also describe weaker conditions.

PROPOSITION 5.39. *If T admits a good cyclic approximation then*:

(1) *T is ergodic. This remains true for a cyclic approximation with speed $(4-\theta)/q$ for any fixed $\theta > 0$* [81,29].

(2) *T is not mixing. This remains true for the speed* $(2-\theta)/q$, $\theta > 0$ [81,29].
(3) *The maximal spectral type of T is singular* [81]. *This remains true for speed* $(1-\theta)/n$ *for any* $\theta > 0$ [78, *Section* 3].
(4) *T is rigid* [81]. *This property implies* (3).
(5) *T is standard*; *this remains true for the speed* $(2-\theta)/q$, $\theta > 0$.
(6) *T is rank one.*
(7) *T has simple spectrum. This follows from* (6). *This property remains true for speed* $(1-\theta)/q$ *for any* $\theta > 0$ [81,29].

Good cyclic approximation does not allow to distinguish between transformations with pure point, mixed or continuous spectrum. In fact, every ergodic transformation with pure point spectrum admits good cyclic approximation [75, Section 8]. Here we give an example of another approximation property which guarantees weak mixing.

The type of periodic approximation is generated by periodic processes equivalent to processes consisting of two substantial towers t_1, t_2 whose heights differ by 1. Equivalently the heights of the two towers are equal to n and $n+1$ and for some $r > 0$,

$$m(t_1) > r/n \quad \text{and} \quad m(t_2) > r/n. \tag{5.3}$$

This type of approximation is said to be of *type* $(n, n+1)$. This type of approximation is related with the rank two property (see Section 5.2.2) and it implies rank two if the speed is sufficiently high; however the extra property that the roof of each tower returns mostly to the base of the same tower makes it stronger. For approximation of type $(n, n+1)$ the choice of the characteristic parameter is ambiguous. There are two natural ways to define it according to what properties of the limit transformation T we want to study. Namely, we can either take the characteristic parameters q as the length of one of the cycles (n or $n+1$), or as the period $n(n+1)$ of the permutation σ. We will call the approximation of type $(n, n+1)$ with speed $o(1/n)$ *good* and the approximation with speed $o(1/n(n+1))$ *excellent*. One some occasions it will be necessary to assume that the two towers involved in the approximation are equivalent. This simply insures that the partitions generated by the union of the bases of the towers and the iterates of this set is fine. The corresponding approximation will be called *linked approximation of type* $(n, n+1)$.

PROPOSITION 5.40 [82, Theorem 5.1]. *If a transformation T admits a good linked approximation of type* $(n, n+1)$ *or if T is ergodic and admits a good approximation of type* $(n, n+1)$ *then T has continuous spectrum.*

SKETCH OF PROOF. The proof is very similar to the proof of weak mixing for the Chacon transformation (Theorem 5.12). Namely an eigenfunction with eigenvalue λ would have to be almost constant on a typical level of the linked towers and hence on the base. But since return to the base happens mostly in two successive moments n and $n+1$ which implies that both λ^n and λ^{n+1} are close to one and hence in the limit $\lambda = 1$ which contradicts ergodicity. □

The property of approximation of type $(n, n+1)$ (linked or not) is compatible with cyclic approximation with arbitrary high speed. This allows to demonstrate in very simple

concrete examples how transformations admitting good periodic approximation may have mixed or continuous spectrum, see, for example, Theorem 5.28. Another historically important two-parametric family of examples is the two point extension of the rotation R_α with the switch of levels on the interval $[0, \beta]$. For almost every (α, β) the spectrum in the space of "odd" functions is simple, singular and continuous.

5.4.3. *Further properties and applications.* Some more elaborate versions of periodic approximation either compatible with fast cyclic approximation or not produce interesting properties.

PROPOSITION 5.41. *If T admits an excellent linked approximation of type $(n, n+1)$ then the maximal spectral multiplicity $M_{T\times T}$ (cf. 4.1) of $T \times T$ is finite and is less than or equal to $2[1/2r(1-r)]$, where r is the constant from* (3.6). *In particular, if $r > 1/2 - \sqrt{3}/6$ then $M_{T\times T} \leqslant 4$.*

SKETCH OF PROOF. The Cartesian square of the periodic process approximating T is a periodic process approximating $T \times T$ which, in this case, includes two substantial towers of height $n(n+1)$ and each of these towers has measure at least $r(1-r)$. For this periodic process, the length of the maximal cycle is equal to the period of the permutation $\sigma \times \sigma$. Furthermore, if the original approximation is excellent then the approximation of $T \times T$ is good when measured against this parameter. We consider the invariant subspace generated by characteristic functions of the bases of two towers of height $n(n+1)$ and apply Theorem 1.21 to this subspace. □

There is a natural generalization of an approximation of type $(n, n+1)$ which is useful for dealing with higher Cartesian powers. It involves several substantial towers whose heights are consecutive integers. A version of this property is also crucial in the proof of the genericity of the following useful property due to Stepin and Oseledec [148]; see also [149].

DEFINITION 5.42. Given $0 \leqslant \alpha \leqslant 1$, a measure-preserving transformation T is called *α-weak mixing* if for some sequence $n_k \to \infty$ and for every set A,

$$\lim_{k\to\infty} \mu\big(T^{n_k} A \cap A\big) = \alpha\mu(A)^2 + (1-\alpha)\mu(A).$$

An equivalent formulation of α-weak mixing is that the operators $U_T^{n_k}$ converge in the weak operator topology to $(1-\alpha)\,\mathrm{Id} + \alpha P_c$ where P_c is the orthogonal projection to the one-dimensional space of constants. 0-weak mixing corresponds to rigidity, whereas 1-weak mixing corresponds to the usual notion of weak mixing. Although the terminology may suggest it, α-weak mixing does not imply β-weak mixing for $\beta < \alpha$. On the contrary, α-weak mixing for any $\alpha > 0$ implies 1-weak mixing.

PROPOSITION 5.43. *If T is α-weak mixing for some $0 < \alpha < 1$ and ρ is the maximal spectral type for $U_T|_{L_2^0(X,\mu)}$, then all of the convolutions $\rho^{(m)}$ for $m = 1, 2, \ldots$ are pairwise singular.*

Now let us show how to derive α-weak mixing from an approximation. We consider a process with s linked towers $t_1, \dots, t_s$ of consecutive heights $q, q+1, q+2, \dots, q+s-1$, where q will serve as the parameter for good approximation. If $m(t_i) = \mu_i$, $i = 1, \dots, s$, we will call such an approximation a *linked approximation of type* $(q, q+1, \dots, q+s-1; \mu_1, \dots, \mu_s)$.

PROPOSITION 5.44. *Given α, $0 \leqslant \alpha \leqslant 1$, if T admits a good linked approximation of type*

$$\left(q, q+1, \dots, q+s-1;\ \frac{1-\alpha}{q}, \frac{\alpha}{(q+1)(s-1)}, \dots, \frac{\alpha}{(q+s-1)(s-1)}\right)$$

for an arbitrary large s, then T is α-weak mixing.

An application of α-weak mixing, given by del Junco and Lemańczyk [32], is that it implies a kind of "rigidity of joinings" property.

THEOREM 5.45. *Let $(X, \mathcal{A}, m, T)$ be α-weakly mixing with $0 < \alpha < 1$. Consider $S = \prod_{i \in N}(X_i, \mathcal{A}_i, m_i, T_i)$, where $(X_i, \mathcal{A}_i, m_i, T_i)$ is a copy of $(X, \mathcal{A}, m, T)$ for each $i \in N$. If $\mathcal{B}$ is an S invariant subalgebra of $\prod_{i \in M} \mathcal{A}_i$ restricted to which S acts isomorphically to a factor of T, then $\mathcal{B}$ is a factor of some $\mathcal{A}_i$.*

The proof uses Proposition 5.43 and the property is already sufficient to produce, with the help of the same techniques, some of the examples which can be obtained using transformations which have minimal self-joinings. The authors have given an extension of the notion of α-weak mixing, $(\alpha_1, \alpha_2, \dots, \alpha_s)$-weak mixing, such that transformations which satisfy it can be used as building blocks to exhibit most of the examples of the "counterexample machine" of D. Rudolph. It is interesting that this can be reached out of purely spectral properties. However $(\alpha_1, \alpha_2, \dots, \alpha_s)$-weak mixing transformations are only produced through constructions involving some grafting of "mixing rank one type" objects, which hinders any simple presentation.

B. Fayad [46] developed a novel concept of periodic approximation where at each given moment only a small part of the space returns close to itself but over the time most points experience this return infinitely many times. The goal was to find a criterion of singular spectrum which is compatible with mixing. Abstract description of the property in purely measurable terms in somewhat cumbersome and in [46] a structure of metric space is assumed. Then the property of *slowly coalescent periodic approximation* involves systems of balls of decreasing size returning to themselves at exponentially growing moments of time with exponentially small relative error in such a way that almost every point belongs to infinitely many such balls.

PROPOSITION 5.46. *Any transformation which admits slowly coalescent periodic approximation has singular spectrum* (*not necessarily continuous*).

5.4.4. *Genericity of periodic approximation* [78, Section 2]. Many important properties generic for measure preserving transformations in weak topology can be deduced for the following result (see [78, Theorem 2.1]).

THEOREM 5.47. *Given a type $T = \{\tau_n\}$ and a speed $g(n)$, the set of all measure-preserving transformations of a Lebesgue space which admit a periodic approximation of type T with speed $g(n)$ is a residual set* (*i.e. it contains a dense G_δ set*) *in the weak topology.*

In particular, all properties discussed earlier in this section which follow form a certain type of periodic approximation belong to this category. For convenience we formulate this as a separate statement.

COROLLARY 5.48. *A generic measure preserving transformation in the weak topology is weakly mixing* (*hence ergodic*), *rigid* (*hence is not mildly mixing*), *has simple singular spectrum such that the maximal spectral type in L_0^2 together with all its convolutions are mutually singular and supported by a thin set on any given scale.*

We will see later that one can add to this list homogeneous spectrum of multiplicity two for the Cartesian square, see Section 5.8.2, and other properties.

5.5. *Approximation by conjugation*

5.5.1. *General scheme.* Approximation by conjugation is a method of producing transformations admitting fast periodic approximation as well as some other transformations with interesting properties by conjugating elements (usually periodic) of actions of compact groups (usually S^1, but sometimes $\mathbb{T}^k$ and others) and taking limits in various topologies. This method is particularly suitable for smooth realizations of measure preserving transformations with various properties. It was first introduced in [18]; this is still the basic source on the subject. For an account of some recent development as well as an up-to-date perspective on the topic see [47]. A purely measurable version of the method and some of its applications are described in [78, Section 8]. Since most applications of the method still deal with the smooth situation, we will present the set-up and results for that case. We present a general overview of the method following [47].

Let M be a differentiable manifold with a non-trivial smooth circle action $\mathcal{S} = \{S_t\}_{t\in\mathbb{R}}$, $S_{t+1} = S_t$, preserving a smooth volume. Every smooth S^1 action preserves a smooth volume ν which can be obtained by taking any volume μ and averaging it with respect to the action: $\nu = \int_0^1 (S_t)_*\mu\, dt$. Similarly $\mathcal{S}$ preserves a smooth Riemannian metric on M obtained by averaging of any smooth Riemannian metric.

Volume preserving maps with various interesting, often surprising, topological and ergodic properties are obtained as limits of volume preserving periodic transformations

$$f = \lim_{n\to\infty} f_n, \quad \text{where } f_n = H_n S_{\alpha_{n+1}} H_n^{-1} \tag{5.4}$$

with $\alpha_n = \frac{p_n}{q_n} \in \mathbb{Q}$ and

$$H_n = h_1 \circ \cdots \circ h_n, \tag{5.5}$$

where every h_n is a volume preserving diffeomorphism of M that satisfies

$$h_n \circ S_{\alpha_n} = S_{\alpha_n} \circ h_n. \tag{5.6}$$

In certain versions of the method the diffeomorphisms h_n are chosen not preserving the volume but distorting it in a controllable way; this, for example, is the only interesting situation when M is the circle (see, e.g., [79, Section 12.6]).

Usually at step n, the diffeomorphism h_n is constructed first, and α_{n+1} is chosen afterwards close enough to α_n to guarantee convergence of the construction. For example, it is easy to see that for the limit in (5.4) to exist in the C^∞ topology it is largely sufficient to ask that

$$|\alpha_{n+1} - \alpha_n| \leqslant \frac{1}{2^n q_n \|H_n\|_{C^n}}. \tag{5.7}$$

The power and fruitfulness of the method depend on the fact that the sequence of diffeomorphisms f_n is made to converge while the conjugates H_n diverge often "wildly" albeit in a controlled (or prescribed) way. Dynamics of the circle actions and of their individual elements is simple and well-understood. In particular, no element of such an action is ergodic or topologically transitive, unless the circle action itself is transitive, i.e. $M = S^1$. To provide interesting asymptotic properties of the limit typically the successive conjugates spread the orbits of the circle action $\mathcal{S}$ (and hence also those of its restriction to the subgroup C_q of order q for any sufficiently large q) across the phase space M making them almost dense, or almost uniformly distributed, or approximate another type of interesting asymptotic behavior. Due to the high speed of convergence this remains true for sufficiently long orbit segments of the limit diffeomorphism. To guarantee an appropriate speed of approximation extra conditions on convergence of approximations in addition to (5.7) may be required.

There are many variations of the construction within this general scheme. In different versions of the approximation by conjugation method one may control the asymptotic behavior of almost all orbits with respect to the invariant volume, or of all orbits. Somewhat imprecisely we will call those versions ergodic and topological.

Ergodic constructions deal with measure-theoretic (ergodic) properties with respect to a given invariant volume, such as the number of ergodic components (in particular ergodicity), rigidity, weak mixing, mixing, further spectral properties. Topological constructions deal with minimality, number of ergodic invariant measures (e.g., unique ergodicity) and their supports, presence of particular invariant sets, and so on.

Control over behavior of the orbits of approximating periodic diffeomorphisms f_n in (5.4) on the nth step of the construction is typically provided by taking an invariant under S_{α_n} (and hence under $S_{\frac{1}{q_n}}$) collection of "kernels", usually smooth balls, and redistributing them in the phase space in a prescribed fashion (also $S_{\frac{1}{q_n}}$ invariant). In ergodic constructions one requires the complement to the union of the kernels to have small volume and hence most orbits of $\mathcal{S}$ (and consequently of any finite subgroup C_q for a sufficiently large q) to spend most of the time inside the kernels. In the topological versions the kernels need to be chosen in such a way that *every* orbit of $\mathcal{S}$ spends most of the time inside the kernels. This requires more care and certain attention to the geometry of orbits.

A natural way of selecting the kernels, their intended images, and constructing a map h_n satisfying (5.6) is by taking a fundamental domain Δ for S_{α_n} (or, equivalently, for $S_{\frac{1}{q_n}}$) choosing kernels and images inside Δ, constructing a diffeomorphism of Δ to itself identical near its boundary which sends kernels into their intended images, and extending the map to the images $S_{\frac{k}{q_n}}$, $k = 1, \dots, q_n - 1$, by commutativity. This method in particular is used in the construction of ergodic diffeomorphisms conjugate to a rotation on manifolds other than the circle as well as in a number of constructions where topological properties are involved. However in order to achieve other ergodic properties, for example weak mixing, it is necessary to use more general constructions.

5.5.2. *Generic constructions.* The first group of results obtained by the approximation by conjugation method deals with realization of certain ergodic properties in the category of C^∞ diffeomorphisms of a compact manifold preserving a smooth volume, i.e. a volume given by a positive C^∞ function in every local C^∞ coordinate system. First recall that all volumes with fixed total volume on a given manifold are conjugate by a C^∞ diffeomorphism [113]. Before we start listing properties which can be produced in the framework of the method it is useful to mention that the constructions come in two different varieties which will be called generic and non-generic; justification for this terminology will become apparent soon.

In the constructions of the first kind (generic) it is sufficient *to control the behavior of approximating and hence resulting diffeomorphisms on a series of growing but unrelated time scales*. To carry out those construction the commutativity condition (5.6) is not necessary. In fact the conjugating maps H_n while formally can be written as products as in (5.5) are not constructed as such. Instead an approximate version of the desired property is achieved by conjugation and care is taken that the sequence f_n converges. The approximate pictures may look quite whimsical (see, e.g., the original weak mixing construction in [18, Section 5] and a modern version in [67]), but as long as a diffeomorphism is close enough to conjugates of rotations appearing in such pictures the property is guaranteed. A natural setting for those constructions is categorical. One considers the space $\mathcal{A}$, the closure of diffeomorphisms of the form gS_tg^{-1} in C^∞ topology. Here we fix a volume ν invariant by the action $\mathcal{S}$ and consider all C^∞ diffeomorphisms g preserving ν. Notice that $\mathcal{A}$ is a complete metrizable space and hence Baire category theorem can be used.

This was first noticed in [18, Section 7] in connection with ergodic properties with respect to the invariant volume and was used in [41] to control topological properties. In fact, for a proof of genericity in $\mathcal{A}$ of a property exhibited by a construction of this sort no actual inductive construction is needed. One just needs to show that an approximate picture at each scale appears for an open dense subset of conjugates of rotations. If appearance in an approximate picture at infinitely many growing scales guarantees the property then by the Baire category theorem the property holds for a dense G_δ subset on $\mathcal{A}$.

THEOREM 5.49. *For any positive function $g(n)$ the space $\mathcal{A}$ contains a dense G_δ subset of weakly mixing diffeomorphisms which admit cyclic approximation with the speed g* [18]. *Furthermore, transformations in that set are α-weak mixing for every α, $0 \leqslant \alpha \leqslant 1$.*

If the action $\mathcal{S}$ is fixed point free then $\mathcal{A}$ contains a dense G_δ subset of uniquely ergodic diffeomorphisms [41].

Even if the action $\mathcal{S}$ has fixed points or if the manifold M has a boundary the number of invariant measures can be controlled and is generically the minimal possible. Here is a nice low-dimensional example.

Let M be one of three manifolds: the disc $\mathbb{D}^2$, the annulus $[0,1]\times S^1$ or the sphere S^2, λ Lebesgue measure and $\mathcal{S}$ action by rotations (uniquely defined on the disc and the annulus and defined by a choice of axis on the sphere). Let us call Lebesgue measure on the manifold, the δ-measures at the fixed points of the rotations and Lebesgue measures on the boundary components the *natural measures*.

THEOREM 5.50 [47, Theorem 3.3]. *Let M be $\mathbb{D}^2$, $[0,1]\times S^1$ or S^2, and S_t be the standard action by rotations. Diffeomorphisms that have exactly three ergodic invariant measures, namely the natural measures on M, form a residual set in the space $\mathcal{A}'$: the closure in the C^∞ topology of the conjugates of rotations with conjugates fixing the fixed points of $\mathcal{S}$ and every point of the boundary.*

5.5.3. *Non-generic constructions.* In the constructions of the second kind approximations at different steps of the construction are linked and hence in principle *the asymptotic behavior of the resulting diffeomorphism is controlled for all times*. Constructions of this kind appear most naturally when the resulting diffeomorphism is constructed to be measure-theoretically conjugate to a map of a particular kind, but they also appear when one constructs transformations with more than one ergodic component [157]. This category also includes mixing constructions which were first introduced for time changes for flows on higher-dimensional tori [43,44] and were developed in [47, Section 6] in the context of the approximation by conjugation method. In the latter case one needs to start from a smooth action of a torus rather than of a circle.

Non-standard realizations of Liouvillean rotations. Recall that a number α is called *Liouvillean* if it allows approximation by rationals better than any negative power of denominators.

THEOREM 5.51. *Let α be an arbitrary Liouvillean number. Then arbitrary close to S_α in C^∞ topology there exists a diffeomorphism preserving the volume ν, ergodic and measurably conjugate to the rotation R_α.*

This result was proved in [18, Section 4] for a dense set of α; the proof for arbitrary Liouvillean α is forthcoming [49].

Let us explain why this result may be considered definitive.

In the case of the disc or the annulus with the standard action by rotations the diffeomorphisms in question act as rotations R_α on boundary component(s).

Numbers other than Liouvillean are called *Diophantine*. For Diophantine rotation numbers such a realization on the disc or annulus (with rotation on the boundary) is impossible since due to M. Herman's "last geometric theorem" (to be published posthumously) any such diffeomorphism has uncountably many invariant circles and hence cannot be ergodic.

Other realization results. Possibilities of realizing of particular transformations or members of particular families within the framework of the approximation by conjugation method has not been explored systematically; see [47, Section 7] for a sample of open questions as well as a discussion of prospects and difficulties. As is the case or rotations it looks that realization is often possible for certain subsets of transformations from finite- or infinite-parameter families for the sets of parameters which are residual but very "thin" in the metric sense. However, unlike the rotation situation it is hard to expect definitive results. We restrict ourselves to a sample of results for that kind.

THEOREM 5.52 [18, Section 6]. *For any natural number n there is a dense set of vectors $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ whose coordinates satisfy no rational relation such that there exist a diffeomorphism $f \in \mathcal{A}$ arbitrary close to S_β for some β and measurably conjugate to the translation T_α on the torus $\mathbb{T}^n$.*

There exists a dense in the product topology set of vectors $\alpha = (\alpha_1, \alpha_2, \dots) \in \mathbb{R}^\infty$ whose coordinates satisfy no rational relation such that there exist a diffeomorphism $f \in \mathcal{A}$ arbitrary close to S_β for some β and measurably conjugate to the translation

$$T_\alpha : x \to x + \alpha \pmod 1$$

on the torus $\mathbb{T}^\infty$.

THEOREM 5.53. *Arbitrary close to any transformation S_β for any β there exists a non-standard ergodic diffeomorphism.*

The proof is based on a smooth realization of a version of Feldman's construction described in [78, Section 8].

5.5.4. *Toral actions and mixing transformations.* The use of approximation type techniques to produce mixing transformations and flows was pioneered by B. Fayad [43]. He used reparametrizations of linear flows on the tori of dimension $\geqslant 3$ to produce mixing by carefully controlling behavior of the sequence of overlapping time scales. See Section 5.6.3 for a brief outline of the method. In [47] the techniques of reparametrization of linear flows on $\mathbb{T}^3$ were combined with the explicit approximation by conjugation methods. The basic setting is a compact smooth manifold M with non-trivial smooth $\mathbb{T}^3$ action $\mathcal{S} = \{S_v\}_{v \in \mathbb{R}^3}$, $S_{v+k} = S_v$ if $k \in \mathbb{Z}^3$ and a smooth volume μ preserved by $\mathcal{S}$.

THEOREM 5.54 [47, Theorem 6.2]. *There exists a sequence $\gamma_n \in \mathbb{Q}^3$ and a sequence H_n of diffeomorphisms preserving μ such that the sequence $H_n S_{t\gamma_n} H_n^{-1}$ converges in the C^∞ topology to a flow preserving μ and mixing for this measure.*

5.6. *Time change*

5.6.1. *General results.* Given a flow, T_t, the operation of *time change* produces a flow with the same orbits as T_t but evolving at a different speed and with an invariant measure accordingly changed with a suitable density.

Time changes can be described in terms of $\mathbb{R}$-valued untwisted one-cocycles over the flow, see [8, Section 1.3m] for a discussion in very general context and [78, Section 9.3] for basic definitions in the specific setting of flows. This operation is important in the topological and differentiable dynamics where the time change is assumed correspondingly continuous and differentiable. We already discussed a specific case of time change in Section 2.2.4.

Since every flow by Ambrose–Kakutani theorem can be represented as a special flow over a measure preserving transformation, the time change produces a special flow over the same transformation with a different roof function. In particular, if the roof functions φ and ψ for special flows over the same measure preserving transformation T are *cohomologous*, i.e.

$$\phi = \varphi + h \circ T - h$$

for a measurable function h then the special flows are isomorphic. The function h which in the case of ergodic T is uniquely defined mod 0 up to a constant is sometimes called *transfer function.*

The basic properties which are preserved by any time change are ergodicity (more generally, the structure of the decomposition into ergodic components) and the property of entropy to be zero, a positive number, or infinity. Other spectral and non-spectral invariants are in general not preserved.

Still it is a meaningful question to ask how the spectral properties of a flow may be modified by a time change.

Two basic general results in this direction show that stochastic properties may be improved by a proper time change. They are due to Kochergin [90] and Ornstein and Smorodinsky [120] correspondingly.

THEOREM 5.55.

(1) *For any ergodic flow there exists a time change which is mixing* [90].
(2) *For any ergodic flow with positive entropy there exists a time change which is a K-flow* [120].

In both cases the time change can be chosen arbitrary close to identity in a variety of senses; *for example, if a flow is represented as a special flow over a transformation the roof function can be changed arbitrary little in the uniform norm.*

5.6.2. *Continuous and almost differentiable time changes.* It is interesting and in fact remarkable that in the continuous category the time changes described in the previous theorem can be made continuous and in differentiable category "almost" differentiable (with derivative discontinuous only at one point). This follows from the analysis of cohomology classes of cocycles which produce time changes. If two cocycles are cohomologous then corresponding time changes are metrically isomorphic by a conjugacy which moves each point along its orbit according to the solution of the cocycle equation. We follow the presentation of [78, Section 10.2].

THEOREM 5.56. *Let $\mathcal{L} \subset L^1(X, \mu)$ be a linear subspace of L^1 dense in the L^1 topology and closed in the L^∞ topology* (*uniform convergence almost everywhere*). *Then for every*

$f \in L^1(X, \mu)$ the set $\mathcal{L}_f = \{h \in \mathcal{L}\colon$ h is cohomologous to $f\}$ is dense in the L^∞ topology in the set $\{h \in \mathcal{L},\ \int h\, d\mu = \int f\, d\mu\}$.

If we put $\mathcal{L} = C(X)$, the space of all continuous functions, we immediately obtain the following statement which was originally proved in [120].

COROLLARY 5.57. *Let X be a compact metric space, μ be a Borel probability nonatomic measure on X, $T : X \to X$ be a measure-preserving transformation* (*not necessarily continuous*). *Then every real-valued cocycle $f \in L^1(X, \mu)$ is cohomologous to a continuous cocycle. Moreover the set of continuous cocycles cohomologous to f is dense in uniform topology in the space of all continuous functions with the same integral as f.*

Corollary 5.57 can be strengthened by specifying the values of a continuous function cohomologous to f on any closed set F so that $\mu(X \setminus F) > 0$. Pushing the method described above a bit further one obtains the result advertized above which looks quite striking at first glance.

THEOREM 5.58. *Let M be a compact differentiable manifold, μ be a Borel probability measure on M, $T : M \to M$ be a measure-preserving transformation. Then every real-valued cocycle $f \in L^1(M, \mu)$ is cohomologous to a continuous cocycle $\bar{f}$ which is continuously differentiable except at a single point.*

SKETCH OF PROOF. First, one finds a continuous cocycle f_1 cohomologous to f which is continuously differentiable outside a ball B_1 of radius, say, $1/2$ and can be extended to a continuously differentiable function. This is possible by a stronger version of Corollary 5.57 mentioned above. Then one approximates f_1 in uniform topology by a continuously differentiable cocycle g_1 which coincides with f outside B_1. If the L^1 norm of $f_1 - g_1$ is small enough one can find a cocycle f_2 cohomologous to f_1 (and hence to f) which coincides with f_1 outside a smaller ball $B_2 \subset B_1$ of radius $1/4$ and extends to a continuously differentiable function and such that the support of the transfer function ψ_1 has measure less than $1/2$. Continuing by induction one constructs on the nth step the cocycle f_n continuously differentiable outside of a ball $B_n \subset B_{n-1}$ of radius 2^{n+1} which coincides with f_{n-1} outside of the ball B_{n-1} and extends to a continuously differentiable function and such that a transfer function ψ_n connecting f_n with f_{n-1} is supported on a set of measure less than 2^n. In the limit the function $\bar{f} = \lim_{n\to\infty} f_n$ is continuous everywhere and continuously differentiable outside of the single point $\bigcap_{n=1}^{\infty} B_n$. By the Borel–Cantelli lemma the series $\sum_{n=1}^{\infty} \psi_n$ converges and hence gives a transfer function between f_1 and $\bar{f}$. Since f_1 is cohomologous to f this finishes the proof. □

5.6.3. *Regular time change in various classes of systems.* Notice that the property of almost differentiability in Theorem 5.58 cannot be replaced by any reasonable *uniform* property stronger than continuity.

Hyperbolic and parabolic systems. For example, Hölder time changes behave quite differently for many classes of dynamical systems such as Anosov flows or special flows

over subshifts of finite type [79, Section 19.2], [78, Sections 11.3–4]. In those cases on the one hand, there are infinitely many moduli for existence of a *measurable* solution of the cohomological equation, and, on the other, robustness of spectral properties. The spectrum is either countable Lebesgue or that plus pure point component with single frequency. The latter is impossible for example for contact Anosov flows. Thus, in the hyperbolic (and to a certain extent partially hyperbolic situation [78, Section 11.5]) spectral properties exhibit robustness under reasonably regular time changes with the countable Lebesgue spectrum prevailing.

A somewhat similar albeit more subtle and less understood situation exists for parabolic systems. Since these effects are in essence different from those produced by combinatorial constructions which dominate this part of the survey we will discuss the topic later in Section 6.3.

Elliptic systems: codimension one. Now we will consider specific situations where interesting effects can be achieved by producing a nice (smooth, analytic Hölder, etc.) time change with interesting properties by means of a construction which successfully controls behavior at various time scales. In this respect this class of constructions fits with the general theme of this part of the survey. We already discussed time changes in a linear flow on the two-dimensional torus in Section 2.2.4. We will discuss the situation in more detail and comment on methods used. First, notice that for any irrational slope and for a sufficiently smooth time change (or, equivalently, the roof function for the special flow) the resulting flow (or the time one transformations) allows sufficiently good cyclic approximation to guarantee simple singular spectrum and the absence of mixing, see Proposition 5.39; the latter property also follows under much weaker assumptions from Theorem 5.61 below. Weak mixing of course does not follow from cyclic approximation. It can be produced by several different methods. To produce genericity one can use perturbation with small sinusoidal waves similar to those described below for producing mixing in higher dimension. A more interesting method deals with the study of special flows with a fixed roof function and varying translation in the base. This method leads to a conclusion that, while other types of behavior are possible, under certain assumptions, weak mixing is the only alternative to at least measurable conjugacy to the linear flow.

First, if the roof function is a trigonometric polynomial of if the translation is Diophantine and the function is C^∞ then the roof function is cohomologous (with the transfer function which is correspondingly itself a trigonometric polynomial of a C^∞ function) to its average and hence the flow is smoothly conjugate to a constant time suspension or, equivalently to a linear flow.

Shklover [146] proved the following converse to the statement about trigonometric polynomials.

THEOREM 5.59. *For any real-analytic function f other than trigonometric polynomials* (*in other words those with infinitely many non-zero Fourier coefficients*) *there is always an α such that the special flow over R_α with the roof function f is weakly mixing.*

A more quantitative statement connecting the approximation in the base with the decay of Fourier coefficients for the roof function is in [78, Theorem 13.7].

THEOREM 5.60. *Let $h(x) = \sum_{n\neq 0} h_n \exp 2\pi inx$ be a C^2 real valued function on S^1 with zero average. Suppose for a certain sequence of rational numbers p_n/q_n,*

$$\frac{q_n|\alpha - p_n/q_n|}{\sum_{k=1}^{\infty}|h_{kq_n}|} \to 0 \quad \textit{and} \quad \frac{|h_{q_n}|}{\sum_{k=1}^{\infty}|h_{kq_n}|} > c > 0.$$

Then for any h_0 and r the cocycle $\exp ir(h_0 + h(z))$ is not a coboundary and consequently the special flow over R_α with the roof function $h_0 + h(z)$ is weakly mixing.

Developing this method and using new ideas involving a central limit theorem to treat the case of intermediate approximation Fayad and Windsor proved in [50] that under stronger conditions on regularity of decay of Fourier coefficients than in Theorem 5.60 (satisfied, for example, when they are close enough to a geometric progression) there is a dichotomy between solvability of the cohomological equation in L^2 (and hence the pure point spectrum with two "right" frequencies) and weak mixing.

The following general criterion for absence of mixing was found in [76].

THEOREM 5.61. *Any special flow with the roof function of bounded variation over an interval exchange transformation is not mixing.*

The proof is based on using the return properties of the base transformation (Theorem 5.22) and the bounded variation of the roof function to show that returns in the base produce returns for the flow within a bounded time. Thus since a bounded from below proportion of measure returns close to itself in the base at a certain sequence of time moments growing to ∞ the same can be said about a fixed proportion of measure for the flow for a sequence of fixed length time segments. This contradicts mixing.

This result has been recently strengthened by Fraczek and Lemańczyk.

THEOREM 5.62 [57]. *Any special flow with the roof function of bounded variation over an ergodic interval exchange transformation is disjoint from any mixing flow.*

Mixing can be produced with a minimal loss of regularity. For example, any Lipschitz time change in a linear flow on $\mathbb{T}^2$ is not mixing by Theorem 5.61. One the other hand Kochergin proved the following converse to that statement.

THEOREM 5.63 [93]. *For any modulus of continuity ω weaker than Lipschitz, i.e. such that $\lim_{t\to 0} \frac{t}{\omega(t)} = 0$ one can find a linear flow on T^2 and a time change with modulus of continuity ω which is mixing.*

Equivalently, one can find a rotation R_α and a function f with modulus of continuity ω such that the special flow over R_α with the roof function f is mixing.

The construction is of inductive character producing approximate mixing on growing but overlapping time scale and is somewhat similar to a more subtle and specialized version of the general construction from [90].

Elliptic systems: higher codimension. An essential new phenomenon for time changes in linear flows on $\mathbb{T}^k$, $k \geqslant 3$, or, equivalently, in special flows over translations on $\mathbb{T}^k$, $k \geqslant 2$, is a possibility of mixing in very regular situations including real analytic [43]. This is a very special situation impossible in the Diophantine context and non-generic in Liouvillean. It is produced by an inductive combinatorial construction which we briefly outline for the case of a special flow over a translation of $\mathbb{T}^2$ with coordinates (x, y).

If we assume that the rotation in the x direction is periodic with period n consider an addition to a given roof function of the form $a \sin 2\pi n x$ then the successive returns will develop sinusoidal waves which at the time scale grater than n will produce approximate mixing for sets transversal to the x direction. Now we add a very small translation in the x direction to keep this effect for the perturbed system for a long enough time until the effect of a similar perturbation in the y direction of much greater frequency but much smaller magnitude takes over. This relies on a proper very special choice of periodic approximations in the x and y direction. The scales when mixing is produced by the stretching in the two directions overlap but because of the independence of the perturbations they do not interfere and cancel each other. Thus genuine albeit fairly slow mixing is achieved for the limit transformation whose base translation has the form (α, β) with $\alpha = \sum_{k=1}^{\infty} \frac{1}{n_k}$, $\beta = \sum_{k=1}^{\infty} \frac{1}{m_k}$ with $n_k \ll m_k \ll n_{k+1}$ and the roof function is of the form

$$\sum_{k=1}^{\infty} a_k \sin n_k x + b_k \sin m_k y \tag{5.8}$$

with $a_k \gg b_k \gg a_{k+1}$. The construction can be carried out in such a way that the function (5.8) is real analytic.

There are variations of this method where instead of sinusoidal waves different more elaborate shapes are used. For example, using some version of Dirichlet kernels one can combine mixing with Fayad's criterion of slowly coalescent periodic approximation for singularity of the spectrum which is compatible with mixing, see Proposition 5.46.

THEOREM 5.64 [46]. *There exists a C^∞ time change of a linear flow on $\mathbb{T}^3$ which is mixing and has singular spectrum.*

5.7. *Inducing*

The operation analogous to time change in the discrete case is the operation of inducing. The natural topology in the space of measurable subsets of a given space (X, μ) is given by the metric

$$d(A, B) = \mu(A \Delta B).$$

Denote the collection of all classes mod 0 of measurable sets provided with this metric by $\mathcal{X}$.

The following result is a counterpart of Theorem 5.55.

THEOREM 5.65.
(1) *Any ergodic transformation induces mixing on a dense in $\mathcal{X}$ class of sets* [58].
(2) *Any ergodic transformation with positive entropy induces K-automorphisms on a dense in $\mathcal{X}$ class of sets* [120].

The method of proving Theorems 5.55 and 5.65 is similar in spirit to cutting and stacking constructions albeit limited to introduction of spacers since the return maps for the towers are fixed.

De la Rue improved the first statement of the previous theorem:

THEOREM 5.66 [35]. *An ergodic transformation induces a transformation with Lebesgue spectrum on a dense in $\mathcal{X}$ class of sets.*

Multiplicity of Lebesgue spectrum in this construction is not known. Thus the following problem is open:

PROBLEM 5.67. Does any ergodic transformation with zero entropy induce a transformation with countable Lebesgue spectrum?

De la Rue in [36] has produced a spectral type which cannot be obtained in a standard transformation, i.e. on any induced of an irrational rotation. We will discuss this result based on the theory of Gaussian dynamical systems in Section 6.4.3.

Positive answer to the following problem would require an essentially new construction.

PROBLEM 5.68. Does any ergodic transformation with zero entropy induce a transformation with simple spectrum?

Conze [27] has proved that it is in fact generic that an induced of an ergodic transformation is weakly mixing. Notice that mixing is not generic.

In [78, Section 7] transformations induced by a standard transformation on various sets are considered. The following result is parallel to Theorem 5.47.

THEOREM 5.69. *Let T be a standard measure-preserving transformation. Given a type $\mathcal{T}$ and a speed $g(n)$, the set of all $A \in \mathcal{X}$ such that the induced transformation T_A admits a periodic approximation of type $\mathcal{T}$ with speed $g(n)$ is a residual set in $\mathcal{X}$.*

All the standard corollaries follow such as simple continuous singular spectrum which is mutually singular with all its convolutions. Since inducing (and the inverse operation of taking a *special transformation* which is the discrete time equivalent of the special flow) involves cohomological equations with integer values, interesting questions related with behavior of regular (analytic, smooth, etc.) real-valued cocycles which played the central role in Section 5.6.3 do not have direct equivalents in this setting.

5.8. *Spectral multiplicity, symmetry and group extensions*

5.8.1. *Introduction.* Most of this section deals with the realization problem for various sets of essential values of spectral multiplicity; see the preview in Section 3.6.2. Success in treating of this problem via appropriate constructions is based on the combination of two principal elements:

(i) *Symmetry* which allows to produce for certain classes of transformations various intertwining operators in L^2 (often but not always coming from commuting measure preserving transformations) which interchanges various subspaces and hence guarantees that certain parts of the spectrum come with multiplicity, and

(ii) *Approximation* which shows that "minimal" multiplicities compatible with the symmetry are actually realized. Approximation properties come from combinatorial constructions. Sometimes it is sufficient to consider generic data within a given class of transformations; in other cases more careful inductive process might be needed.

The subtlety of using approximation techniques is in that it is not always sufficient to produce approximation which allows to obtain an above estimate for the multiplicity using Theorem 1.21 or something similar but (in the case of non-homogeneous spectrum) one needs separate estimates in various subspaces responsible for parts of the spectrum with different values of the multiplicity function.

5.8.2. *Homogeneous spectrum of multiplicity two and Cartesian products.* Ergodic measure preserving transformations with homogeneous spectrum of multiplicity two were found simultaneously and independently by Ryzhikov [144] and Ageev [16]. They used approach of [78] and improved the estimate given by Proposition 5.41.

THEOREM 5.70. *For a generic in the weak topology measure preserving transformation T the Cartesian square $T \times T$ has homogeneous spectrum with multiplicity two.*

PROOF. The symmetry here is the involution $J : (x, y) \mapsto (y, x)$ which guarantees that essential values of the spectral multiplicity are even (see Proposition 4.2) and the approximation is, first, good cyclic approximation for T which insure simple spectrum and hence, multiplicity two for the part of the spectrum coming from functions depending only on one coordinate and, second, a slightly generalized version of good approximation of type $(n, n+1)$ (see Section 5.4.2). Namely, for a given natural number m we will consider a good linked approximation of type $(n, n+m)$ by periodic processes with two towers whose size is bounded away from zero and heights differing by m. Existence of this kind of approximation guarantees that weak limit of powers of T contains a linear combination $\alpha \operatorname{Id} + (1-\alpha)T$. This of course means that the limit of U_{T^n} in the strong operator topology contains $\alpha \operatorname{Id} + (1-\alpha)U_T$.

It is sufficient to prove that the maximal spectral multiplicity of $U_{T\times T}$ is at most two. Thus the theorem will follow from the following lemma

LEMMA 5.71. *If T admits a good cyclic approximation and a good approximation of type $(n, n+m)$ for any natural m and f is a cyclic vector for U_T then the functions $f(x)f(y)$ and $f(x)f(Ty)$ generate L^2.*

PROOF. Since f is a cyclic vector for U_T the functions of the form $f(T^k x) f(T^m y)$ generate L^2 for the Cartesian product. Thus it is sufficient to show that any function of the form $f(x) f(T^m y)$ belongs to the space generated by $f(x) f(y)$ and $f(x) f(Ty)$ which we will denote by H. To simplify notations let us denote $f(T^m x) f(T^k y)$ by $m \times k$ and use similar notation for linear combinations of such functions. From the invariance one gets for every $m \in \mathbb{Z}$,

$$m \times m \in H \quad \text{and} \quad m \times (m+1) \in H.$$

Thus from the approximation criterion $(\alpha 0 + (1-\alpha)m) \times (\alpha 0 + (1-\alpha)m) \in H$ hence by invariance $0 \times m + m \times 0 \in H$. Similarly by taking limits of some iterates of 0×1 we obtain $0 \times m + (m-1) \times 1 \in H$ and hence

$$m \times 0 + (m-1) \times 1 \in H. \tag{5.9}$$

Using these inclusions inductively for $m = 2, 3, \ldots$ we obtain that $m \times 0 \in H$. For $m = 2$ one obtains $0 \times 2 + 1 \times 1 \in H$ and hence $0 \times 2 \in H$. Assuming that $k \times 0 \in H$ for $k \leqslant m$, in particular, $(m-1) \times 0 \in H$ and hence $m \times 1 \in H$ we get from (5.9) that $(m+1) \times 0 \in H$. □

This finishes the proof of the theorem. □

Looking back at the structure of the spectrum for the Cartesian square described in Proposition 4.2 we deduce interesting arithmetic structure of the maximal spectral type for a transformation T whose Cartesian square has spectrum of multiplicity two. First, any measure μ of the maximal spectral type is singular with respect to its convolution $\mu * \mu$ and, second for almost every with respect to $\mu * \mu$ $\lambda \in S^1$ the conditional of $\mu \times \mu$ on the circle $\lambda_1 \lambda_2 = \lambda$ is concentrated in two symmetric points $(\lambda_1^0, \lambda_2^0)$ and $(\lambda_2^0, \lambda_1^0)$.

A more sophisticated analysis allows to describe essential values of spectral multiplicity for the mth Cartesian power of a generic measure preserving transformation where the symmetry is given by the symmetric group S_m of permutations of components and where the maximal spectral multiplicity is at least $m!$.

THEOREM 5.72 [144,16]. *For a generic measure preserving transformation T the mth, $m \geqslant 3$, Cartesian power $T^{(m)}$ has $m-1$ different values of the spectral multiplicity*: $m, m(m-1), m(m-1)(m-2), \ldots, m!$.

5.8.3. *Homogeneous spectrum of arbitrary multiplicity and group actions.* Measure preserving transformations with homogeneous spectrum of arbitrary multiplicity (including new examples with multiplicity two) were recently found by Ageev [17] using a different type of symmetry. His main idea is quite brilliant although in retrospect it looks natural.

Ageev considers the following group G_m. It is a finite extension of $\mathbb{Z}^m$ and has generators $T_1, \ldots, T_m, S$ where $T_1, \ldots, T_m$ commute, $T_1 \cdot T_2 \cdot \cdots \cdot T_m = \mathrm{Id}$ and $T_{i+1} = S \cdot T_i \cdot S^{-1}$ for $i = 1, \ldots, m-1$. Notice that S^m commutes with $T_1, \ldots, T_m$ and thus the group G_m is an m-fold extension of the Abelian group with generators $T_1, \ldots, T_{m-1}, S^m$.

THEOREM 5.73. *For a generic action α of the group G_m by measure preserving transformations of Lebesgue space the transformation $\alpha(S^m)$ has homogeneous spectrum of multiplicity m.*

The upper bound on the spectral multiplicity is provided by simplicity of the spectrum for S; this can be achieved using a proper version of periodic approximation theory for actions of G_m. It is a standard corollary of the Spectral Theorem 1.8 that then the spectrum of the mth power has multiplicity at most m. Spectral theory for this group provides for symmetry. In particular if S is weakly mixing (which can also be guaranteed by approximation arguments) there are m mutually orthogonal S^m invariant subspaces where the restriction of the Koopman operator are unitarily equivalent so by Corollary 1.20 the values of spectral multiplicity are multiples of m.

5.8.4. *Non-homogeneous spectrum, group extension and factors.* These examples which produced successively more general sets of values of spectral multiplicities from $\{1, m\}$ [134], to finite [135] and infinite [65] sets containing 1 and invariant under taking the least common multiple, to arbitrary sets containing 1 [100], are all based on finite and, more generally, compact group extensions of transformations admitting good cyclic approximation with cocycles possessing certain symmetry. The idea actually goes back to the work of Oseledets [123] who was the first to construct an example of a measure preserving transformation with non-simple spectrum of bounded multiplicity. However, his upper estimate based on Theorem 5.24 was very crude. Oseledets' example was the starting point for Robinson who introduced finer methods of estimating the multiplicity from above. Here we will describe Robinson's first construction since it shows both the symmetry and approximation elements in a clear and suggestive way. We follow [78].

We will consider T, the double group extension of a transformation T_0. $T : X \times \mathbb{Z}/m\mathbb{Z} \times \mathbb{F}_p \to X \times \mathbb{Z}/m\mathbb{Z} \times \mathbb{F}_p$ where p is a prime number specified below and $\mathbb{F}_p$ is the finite field with p elements, of the following special form

$$T(x, y, z) = \big(T_0 x, \gamma(x) + y, \phi(y) + z\big). \tag{5.10}$$

Here $\gamma : X \to \mathbb{Z}/m\mathbb{Z}$ is a measurable function which will be specified to provide approximation properties needed to the above estimate of the spectral multiplicity. For any m there exists a prime number p and an isomorphism $\phi : \mathbb{Z}_m \to G \subseteq \mathbb{F}_p$, where G is a subgroup of the multiplicative group $\mathbb{F}_p^*$ of the finite field $\mathbb{F}_p$ with p elements. These are the data which go to the second extension.

THEOREM 5.74. *For a generic in weak topology T_0 and a generic in L_1 set of cocycles γ the transformation T defined by* (5.10) *is weakly mixing and has $\{1, m\}$ as the set of essential values of the spectral multiplicity.*

REMARK. In fact, genericity arguments are not necessary as the proof below shows. The required conditions are certain approximation properties which can be guaranteed by choosing, for example, a certain exchange of three intervals as T_0 and a certain piecewise constant function as γ.

PROOF. Associated with a finite group extension there is a natural orthogonal decomposition of L_2 into U_T-invariant subspaces corresponding to the characters of the group. The additive characters of $\mathbb{F}_p$ are given by $\chi_w(z) = \exp 2\pi i z w/p$ where $w \in \mathbb{F}_p$, so that if T is given by (5.10) we obtain an invariant orthogonal decomposition

$$L_2(X \times \mathbb{Z}/m\mathbb{Z} \times \mathbb{F}_p) = \bigoplus_{w \in \mathbb{F}_p} H_w,$$

where

$$H_w = \{\chi_w(z) f(x, y) \colon\ f \in L_2(X \times \mathbb{Z}/m\mathbb{Z})\}.$$

Let us define a permutation $\sigma : \mathbb{F}_p \to \mathbb{F}_p$ by $\sigma(w) = \phi(1)w$. For $w \neq 0$ we also define the operator

$$S_w : H_w \to H_{\sigma(w)} \quad \text{by} \quad S_w\big(\chi_w(z) f(x, y)\big) = \chi_{\sigma(w)} f(x, y + 1).$$

Since $\chi_{\sigma(w)}(\phi(y)) = \chi_w(\phi(y + 1))$, one has $U_T|_{H_{\sigma(w)}} \cdot S_w = S_w \cdot U_T|_{H_w}$. Now let us examine the permutation σ. It fixes 0 and has $m' = \frac{p-1}{m}$ cycles of length m. This explains how the operators S_w permute the subspaces H_w. We will choose an arbitrary element θ_k, $k = 1, \ldots, m'$, from the kth cycle of σ, and for $j = 0, \ldots, m - 1$ we will define the subspace

$$H^j = H_{\sigma^j(\theta_1)} \oplus H_{\sigma^j(\theta_2)} \oplus \cdots \oplus H_{\sigma^j(\theta_{m'})}.$$

We will also define

$$H^* = H_0.$$

It is clear that $L_2(X \times \mathbb{Z}/m\mathbb{Z} \times \mathbb{F}_p) = H^* \oplus H^0 \oplus \cdots \oplus H^{m-1}$. The linear operator

$$S^j : H^j \to H^{j+1}, \quad j \neq *,$$

is defined in the natural way so that

$$S^j|_{H_w} H_w = H_{\sigma(w)} \subseteq H^{j+1}.$$

It follows that

$$S^j \cdot U_T|_{H^j} = U_T|_{H^{J+1}} \cdot S^j$$

and thus since the spectra in all of the spaces H^j are identical, the maximal spectral multiplicity of T is at least m. To obtain the estimate of the maximal spectral multiplicity for T from above we will need two types of approximation for the first extension T_1. In particular, these will guarantee that U_{T_1}, or equivalently $U_T|_{H^*}$, has simple continuous spectrum.

They are, (i) a good linked approximation of type $(n, n+1)$ (see Proposition 5.40) and (ii) a certain good approximation with m towers of equal height, which are related to each other by shifts $(x, y) \to (x, y+k)$. By extending the approximation for T_1 to the second extension, we obtain from (ii): (iii) a good approximation for T with m towers. Since at least one of the towers has size close to $1/m$, Theorem 5.10 implies that maximal spectral multiplicity for T is no greater than m. This in particular implies ergodicity of T since otherwise there would be invariant functions in every H^j in addition to constants contradicting the above estimate for the spectral multiplicity. This in turn implies weak mixing since otherwise there would be eigenfunctions with the same eigenvalue in every H^j and their ratios would produce non-constant invariant functions. By a combinatorial analysis of the approximating cocycles γ_n, measurable with respect to the partitions involved in the approximation of T_0, one can show that (i), (ii) and (iii) hold for a generic set of cocycles γ in the L_1 topology. Since the maximal spectral types in all H^j are identical the above estimate of the maximal spectral multiplicity by m implies that the spectra in those subspaces are simple and with maximal spectral type singular with respect to that in H^*. This implies that set of essential values of spectral multiplicity is $\{1, m\}$. □

For constructions with many values of spectral multiplicity the algebraic or "symmetry" part is more complicated but similar in principle. For infinite sets of values finite extensions are not sufficient and other compact group extensions are used. The most general case is represented by [100, Algebraic Lemma]. Approximation part has to be done differently though. The above estimate is not sufficient to conclude that all components in the spectrum which come from the algebraic construction are mutually singular and have maximal possible multiplicity. The solution is to consider approximation constructions directly for operators in invariant subspaces, to produce simple spectrum for those operators and guarantee mutual singularities of spectra.

5.8.5. *Finite extensions and spectral properties.* In [69] a construction was found which produced finite extensions of simple systems with certain functions with Lebesgue spectral measure. Based on this work Matthew and Nadkarni [112] have constructed a two points extension of an adding machine which they showed has a Lebesgue component of multiplicity 2. The Matthew–Nadkarni example involves a construction of a cocycle over the adding machine which takes values in $\mathbb{Z}/2\mathbb{Z}$ in such a way that the corresponding two point extension possesses a natural partition in two sets of equal measures whose iterates are pairwise independent. By replacing the adding machine in the base and modifying the construction appropriately Ageev [15] proved

THEOREM 5.75. *For any $n \geqslant 1$ there exists a weakly mixing transformation with essential values of spectral multiplicity $\{1, 2n\}$ where the component of multiplicity $2n$ is Lebesgue.*

The construction is also a finite extension, but this time, of a weakly mixing rank one transformation. See also [103] for examples with Lebesgue component of any given even multiplicity in the spectrum.

6. Key examples outside combinatorial constructions

6.1. *Introduction*

Of the four principal classes of systems which appear in smooth dynamics, two, hyperbolic and (typical) partially hyperbolic, are well understood from the point of view of ergodic theory. Modulo some sufficiently trivial modifications ergodic behavior of such systems with respect to an absolutely continuous invariant measure (as well as some other good invariant measures, such as maximal entropy or more general Gibbs measures) is described by the Bernoulli model which has countable Lebesgue spectrum and is classified up to a measurable isomorphism by the single invariant, entropy [3, Sections 2.3 and 3]. On the other hand, it is worth noticing that certain partially hyperbolic systems exhibit complicated and non-standard ergodic behavior. For example there are partially hyperbolic volume preserving diffeomorphisms which are K but not Bernoulli [77].

Elliptic systems admit in addition to the basic model of the toral translation a variety of behaviors which are well modeled by several kinds of combinatorial constructions discussed above.

The remaining class, parabolic systems, characterized by moderate and more or less uniform growth of orbit complexity do not naturally appear in the context of combinatorial constructions. In fact, it would be fair to say that many of the examples of the greatest intrinsic interest produced by combinatorial construction display phenomena which are difficult to render in the smooth situation.

In the next two sections we briefly review ergodic properties of two classes of parabolic systems which appear most naturally and are best understood. Key results concerning those systems are among the deepest in the field of ergodic theory and they yield remarkable applications outside the field, see [10]. In the last section we discuss another class of examples which came from probability theory and which provide a remarkably flexible and powerful tool for the spectral realization problem; in particular, the first example of a measure preserving transformation with simple continuous spectrum was found among Gaussian systems by Girsanov in 1958 [61] almost a decade earlier than direct methods based on rank and periodic approximation were developed.

6.2. *Unipotent homogeneous systems*

6.2.1. *Definitions and simple examples.* A homogeneous system has naturally defined *linear part* namely the adjoint action on the Lie algebra of G.

If all eigenvalues of the linear part of a homogeneous map are equal to one the map is called *unipotent*. A one-parameter group of unipotent maps is called a *unipotent flow*. If the linear part is semisimple, i.e. linearizable over complex numbers the flow acts by isometries with respect to a Riemannian metric and hence the spectrum is always pure point. Linear flows on the torus are examples; more generally this can happen on Euclidean manifolds (see Section 1.4b and Theorem 2.3.3 in [10]).

More interesting behavior appears when the linear part has non-trivial Jordan blocks. For example, mixture of pure point and countable Lebesgue spectrum appears in homogeneous

flows on nilpotent groups which in many respects are similar to unipotent affine maps on the torus like those in Examples 3.17 and 3.18.

6.2.2. *Horocycle flows and property R.* Horocycle flows which appeared in Section 2.1.3 are the simplest and best understood non-trivial examples among unipotent flows on homogeneous spaces of semisimple Lie groups.

We showed that they have countable Lebesgue spectrum which appears quite often in ergodic theory. However, beyond that horocycle flows possess very striking ergodic properties which imply strong rigidity statements. These properties are summarized in the following theorems due to M. Ratner [130]:

THEOREM 6.1. *If λ is an ergodic self-joining of a horocycle flow which is not the product measure, then it is a finite extension of its two marginals.*

REMARK. This statement is very close to simplicity. Simplicity is saying that $\mathcal{V} = \mathcal{H}$, here we have that $\mathcal{V}$ and $\mathcal{H}$ both have finite fibers in $\mathcal{V} \vee \mathcal{H}$.

THEOREM 6.2. *Horocycle flows have the pairwise independently determined property* (*see Definition* 4.15).

This implies mixing of all orders for the horocycle flows. As a consequence of these theorems, Ratner has obtained the following rigidity results:

THEOREM 6.3. *If two horocycle flows are measure theoretically isomorphic they are algebraically isomorphic.*

THEOREM 6.4. *Every factor of a horocycle flow is algebraic.*

THEOREM 6.5. *The time one transformation of every horocycle flow is a factor of a simple transformation. In case the subgroup γ is maximal and not arithmetic* [4, *Section* 1.5c], *the horocycle flow has minimal self-joinings as an $\mathbb{R}$ action.*

A key property for the understanding of the horocycle flow is the R property of Ratner which can be formulated in a general context.

Let T_t be a flow on a metric space with σ-compact metric d preserving a Borel measure.

DEFINITION 6.6. The flow T_t has the property R_p, $p \neq 0$ if the following is true:

For every $\varepsilon > 0$ and $N > 0$ there exist $\alpha(\varepsilon), \delta(\varepsilon, N) > 0$ and a subset $A(\varepsilon, N) \subset X$ such that $m(A) > 1 - \varepsilon$ with the property that if $x, y \in A$ and $d(x, y) < \delta(\varepsilon, N)$ and y is not on the T_t orbit of x, then there are $L = L(x, y)$ and $M = M(x, y) \geqslant N$ with $M/L \geqslant \alpha$ such that if

$$K^{+}_{-}(x, y) = \big\{n \in Z \cap [L, L + M]\colon d\big(T_{np}(x), T_{(n\pm1)p}(y)\big) < \varepsilon\big\},$$

then

$$|K^+|/M > 1-\varepsilon \quad \text{or} \quad |K^-|/M > 1-\varepsilon.$$

It is remarkable that this property of "slow relative drift of nearby points" is also satisfied by the Chacon transformation.

It is not known how far the R-property is from simplicity.

PROBLEM 6.7. Does there exist a flow satisfying the R-property such that its time one map is disjoint from all simple transformations?

6.2.3. *Ratner theory.* Recall that spectral properties of unipotent homogeneous systems are fairly standard: as for all homogeneous systems in general the mixture of pure point and countable Lebesgue spectrum. In the most interesting case of unipotent maps and flows on homogeneous spaces of semisimple Lie groups the spectrum is countable Lebesgue.

On the other hand, these systems exhibit very interesting ergodic properties beyond spectrum. For example, they provide examples of infinitely many systems with countable Lebesgue spectrum and zero entropy which are pairwise not Kakutani equivalent, namely different Cartesian powers of any horocycle flow [129]. The distinguishing invariant is of "slow entropy" type but adapted to the Kakutani rather than Hamming metrics in the spaces of sequences coding orbit segments; see [75] for the discussion of metrics and [83] for a general discussion of these invariants.

Isomorphisms, factors and joining between unipotent systems can be systematically studied with the powerful tool, the Ratner Measure Rigidity Theorem [132] which basically states that any invariant Borel probability ergodic measure is of algebraic nature. For a detailed exposition of Ratner theory and its applications see [10, Section 3].

It is worth noticing that while great attention has been paid to the number theoretical applications of Ratner's rigidity for unipotent systems there has been no systematic study of its implications to the ergodic theory of such actions, as has been done for the horocycle flows. Certainly it deserves to be looked at.

6.3. *Effects of time change in parabolic systems*

We will now complete the discussion of Section 5.6.3 of known spectral and other ergodic properties which appear under sufficiently nice time change in principal classes of systems.

6.3.1. *Time change in horocycle flows.* Let v be the vector field generating a horocycle flow. In [98] it is proved that if C^1 time change is not too large, namely if $f - \mathcal{L}f > 0$ where $\mathcal{L}$ is the derivative with respect to the geodesic flow then the flow generated by fv is mixing. The idea of the proof is of course to show that there is enough uniform twist across the orbits so that a small piece gets spread sufficiently uniformly across the space. However, the rate of mixing is not controlled well enough to guarantee absolutely continuous or Lebesgue spectrum. Still this looks plausible.

CONJECTURE 6.8. *Any flow obtained by a sufficiently smooth time change from a horocycle flow has countable Lebesgue spectrum.*

Cohomological equations over the horocycle flows has been throughly studied by Flaminio and Forni in [53]; see [78, Section 11.6.2] for a summary. While the results (growing number of invariant distributions of increasing orders) indicate complex structure of measurable isomorphism classes they do not shed direct light on spectral or other ergodic properties of time changes.

In an earlier work Ratner [131] shows that rigidity of isomorphisms between horocycle flows is partly inherited by time changes with very moderate degree of regularity in the sense that isomorphic time changes appear only for isomorphic horocycle flows. A key ingredient in the proof is showing property R for this class of time changes.

6.3.2. *Flows on surfaces of higher genus.* Another class of parabolic systems after unipotent homogeneous systems is represented by area preserving flows on surfaces with finitely many fixed points. In this case the section maps on transversals are one-dimensional, in fact they are interval exchange transformations. On the other hand, the slowdown near a fixed point leads to strong stretching which albeit not uniform in space is somewhat similar in effects with the uniform transverse stretching in unipotent systems.

A model example. The simplest example where it is evident that the slowdown and not transverse dynamics plays the main role in determining the asymptotic behavior is a flow on $\mathbb{T}^2$ obtained from an irrational linear flow by slowing down near a single point. In order to have a absolutely continuous measure preserved the inverse of the velocity change function must be integrable and the measure will still have a singularity. An alternative way is to change the flow in a neighborhood of a point so that in a local linear coordinate system (x, y) in which the linear flow is generated by the vector field $\partial/\partial x$ and hence is Hamiltonian with Hamiltonian function y to have the new flow with the Hamiltonian which locally has the form $y(x^2 + y^2)^k$ and gradually changes to y. One can make the change carefully so that the section map on a circle which still be a rotation and the flow will be isomorphic to the special flow with the roof function smooth except of one point near which it has an integrable singularity of a power type. In contrast with the case when the roof function has bounded variation such a flow is mixing [91]. The method is similar to that of [98] albeit the estimates are more subtle. Notice that unlike the latter case flows here the direction of stretching is different on two sides of the singularity.

Degenerate and non-degenerate saddles. A natural class of systems of this kind consists of area preserving flows on surfaces of genus $\geqslant 2$ with singularities of the saddle type. To include the previous example one may also allow a finite number of stopping points. The section map on a transversal is an interval exchange transformation and return time function has singularities at the endpoints of the intervals. There is an interesting difference between non-degenerate saddles (zeroes of the first order for the vector field) and other degenerate saddles which include stopping points (the latter can be considered as saddles with two separatrices). Non-degenerate saddles produce milder symmetric *logarithmic* singularities of the return time functions whereas others produce power singularities; in the

latter case if the flow is ergodic it is mixing [91]. This in particular implies the following existence result:

THEOREM 6.9. *There is an area preserving mixing flow of class C^∞ on any close surface other than the sphere, projective plane and Klein bottle.*

On the other hand, if the section map happens to be a rotation then any flow with only non-degenerate saddles is not mixing [92].

An interesting phenomenon appears when the singularities of the return time function are logarithmic but asymmetric; this still may produce mixing [87]. This situation appears, for example, on the torus for a flow with a separatrix loop.

Thus sufficiently strong stretching due to power or asymmetric logarithmic singularities of the return time function produces mixing while slightly weaker symmetric logarithmic singularities do not if the base transformation is a rotation (this can be explained from the point of view of Fourier analysis, see [105]). However mixing properties of typical flows on higher genus surfaces, namely flows with zeroes of order one, remain unknown.

PROBLEM 6.10. Does there exist a mixing special flow over an interval exchange transformation with the roof function smooth except for symmetric logarithmic singularities at the interval endpoints?

Also little is known about the spectral properties of mixing flows. Some estimate of correlation decay have been obtained but they are too weak to conclude that the spectrum is absolutely continuous. Nothing is also known about multiplicity of the spectrum.

Cohomological equations. Cohomological equations over interval exchange transformations and related classification of flows on surfaces have been studied by Forni in two very powerful papers [55,56]. Those results contain some of the deepest insights into interplay between ergodic theory and harmonic analysis. There are important applications to the speed of convergence of ergodic averages for various classes of functions. See [5] for an exposition of Forni's work.

However, as is the case with horocycle flows, there are no direct implications for spectral and other invariant under metric isomorphism ergodic properties of the flows.

6.4. *Gaussian and related systems*

6.4.1. *Spectral analysis of Gaussian systems.* For a detailed introduction to the subject see [29, Chapter 14].

Recall that from the "classical" ergodic point of view, given a measure preserving transformation T on a measure space (X, μ) and a measurable function f on X, the sequence $Y_n = f \circ T^n$, $n \in \mathbb{Z}$, defines a stationary stochastic process. A stochastic process can then be considered as a measure preserving transformation together with a measurable function f.

DEFINITION 6.11. A stationary process $X_n, n \in \mathbb{Z}$, with zero mean defined on a probability space $(\Omega, \mathcal{A}, P)$ is called *Gaussian* if for all $n \in \mathbb{Z}$, $m \in \mathbb{N}$ the law of the m-tuple $(X_n, X_{n+1}, \ldots, X_{n+m-1})$ is Gaussian (and independent of n). The shift transformation T_σ defined by $T(X_n)_{n\in\mathbb{Z}} = (X_{n+1})_{n\in\mathbb{Z}}$ is obviously measure preserving. It is often called the *Gaussian dynamical system* generated by the process X_n.

The spectral measure of the Koopman operator associated to T restricted to the closure of the space of linear combinations of X_n is called the *spectral measure of the Gaussian process.*

We will soon see how this measure determines the maximal spectral type of the corresponding Gaussian dynamical system. The covariance matrix of the stationary Gaussian process X_n, $n \in \mathbb{Z}$, $E(X_n X_{n+m})$ is entirely determined by the spectral measure σ:

$$E(X_n X_{n+m}) = \int_{S^1} e^{ixm}\, d\sigma.$$

Conversely, given a positive symmetric measure σ on the circle, there exists a stationary Gaussian process with zero mean X_n^σ, $n \in \mathbb{Z}$, with associated shift transformation T_σ such that

$$E(X_n X_{n+m}) = \int_{S^1} e^{ixm}\, d\sigma.$$

A way to construct X_n^σ is to first consider a probability space $(\Omega, \mathcal{A}, P)$ on which a family Z_n, $n \in \mathbb{Z}$, is defined, consisting of independent Gaussian random variables with law $N(0, 1)$ and with H being the L^2-closure of the linear span of the Z_n, $n \in \mathbb{Z}$. The Z_n are thus an orthonormal basis for H and every element in H is a random variable with zero mean and a Gaussian distribution law. Consider the operator U_σ on H which is isometric to the unitary operator M on $L^2(S^1, d\sigma)$ defined by $g \to e^{ix} g$ (as in Theorem 1.1), by means of an isometry V between H and $L^2(S^1, d\sigma)$. ($U_\sigma = V^{-1} M V$.) Then

$$X_n^\sigma = U_\sigma^n\big(V^{-1}1\big), \quad n \in \mathbb{Z},$$

is a Gaussian process which obviously satisfies

$$E\big(X_n^\sigma X_{n+m}^\sigma\big) = \int_{S^1} e^{ixm}\, d\sigma.$$

Let $\mathcal{B}(H)$ be the smallest σ-algebra which makes all elements in H measurable. Then $L^2(\mathcal{B}(H))$ is the direct sum of orthogonal spaces $H^{(n)}, n \in \mathbb{N}$ (the Wiener chaos) where $H^{(n)}$ is the orthocomplement of the direct sum of the $H^{(k)}$, $1 \leqslant k \leqslant n-1$, in the closure of the linear space generated by the polynomials of degree n in variables which are in H. These spaces are invariant under U_{T_σ} and the spectral measure of U_{T_σ} restricted to $H^{(n)}$ is the n-fold convolution $\sigma^{(n)}$. Thus we can calculate the maximal spectral type of the Gaussian system.

PROPOSITION 6.12. *The maximal spectral type of the Gaussian transformation T_σ is the sum of the spectral measure σ and all it convolutions $\sigma^{(n)}$.*

COROLLARY 6.13. *T_σ is ergodic only when σ is non-atomic, and in that case it is weakly mixing.*

If a symmetric measure σ on S^1 is the sum of two symmetric measures σ_1 and σ_2 which are mutually singular, T_σ is isomorphic to the direct product $T_{\sigma_1} \times T_{\sigma_2}$. Therefore, decomposing σ as the sum of its singular part σ_s and of its absolutely continuous part σ_a, we see that, since a Gaussian process with singular spectral measure has 0 entropy, T_σ is isomorphic to a factor of the product of a zero entropy transformation by an infinite entropy Bernoulli shift, and is itself of this form, as an application of general theorems. If in the preceding construction, we consider the more general situation where the operator U on H has no longer simple spectrum, we still obtain a transformation, which is no longer described by a single Gaussian process, which we call *generalized Gaussian*. Generalized Gaussian processes share many properties with ordinary Gaussian processes.

A version of the generalized Gaussian construction for more general groups provides a general way to construct many spectrally (and hence metrically) non-isomorphic actions by measure preserving transformations [4, Section 4.4]. For such groups as semisimple Lie groups of rank $\geqslant 2$ and lattices in such groups whose actions possess strong rigidity properties which render many standard constructions trivial this is the only known way to produce many non-isomorphic actions.

6.4.2. *Spectral multiplicity for Gaussian systems.* In order for U_{T_σ} to have simple spectrum it is necessary for all $\sigma^{(n)}$ to be pairwise singular.

On the other hand, the spectrum is simple if there is a set K such that (i) $K \cup -K$ has full σ-measure, and (ii) all its elements are independent over the rationals, that is if

$$\lambda_1, \ldots, \lambda_n \in K, \quad \text{and} \quad (m_1, \ldots, m_n) \in \mathbb{Z}^n \setminus \{0\}, \quad \text{then} \quad m_1\lambda_1 + \cdots + m_n\lambda_n \neq 0. \tag{6.1}$$

We use here additive coordinate on the circle $S^1 = \mathbb{R}/\mathbb{Z}$. The first proof of existence of a measure preserving transformation with a simple but not pure point spectrum was given by Girsanov in [61] using the Gaussian system of this kind. A stronger condition which implies (6.1) is the following:

($\mathfrak{K}$) *Every continuous function on the set K of modulus* 1 *is a uniform limit of characters*.

A closed set satisfying condition ($\mathfrak{K}$) is called a *Kronecker set*. D. Newton [116] first used Kronecker sets to construct Gaussian systems with simple spectrum. His examples were Gaussian systems with spectral measures supported by the union of a Kronecker set K and $-K$. Let us call such a measure *Kronecker*. Also using a construction of a mixing measure suggested by Rudin [139] Newton found a mixing Gaussian transformation with simple spectrum.

PROPOSITION 6.14. *For Gaussian systems the multiplicity function is multiplicative almost everywhere with respect to a measure of the maximal spectral type.*

COROLLARY 6.15. *Either the spectrum of a Gaussian transformation is simple or the maximal spectral multiplicity is unbounded.*

PROPOSITION 6.16. *There exists σ such that T_σ has non-simple spectrum and for which the multiplicity function is finite almost everywhere.*

Corollary 6.15 and Proposition 6.16 explain why finding systems with non-simple spectrum of bounded multiplicity was considered an interesting problem when Gaussian systems and their modifications provided the only models with interesting spectral properties. After the initial success in the study of Gaussian systems there was a hope to organize a good part of ergodic theory around a generalized version of the Gaussian model reflected in [147]. One of the original impulses which led to the development of the theory or periodic approximations and similar geometric methods came from attempts to understand how restrictive were the assumptions on which this approach was based. The answer on the occasion was that they almost never held in natural geometric situations.

6.4.3. *Spectrally defined isomorphism in Gaussian and similar systems.* Foias and Stratila in [54] showed that Newton's examples have a remarkable property which makes them similar to transformations with pure point spectrum, in fact like translations on continuum-dimensional tori.

THEOREM 6.17. *Let σ be a Kronecker measure. Then if $(X, \mathcal{A}, m, T)$ is ergodic and if $f \in L^2(X)$ satisfies $\nu_f = \sigma$, the process $T^n f$, $n \in \mathbb{Z}$, is Gaussian.*

One important consequence of this theorem is the following [150].

THEOREM 6.18. *Let σ be a Kronecker measure and T_σ the associated Gaussian transformation. All ergodic self-joinings of T_σ remain generalized Gaussian.*

One can prove that the conclusion of this theorem holds for measures σ such that the associated Gaussian T_σ has simple spectrum. Those processes such that all their ergodic joinings remain generalized Gaussian are called GAG and are the subject of a comprehensive study in [107]. They can be thought of as a limit of a product of pairwise disjoint simple transformations. Let us say that σ for which the conclusion of Theorem 6.17 holds has the *F.S.-property*. There are examples in [107] where measures satisfying the F.S.-property have as support S the union of two disjoint Kronecker sets without S itself being Kronecker. An interesting question is the following:

PROBLEM 6.19. Does there exist a mixing measure which possesses the F.S.-property?

F. Parreau (unpublished) has produced a mildly mixing measure with the F.S.-property. Notice that Kronecker systems are rigid. This is a direct consequence of the property of Kronecker sets that every continuous function is a uniform limit of characters. The rigidity is just this statement applied to the constant function 1. We are now going to show that, with

the use of Gaussian processes, it is easy to produce two transformations which are weakly isomorphic but not isomorphic. Given a Gaussian process X_n^σ to which we associate the shift transformation T_σ (acting on $(X, \mathcal{A}, m)$), if we let H be the L^2-closure of the linear span of X_n^σ, $n \in \mathbb{Z}$, we have seen that every unitary operator U on H gives rise to a measure preserving transformation τ_U (the one coming from the Gaussian processes associated to U when H is decomposed into an orthogonal sum of U-cyclic subspaces). If we take

$$UX = -X,$$

the map τ_U is an involution which commutes with T_σ; the σ-algebra $\mathcal{B}$ of τ_U-invariant sets defines a factor (which we call $\widehat{T_\sigma}$) of T_σ and

$$L^2(\mathcal{B}) = \sum_{n \geqslant 0} H^{(2n)}.$$

Such a factor was first defined by Newton and Parry. We take σ such that σ is continuous and $\sigma^{(n)} \perp \sigma^{(m)}$, $n \neq m$. We define

$$T_1 = \prod_{k \in \mathbb{N}} T_{k,\sigma},$$

where every $T_{k,\sigma}$, $k \in \mathbb{N}$, is a copy of T_σ and

$$T_2 = \widehat{T_\sigma} \times T_1.$$

It is therefore obvious that T_1 and T_2 are weakly isomorphic.

THEOREM 6.20. *T_1 and T_2 are weakly isomorphic but not isomorphic.*

PROOF. In L^2 of the space on which T_1 lives, U_{T_1} (the unitary operator associated to T_1) has spectral measure σ on

$$\sum_{k \in \mathbb{N}} \oplus H_k = \overline{H}.$$

The spectral measure of U_{T_1} on $\overline{H}^\perp$ is singular with respect to σ (because of the hypothesis on σ). Again, because

$$\sigma^{(2n)} \perp \sigma,$$

the spectral measure of U_{T_2} on $\overline{H}_2^\perp$ is singular with respect to σ. ($\overline{H}_2$ is $\sum_{k \in \mathbb{N}} \oplus H_k$ in the space on which T_2 lives.) Therefore, if T_1 and T_2 are isomorphic then the associated isometry must send $\overline{H}$ onto $\overline{H}_2$, which is impossible. □

Note that it is also very easy to construct weakly isomorphic but not isomorphic transformations from a transformation which has MSJ (and therefore from Chacon transformations). This was done by D. Rudolph before the Gaussian example described above. The first example of two weakly isomorphic but not isomorphic transformations is due to S. Polit. Kwiatkowski, Lemańczyk and Rudolph [101] have constructed an example of two smooth dynamical systems which are weakly isomorphic but not isomorphic. The following result by De la Rue shows that Girsanov examples and more general transformations with simple spectrum coming from the Gaussian construction are quite different from transformations with simple spectrum constructed by more geometric methods in earlier parts of this survey.

THEOREM 6.21 [37]. *A Gaussian transformation cannot be locally rank one.*

Another result in a similar vein was proved by del Junco and Lemańczyk [31] who extended an earlier result by Thouvenot [151].

THEOREM 6.22. *Gaussian transformations are disjoint from simple transformations.*

One more striking property of Kronecker Gaussian systems is the De la Rue result [36] mentioned before that there are maximal spectral types which appear for zero entropy ergodic transformations but not for standard ones (Kakutani equivalent to adding machines and irrational rotations). This follows form the Foias–Stratila theorem and the following fact proved by De la Rue.

THEOREM 6.23. *There exists a Kronecker measure such that the corresponding Gaussian transformation is not standard.*

An interesting open problem tying together the themes of this section and Section 5.5 is the following:

PROBLEM 6.24. Does there exist a volume preserving diffeomorphism of a compact differentiable manifold which is measurably conjugate to a Gaussian system?

More specifically, given a non-trivial volume preserving smooth action $\mathcal{S}$ of S^1 on a compact differentiable manifold M, does there exist a diffeomorphism measurably conjugate to a Kronecker Gaussian system in the space $\mathcal{A}$, the closure of conjugates of the elements of $\mathcal{S}$ (see Section 5.5.2)?

Acknowledgements

We would like to thank the referee for a number of valuable suggestions and critical comments and for extremely careful reading of the draft of the survey.

Anatole Katok is partially supported by NSF Grant DMS 0071339.

References

Surveys in volume 1A and this volume

[1] L. Barreira and Ya. Pesin, *Smooth ergodic theory and non-uniformly hyperbolic dynamics*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 57–263.

[2] V. Bergelson, *Combinatorial and Diophantine applications of ergodic theory*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 745–869.

[3] N. Chernov, *Invariant measures for hyperbolic dynamical systems*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 321–407.

[4] R. Feres and A. Katok, *Ergodic theory and dynamics of G-spaces*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 665–763.

[5] G. Forni, *On the Lyapunov exponents of the Kontsvich–Zorich cocycle*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 549–580.

[6] E. Glasner and B. Weiss, *On the interplay between measurable and topological dynamics*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 597–648.

[7] B. Hasselblatt, *Hyperbolic dynamical systems*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 239–319.

[8] B. Hasselblatt and A. Katok, *Principal structures*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 1–203.

[9] B. Hasselblatt and Ya. Pesin, *Partially hyperbolic dynamical systems*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 1–55.

[10] D. Kleinbock, N. Shah and A. Starkov, *Dynamics of subgroup actions on homogeneous spaces of Lie groups and applications to number theory*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 813–930.

[11] H. Masur and S. Tabachnikov, *Rational billiards and flat structures*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 1015–1089.

[12] J.-P. Thouvenot, *Entropy, isomorphism and equivalence in ergodic theory*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2002), 205–238.

Other sources

[13] L.M. Abramov, *Metric automorphisms with quasi-discrete spectrum*, Izvestia Akad. Nauk SSSR **26** (1962), 513–550; Amer. Math. Soc. Transl. **39** (1964), 37–56.

[14] T. Adams, *Smorodinsky's conjecture on rank one systems*, Proc. Amer. Math. Soc. **126** (3) (1998), 739–744.

[15] O.N. Ageev, *Dynamical systems with a Lebesgue component of even multiplicity in their spectrum*, Mat. Sbornik **64** (1989), 305–317.

[16] O.N. Ageev, *On ergodic transformations with homogeneous spectrum*, J. Dynamical Control Systems **5** (1999), 149–152.

[17] O.N. Ageev, *On a homogeneous spectrum in ergodic theory*, Invent. Math. **160** (2005), 417–446.

[18] D.V. Anosov and A.B. Katok, *New examples in smooth ergodic theory. Ergodic diffeomorphisms*, Trans. Moscow Math. Soc. **23** (1970), 1–35.

[19] H. Anzai, Ergodic skew product transformations on the torus, *Osaka J. Math.* **3** (1951), 83–99.

[20] A. Avila and G. Forni, *Weak mixing for interval exchange transformations and translations flows*, Preprint (2004).

[21] P. Billingsley, *Ergodic Theory and Information*, Wiley Inc. (1966).

[22] M. Boshernitzan, *A criterion for interval exchange maps to be uniquely ergodic*, Duke Math. J. **52** (1985), 723–752.

[23] J. Bourgain, *Pointwise ergodic theorems for arithmetic sets*, Inst. Hautes Etudes Sci. Publ. Math. **69** (1989), 5–45.

[24] J. Bourgain, *On the spectral type of Ornstein's class one transformation*, Israel J. Math. **84** (1993), 53–63.
[25] J. Brezin and C.C. Moore, *Flows on homogeneous spaces: A new look*, Amer. J. Math. **103** (1981), 571–613.
[26] R.V. Chacon, *Weakly mixing transformations which are not strongly mixing*, Proc. Amer. Math. Soc. **22** (1969), 559–562.
[27] J.-P. Conze, *Equations fonctionnelles et systemes induits en theorie ergodique*, Z. Wahrsch. Verw. Gebiete **23** (1972), 75–82 (French).
[28] J.-P. Conze, *Entropie d'un groupe abelien de transformations*, Z. Wahrsch. Verw. Gebiete **25** (1972), 11–30 (French).
[29] I.P. Cornfeld, Ya.G. Sinai and S.V. Fomin, *Ergodic Theory*, Grundlehren Math. Wiss., Vol. 245, Springer, New York (1982).
[30] R. de la Llave, *A tutorial in KAM theory*, Smooth Ergodic Theory and its Applications, Proc. Sympos. Pure Math., Vol. 69, Amer. Math. Soc. (2001), 175–293.
[31] A. del Junco and M. Lemańczyk, *Simple systems are disjoint from Gaussian systems*, Studia Math. **133** (1999), 249–256.
[32] A. del Junco and M. Lemańczyk, *Generic spectral properties of measure preserving maps and applications*, Proc. Amer. Math. Soc. **115** (3) (1992), 725–736.
[33] A. del Junco, A. Rahe and L. Swanson, *Chacon's automorphism has minimal self-joinings*, J. Anal. Math. **37** (1980), 276–284.
[34] A. del Junco and D.J. Rudolph, *On ergodic actions whose self-joinings are graphs*, Ergodic Theory Dynamical Systems **7** (1987), 531–557.
[35] T. De la Rue, *L'ergodicité induit un type spectral maximal équivalent à la mesure de Lebesgue*, Ann. Inst. H. Poincaré **34** (1998), 249–263 (French).
[36] T. De la Rue, *L'induction ne donne pas toutes les mesures spectrales*, Ergodic Theory Dynamical Systems **18** (1998), 1447–1466 (French).
[37] T. De la Rue, *Rang des systemés dynamique gaussiens*, Israel J. Math. **104** (1998), 261–283 (French).
[38] J. Diximier, *Les C^*-algèbres et leurs représentations*, Cahiers Scientifiques, Fasc. XXIX, Gauthier-Villars, Paris (1964) (French).
[39] A.H. Dooley and V.Ya. Golodets, *The spectrum of completely positive entropy actions of countable amenable groups*, J. Funct. Anal. **196** (1) (2002), 1–18.
[40] E.H. El Abdalaoui, *La singularité mutuelle presque sure du spectre des transformations d'Ornstein*, Israel J. Math., to appear.
[41] A. Fathi and M.R. Herman, *Existence de difféomorphismes minimaux*, Dynamical Systems, Vol. I, Warsaw, 37–59; Astérisque, No. 49, Soc. Math. France, Paris (1977) (French).
[42] B. Fayad, *Weak mixing for reparameterized linear flows on the torus*, Ergodic Theory Dynamical Systems **22** (1) (2002), 187–201.
[43] B. Fayad, *Analytic mixing reparametrizations of irrational flows*, Ergodic Theory Dynamical Systems **22** (2) (2002), 437–468.
[44] B. Fayad, *Skew products over translations on T^d, $d \geqslant 2$*, Proc. Amer. Math. Soc. **130** (1) (2002), 103–109.
[45] B. Fayad, *Rank one and mixing differentiable flows*, Invent. Math. **160** (2005), 305–340.
[46] B. Fayad, *Smooth singular flows with purely singular spectra*, Preprint (2004).
[47] B. Fayad and A. Katok, *Constructions in elliptic dynamics*, Ergodic Theory Dynamical Systems, Herman memorial issue **24** (2004), 1477–1520.
[48] B. Fayad, A. Katok and A. Windsor, *Mixed spectrum reparametrizations of linear flows on T^2*, Moscow Math. J. **1** (2001), 521–537.
[49] B. Fayad, M. Saprykina and A. Windsor, *Nonstandard smooth realization of Liouvillean rotations*, in preparation.
[50] B. Fayad and A. Windsor, *A dichotomy between discrete and continuous spectrum for a class of special flows over rotations*, Preprint (2003).
[51] J. Feldman, *New K-automorphisms and a problem of Kakutani*, Israel J. Math. **24** (1976), 16–38.
[52] S. Ferenczi, *Systémes de rang un gauche*, Ann. Inst. H. Poincaré **21** (1985), 177–186 (French).
[53] L. Flaminio and G. Forni, *Invariant distributions and time averages for horocycle flows*, Duke Math. J. **119** (3) (2003), 465–526.

[54] C. Foias and S. Stratila, *Ensembles de Kronecker dans la théorie ergodique*, C. R. Acad. Sci. Paris Sér. A-B **267** (1968), A166–A168 (French).
[55] G. Forni, *Solutions of the cohomological equation for area-preserving flows on compact surfaces of higher genus*, Ann. of Math. (2) **146** (2) (1997), 295–344.
[56] G. Forni, *Deviation of ergodic averages for area-preserving flows on surfaces of higher genus*, Ann. of Math. (2) **155** (1) (2002), 1–103.
[57] C. Fraczek and M. Lemańczyk, *On disjointness properties of some smooth flows*, Preprint (2004).
[58] N.A. Friedman and D.S. Ornstein, *Ergodic transformations induce mixing transformations*, Adv. Math. **10** (1973), 147–163.
[59] H. Furstenberg, *Disjointness in ergodic theory, minimal sets and a problem in Diophantine approximation*, Math. Systems Theory **1** (1967), 1–49.
[60] H. Furstenberg, *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Anal. Math. **31** (1977), 204–256.
[61] I.V. Girsanov, *On spectra of dynamical systems generated by Gaussian stationary processes*, Dokl. Acad. Sci. USSR **119** (1958), 851–853 (Russian).
[62] E. Glasner, *Ergodic Theory via Joinings*, Math. Surveys Monographs, Vol. 10, Amer. Math. Soc., Providence, RI (2003).
[63] E. Glasner, B. Host and D. Rudolph, *Simple systems and their higher order self-joinings*, Israel J. Math. **78** (1992), 131–142.
[64] G.R. Goodson, *A survey of recent results in the spectral theory of ergodic dynamical systems*, J. Dynamical Control Systems **5** (1999), 173–226.
[65] G.R. Goodson, J. Kwiatkowski, M. Lemańczyk and P. Liardet, *On the multiplicity function of ergodic group extensions of rotations*, Studia Math. **102** (1992), 157–174.
[66] G.R. Goodson and M. Lemańczyk, *Transformations conjugate to their inverses have even essential values*, Proc. Amer. Math. Soc. **124** (1996), 2703–2710.
[67] R. Gunesch and A. Katok, *Construction of weakly mixing diffeomorphisms preserving measurable Riemannian metric and smooth measure*. With an appendix by Alex Furman, Discrete Contin. Dynam. Systems **6** (1) (2000), 61–88.
[68] P.R. Halmos, *Lectures on Ergodic Theory*, Chelsea Publishing Company, New York (1956).
[69] H. Helson and W. Parry, *Cocycles and spectra*, Ark. Mat. **16** (1978), 195–206.
[70] M. Herman, *On the dynamics of Lagrangian tori invariant by symplectic diffeomorphisms*, Progress in Variational Methods in Hamiltonian Systems and Elliptic Equations (L'Aquila, 1990), Pitman Res. Notes Math. Ser., Vol. 243, Longman Sci. Tech., Harlow (1992), 92–112.
[71] C. Hoffman, *A K counterexample machine*, Trans. Amer. Math. Soc. **351** (10) (1999), 4263–4280.
[72] B. Host, *Mixing of all orders and pairwise independent joinings of systems with singular spectrum*, Israel J. Math. **76** (1991), 289–298.
[73] S. Kalikow, *Two fold mixing implies three fold mixing for rank one transformations*, Ergodic Theory Dynamical Systems **4** (1984), 237–259.
[74] T. Kamae, Spectral properties of automata generating sequences, unpublished.
[75] A. Katok, *Monotone equivalence in ergodic theory*, Math. USSR-Izv. **10** (1977), 99–146.
[76] A. Katok, *Interval exchange transformations and some special flows are not mixing*, Israel J. Math. **35** (1980), 301–310.
[77] A. Katok, *Smooth non-Bernoulli K-automorphisms*, Invent. Math. **61** (1980), 291–300.
[78] A. Katok, *Combinatorial Constructions in Ergodic Theory and Dynamics*, University Lecture Series, Vol. 30, Amer. Math. Soc. (2003).
[79] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press (1995).
[80] A.B. Katok and E.A. Sataev, *Standardness of automorphisms of transpositions of intervals and fluxes on surfaces*, Math. Notes Acad. Sci. USSR **20** (1977), 826–831.
[81] A.B. Katok and A.M. Stepin, *Approximations in ergodic theory*, Russian Math. Surveys **22** (1967), 77–102.
[82] A.B. Katok and A.M. Stepin, *Metric properties of measure preserving homeomorphisms*, Russian Math. Surveys **25** (1970), 191–220.

[83] A. Katok and J.-P. Thouvenot, *Slow entropy type invariants and smooth realization of commuting measure preserving transformations*, Ann. Inst. H. Poincaré Probab. Statist. **33** (3) (1997), 323–338.
[84] Y. Katznelson and B. Weiss, *Commuting measure preserving transformations*, Israel J. Math. **12** (1972), 16–23.
[85] M. Keane, *Interval exchange transformations*, Math. Z. **141** (1975), 25–31.
[86] A. Kechris, *Classical Descriptive Set Theory*, Graduate Texts in Math., Vol. 156, Springer, New York (1995).
[87] K.M. Khanin and Ya.G. Sinai, *Mixing of some classes of special flows over rotations of the circle*, Funct. Anal. Appl. **26** (3) (1992), 155–169.
[88] J. King, *Joining rank and the structure of finite rank mixing transformations*, J. Anal. Math. **51** (1988), 182–227.
[89] I. Klemes, *The spectral type of the staircase transformation*, Tohôku Math. J. **48** (1996), 247–248.
[90] A.V. Kočergin, *Change of time in flows, and mixing*, Math. USSR-Izv. **7** (1973), 1273–1294.
[91] A.V. Kočergin, *Mixing in special flows over a rearrangement of segments and in smooth flows on surfaces*, Math. USSR-Sb. **25** (1975), 441–469.
[92] A.V. Kočergin, *Nondegenerate saddles and the absence of mixing*, Math. Notes USSR Acad. Sci. **19** (1976), 277–286.
[93] A.V. Kochergin, *A mixing special flow over a rotation of the circle with an almost Lipschitz function*, Sb. Math. **193** (3–4) (2002), 359–385.
[94] A.N. Kolmogorov, *On dynamical systems with integral invariant on the torus*, Dokl. Akad. Nauk SSSR **93** (1953), 763–766 (Russian).
[95] B.O. Koopman, *Hamiltonian systems and transformations in Hilbert space*, Proc. Nat. Acad. Sci. U.S.A. **17** (1931), 315–318.
[96] W. Krieger, *On entropy and generators of measure preserving transformations*, Trans. Amer. Math. Soc. **119** (1970), 89–119.
[97] A.G. Kuschnirenko, *On metric invariants of entropy type*, Russian Math. Surveys **22** (5) (1967), 53–61.
[98] A.G. Kuschnirenko, *Spectral properties of some dynamical systems with polynomial divergence of orbits*, Moscow Univ. Math. Bull. **29** (1974), 82–87.
[99] J. Kwiatkowski and Y. Lacroix, *Multiplicity, rank pairs*, J. Anal. Math. **71** (1997), 205–235.
[100] J. Kwiatkowski and M. Lemańczyk, *On the multiplicity function of ergodic group extensions. II*, Studia Math. **116** (1995), 207–215.
[101] J. Kwiatkowski, M. Lemańczyk and D. Rudolph, *Weak isomorphism of measure preserving diffeomorphisms*, Israel J. Math. **80** (1992), 33–64.
[102] F. Ledrappier, *Des prodits de Riesz comme measures spectrale*, Ann. Inst. H. Poincaré **6** (1970), 335–344 (French).
[103] M. Lemańczyk, *Toeplitz Z_2-extensions*, Ann. Inst. H. Poincaré **24** (1988), 1–43.
[104] M. Lemańczyk, *Introduction to ergodic theory from the point of view of spectral theory*, Lecture Notes on the Tenth Kaisk Mathematics Workshop, Geon Ho Choe, ed., Korea Advanced Institute of Science and Technology, Math. Res. Center, Taejon, Korea (1995).
[105] M. Lemańczyk, *Sur l'absence de melange pour le flot special au dessus d'une rotation irrationnelle*, Colloq. Math. **84/85** (1) (2000), 29–41 (French).
[106] M. Lemańczyk and M.K. Mentzen, *Compact subgroups in the centralizers of natural factors of an ergodic group extension of a rotation determine all factors*, Ergodic Theory Dynamical Systems **10** (1990), 763–776.
[107] M. Lemańczyk, F. Parreau and J.-P. Thouvenot, *Gaussian automorphisms whose self joinings are Gaussian*, Fund. Math. **164** (2000), 253–293.
[108] M. Lemańczyk and A. Sikorski, *A class of not local rank one automorphisms arising from continuous substitutions*, Probab. Theory Related Fields **76** (4) (1987), 421–428.
[109] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, Cambridge (1995).
[110] S.A. Malkin, *An example of two metrically non-isomorphic ergodic automorphisms with identical simple spectra*, Izv. Vyssh. Uchebn. Zaved. Mat. **6** (1968), 69–74 (Russian).
[111] H. Masur, *Interval exchange transformations and measured foliations*, Ann. Math. **115** (1982), 169–200.

[112] J. Mathew and M.G. Nadkarni, *A measure preserving transformation whose spectrum has a Lebesgue component of multiplicity two*, Bull. London Math. Soc. **16** (1984), 402–406.
[113] J. Moser, *On the volume elements on a manifold*, Trans. Amer. Math. Soc. **120** (1965), 286–294.
[114] M.G. Nadkarni, *Spectral Theory of Dynamical Systems*, Birkhäuser, Berlin (1998).
[115] M.A. Naimark, *Normed Algebras*, Translated from the second Russian edition, Wolters-Noordhoff Series of Monographs and Textbooks on Pure and Applied Mathematics, Wolters-Noordhoff Publishing, Groningen (1972).
[116] D. Newton, *On Gaussian processes with simple spectrum*, Z. Wahrsch. Verw. Gebiete **5** (1966), 207–209.
[117] D.S. Ornstein, *On the root problem in ergodic theory*, Proc. 6th Berkeley Symposium on Math. Statist. Probab., University of California Press, Berkeley (1970), 348–356.
[118] D.S. Ornstein, D.J. Rudolph and B. Weiss, *Equivalence of measure preserving transformations*, Mem. Amer. Math. Soc. **37** (262) (1982).
[119] D.S. Ornstein and P.S. Shields, *An uncountable family of K-automorphisms*, Adv. Math. **10** (1973), 63–88.
[120] D.S. Ornstein and M. Smorodinsky, *Ergodic flows of positive entropy can be time changed to become K-flows*, Israel J. Math. **26** (1) (1977), 75–83.
[121] D.S. Ornstein and B. Weiss, *Ergodic theory of amenable group actions. I. The Rohlin lemma*, Bull. Amer. Math. Soc. **2** (1980), 161–164.
[122] D.S. Ornstein and B. Weiss, *Entropy and isomorphism theorems for actions of amenable groups*, J. Anal. Math. **48** (1987), 1–141.
[123] V.I. Oseledec, *On the spectrum of ergodic automorphisms*, Soviet Math. Dokl. **7** (1966), 776–779.
[124] W. Parry, *Topics in Ergodic Theory*, Cambridge University Press, Cambridge (1981).
[125] J. Peyriére, *Étude de quelques propriétés de produits de Riesz*, Ann. Inst. Fourier (2) **25** (1975), 127–196 (French).
[126] L.S. Pontryagin, *Topological Groups*, Translated from the second Russian edition, Gordon and Breach Science Publishers, Inc., New York (1966).
[127] A.A. Prikhod'ko and V.V. Ryzhikov, *Disjointness of the convolutions for Chacon's automorphism*, Colloq. Math. **84/85** (1) (2000), 67–74.
[128] M. Queffelec, *Substitution Dynamical Systems: Spectral Analysis*, Lecture Notes in Math., Vol. 1294, Springer (1987).
[129] M. Ratner, *Some invariants of Kakutani equivalence*, Israel J. Math. **38** (3) (1981), 231–240.
[130] M. Ratner, *Horocycle flows, joinings and rigidity of products*, Ann. Math. **118** (1983), 277–313.
[131] M. Ratner, *Rigidity of time changes for horocycle flows*, Acta Math. **156** (1986), 1–32.
[132] M. Ratner, *On Raghunathan mesure conjecture*, Ann. Math. **134** (1991), 545–607.
[133] G. Rauzy, *Echanges d'intervalles et transformations induites*, Acta Arith. **34** (1979), 315–328 (French).
[134] E.A. Robinson, *Ergodic measure preserving transformations with arbitrary finite spectral multiplicities*, Invent. Math. **72** (1983), 299–314.
[135] E.A. Robinson, *Transformations with highly non-homogeneous spectrum of finite multiplicity*, Israel J. Math. **56** (1986), 75–88.
[136] V.A. Rokhlin, *Unitary rings*, Dokl. Akad. Nauk SSSR **59** (1948), 643–646 (Russian).
[137] V.A. Rokhlin, *A "general" measure preserving transformation is not mixing*, Dokl. Akad. Nauk SSSR **60** (1948), 349–351 (Russian).
[138] A. Rothstein, *Veršik processes: First steps*, Israel J. Math. **36** (3–4) (1980), 205–224.
[139] W. Rudin, *Fourier–Stieltjes transforms of measures on independent sets*, Bull. Amer. Math. Soc. **66** (1960).
[140] D.J. Rudolph, *An example of a measure preserving transformation with minimal self-joinings and applications*, J. Anal. Math. **35** (1979), 97–122.
[141] D.J. Rudolph, *Fundamentals of Measurable Dynamics*, Oxford University Press, Oxford (1990).
[142] V.V. Ryzhikov, *Joinings and multiple mixing of finite rank actions*, Funktsional. Anal. i Prilozhen. **26** (1993), 63–78.
[143] V.V. Ryzhikov, *Joinings, intertwining operators, factors, and mixing properties of dynamical systems*, Russian Acad. Sci. Izv. Math. **42** (1994), 91–114.
[144] V.V. Ryzhikov, *Transformations having homogeneous spectra*, J. Dynamical Control Systems **5** (1999), 145–148.
[145] E.A. Satayev, *On the number of invariant measures for flows on orientable surfaces*, Math. USSR-Izv. **9** (1975), 813–830.

[146] M. Shklover, *Classical dynamical systems on the torus with continuous spectrum*, Izv. Vyssh. Uchebn. Zaved. Mat. **10** (65) (1967), 113–124 (Russian).
[147] Ya. Sinai, *Some remarks on spectral properties of ergodic dynamical systems*, Uspekhi Math. Nauk **18** (5) (1963), 41–54 (Russian).
[148] A.M. Stepin, *Les spectres des systemes dynamique*, Actes, Congress Intern. Math. (1970), 941–946 (French).
[149] A.M. Stepin, *Spectral properties of generic dynamical systems*, Math. USSR-Izv. **29** (1987), 159–282.
[150] J.-P. Thouvenot, *The metrical structure of some Gaussian processes*, Proceedings of the Conference: Ergodic Theory and Related Topics II (Georgenthal, 1986), Teubner-Texte Math., Vol. 94, Leipzig (1987), 195–198.
[151] J.-P. Thouvenot, *Some properties and applications of joinings in ergodic theory*, Proceedings of the 1993 Alexandria Conference, Ergodic Theory and its Connections with Harmonic Analysis, Cambridge University Press (1995), 207–238.
[152] W. Veech, *A criterion for a process to be prime*, Monatsh. Math. **94** (1982), 335–341.
[153] W. Veech, *Gauss measures for transformations on the space of interval exchange maps*, Ann. Math. **115** (1982), 201–242.
[154] W. Veech, *The metric theory of interval exchange transformations. I, II, III*, Amer. J. Math. **106** (1984), 1331–1421.
[155] J. von Neumann, *Zur operatorenmethode in der klassichen mechanik*, Ann. Math. **33** (1932), 587–642.
[156] P. Walters, *An Introduction to Ergodic Theory*, Graduate Texts in Math., Vol. 79, Springer, New York (1982).
[157] A. Windsor, *Minimal but not uniquely ergodic diffeomorphisms*, Smooth Ergodic Theory and its Applications (Seattle, WA, 1999), Proc. Sympos. Pure Math., Vol. 69, Amer. Math. Soc., Providence, RI (2001) 809–824.

This page intentionally left blank

CHAPTER 12

Combinatorial and Diophantine Applications of Ergodic Theory

Vitaly Bergelson[1]
The Ohio State University, Columbus, OH 43210, USA
E-mail: *vitaly@math.ohio-state.edu*

With Appendix A by A. Leibman[2]
The Ohio State University, Columbus, OH 43210, USA
E-mail: *leibman@math.ohio-state.edu*

With Appendix B by Anthony Quas[3] and Máté Wierdl[4]
Department of Mathematical Sciences, University of Memphis, 373 Dunn Hall, Memphis, TN 38152-3240, USA
E-mail: *quasa@msci.memphis.edu, mw@csi.hu*

Contents

[1]The author acknowledges support received from the National Science Foundation (USA) via grant DMS-0345350.
[2]Supported by NSF, grant DMS-0345350.
[3]A. Quas' research is partially supported by NSF grant #DMS-0200703.
[4]M. Wierdl's research is partially supported by NSF grant #DMS-0100577.

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

1. Introduction

The main focus of this survey is the mutually perpetuating interplay between ergodic theory, combinatorics and Diophantine analysis.

Ergodic theory has its roots in statistical and celestial mechanics. In studying the long time behavior of dynamical systems, ergodic theory deals first of all with such phenomena as recurrence and uniform distribution of orbits.

Ramsey theory, a branch of combinatorics, is concerned with the phenomenon of preservation of highly organized structures under finite partitions.

Diophantine analysis concerns itself with integer and rational solutions of systems of polynomial equations.

To get a feeling about possible connections between these three quite distinct areas of mathematics, let us consider some examples.

1.1. *Fermat's theorem over finite fields*

Our first example is related to Fermat's last theorem. Given $n \in \mathbb{N}$, where $\mathbb{N}$, here and throughout this survey, represents the set of positive integers, and a prime p, consider the equation $x^n + y^n \equiv z^n \pmod p$. This equation (as well as its more general version $ax^n + by^n + cz^n \equiv 0 \pmod p$) was extensively studied in the 19th and early 20th centuries. (See [50, Chapter 26] for information on the early work and [118, Chapter XII] for more recent developments and extensions.) We are going to prove, with the help of ergodic and combinatorial considerations, the following theorem.

THEOREM 1.1. *For fixed $n \in \mathbb{N}$ and a large enough prime p, the polynomial $f(z, y) = z^n - y^n$ represents the finite field $\mathbb{Z}_p = \mathbb{Z}/p\mathbb{Z}$. In other words, for any $c \in \mathbb{Z}_p$ there exist $z, y \in \mathbb{Z}_p^* = \mathbb{Z}_p \setminus \{0\}$, such that $c = z^n - y^n$.*

Putting $c = x^n$ immediately gives the following result, which was proved by Schur in 1916. (See also [49].)

COROLLARY 1.2 [126]. *For fixed $n \in \mathbb{N}$ and large enough prime p, the equation $x^n + y^n \equiv z^n \pmod p$ has nontrivial solutions.*

In the course of the proof of Theorem 1.1 we shall utilize the following classical fact due to F. Ramsey [117]. For a nice discussion which puts Ramsey's theorem into the perspective of *Ramsey theory*, see [72]. In what follows, $|A|$ denotes the cardinality of a set A.

THEOREM 1.3. *For any $n, r \in \mathbb{N}$ there exists a constant $c = c(n, r)$ such that if a set A satisfies $|A| \geqslant c$ and the set $[A]^2$ of two-element subsets of A is partitioned into r cells (or, as we will often say, is r-colored): $[A]^2 = \bigcup_{i=1}^r C_i$, then there exists a subset $B \subset A$ satisfying $|B| > n$ and such that for some i, $1 \leqslant i \leqslant r$, $[B]^2 \subset C_i$. (In this case we say that $[B]^2$ is* monochromatic.)

We shall also be using the following abstract version of the Poincaré recurrence theorem (cf. [115, pp. 69–72]).

THEOREM 1.4. *Assume that μ is a finitely-additive probability measure on a measurable space $(X, \mathcal{B})$, and let a group G (which is not necessarily infinite or commutative) act on $(X, \mathcal{B}, \mu)$ by measure preserving transformations T_g, $g \in G$. Let $A \in \mathcal{B}$ with $\mu(A) = a > 0$ and let an integer k satisfy $k > \lfloor \frac{1}{a} \rfloor$. If $|G| \geqslant k$, then for any k distinct elements $g_1, g_2, \ldots, g_k \in G$ there exist $1 \leqslant i < j \leqslant k$ such that $\mu(A \cap T_{g_i g_j^{-1}} A) > 0$.*

PROOF. If the statement does not hold, then for any $i \neq j$, $\mu(T_{g_i} A \cap T_{g_j} A) = 0$. But then $\mu(\bigcup_{i=1}^k T_{g_i} A) = \sum_{i=1}^k \mu(T_{g_i} A) = ka > 1$, in contradiction with $\mu(X) = 1$. □

REMARK 1.5. If one measures the triviality of a mathematical statement by the triviality of its proof, one can only wonder how and why a statement as trivial as Theorem 1.4 can lead to interesting applications. Yet it does! In particular, we shall utilize it in the proof of Theorem 1.1 and, at least implicitly, on few more occasions. (See Theorems 1.11 and 1.12 below. See also [10] for additional examples and more discussion.)

PROOF OF THEOREM 1.1. Let ν be the normalized counting measure on $\mathbb{Z}_p$. Noting that the index r of the multiplicative subgroup $\Gamma = \{x^n\colon\ x \in \mathbb{Z}_p^*\}$ in $\mathbb{Z}_p^*$ is at most n, we get, for p sufficiently large, $\nu(\Gamma) \geqslant \frac{1}{n+1}$. Let $a_1, \ldots, a_r \in \mathbb{Z}_p^*$, where $r \leqslant n$, be such that $\mathbb{Z}_p^* = \bigcup_{i=1}^r \Gamma a_i$ is the partition of $\mathbb{Z}_p^*$ into disjoint cosets of Γ. Let $A = \{2^j,\ 1 \leqslant j < \log_2 p\}$. Interpreting A as a subset of $\mathbb{Z}_p^*$, we note that since all the differences $2^j - 2^i$, $1 \leqslant i < j < \log_2 p$ are distinct, there is a natural bijection between the set $[A]^2$ of two-element subsets of A and the set $\Delta(A) = \{2^j - 2^i,\ 1 \leqslant i < j < \log_2 p\} \subseteq \mathbb{Z}_p$. The partition $\mathbb{Z}_p^* = \bigcup_{i=1}^r \Gamma a_i$ naturally induces a partition (coloring) of $\Delta(A)$. Assuming that p is large enough, we get by Theorem 1.3 a subset $B \subset A$ with the property that $|B| > n$ and such that the set of differences of distinct elements from B, $\Delta(B)$, is monochromatic, i.e. for some $i_0 \in \{1, 2, \ldots, r\}$, $\Delta(B) \subset \Gamma a_{i_0}$. But then Γ itself also contains a set of differences, namely $\Delta(B a_{i_0}^{-1})$.

Let us apply now Theorem 1.4 to the action of $\mathbb{Z}_p$ on itself by translations: $x \to x + g$, $g \in \mathbb{Z}_p$. Let $c \in \mathbb{Z}_p^*$ be arbitrary. Consider the set $B a_{i_0}^{-1} c \subset \Gamma c$. Since $|B a_{i_0}^{-1} c| = |B| > n$, we have by Theorem 1.4 that there is an element x in the set of differences $\Delta(B a_{i_0}^{-1} c)$, such that $\nu(\Gamma \cap \Gamma - x) > 0$. Noting that x is of the form gc, where $g \in \Delta(B a_{i_0}^{-1} c) \subset \Gamma$, we have $(\Gamma \cap \Gamma - gc) \neq 0$, which implies $gc \in \Gamma - \Gamma = \{z^n - y^n\colon\ z, y \in \mathbb{Z}_p^*\}$. Utilizing the fact that $g \in \Gamma$, we get $c \in \Gamma - \Gamma$. Since $c \in \mathbb{Z}_p^*$ was arbitrary (and since, trivially, $0 \in \Gamma - \Gamma$) we finally get $\mathbb{Z}_p = \Gamma - \Gamma$. □

We leave it to the reader to check that routine adaptation of the proof above allows one to show that for fixed n the polynomial $f(z, y) = z^n - y^n$ represents any large enough finite field. While this result has also a more traditional number-theoretical proof (see [125]), the "soft" method utilized in the proof of Theorem 1.1, gives, after appropriate modifications,

the following more general result, which so far has no conventional proof. We shall provide the proof at the end of Section 5.

THEOREM 1.6 [35]. *Let F be an infinite field and let Γ be a multiplicative subgroup of finite index in $F^* = F \setminus \{0\}$. Then*

$$\Gamma - \Gamma = \{x - y\colon\ x, y \in \Gamma\} = F.$$

While Theorem 1.1 is stronger than Schur's result (Corollary 1.2), the following key lemma from [126] is of independent interest as one of the earliest results of Ramsey theory.

THEOREM 1.7. *For any $r \in \mathbb{N}$, there exists a positive constant $c = c(r)$ such that for any integer $N \geqslant c$, any r-coloring $\{1, 2, \ldots, N\} = \bigcup_{i=1}^{r} C_i$ yields a monochromatic solution of the equation $x + y = z$.*

PROOF. The result almost immediately follows from Ramsey's theorem (Theorem 1.3 above) via an argument similar to the one utilized in the proof of Theorem 1.1. (Schur's original proof was somewhat longer, but completely elementary.) Observe that if r is fixed and N is sufficiently large, then one of the C_i contains the set of differences of a 3-element set $A = \{a_1, a_2, a_3\}$. The desired result then follows by setting $x = a_3 - a_2$, $y = a_2 - a_1$, $z = a_3 - a_1$. □

To derive Corollary 1.2 from Theorem 1.7, one considers the partition of $\{1, 2, \ldots, p - 1\}$ induced by the partition of $\mathbb{Z}_p^*$ into disjoint cosets of the multiplicative group $\Gamma = \{x^n\colon\ x \in \mathbb{Z}_p^*\}$. It then follows from Theorem 1.7 that there exists a coset Γc and $x, y, z \in \Gamma c$ such that (both in $\mathbb{N}$ and in $\mathbb{Z}_p$) $x + y = z$. Writing, for some $x_1^n, y_1^n, z_1^n \in \Gamma$, $x = x_1^n c$, $y = y_1^n c$, $z = z_1^n c$, we get, after the cancellation, $x_1^n + y_1^n \equiv z_1^n \pmod{p}$.

1.2. *Hilbert's theorem*

Arguably, the earliest nontrivial result of Ramsey theory is the following theorem which D. Hilbert utilized in [82] in order to show that if the polynomial $p(x, y) \in \mathbb{Z}[x, y]$ is irreducible, then there exists $n \in \mathbb{N}$ such that $p(x, n) \in \mathbb{Z}[x]$ is also irreducible. Given d distinct integers $x_1, \ldots, x_d$, define the *d-cube generated by $x_1, \ldots, x_d$* by $Q(x_1, \ldots, x_d) = \{\sum_{i=1}^{d} \varepsilon_i x_i, \varepsilon_i \in \{0, 1\}\}$.

THEOREM 1.8 [82]. *For any $d, r \in \mathbb{N}$ and any partition $\mathbb{N} = \bigcup_{i=1}^{r} C_i$, one of the C_i contains infinitely many translates of a d-cube.*

We shall see below that Hilbert's theorem admits a very simple proof based on a version of Poincaré recurrence theorem. But first we are going to formulate and discuss Hindman's classical Finite Sums Theorem, proved in [83], which contains both Schur's and Hilbert's theorems as very special cases.

DEFINITION 1.9. Let $(x_i)_{i=1}^{\infty} \subset \mathbb{N}$. The IP set generated by the sequence $(x_i)_{i=1}^{\infty}$ is the set $FS(x_i)_{i=1}^{\infty}$ of finite sums of elements of $(x_i)_{i=1}^{\infty}$ with distinct indices:

$$FS(x_i)_{i=1}^{\infty} = \left\{ x_\alpha = \sum_{i \in \alpha} x_i, \alpha \subset \mathbb{N}, 1 \leqslant |\alpha| < \infty \right\}.$$

1.3. *IP sets and Hindman's finite sum theorem*

IP sets can be viewed as a natural generalization of the notion of a d-cube (if one disregards the following subtle distinction: while the vertices x_i of the d-cube are supposed to be distinct, no such assumption is made in Definition 1.9). This explains the term IP (coined by H. Furstenberg and B. Weiss in [65]): Infinite-dimensional Parallelepiped.

THEOREM 1.10 [83]. *For any finite partition of $\mathbb{N}$, one of the cells of the partition contains an IP set.*

The original proof of Theorem 1.10 in [83] was, in Hindman's own words, "horrendously complicated." It therefore comes as a pleasant surprise that Hindman's theorem admits a short and easy proof. The following simple proposition is the key to proofs of Hindman's and many other results of a similar nature.

THEOREM 1.11. *Let $\mathcal{S}$ be a family of nonempty sets in $\mathbb{N}$. If $\mathcal{S}$ has the following property*:

(i) *for any $A \in \mathcal{S}$ there exist arbitrarily large $t \in \mathbb{N}$ such that*

$$A \cap (A - t) \in \mathcal{S},$$

then for any $A \in \mathcal{S}$ and any $d \in \mathbb{N}$, there exist $t_1 < t_2 < \cdots < t_d$ such that A contains infinitely many translates of the d-cube $Q(t_1, t_2, \ldots, t_d)$. If the following stronger property holds:

(ii) *for any $A \in \mathcal{S}$ there exist arbitrarily large $t \in A$ such that*

$$A \cap (A - t) \in \mathcal{S},$$

then each $A \in \mathcal{S}$ contains an IP set.

PROOF. Let $A \in \mathcal{S}$ and let t_1 be such that $A_1 = A \cap (A - t_1) \in \mathcal{S}$. By assumption, there exists $t_2 > t_1$ such that $A_2 = A_1 \cap (A_1 - t_2) \in \mathcal{S}$. But $A_2 = A \cap (A - t_1) \cap (A - t_2) \cap (A - (t_1 + t_2))$ and so it is clear that, for any $a \in A_2$, one has $a + Q(t_1, t_2) \subset A$. Continuing in this fashion one gets, after d steps, $t_1 < t_2 < \cdots < t_d$ such that $A_d = \bigcap_{\alpha \in \mathcal{F}_d} (A - t_\alpha) \in \mathcal{S}$, where $\mathcal{F}_d$ is the set of all subsets of $\{1, 2, \ldots, d\}$ and $t_\alpha = \sum_{\alpha \in \mathcal{F}_d} t_i$. Then any $a \in A_d$ has the property that $a + Q(t_1, t_2, \ldots, t_d) \subset A_d \subset A$, which proves the first assertion of the theorem. Now, let us assume that property (ii) holds. It is easy to see that by choosing at each step $t_i \in A_{i-1}$, where $A_0 = A$, one gets, for any $d \in \mathbb{N}$, $Q(t_1, t_2, \ldots, t_d) \subset A$. This clearly implies that $FS(t_i)_{i=1}^{\infty} \subset A$ and we are done. □

Recall that, for a set $A \subset \mathbb{N}$, the upper density $\bar{d}(A)$ is defined by $\bar{d}(A) = \limsup_{N\to\infty} \frac{|A\cap\{1,2,\ldots,N\}|}{N}$. It is easy to see, by trivial adaptation of the proof of Theorem 1.4 above, that if $\bar{d}(A) > 0$ then there exist arbitrarily large $t \in \mathbb{N}$ such that $\bar{d}(A \cap (A - t)) > 0$. Applying Theorem 1.11, we have now the following result which, in view of the fact that for any finite partition $\mathbb{N} = \bigcup_{i=1}^{r} C_i$ at least one of the C_i has positive upper density, may be considered as a strengthening of Hilbert's Theorem 1.8.

THEOREM 1.12. *Let $A \subset \mathbb{N}$ have positive upper density. Then for any $d \in \mathbb{N}$, there exist $t_1 < t_2 < \cdots < t_d$ such that the set*

$$\{a \in A\colon\ a + Q(t_1, t_2, \ldots, t_d) \subset A\}$$

has positive upper density. In particular, A contains infinitely many translates $a + Q(t_1, t_2, \ldots, t_d)$ with $a \in A$.

REMARK 1.13. One says that Theorem 1.12 is a *density* version of Theorem 1.8, which is a result about *partitions*. While we were lucky to produce a rather trivial proof of this density result, usually this is not the case. As we shall see in detail in Section 4, the density versions of partition results are much deeper and have rather involved and sophisticated proofs.

As we shall momentarily see, Hindman's theorem also follows from Theorem 1.11. To make the derivation possible, one needs only to find a family $\mathcal{S}$ of subsets of $\mathbb{N}$ which satisfies condition (ii) and has the property that for any finite partition $\mathbb{N} = \bigcup_{i=1}^{r} C_i$, one of the C_i belongs to $\mathcal{S}$. This is best achieved by utilizing $\beta\mathbb{N}$, the Stone–Čech compactification of $\mathbb{N}$ interpreted as the space of ultrafilters on $\mathbb{N}$. To be more precise, one utilizes the fact that, with respect to a naturally inherited operation extending the addition in $\mathbb{N}$, $\beta\mathbb{N}$ is a compact semitopological semigroup and, as such, has an idempotent. Any such idempotent allows one to introduce a certain $\{0, 1\}$-valued measure μ on the power set $\mathcal{P}(\mathbb{N})$ which, in turn, provides the sought after family $\mathcal{S}$ by the rule $A \in \mathcal{S} \Leftrightarrow \mu(A) = 1$.

The properties of such measures are described in the following proposition, the proof of which will be given in Section 3. (See Theorem 3.3 below.)

PROPOSITION 1.14. *There exists a finitely additive $\{0, 1\}$-valued probability measure μ on the space $\mathcal{P}(\mathbb{N})$ of all subsets of $\mathbb{N}$ which is "almost shift-invariant" in the following sense. For any $C \subset \mathbb{N}$ with $\mu(C) = 1$, the set*

$$T_C = \{n \in \mathbb{N}\colon\ \mu(C - n) = 1\} \tag{1.1}$$

satisfies $\mu(T_C) = 1$.

We are now in a position to give a proof of Hindman's theorem.

PROOF OF THEOREM 1.10. Let μ be an almost shift-invariant measure as described in Proposition 1.14, and let $\mathbb{N} = \bigcup_{i=1}^{r} C_i$ be a finite partition. Since μ is a probability measure, $\mu(\bigcup_{i=1}^{r} C_i) = 1$, which, by finite additivity and $\{0, 1\}$-valuedness, implies that one

of the C_i, call it C, satisfies $\mu(C) = 1$. By (1) we have $\mu(T_C) = 1$, which implies that $\mu(C \cap T_C) = 1$. It follows that the set $\{n \in C\colon\ \mu(C - n) = 1\}$ is of full measure and, in particular, that property (ii) in Theorem 1.11 is satisfied. Hence C contains an IP set and we are done. □

REMARK 1.15. Hindman's theorem finds numerous applications in ergodic theory, topological dynamics, and Diophantine analysis. Some of these will be discussed in this survey. Before moving on with our discussion, we want to record here the following equivalent version of Hindman's theorem, which can be interpreted as "indestructibility" of IP sets under finite partitions.

THEOREM 1.16. *For any finite partition of an IP set, one of the cells of the partition contains an IP set.*

We leave the elementary derivation of Theorem 1.16 from Hindman's theorem to the reader. (The other direction is trivial due to the fact that $\mathbb{N} = FS(2^i)_{i=1}^{\infty}$.) On a more sophisticated level, offered by the familiarity with $\beta\mathbb{N}$, Theorem 1.16 becomes an immediate consequence of the proof of Hindman's theorem given above. Indeed, one can show that any IP set in $\mathbb{N}$ is the support of an almost shift-invariant measure. (See Theorem 3.4 below.)

1.4. *Van der Waerden theorem: combinatorial and dynamical versions*

Our next example is the celebrated van der Waerden theorem.

THEOREM 1.17 [130,131]. *For any $r \in \mathbb{N}$ and any finite partition $\mathbb{N} = \bigcup_{i=1}^{r} C_i$, one of the C_i contains arbitrarily long arithmetic progressions.*

We remark that one cannot, in general, expect to get in Theorem 1.17 an infinite arithmetic progression in one of the C_i. Indeed, let us represent $\mathbb{N}$ as the union of disjoint intervals of increasing length and alternately color them red and blue. This obviously gives a two-coloring $\mathbb{N} = R \cup B$ without an infinite monochromatic progression.

Another remark is that Theorem 1.17 implies the following, ostensibly stronger, finitistic version.

THEOREM 1.18. *For any $r, l \in \mathbb{N}$ there exists $c = c(r, l)$ such that if $N \geqslant c$, then for any partition $\{1, 2, \ldots, N\} = \bigcup_{i=1}^{r} C_i$, one of the C_i contains an arithmetic progression of length l.*

PROOF OF THEOREM 1.18 VIA THEOREM 1.17. Assume by way of contradiction that Theorem 1.18 fails. Then there exist natural numbers r, l and, for any $N \in \mathbb{N}$, an interval I with $|I| \geqslant N$ and an r-coloring of I, which we will find convenient to view as a mapping $f : I \to \{1, 2, \ldots, r\}$, such that I contains no monochromatic progression of length l. Calling such r-colorings (and the corresponding intervals) AP_l-free, we may

assume without loss of generality that AP_l-free intervals I_n, $n \in \mathbb{N}$ tile $\mathbb{N}$ and satisfy $|I_{n+1}| \geqslant 2|I_n|$. Given an r-coloring $f : I \to \{1, 2, \ldots, r\}$ of an interval I, let us call the r-coloring defined by $\tilde{f} : I \to \{r+1, r+2, \ldots, 2r\}$ a *disjoint copy of* f if for all $k \in I$, $f(k) = \tilde{f}(k) - r$. To finish the argument, let us replace, for every $n \in \mathbb{N}$, the AP_l-free colorings $f_{2n} : I_{2n} \to \{1, 2, \ldots, r\}$ by their disjoint copies $\tilde{f_{2n}} : I_{2n} \to \{r+1, r+2, \ldots, 2r\}$. This results in a $2r$-coloring of $\mathbb{N}$ which has no monochromatic arithmetic progressions of length l, which contradicts Theorem 1.17. □

While in Khintchine's book [91] van der Waerden's theorem is called a "pearl of number theory," it should, perhaps, be more properly called a pearl of geometry. Indeed, it is not hard to see that van der Waerden's theorem is equivalent to the following result, which not only has an apparent geometric flavor, but also is suggestive of natural multidimensional extensions.

THEOREM 1.19. *For any finite partition $\mathbb{Z} = \bigcup_{i=1}^{r} C_i$, one of the C_i has the property that for any finite set $F \subset \mathbb{Z}$, there exist $a \in \mathbb{Z}$, and $b \in \mathbb{N}$ such that $aF + b = \{ax + b\colon\ x \in F\} \subset C_i$. In other words, one of the C_i contains a homothetic copy of any finite set.*

Here is the formulation of the multidimensional analogue of Theorem 1.19. It was first proved by Grünwald (Gallai), who apparently never published his proof. (Grünwald's authorship is acknowledged in [116, p. 123].)

THEOREM 1.20. *For any $d \in \mathbb{N}$ and any finite partition $\mathbb{Z}^d = \bigcup_{i=1}^{r} C_i$, one of the C_i has the property that for any finite set $F \subset \mathbb{Z}^d$, there exist $n \in \mathbb{N}$, and $v \in \mathbb{Z}^d$ such that $nF + v = \{nx + v\colon\ x \in F\} \subset C_i$.*

We shall now formulate yet another, dynamical, version of the (multidimensional) van der Waerden theorem. The idea to apply the methods of topological dynamics to partition results is due to H. Furstenberg and B. Weiss. (See [65].)

THEOREM 1.21. (Cf. [65, Theorem 1.4].) *Let $d \in \mathbb{N}$, $\varepsilon > 0$, and let X be a compact metric space. For any finite set of commuting homeomorphisms $T_i : X \to X$, $i = 1, 2, \ldots, k$, there exist $x \in X$ and $n \in \mathbb{N}$ such that*

$$\operatorname{diam}\left\{x, T_1^n x, T_2^n x, \ldots, T_k^n x\right\} < \varepsilon.$$

The reader will find various proofs of Theorem 1.21 in Sections 2 and 3. For now, we shall confine ourselves to the proof of the equivalence of Theorems 1.20 and 1.21.

THEOREM 1.20 $\Rightarrow$ THEOREM 1.21. Let $y \in X$ be arbitrary. For a vector $m = (m_1, m_2, \ldots, m_k) \in \mathbb{Z}^k$, write $T^m y = T_1^{m_1} T_2^{m_2} \cdots T_k^{m_k} y$. Since X is compact, there exists a finite family of open balls of radius $\varepsilon/2$, call it $\{B_i\}_{i=1}^{r}$, which covers X. Assign to each $m \in \mathbb{Z}^k$ the minimal i for which $T^m y \in B_i$. This produces a finite coloring $\mathbb{Z}^k = \bigcup_{i=1}^{r'} C_i$ (where $r' \leqslant r$). Let $S = \{0, e_1, \ldots, e_k\}$, where e_i are the standard unit vectors. By Theorem 1.20, there exist $n \in \mathbb{N}$ and $v \in \mathbb{Z}^k$ such that $nS + v$ is monochromatic. But this means

that $T^v y, T^{v+ne_1} y, \ldots, T^{v+ne_k} y$ all belong to the same $\frac{\varepsilon}{2}$-ball. Writing $x = T_y^v$ and noting that $T^{ne_i} = T_i^n$, $i = 1, 2, \ldots, k$, we get $\operatorname{diam}\{x, T_1^n x, T_2^n x, \ldots, T_k^n x\} < \varepsilon$. □

THEOREM 1.21 ⇒ THEOREM 1.20. The r-colorings of $\mathbb{Z}^d$ (viewed as mappings from $\mathbb{Z}^d$ to $\{1, 2, \ldots, r\}$) are naturally identified with the points of the compact product space $\Omega = \{1, 2, \ldots, r\}^{\mathbb{Z}^d}$. For $m = (m_1, m_2, \ldots, m_d) \in \mathbb{Z}^d$, let $|m| = \max_{1 \leqslant i \leqslant d} |m_i|$. Introduce a metric on Ω by defining, for any pair $x, y \in \Omega$, $\rho(x, y) = \inf\{\frac{1}{n}\colon\ x(m) = y(m)$ for m with $|m| < n\}$. It is easy to see that the metric ρ is compatible with the product topology and that $\rho(x, y) < 1 \Leftrightarrow x(0) = y(0)$. Let $F = \{a_1, a_2, \ldots, a_k\} \subset \mathbb{Z}^d$. Define the homeomorphisms $T_i : \Omega \to \Omega$, $i = 1, 2, \ldots, k$, by $(T_i x)(m) = x(m + a_i)$, and set, for $n = (n_1, n_2, \ldots, n_k) \in \mathbb{Z}^k$, $T^n = T_1^{n_1} T_2^{n_2} \cdots T_k^{n_k}$. Let now $x(m)$ be the element of Ω corresponding to the coloring $\mathbb{Z}^d = \bigcup_{i=1}^r C_i$ (in other words, for any $m \in \mathbb{Z}^d$, $x(m) = i$ iff $m \in C_i$). Let $X = \overline{\{T^n x\}}_{n \in \mathbb{Z}^k}$ be the orbital closure of x in Ω. Note that for any $\delta > 0$ and any $y \in X$, there exists $m \in \mathbb{Z}^k$ with $\rho(T^m x, y) < \delta$. Setting $\varepsilon = 1$ in Theorem 1.21, we can find $y \in X$ and $n \in \mathbb{N}$ such that $\operatorname{diam}\{y, T_1^n y, \ldots, T_k^n y\} < 1$. Choosing $u = (u_1, \ldots, u_k) \in \mathbb{Z}^k$ so that the element of the orbit $T^u x$ is close enough to y, and also such that $T_i^n (T^m x)$ are close enough to $T_i^n y$ for $i = 1, 2, \ldots, k$, we shall still have $\operatorname{diam}\{T^u x, T^u T_1^n x, \ldots, T^u T_k^n x\} < 1$. This implies that, for $v = u_1 a_1 + u_2 a_2 + \cdots + u_k a_k$, $x(v) = x(v + na_1) = \cdots = x(v + na_k)$, which means that the set $v + nF$ is monochromatic. □

1.5. *Density Ramsey theory*

In accordance with the general philosophy of Ramsey Theory (see [9] for more discussion), one should expect the density version of Theorem 1.20 to hold true as well. While the proof of this density version is far from being trivial, its formulation is easily guessable (see Theorem 1.23 below). It is also natural to expect that the dynamical form of the multidimensional van der Waerden theorem, our Theorem 1.21, can be "upgraded" in such a way that it gives a dynamical equivalent to the density version of Theorem 1.20. Theorem 1.24 below, proved in [60], confirms these expectations. To present the historical development in its natural order, we should mention that already the density version of the one-dimensional van der Waerden theorem, conjectured by Erdős and Turán in the mid-thirties in [53], proved quite recalcitrant and was settled only in 1975 by Szemerédi [127]. A few years later, Furstenberg [57] gave a completely different, ergodic-theoretical proof of Szemerédi's theorem, thereby starting a new area of dynamics which is today called Ergodic Ramsey Theory. The multidimensional Szemerédi theorem proved in [60] was the first result in the long and impressive line of dynamical proofs of various combinatorial and number-theoretical results, most of which still do not have a conventional proof. Many of these results will be discussed in the subsequent sections. See [65,60,63,63,64,97,98, 23–25,29,31,30,67]. We note that recently purely combinatorial proofs have been given for the multidimensional Szemerédi theorem based on an extension of the Szemerédi regularity lemma [71,106,119,120,128]. See also the remarkable recent preprint by B. Green and T. Tao [73] where (a modification of) the ergodic approach to Szemerédi's theorem is

blended with the techniques of the analytic number theory to establish the striking fact that the primes contain arbitrarily long arithmetic progressions.

DEFINITION 1.22. Let $d \in \mathbb{N}$ and $E \subset \mathbb{Z}^d$.

(i) The upper density of E, $\bar{d}(E)$, is defined by

$$\bar{d}(E) = \limsup_{N\to\infty} \frac{|E \cap [-N, N]^d|}{(2N+1)^d}.$$

(ii) The upper Banach density of E, $d^*(E)$, is defined by

$$d^*(E) = \limsup_{N_i - M_i \to \infty, 1 \leqslant i \leqslant d} \frac{\left|E \cap \prod_{i=1}^d [M_i, N_i - 1]\right|}{\prod_{i=1}^d (N_i - M_i)}.$$

Here are now combinatorial and dynamical formulations of the density version of the multidimensional van der Waerden theorem. (Cf. [60].)

THEOREM 1.23 (Multidimensional Szemerédi Theorem). *Let $d \in \mathbb{N}$, and let $E \subset \mathbb{Z}^d$ have positive upper Banach density. For any finite set $F \subset \mathbb{Z}^d$, there exist $n \in \mathbb{N}$ and $v \in \mathbb{Z}^d$ such that $nF + v \subset E$.*

THEOREM 1.24. *Let $(X, \mathcal{B}, \mu)$ be a probability measure space. For any finite set $\{T_1, \ldots, T_k\}$ of commuting measure preserving transformations of X and for any $A \in \mathcal{B}$ with $\mu(A) > 0$, there exists $n \in \mathbb{N}$ such that*

$$\mu\left(A \cap T_1^{-n} A \cap T_2^{-n} A \cap \cdots \cap T_k^{-n} A\right) > 0.$$

1.6. *Furstenberg's correspondence principle*

To see that Theorem 1.23 follows from Theorem 1.24, one can use a correspondence principle, introduced by Furstenberg in [57] in order to derive Szemerédi's theorem from an ergodic multiple recurrence result which he established in [57] and which corresponds to taking $T_i = T^i$, $i = 1, 2, \ldots, k$, in Theorem 1.24.

For the proof of the following version of Furstenberg's correspondence principle, see [31, Proposition 7.2]. See also Theorem 5.8 in Section 5 for a general form of Furstenberg's correspondence principle for amenable (semi)groups.

THEOREM 1.25. *Let $d \in \mathbb{N}$. For any set $E \subset \mathbb{Z}^d$ with $d^*(E) > 0$, there exists a probability measure preserving system $(X, \mathcal{B}, \mu, \{T^{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathbb{Z}^d})$ and a set $A \in \mathcal{B}$ with $\mu(A) = d^*(E)$ such that for all $k \in \mathbb{N}$ and $\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_k \in \mathbb{Z}^d$ one has*

$$d^*\left(E \cap (E - \boldsymbol{n}_1) \cap \cdots \cap (E - \boldsymbol{n}_k)\right) \geqslant \mu\left(A \cap T^{\boldsymbol{n}_1} A \cap \cdots \cap T^{\boldsymbol{n}_k} A\right).$$

We leave it to the reader to verify (with the help of Theorem 1.25) that Theorem 1.24 implies Theorem 1.23. Here is a simple proof of the other implication (the idea behind this proof will also be used in the proof of Lemma 5.10 in Section 5).

THEOREM 1.23 $\Rightarrow$ THEOREM 1.24. Assume, by way of contradiction, that there exist a probability measure space $(X, \mathcal{B}, \mu)$, commuting measure preserving transformations $T_1, T_2, \ldots, T_k$ of X, and a set $A \in \mathcal{B}$ with $\mu(A) > 0$ such that for all $n \in \mathbb{N}$, $\mu(A \cap T_1^{-n}A \cap \cdots \cap T_k^{-n}A) = 0$. Deleting, if needed, a set of measure zero from A, we may and will assume that one actually has $A \cap T_1^{-n}A \cap \cdots \cap T_k^{-n}A = \emptyset$ for all $n \in \mathbb{N}$. For $\boldsymbol{n} = (n_1, n_2, \ldots, n_k)$, write $T^{\boldsymbol{n}} = T_1^{n_1} T_2^{n_2} \cdots T_k^{n_k}$ and let

$$f_N(x) = \frac{1}{(2N+1)^k} \sum_{\boldsymbol{n} \in [-N,N]^k} 1_A(T^{\boldsymbol{n}} x), \quad N = 1, 2, \ldots.$$

Note that $0 \leqslant f_N(x) \leqslant 1$ for all $x \in X$ and $N \in \mathbb{N}$ and that $\int f_N \, d\mu = \mu(A)$. Let $f(x) = \limsup_{N\to\infty} f_N(x)$. By Fatou's lemma, we have

$$\int f \, d\mu = \int \limsup_{N\to\infty} f_N \, d\mu \geqslant \limsup_{N\to\infty} \int f_N \, d\mu = \mu(A).$$

It follows that there exists $x_0 \in X$ such that $\limsup f_N(x_0) = f(x_0) \geqslant \mu(A)$. Hence for some increasing sequence $N_i \to \infty$, one has

$$\lim_{i\to\infty} f_{N_i}(x_0) = \lim_{i\to\infty} \frac{1}{(2N_i+1)^k} \sum_{\boldsymbol{n} \in [-N_i,N_i]^k} 1_A(T^{\boldsymbol{n}} x_0) = f(x_0) \geqslant \mu(A).$$

This implies that the set $E = \{\boldsymbol{n} \in \mathbb{Z}^k\colon\ T^{\boldsymbol{n}} x_0 \in A\}$ has positive upper density. (We actually showed that $\bar{d}(E) \geqslant \mu(A)$.) By Theorem 1.23, the set E contains a configuration of the form $nF + v$, where $F = \{0, e_1, e_2, \ldots, e_k\}$, $n \in \mathbb{N}$, and $v \in \mathbb{Z}^k$ (where e_i are the standard unit vectors). This implies that $A \cap T_1^{-n}A \cap \cdots \cap T_k^{-n}A \neq \emptyset$, which contradicts the assumption made above. □

It is perhaps of interest to observe that while the combinatorial version of the multidimensional van der Waerden theorem follows immediately from Theorem 1.23 by the observation that for any finite partition $\mathbb{Z}^d = \bigcup_{i=1}^r C_i$, at least one of the C_i satisfies $\bar{d}(C_i) \geqslant 1/r$, the derivation of Theorem 1.21 from Theorem 1.24 is less trivial, and depends on the fact that for any $\mathbb{Z}^d$-action by homeomorphisms of a compact space, there exists an invariant measure.

1.7. *Hales–Jewett theorem*

We would like to formulate still another important extension of van der Waerden's theorem, the powerful Hales–Jewett theorem.

Consider the following generalization of tic-tac-toe: there are r players which are taking turns in placing the symbols $s_1, \dots, s_r$ in the $k \times k \times \cdots \times k$ (n times) array, which one views as the nth Cartesian power A^n of a k-element set $A = \{a_1, a_2, \dots, a_k\}$. (In the classical tic-tac-toe, we have $r = 2$, $k = 3$, $n = 2$.) It is convenient to think of the symbols $s_1, \dots, s_r$ as colors, and to identify the elements of the array A^n as the set $W_n(A)$ of words of length n over the alphabet A. We are going to define now the notion of a *combinatorial line* in A^n. Let $\tilde{A} = A \cup \{t\}$ be an extension of the alphabet A, obtained by adding a new symbol t. Let $W_n(t)$ be the set of words of length n over $\tilde{A}$ in which the symbol t occurs. Given a word $w(t) \in W_n(t)$, let us define a combinatorial line as a set $\{w(a_1), w(a_2), \dots, w(a_k)\}$ obtained by substituting for t the elements of A. For example, the word $13t241t2$ over the alphabet $\{1, 2, 3, 4, 5\} \cup \{t\}$ gives rise to the combinatorial line $\{13124112, 13224122, 13324132, 13424142, 13524152\}$. The goal of the players is to obtain a monochromatic combinatorial line. The following celebrated theorem of Hales and Jewett [75] implies that for fixed r, k and large enough n, the first player can always win.

THEOREM 1.26. *Let $r, k \in \mathbb{N}$. There exists $c = c(k, r)$ such that if $n \geqslant c$, then, for any r-coloring of the set $W_n(A)$ of words of length n over the k-letter alphabet $A = \{a_1, a_2, \dots, a_k\}$, there is a monochromatic combinatorial line.*

Taking $A = \{0, 1, \dots, l-1\}$ and interpreting $W_n(A)$ as integers in base l having at most n digits in their base l expansion, we see that in this situation, the elements of a combinatorial line form an arithmetic progression of length l (with difference of the form $d = \sum_{i=0}^{k-1} \varepsilon_i l^i$, where $\varepsilon_i = 0$ or 1). Thus van der Waerden's theorem is a corollary of Theorem 1.26.

Take now A to be a finite field F. Then $W_n(F) = F^n$ has the natural structure of an n-dimensional vector space over F. It is easy to see that, in this case, a combinatorial line is an affine linear one-dimensional subspace of F^n. We have therefore the following corollary of the Hales–Jewett theorem.

THEOREM 1.27. *Let F be a finite field. For any $r \in \mathbb{N}$ there exists $c = c(r)$ such that if V is a vector space over F having dimension at least c, then for any r-coloring $V = \bigcup_{i=1}^{r} C_i$, one of the C_i contains an affine line.*

One of the signs of the fundamental nature of the Hales–Jewett theorem is that one can easily derive from it its multidimensional version. (This fact will be especially appreciated by anyone who tried to derive from van der Waerden's theorem its multidimensional version.) Let $t_1, t_2, \dots, t_m$ be m variables and let $w(t_1, t_2, \dots, t_m)$ be a word of length n over the alphabet $A \cup \{t_1, \dots, t_m\}$. (We assume, of course, that the letters t_i do not belong to A.) If, for some n, $w(t_1, \dots, t_m)$ is a word of length n in which all of the variables $t_1, t_2, \dots, t_m$ occur, the result of the substitution $\{w(t_1, t_2, \dots, t_m)\}_{(t_1, t_2, \dots, t_m) \in A^m} = \{w(a_{i_1}, a_{i_2}, \dots, a_{i_m})\colon\ a_{i_j} \in A,\ j = 1, 2, \dots, m\}$ is called a combinatorial m-space.

Observe now that if we replace the original alphabet A by A^m, then a combinatorial line in $W_n(A^m)$ can be interpreted as an m-space in $W_{nm}(A)$. Thus, we have the following ostensibly stronger theorem as a corollary of Theorem 1.26.

THEOREM 1.28. *Let $r, k, m \in \mathbb{N}$. There exists $c = c(r, k, m)$ such that if $n \geqslant c$, then for any r-coloring of the set $W_n(A)$ of words of length n over the k-letter alphabet A, there exists a monochromatic m-space.*

Theorem 1.28 obviously implies the following multidimensional extension of Theorem 1.27.

THEOREM 1.29. *Let F be a finite field. For any $r, m \in \mathbb{N}$, there exists $c = c(r, m)$ such that if V is a vector space over F having dimension at least c, then for any r-coloring $V = \bigcup_{i=1}^{r} C_i$, one of the C_i contains an m-dimensional affine space.*

We leave it to the reader to derive from Theorem 1.28 the multidimensional van der Waerden theorem and an extension of Theorem 1.27 pertaining to m-dimensional affine subspaces of F^n. See Section 2 for a proof of the Hales–Jewett theorem and for more discussion and applications.

1.8. *Sárközy–Furstenberg theorem*

Our next example is the following surprising theorem, which was proved independently by Sárközy and Furstenberg, and which has interesting links with spectral theory, Diophantine approximations, combinatorics, and dynamical systems. (See [124,57–59,90].)

THEOREM 1.30. *Let $E \subset \mathbb{N}$ be a set of positive upper density, and let $p(n) \in \mathbb{Z}[n]$ be a polynomial with $p(0) = 0$. Then there exist $x, y \in E$ and $n \in \mathbb{N}$ such that $x - y = p(n)$.*

This result is perhaps more surprising than any of the theorems formulated above. One can surely expect the set of differences of a large set to be even larger. For example, if, for $E \subset \mathbb{N}$, $d^*(E) > 0$, then it is not hard to show that the set of differences $E - E = \{x - y\colon\ x, y \in E\}$ is syndetic, i.e. has bounded gaps. (See, for example, [58, Proposition 3.19], or [9, pp. 8–9].) But there is, a priori, no obvious reason for the set $E - E$ to be so "well spread" as to nontrivially intersect the set of values taken by any integer-valued polynomial vanishing at zero. The following dynamical counterpart of Theorem 1.30, due to Furstenberg, is just as striking. (See [57, Proposition 1.3], and [58, Theorem 3.16].)

THEOREM 1.31. *For any invertible probability measure preserving system $(X, \mathcal{B}, \mu, T)$, any $A \in \mathcal{B}$ with $\mu(A) > 0$, and any polynomial $p(n) \in \mathbb{Z}[n]$ with $p(0) = 0$, there exists $n \in \mathbb{N}$ such that $\mu(A \cap T^{p(n)} A) > 0$.*

REMARKS.

(1) One can derive Theorem 1.30 from Theorem 1.31 by utilizing Furstenberg's correspondence principle. In the other direction, one can, for example, mimic the argument that was used above to derive Theorem 1.24 from Theorem 1.23.
(2) One should, of course, view Theorem 1.31 as a refinement of the Poincaré recurrence theorem. While the classical Poincaré recurrence theorem only tells us that

a typical point returns, under the evolution laws of the dynamical system, to a set of positive volume in the phase space, Theorem 1.31 tells us that this will happen along any prescribed in advance sequence of "polynomial" times. However, when compared with the Poincaré recurrence theorem, Theorem 1.31 is a rather deep result. This is, in particular, manifested by the fact that all the known proofs of the Theorem 1.31 prove actually more than stated.

Furstenberg's proof of Theorem 1.31 utilizes the spectral theorem for unitary operators. The proof that we have chosen to present here is "softer" in the sense that it avoids the usage of the spectral theorem and thereby is susceptible to further generalizations. (See Theorems 4.27 and 4.30 below.)

We shall need the following useful result, which can be viewed as a Hilbert space version of the classical van der Corput difference theorem in the theory of uniform distribution.

THEOREM 1.32 (van der Corput trick). *Let $(u_n)_{n\in\mathbb{N}}$ be a bounded sequence in a Hilbert space $\mathcal{H}$. If for every $h \in \mathbb{N}$ it is the case that*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \langle u_{n+h}, u_n \rangle = 0$$

(*where $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathcal{H}$*), *then* $\lim_{N\to\infty} \| \frac{1}{N} \sum_{n=1}^{N} u_n \| = 0$.

PROOF. Observe that for any $\varepsilon > 0$ and any $H \in \mathbb{N}$, if N is large enough then

$$\left\| \frac{1}{N} \sum_{n=1}^{N} u_n - \frac{1}{N} \frac{1}{H} \sum_{n=1}^{N} \sum_{h=0}^{H-1} u_{n+h} \right\| < \varepsilon.$$

But,

$$\begin{aligned} &\limsup_{N\to\infty} \left\| \frac{1}{N} \frac{1}{H} \sum_{n=1}^{N} \sum_{h=0}^{H-1} u_{n+h} \right\|^2 \\ &\quad \leqslant \limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left\| \frac{1}{H} \sum_{h=0}^{H-1} u_{n+h} \right\|^2 \\ &\quad = \limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \frac{1}{H^2} \sum_{h_1,h_2=0}^{H-1} \langle u_{n+h_1}, u_{n+h_2} \rangle \leqslant \frac{B}{H}, \end{aligned}$$

where $B = \sup_{n\in\mathbb{N}} \|u_n\|^2$. Since H was arbitrary, we are done. □

PROOF OF THEOREM 1.31. Let $\mathcal{H} = L^2(X, \mathcal{B}, \mu)$ and let $U : \mathcal{H} \to \mathcal{H}$ be the unitary operator induced by T:

$$(Uf)(x) = f(Tx), \quad f \in L^2(X, \mathcal{B}, \mu).$$

Let, for each $a \in \mathbb{N}$,

$$\mathcal{H}_a = \{f\colon U^a f = f\},$$

$$\mathcal{H}_{\text{erg}}^{(a)} = \left\{ f\colon \left\| \frac{1}{N} \sum_{n=0}^{N-1} U^{an} f \right\| \to 0 \right\}.$$

The classical ergodic splitting (with respect to U^a), $\mathcal{H} = \mathcal{H}_a \oplus \mathcal{H}_{\text{erg}}^{(a)}$, leads to the following, more suitable for our goals, splitting of $\mathcal{H}$ into "rational spectrum" and "totally ergodic" parts. Let

$$\mathcal{H}_{\text{rat}} = \overline{\{f\colon \exists a \in \mathbb{N}\colon U^a f = f\}} = \overline{\bigcup_{a=1}^{\infty} \mathcal{H}_a},$$

$$\mathcal{H}_{\text{tot.erg.}} = \left\{ f\colon \forall a \in \mathbb{N}, \left\| \frac{1}{N} \sum_{n=0}^{N-1} U^{an} f \right\| \to 0 \right\} = \bigcap_{a=1}^{\infty} \mathcal{H}_{\text{erg}}^{(a)}.$$

It is easy to check now that $\mathcal{H}_{\text{rat}}^{\perp} = \mathcal{H}_{\text{tot.erg.}}$ and that $\mathcal{H} = \mathcal{H}_{\text{rat}} \oplus \mathcal{H}_{\text{tot.erg.}}$ Let $1_A = f + g$, where $f \in \mathcal{H}_{\text{rat}}$, $g \in \mathcal{H}_{\text{tot.erg.}}$. We remark that since $1_A \geqslant 0$ and $\int 1_A \, d\mu = \mu(A) > 0$, one has $f \geqslant 0$, $f \neq 0$. Indeed, f minimizes the distance from $\mathcal{H}_{\text{rat}}$ to 1_A, and the function $\max\{f, 0\}$ (which, as is not too hard to check, also belongs to $\mathcal{H}_{\text{rat}}$) would do at least as well in minimizing this distance. This remark equally applies, for any $a \in \mathbb{N}$, to the orthogonal projection f_a of 1_A onto $\mathcal{H}_a$. Note also that $\int f_a \, d\mu = \mu(A)$.

We are going to show that $\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu(A \cap T^{p(n)} A)$ exists and is positive. Note that, in view of the orthogonal decomposition $1_A = f + g$, we have:

$$\begin{aligned} \mu\big(A \cap T^{p(n)} A\big) &= \int (f+g) U^{p(n)} (f+g) \, d\mu \\ &= \int f U^{p(n)} f \, d\mu + \int g U^{p(n)} g \, d\mu. \end{aligned}$$

We shall show first that $\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \int g U^{p(n)} g \, d\mu = 0$ (and hence it will remain to show that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu\big(A \cap T^{p(n)} A\big) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \int f U^{p(n)} f \, d\mu > 0).$$

Note that since $g \in \mathcal{H}_{\text{tot.erg.}}$, one has, for any linear polynomial $p(n)$ with integer coefficients, $\lim_{N\to\infty} \|\frac{1}{N}\sum_{n=0}^{N-1} U^{p(n)} g\| = 0$.

We shall use the van der Corput trick to inductively reduce the situation to this linear case. Let $u_n = U^{p(n)} g$, $n \in \mathbb{N}$. We have:

$$\langle u_{n+h}, u_n\rangle = \big\langle U^{p(n+h)} g, U^{p(n)} g\big\rangle = \big\langle U^{p(n+h)-p(n)} g, g\big\rangle.$$

Notice that, for any fixed $h \in \mathbb{N}$, the degree of the polynomial $p(n+h) - p(n)$ equals $\deg p(n) - 1$. Using the fact that strong convergence implies weak convergence, we have by the induction hypothesis:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \langle u_{n+h}, u_n\rangle = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \big\langle U^{p(n+h)-p(n)} g, g\big\rangle = 0.$$

It follows from Theorem 1.32 that $\lim_{N\to\infty} \|\frac{1}{N}\sum_{n=0}^{N-1} U^{p(n)} g\| = 0$ and hence

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \int g U^{p(n)} g \, d\mu = 0.$$

It remains now to prove that $\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \int f U^{p(n)} f \, d\mu > 0$. Note first that the existence of this limit is almost obvious. Indeed, since $f \in \mathcal{H}_{\text{rat}}$, it is enough to check only the case when f belongs to one of the $\mathcal{H}_a$, in which case there is practically nothing to check since, for such f, the sequence $U^{p(n)} f$, $n \in \mathbb{N}$, is periodic.

To see that the limit in question is strictly positive, choose $a \in \mathbb{N}$ so that $\|f - f_a\|$ is close to zero (where f_a is the orthogonal projection of f on $\mathcal{H}_a$). Note now that if $n \in a\mathbb{N}$ then $p(n)$ is divisible by a, and hence $\int f_a U^{p(n)} f_a \, d\mu = \int f_a^2 \, d\mu \geqslant (\mu(A))^2$. This clearly implies the positivity of the limit

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \int f U^{p(n)} f \, d\mu = \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \mu\big(A \cap T^{p(n)} A\big). \qquad \square$$

We will conclude the introductory section here. Each of the examples above is a small fragment of a much bigger picture. In the subsequent sections, we shall try to supply more facts and details so that the reader will be able to see better both the multiple interconnections between various theorems of Ergodic Ramsey Theory and the general direction of the flow of current developments.

2. Topological dynamics and partition Ramsey theory

2.1. *Introduction*

In various applications, one would like not only to be able to find certain types of monochromatic configurations for any finite coloring of a highly organized structure, such as $\mathbb{N}$ or an infinite vector space over a finite field, but also to know that these configurations are plentiful. There are many notions of largeness which one can use to measure the abundance of sought-after configurations. One of these notions is that of *syndeticity*. A subset S in $\mathbb{N}$ is syndetic if finitely many translates of S cover $\mathbb{N}$, i.e. for some k and $a_1, a_2, \ldots, a_k \in \mathbb{N}$, one has $\bigcup_{i=1}^{k}(S - a_i) = \mathbb{N}$. (This definition can be easily adapted to make sense in any semigroup.)

A stronger notion of largeness which we will presently introduce with the help of IP sets (see Definition 1.9) not only implies syndeticity, but has also the finite intersection property.

DEFINITION 2.1. A set $E \subseteq \mathbb{N}$ is said to be IP* if it has nontrivial intersection with any IP set.

It is not too hard to see that any IP* set is syndetic. Indeed, if an IP* set S were not syndetic, then its complement would contain an infinite union of intervals $[a_n, b_n]$ with $b_n - a_n \to \infty$, and it is not hard to show that any such union of intervals contains an IP set, which leads to a contradiction with the assumption that S is an IP* set.

Let us show now that the family of IP* sets has the finite intersection property. It is enough to prove that if S_1, S_2 are IP* sets, then $S_1 \cap S_2$ is as well. Let E be an arbitrary IP set and consider the partition $E = (E \cap S_1) \cup (E \cap S_1^c)$. By Hindman's theorem (see Theorem 1.10), either S_1 or S_1^c has to contain an IP set E_1. But, it is clear that it has to be S_1, since S_1 is IP*, and hence $S_1 \cap E_1 \neq \emptyset$, which implies that $E_1 \subset E \cap S_1$. Now, since S_2 is also an IP* set, we have $E_1 \cap S_2 \neq \emptyset$, which implies that $(E \cap S_1) \cap S_2 = E \cap (S_1 \cap S_2) \neq \emptyset$.

2.2. *IP van der Waerden theorem*

We are going to formulate and prove now the so-called IP van der Waerden theorem (proved first in [65]) which, in particular, will tell us that the set of differences of monochromatic arithmetic progressions always to be found in any finite coloring of $\mathbb{Z}$ is an IP* set. At the same time, this IP van der Waerden theorem is powerful enough to imply not only Theorem 1.20, but also Theorem 1.29. The proof presented below is taken from [11] and is based on the proof of the multidimensional van der Waerden's theorem in [37]. We need first to introduce a few more definitions, and some notation.

An *$\mathcal{F}$-sequence* in an arbitrary space Y is a sequence $\{y_\alpha\}_{\alpha \in \mathcal{F}}$ indexed by the set $\mathcal{F}$ of the finite nonempty subsets of $\mathbb{N}$. If Y is a (multiplicative) semigroup, one says that an $\mathcal{F}$-sequence defines an *IP-system* if for any $\alpha = \{i_1, i_2, \ldots, i_k\} \in \mathcal{F}$, one has $y_\alpha = y_{i_1} y_{i_2} \cdots y_{i_k}$. IP-systems should be viewed as generalized semigroups. Indeed, if

$\alpha \cap \beta = \emptyset$, then $y_{\alpha\cup\beta} = y_\alpha y_\beta$. We shall often use this formula for sets α, β satisfying $\alpha < \beta$.

We will be working with IP-systems generated by homeomorphisms belonging to a commutative group G acting minimally on a compact space X. (Recall that (X, G) is a *minimal* dynamical system if for each nonempty open set $V \subset X$ there exist $S_1, \ldots, S_r \in G$ so that $\bigcup_{i=1}^r S_i V = X$.)

THEOREM 2.2. *Let X be a compact topological space and G a commutative group of its homeomorphisms such that the dynamical system (X, G) is minimal. For any nonempty open set $V \subseteq X$, any $k \in \mathbb{N}$, any IP-systems $\{T_\alpha^{(1)}\}_{\alpha\in\mathcal{F}}, \ldots, \{T_\alpha^{(k)}\}_{\alpha\in\mathcal{F}}$ in G and any $\alpha_0 \in \mathcal{F}$, there exists $\alpha \in \mathcal{F}, \alpha > \alpha_0$, such that $V \cap T_\alpha^{(1)} V \cap \cdots \cap T_\alpha^{(k)} V \neq \emptyset$.*

PROOF. We fix a nonempty open $V \subseteq X$ and $S_1, \ldots, S_r \in G$ with the property that $S_1 V \cup S_2 V \cup \cdots \cup S_r V = X$. (The existence of $S_1, \ldots, S_r$ is guaranteed by the minimality of (X, G).) The proof proceeds by induction on k. The case $k = 1$ is almost trivial, but we shall do it in detail to set up the notation in a way that indicates the general idea.

So, let $\{T_i\}_{i=1}^\infty$ be a fixed sequence of elements in G and $\{T_\alpha\}_{\alpha\in\mathcal{F}}$ the IP-system generated by $\{T_i\}_{i=1}^\infty$. (This means of course that for any finite nonempty set $\alpha = \{i_1, i_2, \ldots, i_m\} \subset \mathbb{N}$, one has $T_\alpha = T_{i_1} T_{i_2} \cdots T_{i_m}$.)

Now we construct a sequence $W_0, W_1, \ldots$ of nonempty open sets in X so that:

(i) $W_0 = V$;

(ii) $T_n^{-1} W_n \subseteq W_{n-1}, \forall n \geqslant 1$;

(iii) each $W_n, n \geqslant 1$, is contained in one of the sets $S_1 V, S_2 V, \ldots, S_r V$. (We recall that $S_1 V \cup S_2 V \cup \cdots \cup S_r V = X$.)

To define W_1, let t_1, $1 \leqslant t_1 \leqslant r$, be such that $T_1 V \cap S_{t_1} V = T_1 W_0 \cap S_{t_1} V \neq \emptyset$; let $W_1 = T_1 W_0 \cap S_{t_1} V$. If W_n was already defined, then let t_{n+1} be such that $1 \leqslant t_{n+1} \leqslant r$ and $T_{n+1} W_n \cap S_{t_{n+1}} V \neq \emptyset$, and let $W_{n+1} = T_{n+1} W_n \cap S_{t_{n+1}} V$. By the construction, each W_n is contained in one of the $S_1 V, \ldots, S_r V$, so there will necessarily be two natural numbers $i < j$ and $1 \leqslant t \leqslant r$ such that $W_i \cup W_j \subseteq S_t V$ (pigeonhole principle!). Let $U = S_t^{-1} W_j$ and $\alpha = \{i + 1, i + 2, \ldots, j\}$. We have

$$T_\alpha^{-1} U = T_{i+1}^{-1} T_{i+2}^{-1} \cdots T_j^{-1} S_t^{-1} W_j = S_t^{-1} T_{i+1}^{-1} T_{i+2}^{-1} \cdots T_j^{-1} W_j$$
$$\subseteq S_t^{-1} T_{i+1}^{-1} T_{i+2}^{-1} \cdots T_{j-1}^{-1} W_{j-1} \subseteq \cdots \subseteq S_t^{-1} T_{i+1}^{-1} W_{i+1} \subseteq S_t^{-1} W_i \subseteq V.$$

So, $U \subseteq T_\alpha V$ and $U \subseteq V$ which implies $V \cap T_\alpha V \neq \emptyset$.

Notice that since the pair $i < j$ for which there exists t with the property $W_i \cup W_j \subseteq S_t V$ could be chosen with arbitrarily large i, it follows that the set $\alpha = \{i + 1, \ldots, j\}$ for which $V \cap T_\alpha V \neq \emptyset$ could be chosen so that $\alpha > \alpha_0$.

Assume now that the theorem holds for any k IP-systems in G. Fix a nonempty set V and $k + 1$ IP-systems $\{T_\alpha^{(1)}\}_{\alpha\in\mathcal{F}}, \ldots, \{T_\alpha^{(k+1)}\}_{\alpha\in\mathcal{F}}$. We shall also fix the homeomorphisms $S_1, \ldots, S_r \in G$ (whose existence is guaranteed by minimality) satisfying $S_1 V \cup \cdots \cup S_r V = G$. We shall inductively construct a sequence $W_0, W_1, \ldots$ of nonempty open sets in X and an increasing sequence $\alpha_1 < \alpha_2 < \cdots$ in $\mathcal{F}$ so that

(a) $W_0 = V$,
(b) $(T^{(1)}_{\alpha_n})^{-1}W_n \cup (T^{(2)}_{\alpha_n})^{-1}W_n \cup \cdots \cup (T^{(k+1)}_{\alpha_n})^{-1}W_n \cup \subseteq W_{n-1}$ for all $n \geqslant 1$, and
(c) each W_n, $n \geqslant 1$ is contained in one of the sets $S_1V, \ldots, S_rV$.

To define W_1, apply the induction assumption to the nonempty open set $W_0 = V$ and IP-systems

$$\left\{\left(T^{(k+1)}_{\alpha}\right)^{-1}T^{(1)}_{\alpha}\right\}_{\alpha\in\mathcal{F}}, \ldots, \left\{\left(T^{(k+1)}_{\alpha}\right)^{-1}T^{(k)}_{\alpha}\right\}_{\alpha\in\mathcal{F}}.$$

There exists $\alpha_1 \in \mathcal{F}$ such that

$$\begin{aligned} &V \cap \left(T^{(k+1)}_{\alpha_1}\right)^{-1}T^{(1)}_{\alpha_1}V \cap \cdots \cap \left(T^{(k+1)}_{\alpha_1}\right)^{-1}T^{(k)}_{\alpha_1}V \\ &\quad = W_0 \cap \left(T^{(k+1)}_{\alpha_1}\right)^{-1}T^{(1)}_{\alpha_1}W_0 \cap \cdots \cap \left(T^{(k+1)}_{\alpha_1}\right)^{-1}T^{(k)}_{\alpha_1}W_0 \neq \emptyset. \end{aligned}$$

Applying $T^{(k+1)}_{\alpha_1}$, we get

$$T^{(k+1)}_{\alpha_1}W_0 \cap T^{(1)}_{\alpha_1}W_0 \cap \cdots \cap T^{(k)}_{\alpha_1}W_0 \neq \emptyset.$$

It follows that for some $1 \leqslant t_1 \leqslant r$

$$W_1 := T^{(1)}_{\alpha_1}W_0 \cap T^{(2)}_{\alpha_1}W_0 \cap \cdots \cap T^{(k+1)}_{\alpha_1}W_0 \cap S_{t_1}V \neq \emptyset.$$

Clearly, W_0 and W_1 satisfy (b) and (c) above for $n = 1$.

If W_{n-1} and $\alpha_{n-1} \in \mathcal{F}$ have already been defined, apply the induction assumption to the nonempty open set W_{n-1} (and the IP-systems $\{(T^{(k+1)}_{\alpha})^{-1}T^{(1)}_{\alpha}\}_{\alpha\in\mathcal{F}}, \ldots,$ $\{(T^{(k+1)}_{\alpha})^{-1}T^{(k)}_{\alpha}\}_{\alpha\in\mathcal{F}}$) to get $\alpha_n > \alpha_{n-1}$ such that

$$W_{n-1} \cap \left(T^{(k+1)}_{\alpha_n}\right)^{-1}T^{(1)}_{\alpha_n}W_{n-1} \cap \cdots \cap \left(T^{(k+1)}_{\alpha_n}\right)^{-1}T^{(1)}_{\alpha_n}W_{n-1} \neq \emptyset,$$

and hence, for some $1 \leqslant t_n \leqslant r$,

$$W_n := T^{(1)}_{\alpha_n}W_{n-1} \cap \cdots \cap T^{(k+1)}_{\alpha_n}W_{n-1} \cap S_{t_n}V \neq \emptyset.$$

Again, this W_n clearly satisfies the conditions (b) and (c).

Since, by the construction, each W_n is contained in one of the sets $S_1V, \ldots, S_rV$, there is $1 \leqslant t \leqslant r$ such that infinitely many of the W_n are contained in S_tV. In particular, there exists i as large as we please and $j > i$ so that $W_i \cup W_j \subseteq S_tV$. Let $U = S_t^{-1}W_j$ and $\alpha = \alpha_{i+1} \cup \cdots \cup \alpha_j$.

Notice that $U \subseteq V$, and for any $1 \leqslant m \leqslant k+1$, $(T^{(m)}_{\alpha})^{-1}U \subseteq V$. Indeed,

$$\begin{aligned} \left(T^{(m)}_{\alpha}\right)^{-1}U &= \left(T^{(m)}_{\alpha_{i+1}\cup\cdots\cup\alpha_j}\right)^{-1}S_t^{-1}W_j \\ &= S_t^{-1}\left(T^{(m)}_{\alpha_{i+1}}\right)^{-1}\cdots\left(T^{(m)}_{\alpha_j}\right)^{-1}W_j \\ &\subseteq S_t^{-1}\left(T^{(m)}_{\alpha_{i+1}}\right)^{-1}\cdots\left(T^{(m)}_{\alpha_{j-1}}\right)^{-1}W_{j-1} \subseteq \cdots \\ &\subseteq S_t^{-1}\left(T^{(m)}_{\alpha_{i+1}}\right)^{-1}W_{i+1} \subseteq S_t^{-1}W_i \subseteq V. \end{aligned}$$

It follows that $U \cup (T_\alpha^{(1)})^{-1}U \cup \cdots \cup (T_\alpha^{(n+1)})^{-1}U \subseteq V$, and this, in turn, implies $V \cap T_\alpha^{(1)}V \cap \cdots \cap T_\alpha^{(k+1)}V \neq \emptyset$. □

COROLLARY 2.3. *If X is a compact metric space and G a commutative group of its homeomorphisms, then for any k IP-systems $\{T_\alpha^{(1)}\}_{\alpha\in\mathcal{F}}, \ldots, \{T_\alpha^{(k)}\}_{\alpha\in\mathcal{F}}$ in G, any $\alpha_0 \in \mathcal{F}$, and any $\varepsilon > 0$ there exist $\alpha > \alpha_0$ and $x \in X$ such that the diameter of the set $\{x, T_\alpha^{(1)}x, \ldots, T_\alpha^{(k)}x\}$ is smaller than ε.*

PROOF. If (X, G) is minimal, then the claim follows immediately from Theorem 2.2. If not, then pass to a minimal, nonempty, closed G-invariant subset of X. (Such a subset always exists by Zorn's lemma.) □

COROLLARY 2.4. *Under the conditions of Corollary* 2.3, *one can find, for any $m \in \mathbb{N}$, finite sets $\alpha_1 < \alpha_2 < \cdots < \alpha_m$ and $x \in X$ such that x and all the points $T_{\alpha_1}^{(i_1)}x, T_{\alpha_2}^{(i_2)}x, \ldots, T_{\alpha_m}^{(i_m)}x$, $i_1, \ldots, i_m \in \{1, 2, \ldots, k\}$, belong to the same open ball of radius ε.*

PROOF. The result follows by simple iteration. Assume that the group generated by $T_\alpha^{(i)}$, $i = 1, 2, \ldots, k$, acts on X in a minimal fashion. Let V be an open ball of radius ε. By Theorem 2.2, for any $\alpha_0 \in \mathcal{F}$, there exists $\alpha_1 > \alpha_0$ such that

$$V_1 = V \cap \bigcap_{i=1}^{k} \left(T_{\alpha_1}^{(i)}\right)^{-1} V \neq \emptyset.$$

Applying Theorem 2.2 again, one gets $\alpha_2 > \alpha_1$ such that

$$V_2 = V_1 \cap \bigcap_{i=1}^{k} \left(T_{\alpha_1}^{(i)}\right)^{-1} V_1 = V \cap \bigcap_{i_1,i_2=1}^{k} \left(T_{\alpha_1}^{i_1} T_{\alpha_2}^{i_2}\right)^{-1} V \neq \emptyset.$$

Let V_k be the nonempty set obtained as the result of k iterations of this procedure. It is easy to see that any $x \in V_k$ satisfies the claim of the corollary. □

The following corollary of Theorem 2.2 is a refinement of the multidimensional van der Waerden theorem.

COROLLARY 2.5. *For any $r, d, k \in \mathbb{N}$, any IP sets $(n_\alpha^{(1)})_{\alpha\in\mathcal{F}}$, $(n_\alpha^{(2)})_{\alpha\in\mathcal{F}}, \ldots, (n_\alpha^{(k)})_{\alpha\in\mathcal{F}}$ in $\mathbb{N}$, any finite set $F = \{u_1, u_2, \ldots, u_k\} \subset \mathbb{Z}^d$ and any partition $\mathbb{Z}^d = \bigcup_{i=1}^r C_i$, there exist $i \in \{1, 2, \ldots, r\}$, $\alpha \in \mathcal{F}$ and $v \in \mathbb{Z}^d$ such that*

$$v + \left\{n_\alpha^{(1)}u_1, n_\alpha^{(2)}u_2, \ldots, n_\alpha^{(k)}u_k\right\} \subset C_i.$$

REMARK. Taking $d = 1$, all $(n_\alpha^{(i)})_{\alpha\in\mathcal{F}}$ identical and $F = \{0, 1, \ldots, k\}$, one obtains the fact that for any finite coloring $\mathbb{N} = \bigcup_{i=1}^r C_i$ and any $k \in \mathbb{N}$, the set $\{n \in \mathbb{N}$: for some $a \in \mathbb{Z}$, $\{a, a+n, \ldots, a+(k-1)n\}$ is monochromatic$\}$ is IP*.

PROOF. Notice that Corollary 2.3 implies that for any commuting homeomorphisms $T_1, T_2, \ldots, T_k$ of a compact space X, any $\varepsilon > 0$ and any IP sets $(n_\alpha^{(1)})_{\alpha\in\mathcal{F}}, (n_\alpha^{(2)})_{\alpha\in\mathcal{F}}, \ldots, (n_\alpha^{(k)})_{\alpha\in\mathcal{F}}$ in $\mathbb{N}$, there exists $x \in X$ and $\alpha \in \mathcal{F}$ such that $\operatorname{diam}\{x, T_1^{n_\alpha^{(1)}} x, \ldots, T_k^{n_\alpha^{(k)}} x\} < \varepsilon$. The desired combinatorial result follows now by the argument which is practically identical to one used in the Introduction in the derivation of Theorem 1.20 from Theorem 1.21. □

Let us show that Theorem 1.29 is also derivable from Theorem 2.2. We shall find it more convenient to deal with the following equivalent form of Theorem 1.29.

THEOREM 2.6. *Let F be a finite field and V_F an infinite vector space over F. For any finite coloring $V_F = \bigcup_{i=1}^r C_i$ and any $m \in \mathbb{N}$, there exists a monochromatic affine m-space, that is, an m-dimensional affine subspace.*

Before embarking on the proof of Theorem 2.6, let us briefly explain why Theorems 1.29 and 2.6 are equivalent. Clearly one has only to show that Theorem 2.6 implies Theorem 1.29. This follows from the compactness of the space of r-colorings of V_F. Assuming without loss of generality that V_F is countably infinite, observe that, as an Abelian groups, V_F is isomorphic to the direct sum F_∞ of countably many copies of F:

$$F_\infty = \big\{g = (a_1, a_2, \ldots)\colon\ a_i \in F \text{ and all but finitely many } a_i = 0\big\} = \bigcup_{n=1}^{\infty} F_n,$$

where $F_n = \{g = (a_1, a_2, \ldots)\colon\ a_i = 0 \text{ for } i > n\} \cong F \oplus \cdots \oplus F$ (n times).

For $g = (a_1, a_2, \ldots) \in F_\infty$, let $|g|$ be the minimal natural number such that $a_i = 0$ for all $i \geqslant |g|$. Note that $|g| = 0$ if and only if $g = \mathbf{0} = (0, 0, \ldots)$. We will identify the space of r-colorings of V_F with $\Omega = \{1, 2, \ldots, r\}^{F_\infty}$. For any pair $x = x(g)$, $y = y(g)$, $g \in F_\infty$, of elements of Ω, let

$$\rho(x, y) = \inf_{n\in\mathbb{N}} \left\{\frac{1}{n}\colon\ x(g) = y(g) \text{ for } g \text{ with } |g| < n\right\}.$$

One readily checks that ρ is a metric on Ω with the property $\rho(x, y) = 1 \Leftrightarrow x(\mathbf{0}) \neq y(\mathbf{0})$. Moreover, (Ω, ρ) is a compact space, and it is the compactness of (Ω, ρ) which, as we shall now see, is behind the fact that Theorem 2.6 implies Theorem 1.29.

Assume that Theorem 2.6 holds true but Theorem 1.29 does not. Then, there exist $r, m \in \mathbb{N}$ such that for any $n \in \mathbb{N}$, there exists an r-coloring $F_n = \bigcup_{i=1}^r C_i$ with no monochromatic affine m-subspace. Viewing each such coloring as a map $f_n : F_n \to \{1, 2, \ldots, r\}$ and extending f_n, for each $n \in \mathbb{N}$, arbitrarily to a map $g_n : F_\infty \to \{1, 2, \ldots, r\}$, we obtain the sequence $(g_n)_{n\in\mathbb{N}}$ of elements of the compact space $\{1, 2, \ldots, r\}^{F_\infty}$, which, by compactness, has a convergent subsequence $(g_{n_i})_{i\in\mathbb{N}}$. The limiting coloring $g = \lim_{i\to\infty} g_{n_i}$ will also not have monochromatic affine m-subspaces, which contradicts Theorem 2.6.

PROOF OF THEOREM 2.6. Fix m IP-systems $\{g_\alpha^{(i)}\}_{\alpha\in\mathcal{F}}$, $i = 1, 2, \ldots, m$, such that, for each i, $\operatorname{Span}\{g_\alpha^{(i)}, \alpha \in \mathcal{F}\}$ is an infinite subset in V_F. We will show a stronger fact that,

for any partition $V_F = \bigcup_{i=1}^{r} C_i$, one of the C_i contains an affine m-space of the form $h + \mathrm{Span}\{g_1, \dots, g_m\}$ with $g_j \in \{g_\alpha^{(j)}\}_{\alpha\in\mathcal{F}}$. In other words, we will show that the set of ordered m-tuples $(g_1, \dots, g_m)$ such that, for some h, $h + \mathrm{Span}\{g_1, \dots, g_m\} \subset C_i$ is an IP* set in the m-fold direct sum $F_\infty \oplus \cdots \oplus F_\infty$ (where the notion of IP* is defined in the obvious sense).

We start with showing that one can always find $\alpha_1 \in \mathcal{F}$ and $i \in \{1, 2, \dots, r\}$ so that the one-dimensional affine subspace $h + \{cg_{\alpha_1}^{(i)}, c \in F\}$ is contained in C_i. For $h \in V_F$, let $T_h : \Omega \to \Omega$ be defined by $(T_h x)(g) = x(g + h)$. Clearly T_h is a homeomorphism of Ω for every $h \in V_F$. Let $\xi \in \Omega$ be the element in Ω corresponding to the partition $V_F = \bigcup_{i=1}^{r} C_i$, i.e. $\xi(g) = i \Leftrightarrow g \in C_i$. Finally, let $X \subseteq \Omega$ be the orbital closure of $\xi(g)$: $X = \overline{\{T_h\xi, h \in V_f\}}$.

Use now the IP system $\{g_\alpha^{(1)}\}_{\alpha\in\mathcal{F}}$ to define, for every $c \in F$, $c \neq 0$, an IP system of homeomorphisms $T_\alpha^{(c)} := T_{cg_\alpha^{(1)}}$, $\alpha \in \mathcal{F}$. In this way, we get $|F| - 1$ IP systems of commuting homeomorphisms of X. Applying Corollary 2.3 to the space X and the IP systems $T_\alpha^{(c)}$ and taking $\varepsilon < 1$, we get a point $x_1 \in X$ and $\alpha_1 \in \mathcal{F}$ such that the diameter of $\{T_{cg_{\alpha_1}^{(1)}}, c \in F\}$ is less than 1. This implies $x_1(0) = x_1(cg_{\alpha_1}^{(1)})$ for every $c \in F$. Since the orbit $\{T_h\xi, h \in V_F\}$ is dense in X, there exists $h_1 \in V_F$ such that $(T_{h_1}\xi)(g)$ and $x_1(g)$ agree on all g satisfying $|g| \leqslant |g_{\alpha_1}^{(1)}|$. If $\xi(h_1) = i$, then C_i contains the affine line $h_1 + \{cg_{\alpha_1}^{(1)}, c \in F\}$. (We, of course, took care in choosing α_1 so that $g_{\alpha_1}^{(1)} \neq \mathbf{0}$, which is possible in view of our assumptions on $\{g_\alpha^{(i)}\}_{\alpha\in\mathcal{F}}$.) Introducing now the IP systems $T_{cg_\alpha^{(2)}}$, $c \in F$, $c \neq 0$, and applying Corollary 2.4, we will find $x_2 \in X$ and $\alpha_2 > \alpha_1$ such that

$$\mathrm{diam}\{T_{c_1 g_{\alpha_1}^{(1)} + c_2 g_{\alpha_2}^{(2)}} x_2 \colon c_1, c_2 \in F\} < 1$$

(again, our assumption allows us to choose α_2 so that $g_{\alpha_1}^{(1)}$ and $g_{\alpha_2}^{(2)}$ are linearly independent in V_F). Similarly to the argument above, it follows now that for some $h_2 \in V_F$ the affine 2-space $h_2 + \mathrm{Span}\{g_{\alpha_1}^{(1)}, g_{\alpha_2}^{(2)}\}$ is monochromatic. After repeating this procedure $m - 2$ more times, we will get the desired monochromatic affine m-space. □

2.3. *A simultaneous proof of van der Waerden and Hales–Jewett theorems*

We are going now to give still another proof of van der Waerden's theorem. This proof has the advantage that, when properly interpreted, it gives also a proof of the Hales–Jewett theorem. To stress the affinity between the van der Waerden theorem and that of Hales–Jewett, this "double" proof is given in two parallel columns having many identical portions. To ease the presentation and to emphasize the correspondence between the number-theoretical and set-theoretical notions, we will abide by the following notational agreement: "$+$" will be used both for addition in $\mathbb{N}$ and for operation of taking (disjoint) unions of sets, "$-$" will be used not only for subtraction in $\mathbb{N}$ (when the minuend is not smaller than the subtrahend) but also instead of the set-theoretical difference "$\setminus$" in expressions of the form $A \setminus B$ where $B \subseteq A$. The sign "$\cdot$" will be used for both the multiplication in $\mathbb{N}$ and for the operation of

taking the Cartesian products (and will often be omitted). The sign "$\preceq$" will mean either the usual inequality "$\leqslant$" or the set-theoretical containment "$\subseteq$". "0" will mean either zero or the empty set $\emptyset$. For any set E, $\mathcal{F}(E)$ will mean the set of finite subsets (including the empty set) of E.

Let (X, ρ) be a compact metric space. Let $q \in \mathbb{N}$.

Denote the set of nonnegative integers by $\mathbb{F}$. Let T be a continuous self-mapping of X. Let A be a set consisting of q pairwise distinct natural numbers

$$A = \{p_i \in \mathbb{N}\colon\ i = 1, \dots, q\};$$

assume without loss of generality that $p_1 < p_2 < \cdots < p_q$.

Let S be an infinite set, denote $\mathcal{F}(S)$ by $\mathbb{F}$. Let $V = \{1, \dots, q\} \times S$ and let $(T^a)_{a \in \mathcal{F}(V)}$ be an action of $\mathcal{F}(V)$ on X. (That is, T is a mapping from $\mathcal{F}(V)$ into the set of continuous self-mappings of X satisfying the following condition: if $a \cap b = \emptyset$, then $T^{a \cup b} = T^a T^b$.) Put $p_i = \{1, \dots, i\}$, $i = 1, \dots, q$, and $a = \{p_1, \dots, p_q\}$.

We are going to prove the following (two) proposition(s):

PROPOSITION 2.7 [24, Proposition L]. *For any $\varepsilon > 0$ there exists $N \in \mathbb{F}$, such that for any $x \in X$ there exist $n \preceq N$, $n \neq 0$, and $a \preceq p_q(N - n)$ such that for any $p \in A$,*

$$\rho(T^{a+pn}x, T^a x) < \varepsilon.$$

REMARK 2.8. Let us show how Proposition 2.7 implies the "classical" Hales–Jewett theorem. (See Theorem 1.26 above.) First, we pass to the combinatorial version of Proposition 2.7: let $r, q \in \mathbb{N}$; there exists $M \in \mathbb{N}$ such that for $N = \{1, \dots, M\}$ and $V = \{1, \dots, q\} \times N$, given an r-coloring of $\mathcal{F}(V)$ one can find a nonempty $n \preceq N$ and $a \preceq \{1, \dots, q\} \times (N - n)$ such that the set $L = \{a \cup (\{1\} \times n), a \cup (\{1, 2\} \times n), \dots, a \cup (\{1, \dots, q\} \times n)\}$ is monochromatic. Second, we identify $\mathcal{F}(V)$ with $\mathcal{F}(\{1, \dots, q\})^M$, $D \leftrightarrow (D_1, \dots, D_M)$ where $D_i = D \cap (\{1, \dots, q\} \times \{j\})$, $j = 1, \dots, M$, and define a mapping φ from $\mathcal{F}(V)$ to the "M-dimensional cube" $Q = \{0, \dots, q\}^M$ by $\varphi(D_1, \dots, D_M) = (|D_1|, \dots, |D_M|)$. Now, any r-coloring of Q induces an r-coloring of $\mathcal{F}(V)$, and the φ-image of a monochromatic set L as in the assertion above is just a monochromatic line in Q.

PROOF. We will prove this proposition by induction on q. Define B by

$$B = \{p_i - p_1,\ i = 2, \dots, q\}.$$

Since B contains $q - 1$ elements, we may assume that the statement to prove is valid for B, that is, for any $\varepsilon > 0$, there exists $N \in \mathbb{F}$ such that for any $x \in X$ there exist $n \preceq N$, $a \preceq (p_q - p_1)(N - n)$, such that $n \neq 0$ and for every $r \in B$ one has $\rho(T^{a+rn}x, T^a x) < \varepsilon$.

Let $\varepsilon > 0$. Let $k \in \mathbb{N}$ be such that among any $k + 1$ points of X there are two points at a distance less than $\varepsilon/2$.

Put $\varepsilon_0 = \varepsilon/2k$. By the induction hypothesis, there exists $N_0 \in \mathcal{F}$ such that for any $x \in X$ there exist $n \preceq N_0$ and $a \preceq (p_q - p_1) \cdot (N_0 - n)$ such that $n \neq 0$ and for every $r \in B$ one has $\rho(T^{a+rn}x, T^a x) < \varepsilon_0$.

Let $\varepsilon_1 > 0$ be such that the inequality $\rho(y_1, y_2) < \varepsilon_1$ implies the inequality

$$\rho\big(T^b y_1, T^b y_2\big) < \varepsilon/2k$$

for any $b \preceq p_q N_0$. Let $N_1 \in \mathcal{F}$ be such that $N_1 \cap N_0 = 0$ (this disjointness condition concerns the part of Proposition 2.7 dealing with the Hales–Jewett theorem only) and for any $x \in X$ there exist $n \preceq N_1$ and $a \preceq (p_q - p_1)(N_1 - n)$ such that $n \neq 0$ and for every $r \in B$ one has $\rho(T^{a+rn}x, T^a x) < \varepsilon_1$.

Continue this process: assume that $\varepsilon_0, \dots, \varepsilon_i$ and $N_0, \dots, N_i \in \mathcal{F}$ have been already chosen. Let $\varepsilon_{i+1} > 0$ be such that the inequality $\rho(y_1, y_2) < \varepsilon_{i+1}$ implies the inequality

$$\rho\big(T^b y_1, T^b y_2\big) < \varepsilon/2k$$

for any $b \preceq p_q(N_0 + \cdots + N_i)$. Let $N_{i+1} \in \mathcal{F}$ be such that $N_{i+1} \cap (N_0 \cup \cdots \cup N_i) = 0$ (again, this disjointness condition is relevant for the Hales–Jewett part of Proposition 2.7 only) and for any $x \in X$ there exist $n \preceq N_{i+1}$ and $a \preceq (p_q - p_1)(N_{i+1} - n)$, such that $n \neq 0$ and for every $r \in B$ one has $\rho(T^{a+rn}x, T^a x) < \varepsilon_{i+1}$.

Continue the process of choosing ε_i, N_i up to $i = k$, and put $N = N_0 + \cdots + N_k$.

Now fix an arbitrary point $x \in X$.

Applying the definition of N_k to the point

$$y_k = T^{p_1 N_k} x,$$

find $n_k \preceq N_k$, $n_k \neq 0$, and $a_k \preceq (p_q - p_1)(N_k - n_k)$ such that for every $r \in B$ we have

$$\rho\big(T^{rn_k + a_k} y_k, T^{a_k} y_k\big) < \varepsilon_k.$$

Then, applying the definition of N_{k-1} to the point

$$y_{k-1} = T^{p_1(N_k + N_{k-1}) + a_k} x,$$

find $n_{k-1} \preceq N_{k-1}$, $n_{k-1} \neq 0$, and $a_{k-1} \preceq (p_q - p_1)(N_{k-1} - n_{k-1})$ such that for every $r \in B$ we have

$$\rho\big(T^{rn_{k-1} + a_{k-1}} y_{k-1}, T^{a_{k-1}} y_{k-1}\big) < \varepsilon_{k-1}.$$

Continue this process: suppose that we have already found $n_k, \dots, n_i$, $a_k, \dots, a_i$. Applying the definition of N_{i-1} to the point

$$y_{i-1} = T^{p_1(N_k + \cdots + N_{i-1}) + a_k + \cdots + a_i} x,$$

find $n_{i-1} \preceq N_{i-1}$, $n_{i-1} \neq 0$, and $a_{i-1} \preceq (p_q - p_1)(N_{i-1} - n_{i-1})$ such that for every $r \in B$ we have

$$\rho\big(T^{rn_{i-1}+a_{i-1}} y_{i-1}, T^{a_{i-1}} y_{i-1}\big) < \varepsilon_{i-1}.$$

Continue the process of choosing n_i, a_i up to $i = 0$.

For every $0 \leqslant i \leqslant k$ we have $0 \neq n_i \preceq N_i$ and $a_i \preceq (p_q - p_1)(N_i - n_i)$; besides, $N_i \cap N_l = 0$ for $i \neq l$. Therefore, for any $0 \leqslant i \leqslant j \leqslant k$ we have

$$\begin{aligned}
&p(n_j + \cdots + n_i) + a_j + \cdots + a_0 \\
&\quad + p_1(N_j + \cdots + N_0 - n_j - \cdots - n_0) \\
&\preceq p_q(n_j + \cdots + n_{i+1}) \\
&\quad + (p_q - p_1)(N_j - n_j + \cdots \\
&\quad + N_0 - n_0) \\
&\quad + p_1(N_j + \cdots + N_0 - n_j - \cdots - n_0) \\
&= p_q(N_0 + \cdots + N_j).
\end{aligned} \tag{2.1}$$

$$\begin{aligned}
&\big(p(n_j + \cdots + n_i) + a_j + \cdots + a_0 \\
&\quad + p_1(N_j + \cdots + N_0 - n_j - \cdots - n_0)\big) \\
&\quad \cap \big(a_{j+1} + (p - p_1)n_{j+1}\big) = 0.
\end{aligned} \tag{2.2}$$

And, for any $0 \leqslant j \leqslant k$,

$$\begin{aligned}
&a_k + \cdots + a_0 + p_1(N_k + \cdots + N_0 - n_j - \cdots - n_0) \\
&\quad \preceq (p_q - p_1)(N_k - n_k + \cdots + N_0 - n_0) \\
&\qquad + p_1(N_k + \cdots + N_0 - n_j - \cdots - n_0) \\
&\quad \preceq p_q(N_k + \cdots + N_0 - n_j - \cdots - n_0).
\end{aligned} \tag{2.3}$$

Define points x_i, $i = 0, \ldots, k$, by

$$x_i = T^{a_k + \cdots + a_0 + p_1(N_k + \cdots + N_0 - n_i - \cdots - n_0)} x.$$

We are going to show that for any $0 \leqslant i \leqslant j \leqslant k$ and any $p \in A$,

$$\rho\big(T^{p(n_j + \cdots + n_{i+1})} x_j, x_i\big) \leqslant \frac{\varepsilon}{2k}(j - i). \tag{2.4}$$

We will prove this by induction on $j - i$; when $j = i$ the statement is trivial. We will derive the validity of (2.4) for i, j, where $i < j$, from its validity for i, $j - 1$.

By the definition of n_j,

$$\rho\big(T^{a_j + (p - p_1)n_j} y_j, T^{a_j} y_j\big) < \varepsilon_j,$$

where

$$y_j = T^{p_1(N_k+\cdots+N_j)+a_k+\cdots+a_{j+1}}x.$$

So, by the choice of ε_j (and (2.1)),

$$\rho\big(T^{p(n_{j-1}+\cdots+n_{i+1})+a_{j-1}+\cdots+a_0+p_1(N_{j-1}+\cdots+N_0-n_{j-1}-\cdots-n_0)}T^{a_j+(p-p_1)n_j}y_j,$$
$$T^{p(n_{j-1}+\cdots+n_{i+1})+a_{j-1}+\cdots+a_0+p_1(N_{j-1}+\cdots+N_0-n_{j-1}-\cdots-n_0)}T^{a_j}y_j\big) < \varepsilon/2k.$$

Using the definition of y_j, x_j (and (2.2)), we see

$$\begin{aligned}&T^{p(n_{j-1}+\cdots+n_{i+1})+a_{j-1}+\cdots+a_0+p_1(N_{j-1}+\cdots+N_0-n_{j-1}-\cdots-n_0)}T^{a_j+(p-p_1)n_j}y_j\\ &\quad= T^{p(n_j+\cdots+n_{i+1})+a_k+\cdots+a_0+p_1(N_k+\cdots+N_0-n_j-\cdots-n_0)}x\\ &\quad= T^{p(n_j+\cdots+n_{i+1})}x_j\end{aligned}$$

and

$$\begin{aligned}&T^{p(n_{j-1}+\cdots+n_{i+1})+a_{j-1}+\cdots+a_0+p_1(N_{j-1}+\cdots+N_0-n_{j-1}-\cdots-n_0)}T^{a_j}y_j\\ &\quad= T^{p(n_{j-1}+\cdots+n_{i+1})+a_k+\cdots+a_0+p_1(N_k+\cdots+N_0-n_{j-1}-\cdots-n_0)}x\\ &\quad= T^{p(n_{j-1}+\cdots+n_{i+1})}x_{j-1}.\end{aligned}$$

Since, by the induction hypothesis,

$$\rho\big(T^{p(n_{j-1}+\cdots+n_{i+1})}x_{j-1}, x_i\big) \leqslant \frac{\varepsilon}{2k}(j-i-1),$$

we obtain (2.4).

By the choice of k, among the $k+1$ points $x_0, \ldots, x_k$ there are two, say x_i, x_j, $0 \leqslant i < j \leqslant k$, for which $\rho(x_i, x_j) < \varepsilon/2$. Put

$$\begin{aligned}n &= n_j + \cdots + n_{i+1},\\ a &= a_k + \cdots + a_0 + p_1(N_k + \cdots + N_0 - n_j - \cdots - n_0).\end{aligned}$$

Then $x_j = T^a x$ and

$$\begin{aligned}\rho(T^{a+pn}x, T^a x) &= \rho(T^{pn}x_j, x_j)\\ &\leqslant \rho(T^{pn}x_j, x_i) + \rho(x_j, x_i) < \varepsilon(j-i)/2k + \varepsilon/2 \leqslant \varepsilon.\end{aligned}$$

Furthermore, $n \preceq N$, $n \neq 0$ and $a \preceq p_q(N-n)$ by (2.3). This proves Proposition 2.7. □

2.4. *Polynomial van der Waerden theorem*

We shall formulate now a polynomial extension of the multidimensional van der Waerden's theorem which was obtained in [23]. We leave it to the reader to formulate the combinatorial equivalent of this result. (See also Theorems 2.12 and 2.14 below.)

THEOREM 2.9. *Let (X, ρ) be a compact metric space, let $T_1, \dots, T_t$ be commuting homeomorphisms of X and let $p_{i,j}$, $i = 1, \dots, k$, $j = 1, \dots, t$, be polynomials taking on integer values on the integers and vanishing at zero. Then, for any positive ε, there exist $x \in X$ and $n \in \mathbb{N}$ such that*

$$\rho\big(T_1^{p_{i,1}(n)} T_2^{p_{i,2}(n)} \cdots T_t^{p_{i,t}(n)} x, x\big) < \varepsilon \tag{2.5}$$

for all $i = 1, \dots, k$ simultaneously. Moreover, the set $\{n \in \mathbb{Z}$: $\forall \varepsilon > 0, \exists x \in X$ such that $\forall i \in \{1, 2, \dots, k\}$, (2.5) is satisfied$\}$ is an IP* set.

We provide now a proof of a special case. Let (X, ρ) be a compact metric space and let T be a homeomorphism of X. Let $\varepsilon > 0$; we will find $x \in X$ and $n \in \mathbb{N}$ such that $\rho(T^{n^2} x, x) < \varepsilon$.

Without loss of generality we will assume that the system (X, T) is minimal. We shall find a sequence $x_0, x_1, x_2, \dots$ of points of X and a sequence $n_1, n_2, \dots$ of natural numbers such that

$$\rho\big(T^{(n_m + \cdots + n_{l+1})^2} x_m, x_l\big) < \varepsilon/2 \quad \text{for every } l, m \in \mathbb{Z}_+,\ l < m \tag{2.6}$$

(where $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$). Since X is compact, for some $l < m$ one will have $\rho(x_m, x_l) < \varepsilon/2$; together with (2.6) this will give $\rho(T^{(n_m + \cdots + n_{l+1})^2} x_m, x_m) < \varepsilon$.

Choose $x_0 \in X$ arbitrarily and put $n_1 = 1$, $x_1 = T^{-n_1^2} x_0$. Let $\varepsilon_1 < \varepsilon/2$ be such that $\rho(T^{n_1^2} y, x_0) < \varepsilon/2$ for every y for which $\rho(y, x_1) < \varepsilon_1$. Using the "linear" van der Waerden theorem, find $y_1 \in X$ and $n_2 \in \mathbb{N}$ such that $\rho(y_1, x_1) < \varepsilon_1/2$ and $\rho(T^{2n_1 n_2} y_1, y_1) < \varepsilon_1/2$. Put $x_2 = T^{-n_2^2} y_1$; then

$$\rho\big(T^{n_2^2} x_2, x_1\big) = \rho(y_1, x_1) < \varepsilon_1/2 < \varepsilon/2;$$

also,

$$\rho\big(T^{2n_1 n_2 + n_2^2} x_2, x_1\big) \leqslant \rho\big(T^{2n_1 n_2} y_1, y_1\big) + \rho(y_1, x_1) < \varepsilon_1$$

and, hence, by the choice of ε_1,

$$\rho\big(T^{(n_1 + n_2)^2} x_2, x_0\big) = \rho\big(T^{n_1^2} T^{2n_1 + n_2^2} x_2, x_0\big) < \varepsilon/2.$$

Suppose that x_m, n_m have been found; let us find x_{m+1}, n_{m+1}. Choose ε_m, $0 < \varepsilon_m < \varepsilon/2$, guaranteeing the implication

$$\rho(y, x_m) < \varepsilon_m \Longrightarrow \rho\big(T^{(n_m + \cdots + n_{l+1})^2} y, x_l\big) < \varepsilon/2, \quad l = 0, \dots, m-1,$$

and (using the linear van der Waerden theorem) find y_m, n_{m+1} such that

$$\rho(y_m, x_m) < \varepsilon_m/2,$$
$$\rho\big(T^{2(n_m+\cdots+n_{l+1})n_{m+1}} y_m, y_m\big) < \varepsilon_m/2, \quad l = 0, \ldots, m-1.$$

Putting $x_{m+1} = T^{-n_{m+1}^2} y_m$, we obtain

$$\rho\big(T^{2(n_m+\cdots+n_{l+1})n_{m+1}+n_{m+1}^2} x_{m+1}, x_m\big)$$
$$\leqslant \rho\big(T^{2(n_m+\cdots+n_{l+1})n_{m+1}} y_m, y_m\big) + \rho(y_m, x_m) < \varepsilon_m, \quad l = 0, \ldots, m-1,$$

and, hence, by the choice of ε_m,

$$\rho\big(T^{n_{m+1}^2} x_{m+1}, x_m\big) < \varepsilon/2$$

and

$$\rho\big(T^{(n_{m+1}+\cdots+n_{l+1})^2} x_{m+1}, x_l\big) < \varepsilon/2 \quad \text{for } l = 0, \ldots, m-1.$$

REMARK. We leave it to the reader to check that the proof above shows actually that a number n with the property that, for some x, $\rho(T^{n^2}x, x) < \varepsilon$ can be chosen from any IP set.

2.5. *Polynomial Hales–Jewett theorem*

We are going to formulate now the polynomial extension of the Hales–Jewett theorem which was obtained in [24]. Like its "linear" special case, the polynomial Hales–Jewett theorem has many equivalent formulations. The one we have chosen to present here is a natural extension of Proposition 2.7.

THEOREM 2.10. *Let (X, ρ) be a compact metric space. For fixed $d, q \in \mathbb{N}$, let $\mathcal{P}_N$ be the set of subsets of $\{1, 2, \ldots, N\}^d \times \{1, 2, \ldots, q\}$. Let $T(c)$, $c \in \mathcal{P}_N$, be a family of self-mappings of X such that $a \cap b = \emptyset$ implies $T(a \cup b) = T(a)T(b)$. Then for any $x \in X$ and any $\varepsilon > 0$ there exist $N \in \mathbb{N}$, $a \in \mathcal{P}_N$ and a nonempty set $\gamma \subseteq \{1, 2, \ldots, N\}$ such that $a \cap (\gamma^d \times \{1, 2, \ldots, q\}) \neq \emptyset$ and $\rho(T(a \cup (\gamma^d \times \{i\}))x, T(a)x) < \varepsilon$ for every $i = 1, 2, \ldots, q$.*

Here is the combinatorial version of Theorem 2.10. We leave it to the reader to verify that it is indeed equivalent to Theorem 2.10. (In one direction, the argument is similar to that in Remark 2.8 above. See also [9, pp. 45–47], and [24, Proposition 3.3].)

THEOREM 2.11 [24, Theorem PHJ]. *For any $r, d, q \in \mathbb{N}$ there exists $N = N(r, d, q)$ such that for any r-coloring of the set $\mathcal{P}_N$ of subsets of $\{1, 2, \ldots, N\}^d \times \{1, 2, \ldots, q\}$ there exist*

$a \in \mathcal{P}_N$ and a nonempty set $\gamma \subseteq \{1, 2, \dots, N\}$ with $a \cap (\gamma^d \times \{1, 2, \dots, q\}) = \emptyset$ and such that the sets

$$a, a \cup \big(\gamma^d \times \{1\}\big), a \cup \big(\gamma^d \times \{2\}\big), \dots, a \cup \big(\gamma^d \times \{q\}\big)$$

are all of the same color.

We will formulate now some corollaries of the polynomial Hales–Jewett theorem. The density versions of these results (or, rather, the ergodic counterparts of these density versions) will be discussed in Section 4. The following result extends and refines the polynomial van der Waerden theorem.

THEOREM 2.12 [24, Theorem 0.14]. *For any $t, m \in \mathbb{N}$, any polynomial mapping $P : \mathbb{Z}^t \to \mathbb{Z}^m$ satisfying $P(0) = 0$, any finite set $F \subset \mathbb{Z}^t$ and any finite coloring $\chi : \mathbb{Z}^m \to \{1, 2, \dots, r\}$, there is $l \in \{1, 2, \dots, r\}$ such that the set*

$$\big\{(n_1, \dots, n_t)\colon \text{ there is } a \in \mathbb{Z}^m \text{ such that } \chi\big(a + P(n_1 v_1, \dots, n_t v_t)\big) = l \text{ for all } v = (v_1, \dots, v_t) \in F\big\}$$

is an IP set in $\mathbb{Z}^t$.*

The following proposition corresponds to the special case $t = 2q$, $m = 1$, $P(n_1, k_1, n_2, k_2, \dots, n_q, k_q) = \sum_{i=1}^{q} n_i k_i$ and $F = \{(1, 1, 0, \dots, 0), (0, 0, 1, 1, 0, \dots, 0), \dots, (0, 0, \dots, 0, 0, 1, 1)\}$.

PROPOSITION 2.13. *Let $(n_i^{(1)})_{i\in\mathbb{N}}, (k_i^{(1)})_{i\in\mathbb{N}}, \dots, (n_i^{(q)})_{i\in\mathbb{N}}, (k_i^{(q)})_{i\in\mathbb{N}}$ be sequences in $\mathbb{Z}$ and let $(n_\alpha^{(1)})_{\alpha\in\mathcal{F}}, (k_\alpha^{(1)})_{\alpha\in\mathcal{F}}, \dots, (n_\alpha^{(q)})_{\alpha\in\mathcal{F}}, (k_\alpha^{(q)})_{\alpha\in\mathcal{F}}$ be (additive) IP sets generated by these sequences. Then for any finite coloring of $\mathbb{Z}$ there exists a monochromatic set of the form*

$$\big\{a, a + n_\gamma^{(1)} k_\gamma^{(1)}, a + n_\gamma^{(2)} k_\gamma^{(2)}, \dots, a + n_\gamma^{(q)} k_\gamma^{(q)}\big\}$$

for some $a \in \mathbb{N}$ and a finite nonempty $\gamma \subset \mathbb{N}$.

One can also derive from the polynomial Hales–Jewett theorem an analogue of the polynomial van der Waerden theorem which is valid in any commutative ring. The following corollary of Theorem 2.11 contains the combinatorial counterpart of Theorem 2.9 as a special case.

THEOREM 2.14 [24, Theorem 0.17]. *Let W and V be vector spaces over an infinite field K, let $P : W \to V$ be a polynomial mapping with $P(0) = 0$, and let $F \subset W$ be a finite set. If $V = \bigcup_{i=1}^{r} C_i$ is a finite coloring of V, then*:

(i) *There exist $a \in V$ and $n \in K$, $n \neq 0$ such that the set*

$$a + P(nF) = \big\{a + P(nv)\colon v \in F\big\}$$

is monochromatic.

(ii) *For some $l \in \{1, 2, \ldots, r\}$ the set*

$$\{n \in K\colon \text{there exists } a \in V \text{ such that } a + P(nv) \in C_i \text{ for all } v \in F\}$$

is an IP* set.

2.6. *Nilpotent van der Waerden theorem*

Observing that the various versions of van der Waerden's theorem are linked with recurrence theorems for commuting homeomorphisms of a compact metric space, one is naturally inclined to inquire whether these recurrence results can be generalized to a noncommutative situation. The answer, in general, is NO. (See [58, p. 40], and [20].) The following theorem, due to A. Leibman [97], shows that, when the homeomorphisms generate a nilpotent group, the answer is YES. Note that Leibman's theorem is, at the same time, a generalization of the polynomial van der Waerden theorem, Theorem 2.9 above.

THEOREM 2.15. *Let (X, ρ) be a compact metric space, let homeomorphisms $T_1, \ldots, T_t$ of X generate a nilpotent group and let $p_{i,j}$, $i = 1, \ldots, k$, $j = 1, \ldots, t$, be polynomials taking on integer values on the integers and vanishing at zero. Then, for any positive ε, there exist $x \in X$ and $n \in \mathbb{N}$ such that $\rho(T_1^{p_{i,1}(n)} T_2^{p_{i,2}(n)} \cdots T_t^{p_{i,t}(n)} x, x) < \varepsilon$ for all $i = 1, \ldots, k$ simultaneously.*

To get the feeling of some of the ideas behind the proof of Theorem 2.15 let us consider the simplest noncommutative situation. Let (X, ρ) be a compact metric space and let homeomorphisms T and S of X do not commute but be such that $R = [T, S]$ commute with both T and S; the group G generated by T and S is then two-step nilpotent. Let $\varepsilon > 0$; our goal is to find $x \in X$ and $n \in \mathbb{N}$ such that both $\rho(T^n x, x) < \varepsilon$ and $\rho(S^n x, x) < \varepsilon$.

Without loss of generality we will assume that X is minimal with respect to the action of the group G. The sequence $S^n T^{-n}$, $n \in \mathbb{N}$, in G can be written as a "polynomial sequence" $(ST^{-1})^n R^{n(n-1)/2}$. Since the homeomorphisms ST^{-1} and R commute, we have, by the polynomial van der Waerden theorem that, for any $\delta > 0$ there exists $y \in X$ such that, for some $n \in \mathbb{N}$, one has $\rho(S^n T^{-n} y, y) < \delta$. We will first show that the set of points with this property is dense in X: for any open set $V \subseteq X$ we will find $y \in V$ such that, for some $n \in \mathbb{N}$, both $T^n y, S^n y \in V$. Since X is minimal with respect to the action of G and compact, there exist $P_1, \ldots, P_k \in G$ such that $\bigcup_{i=1}^k P_i^{-1} V = X$. Let δ be a Lebesgue number for this cover. For any $P = T^a S^b R^c \in G$ we have

$$\begin{aligned} P^{-1} S^n T^{-n} P = S^n T^{-n} [S^n T^{-n}, P] &= \left(ST^{-1}\right)^n R^{n(n-1)/2} [S, T]^{an} [T, S]^{-bn} \\ &= \left(ST^{-1}\right)^n R^{n(n-1)/2 - (a+b)n}, \quad n \in \mathbb{N}. \end{aligned}$$

All these "polynomial sequences" lie in the commutative group generated by ST^{-1} and R. Thus, by the polynomial van der Waerden theorem, there exist $z \in X$ and $n \in \mathbb{N}$ such that

$\rho(P_i^{-1}S^nT^{-n}P_iz, z) < \delta$ for all $i = 1, \dots, k$. Hence, there exists $i \in \{1, \dots, k\}$ such that $z \in P_i^{-1}V$ and $P_i^{-1}S^nT^{-n}P_iz \in P_i^{-1}V$, $i = 1, \dots, k$. It remains to put $y = P_iz$.

We will now construct a sequence $x_0, x_1, x_2, \dots$ of points of X and a sequence $n_1, n_2, \dots$ of positive integers such that

$$\rho\big(T^{n_m+\dots+n_{l+1}}x_m, x_l\big) < \varepsilon/2 \quad \text{and} \quad \rho\big(S^{n_m+\dots+n_{l+1}}x_m, x_l\big) < \varepsilon/2 \quad \text{for every } l, m \in \mathbb{Z}_+, l < m. \tag{2.7}$$

Since X is compact, for some $l < m$ we will have $\rho(x_m, x_l) < \varepsilon/2$; together with (2.7) this implies $\rho(T^{n_m+\dots+n_{l+1}}x_m, x_m) < \varepsilon$ and $\rho(S^{n_m+\dots+n_{l+1}}x_m, x_m) < \varepsilon$.

Using the polynomial van der Waerden theorem find a point $x_0 \in X$ and an integer $n_1 \in \mathbb{N}$ such that $\rho(S^{n_1}T^{-n_1}x_0, x_0) < \varepsilon/2$. Put $x_1 = T^{-n_1}x_0$, then $\rho(T^{n_1}x_1, x_0) = 0 < \varepsilon/2$ and $\rho(S^{n_1}x_1, x_0) < \varepsilon/2$.

Suppose now that x_m, n_m satisfying (2.7) have been found for all $m \leqslant k$; we will find x_{k+1}, n_{k+1}. Choose δ, $0 < \delta < \varepsilon/2$, such that $\rho(y, x_k) < \delta$ implies $\rho(T^{n_k+\dots+n_{l+1}}y, x_l) < \varepsilon/2$ and $\rho(S^{n_k+\dots+n_{l+1}}y, x_l) < \varepsilon/2$ for all $l = 0, \dots, k-1$. Find y_k in the $\delta/2$-neighborhood of x_k and $n_{k+1} \in \mathbb{N}$ such that $\rho(S^{n_{k+1}}T^{-n_{k+1}}y_k, y_k) < \delta/2$. Put $x_{k+1} = T^{-n_{k+1}}y_k$, then $\rho(T^{n_{k+1}}x_{k+1}, x_k) < \delta/2 < \varepsilon/2$ and $\rho(S^{n_{k+1}}x_{k+1}, x_k) < \delta < \varepsilon/2$. By the choice of δ this implies $\rho(T^{n_{k+1}+\dots+n_{l+1}}x_{k+1}, x_l) < \varepsilon/2$ and $\rho(S^{n_{k+1}+\dots+n_{l+1}}x_{k+1}, x_l) < \varepsilon/2$ for all $l = 0, \dots, k-1$.

2.7. *Nilpotent Hales–Jewett theorem*

We want to conclude this section by discussing a nilpotent version of the polynomial Hales–Jewett theorem, which was obtained in [26]. But first we want to give a formulation of a corollary of the polynomial Hales–Jewett theorem, which will be suggestive of the further, nilpotent generalization.

Write $\mathcal{F}' = \mathcal{F} \cup \emptyset$ (where $\mathcal{F}$, as before, denotes the set of finite nonempty subsets of $\mathbb{N}$). Let G be a commutative (semi)group. A mapping $P : \mathcal{F}' \to G$ is an *IP polynomial of degree* 0 if P is constant, and, inductively, is an *IP polynomial of degree* $\leqslant d$ if for any $\beta \in \mathcal{F}'$ there exists an IP polynomial $D_\beta P : \mathcal{F}'(\mathbb{N} \setminus \beta) \to G$ of degree $\leqslant d-1$ (where $\mathcal{F}'(\mathbb{N} \setminus \beta)$ is the set of finite subsets of $\mathbb{N} \setminus \beta$), such that $P(\alpha \cup \beta) = P(\alpha) \cup (D_\beta P)(\alpha)$ for every $\alpha \in \mathcal{F}'$ with $\alpha \cap \beta = \emptyset$. We have the following theorem.

THEOREM 2.16 [24]. *Let G be an Abelian group of self-homeomorphisms of a compact metric space (X, ρ) and let $P_1, P_2, \dots, P_k$ be IP polynomials mapping $\mathcal{F}'$ into G and satisfying $P_i(\emptyset) = 1_G$ for all $i \in \{1, \dots, k\}$. Then for any $\varepsilon > 0$ there exist $x \in X$ and a nonempty $\alpha \in \mathcal{F}'$ such that $\rho(P_i(\alpha)x, x) < \varepsilon$ for $i = 1, \dots, k$.*

It is proved in [24, Theorem 8.3], that if G is an Abelian group then a mapping $P : \mathcal{F}' \to G$ is an IP polynomial of degree $\leqslant d$ if and only if there exists a family $\{g_{(j_1,\dots,j_d)}\}_{(j_1,j_2,\dots,j_d)\in\mathbb{N}^d}$ of elements of G such that for any $\alpha \in \mathcal{F}'$ one has $P(\alpha) =$

$\prod_{(j_1,\dots,j_d)\in\alpha^d} g_{(j_1,\dots,j_d)}$. This characterization of IP polynomials makes sense in the nilpotent setup as well. Given a nilpotent group G, let us call a mapping $P:\mathcal{F}'\to G$ an IP polynomial if for some $d\in\mathbb{N}$ there exists a family $\{g_{(j_1,\dots,j_d)}\}_{(j_1,\dots,j_d)\in\mathbb{N}^d}$ of elements of G and a linear order $<$ on $\mathbb{N}^d$ such that for any $\alpha\in\mathcal{F}'$ one has $P(\alpha)=\prod^{<}_{(j_1,\dots,j_d)\in\alpha^d} g_{(j_1,\dots,j_d)}$. (The entries in the product $\prod^{<}$ are multiplied in accordance with the order $<$.) We can formulate now the nilpotent version of the polynomial Hales–Jewett theorem, which contains many results formulated above, in particular Theorems 1.29, 2.9 and 2.15, as special cases.

THEOREM 2.17 [26, Theorem 0.24]. *Let G be a nilpotent group of self-homeomorphisms of a compact metric space (X,ρ) and let $P_1,\dots,P_k:\mathcal{F}'\to G$ be polynomial mappings satisfying $P_1(\emptyset)=\cdots=P_k(\emptyset)=1_G$. Then, for any $\varepsilon>0$, there exist $x\in X$ and a nonempty $\alpha\in\mathcal{F}'$ such that $\rho(P_i(\alpha)x,x)<\varepsilon$ for all $i=1,2,\dots,k$.*

The following corollary of Theorem 2.17 can be viewed as the nilpotent generalization of Hilbert's theorem (see Theorems 1.8 and 1.12). It is worth noting that, unlike Hilbert's theorem which had an easy proof, the nilpotent version of it is far from being trivial.

THEOREM 2.18 [26, Theorem 5.5]. *Let G be an infinite nilpotent group. For any $k,r\in\mathbb{N}$ there exist $N\in\mathbb{N}$ such that for any $g_j^{(i)}\in G$, $1\leqslant i\leqslant k$, $1\leqslant j\leqslant N$, and any r-coloring of G there exist a nonempty set $\alpha\subseteq\{1,2,\dots,N\}$ and infinitely many $h\in G$ such that for $h_i=\prod_{j\in\alpha} g_j^{(i)}$, $i=1,\dots,k$ (where the entries are multiplied in the natural order of $j\in\mathbb{N}$), all the products $hh_{i_1}h_{i_2}\cdots h_{i_l}$ with $0\leqslant l\leqslant k$ and distinct $i_1,i_2,\dots,i_l$ are of the same color.*

Finally, we formulate a corollary of the nilpotent Hales–Jewett theorem which may be viewed as an extension of Theorem 1.29. See [26, Theorem 5.9], for yet another nilpotent extension of Theorem 1.29.

THEOREM 2.19 [26, Theorem 5.8]. *For any $r,q,c\in\mathbb{N}$ and prime integer p there exists $k\in\mathbb{N}$ such that if F is a field of characteristic p and of cardinality at least k, then for any r-coloring of the group G of $q\times q$ upper triangular matrices over F with unit diagonal, there exist a subgroup H of G with $|H_q|\geqslant c$ and $h\in G$ such that the coset hH is monochromatic.*

3. Dynamical, combinatorial, and Diophantine applications of $\beta\mathbb{N}$

In this section, we shall discuss briefly the Stone–Čech compactification of the natural numbers, $\beta\mathbb{N}$, and indicate some of its connections with an applications to topological dynamics, combinatorics, and the theory of Diophantine approximations.

We start with some general definitions and facts. The reader will find the missing details in [9, Section 3], and [12]. (See also [85] for a comprehensive treatment of topological algebra in Stone–Čech compactifications and applications thereof.)

3.1. *Definition and properties of* $\beta\mathbb{N}$

An *ultrafilter* p on $\mathbb{N}$ is a maximal filter, namely a family of subsets of $\mathbb{N}$ satisfying the following conditions (the first three of which constitute, for a nonempty family of sets, the definition of a *filter*).

(i) $\emptyset \notin p$,
(ii) $A \in p$ and $A \subset B$ imply $B \in p$,
(iii) $A \in p$ and $B \in p$ imply $A \cap B \in p$,
(iv) (maximality) if $r \in \mathbb{N}$ and $\mathbb{N} = \bigcup_{i=1}^{r} A_i$ then, for some $i \in \{1, 2, \ldots, r\}$, $A_i \in p$.

The space of ultrafilters on $\mathbb{N}$ is denoted by $\beta\mathbb{N}$; it has a natural topology which turns it into a universal compactification of $\mathbb{N}$, the so-called Stone–Čech compactification. (See more on that in [85].)

A convenient way of looking at ultrafilters is to identify each ultrafilter p with a finitely additive $\{0, 1\}$-valued probability measure μ_p on the power set $\mathcal{P}(\mathbb{N})$. This measure μ_p is naturally defined by the requirement $\mu_p(A) = 1$ iff $A \in p$ and it follows immediately from the conditions (i) through (iv) above that $\mu_p(\emptyset) = 0$, $\mu_p(\mathbb{N}) = 1$ and that for any finite disjoint collection $A_1, \ldots, A_r$ of subsets of $\mathbb{N}$, one has $\mu_p(\bigcup_{i=1}^{r} A_i) = \sum_{i=1}^{r} \mu_p(A_i)$. Without saying so explicitly, we will always think of ultrafilters as such measures, but we will prefer to write $A \in p$ rather that $\mu_p(A) = 1$.

Any $n \in \mathbb{N}$ naturally defines an ultrafilter $\{A \subset \mathbb{N}\colon\ n \in A\}$. Such ultrafilters, which can be viewed as "delta measures" concentrated at points of $\mathbb{N}$, are called *principal* and, alas, are the only ones which can be constructed without the use of Zorn's lemma (see [45, pp. 161–162]). Since many of the constructions in topological dynamics and ergodic theory use this or that equivalent of Zorn's lemma, we will not be bothered by this, notwithstanding the fact that there is certainly some importance in knowing which mathematical results are Zorn lemma free.

Suppose that $\mathcal{C}$ is a family of subsets of $\mathbb{N}$ which has the finite intersection property. Then there is some $p \in \beta\mathbb{N}$ such that $C \in p$ for each $C \in \mathcal{C}$. Indeed, let

$$\widetilde{\mathcal{C}} = \big\{\mathcal{B} \subset \mathcal{P}(\mathbb{N})\colon\ \mathcal{B} \text{ has the finite intersection property and } \mathcal{C} \subset \mathcal{B}\big\}.$$

Clearly, $\widetilde{\mathcal{C}} \neq \emptyset$ (since $\mathcal{C} \in \widetilde{\mathcal{C}}$). Also, the union of any chain in $\widetilde{\mathcal{C}}$ is a member of $\widetilde{\mathcal{C}}$. By Zorn's lemma there is a maximal member p of $\widetilde{\mathcal{C}}$, which is actually maximal with respect to the finite intersection property and hence a member of $\beta\mathbb{N}$.

To see that nonprincipal ultrafilters exist, take for example

$$\mathcal{C} = \{A \subset \mathbb{N}\colon\ A^c = \mathbb{N} \setminus A \text{ is finite}\}.$$

Clearly $\mathcal{C}$ has the finite intersection property, so there is an ultrafilter $p \in \beta\mathbb{N}$ such that $C \in p$ for all $C \in \mathcal{C}$. It is easy to see that such p cannot be principal.

For another example, take

$$\mathcal{D} = \left\{A \subset \mathbb{N}\colon\ d(A) = \lim_{n\to\infty} \frac{|A \cap \{1, 2, \ldots, n\}|}{n} = 1\right\}.$$

Again, $\mathcal{D}$ clearly satisfies the finite intersection property. If p is any ultrafilter for which $\mathcal{D} \subset p$, then any member of p has positive upper density. (If $d(A) = 0$, then $A^c = (\mathbb{N} \setminus A) \in \mathcal{D}$.)

These examples hint that the space $\beta\mathbb{N}$ is quite large. It is indeed: the cardinality of $\beta\mathbb{N}$ equals that of $\mathcal{P}(\mathcal{P}(\mathbb{N}))$ ([69, 6.10(a)]).

Let us say now a few words about topology in $\beta\mathbb{N}$. Given $A \subset \mathbb{N}$, let $\overline{A} = \{p \in \beta\mathbb{N}: A \in p\}$. The set $\mathcal{G} = \{\overline{A}: A \subset \mathbb{N}\}$ forms a basis for the open sets (and a basis for the closed sets). To see that $\mathcal{G}$ is indeed a basis for a topology on $\beta\mathbb{N}$ observe that if $A, B \subset \mathbb{N}$, then $\overline{A} \cap \overline{B} = \overline{A \cap B}$. Also, $\overline{\mathbb{N}} = \beta\mathbb{N}$ and hence $\bigcup_{\overline{A} \in \mathcal{G}} \overline{A} = \beta\mathbb{N}$. (Notice also that $\overline{A} \cup \overline{B} = \overline{A \cup B}$.) With this topology, $\beta\mathbb{N}$ satisfies the following.

THEOREM 3.1. *$\beta\mathbb{N}$ is a compact Hausdorff space.*

PROOF. Let $\mathcal{K}$ be a cover of $\beta\mathbb{N}$ by sets belonging to the base $\mathcal{G} = \{\overline{A}: A \subset \mathbb{N}\}$. Let $\mathcal{C} \subset \mathcal{P}(\mathbb{N})$ be such that $\mathcal{K} = \{\overline{A}: A \in \mathcal{C}\}$. Assume that $\mathcal{K}$ has no finite subcover. Consider the family $\mathcal{D} = \{A^c: A \in \mathcal{C}\}$. There are two possibilities (each leading to a contradiction):

(i) $\mathcal{D}$ has the finite intersection property. Then, as shown above, there exists an ultrafilter p such that $A^c \in p$ for each $A^c \in \mathcal{D}$. Since p is an ultrafilter, $A^c \in p$ if and only if $A \notin p$. On the other hand, since $\mathcal{K}$ covers $\beta\mathbb{N}$, for some element $\overline{A}$ of the cover $p \in \overline{A}$, or equivalently $A \in p$, a contradiction.

(ii) $\mathcal{D}$ does not have the finite intersection property. Then for some $A_1, \ldots, A_r \in \mathcal{C}$ one has $\bigcap_{i=1}^r A_i^c = \emptyset$, or $\bigcup_{i=1}^r A_i = \mathbb{N}$, which implies that $\bigcup_{i=1}^r \overline{A}_i = \beta\mathbb{N}$. Again, this is a contradiction, as we assumed that $\mathcal{K}$ has no finite subcover.

As for the Hausdorff property, notice that if $p, q \in \beta\mathbb{N}$ are distinct ultrafilters then since each of them is maximal with respect to the finite intersection property, neither of them is contained in the other. If $A \in p \setminus q$, then $A^c \in q \setminus p$, which means that $\overline{A}$ and $\overline{A^c}$ are disjoint neighborhoods of p and q. □

REMARK. Being a nice compact Hausdorff space, $\beta\mathbb{N}$ is in many respects quite a strange object. We mentioned already that its cardinality is that of $\mathcal{P}(\mathcal{P}(\mathbb{N}))$. It follows that $\beta\mathbb{N}$ is not metrizable, as otherwise, being a compact and hence separable metric space, it would have cardinality not exceeding that of $\mathcal{P}(\mathbb{N})$. Another curious feature of $\beta\mathbb{N}$ is that any infinite closed subset of $\beta\mathbb{N}$ contains a homeomorphic copy of all of $\beta\mathbb{N}$. (See [69, ex. 6O6, p. 97], or [85, Theorem 3.59, p. 67].)

3.2. *The semigroup operation in $\beta\mathbb{N}$*

Since $\overline{\mathbb{N}} = \beta\mathbb{N}$, it is natural to attempt to extend the operation of addition from (the densely embedded) $\mathbb{N}$ to $\beta\mathbb{N}$. Since ultrafilters are measures (principal ultrafilters being just the *point measures* corresponding to the elements of $\mathbb{N}$), it comes as no surprise that the extension we look for takes the form of a *convolution*. What is surprising, however, is that the algebraic structure of $\beta\mathbb{N}$ was explicitly introduced only relatively recently (in [44]). In the

following definition, $A - n$ (where $A \subset \mathbb{N}$, $n \in \mathbb{N}$) is the set of all m for which $m + n \in A$. For $p, q \in \mathbb{N}$, define

$$p + q = \{A \subset \mathbb{N}\colon \{n \in \mathbb{N}\colon (A - n) \in p\} \in q\}.$$

REMARKS.

(1) Note that in much of the literature, including [85], what we have written as $p + q$ is denoted as $q + p$.

(2) It is not hard to check that for principal ultrafilters the operation $+$ corresponds to addition in $\mathbb{N}$.

(3) Despite the somewhat forbidding phrasing of the operation just introduced in set-theoretical terms, the perspicacious reader will notice the direct analogy between this definition and the usual formulas for convolution of measures μ, ν on a locally compact group G (cf. [81, 19.11]):

$$\mu * \nu(A) = \int_G \nu\left(x^{-1}A\right) d\mu(x) = \int_G \mu\left(Ay^{-1}\right) d\nu(y).$$

(4) Before checking the correctness of the definition, a word of warning: the operation $+$ just introduced (which will turn out to be well defined and associative) is badly noncommutative. This seems to contradict our intuition since $(\mathbb{N}, +)$ is commutative and in the case of σ-additive measures on Abelian semi-groups convolution *is* commutative. The explanation: our ultrafilters, being only *finitely* additive measures, do not obey the Fubini theorem, which is behind the commutativity of the usual convolution.

Let us show that $p + q$ is an ultrafilter. Clearly $\emptyset \notin p + q$. Let $A, B \in p + q$. This means that $\{n \in \mathbb{N}\colon (A - n) \in p\} \in q$ and $\{n \in \mathbb{N}\colon (B - n) \in p\} \in q$. Since p and q are ultrafilters, we have:

$$\begin{aligned}&\{n \in \mathbb{N}\colon (A \cap B) - n \in p\}\\ &\quad = \{n \in \mathbb{N}\colon (A - n) \in p\} \cap \{n \in \mathbb{N}\colon (B - n) \in p\} \in q.\end{aligned}$$

Assume now that $A \subset \mathbb{N}$, $A \notin p + q$. We want to show that $A^c \in p + q$. Since $A \notin p + q$, we know that $\{n \in \mathbb{N}\colon (A - n) \in p\} \notin q$, or, equivalently, $\{n \in \mathbb{N}\colon (A - n) \in p\}^c \in q$. But this is true precisely when $\{n \in \mathbb{N}\colon (A^c - n) \in p\} \in q$, which is the same as $A^c \in p + q$. It follows that $p + q \in \beta\mathbb{N}$.

Let us now check the associativity of the operation $+$. Let $A \subset \mathbb{N}$ and $p, q, r \in \beta\mathbb{N}$. One has:

$$\begin{aligned}A \in p + (q + r) &\Leftrightarrow \{n \in \mathbb{N}\colon (A - n) \in p\} \in q + r\\ &\Leftrightarrow \{m \in \mathbb{N}\colon (\{n \in \mathbb{N}\colon (A - n) \in p\} - m) \in q\} \in r\\ &\Leftrightarrow \{m \in \mathbb{N}\colon \{n \in \mathbb{N}\colon (A - m - n) \in p\} \in q\} \in r\\ &\Leftrightarrow \{m \in \mathbb{N}\colon (A - m) \in p + q\} \in r \Leftrightarrow A \in (p + q) + r.\end{aligned}$$

THEOREM 3.2. *For any fixed $p \in \beta\mathbb{N}$ the function $\lambda_p(q) = p + q$ is a continuous self map of $\beta\mathbb{N}$.*

PROOF. Let $q \in \beta\mathbb{N}$ and let $\mathcal{U}$ be a neighborhood of $\lambda_p(q)$. We will show that there exists a neighborhood $\overline{B}$ of q such that for any $r \in \overline{B}$, $\lambda_p(r) \in \mathcal{U}$. Let $A \subset \mathbb{N}$ be such that $\lambda_p(q) = p + q \in \overline{A} \subset \mathcal{U}$. Then $A \in p + q$. Let us show that the set

$$B = \{n \in \mathbb{N}\colon\ (A - n) \in p\}$$

will do for our purposes. Indeed, by the definition of $p + q$, $B \in q$, or, in other words, $q \in \overline{B}$. If $r \in \overline{B}$ then $B = \{n \in \mathbb{N}\colon\ (A - n) \in p\} \in r$. This means that $A \in p + r = \lambda_p(r)$, or $\lambda_p(r) \in \overline{A} \in \mathcal{U}$. □

With the operation $+$, $\beta\mathbb{N}$ becomes, in view of Theorem 3.2, a compact *left topological* semigroup.

THEOREM 3.3. *If $(G, *)$ is a compact left topological semigroup (i.e. for any $x \in G$ the function $\lambda_x(y) = x * y$ is continuous) then G has an idempotent.*

REMARK. For compact *topological semigroups* (i.e. with an operation which is continuous in both variables), this result is due to Numakura [108]; for left topological semigroups the result is due to Ellis [51].

PROOF. Let

$$\mathcal{G} = \{A \subset G\colon\ A \neq \emptyset, A \text{ is compact},\ A * A = \{x * y\colon\ x, y \in A\} \subset A\}.$$

Since $G \in \mathcal{G}$, $\mathcal{G} \neq \emptyset$. By Zorn's lemma, there exists a minimal element $A \in \mathcal{G}$. If $x \in A$, then $x * A$ is compact and satisfies

$$(x * A) * (x * A) \subset (x * A) * (A * A) \subset (x * A) * A \subset x * (A * A) \subset x * A.$$

Hence $x * A \in \mathcal{G}$. But $x * A \subset A * A \subset A$, which implies that $x * A = A$. Thus $x \in x * A$, which implies that $x = x * y$ for some $y \in A$. Now consider $B = \{z \in A\colon\ x * z = x\}$. The set B is closed (since $B = \lambda_x^{-1}(\{x\})$), and we have just shown that B is nonempty. If $z_1, z_2 \in B$ then $z_1 * z_2 \in A * A \subset A$ and $x * (z_1 * z_2) = (x * z_1) * z_2 = x * z_2 = x$. So $B \in \mathcal{G}$. But $B \subset A$ and hence $B = A$. So $x \in B$ which gives $x * x = x$. □

For a fixed $p \in \beta\mathbb{N}$ we shall call a set $C \subset \mathbb{N}$ *p-big* if $C \in p$. The notion of largeness induced by idempotent ultrafilters is special (and promising) in that it inherently has a shift-invariance property. Indeed, if $p \in \beta\mathbb{N}$ with $p + p = p$ then

$$A \in p \Leftrightarrow A \in p + p \Leftrightarrow \{n \in \mathbb{N}\colon\ (A - n) \in p\} \in p.$$

A way of interpreting this is that if p is an idempotent ultrafilter, then A is p-big if and only if for p-many $n \in \mathbb{N}$ the shifted set $(A - n)$ is p-big. Or, still somewhat differently:

$A \subset \mathbb{N}$ is p-big if for p-almost all $n \in \mathbb{N}$ the set $(A - n)$ is p-big. This is the reason why specialists in ultrafilters called such idempotent ultrafilters "almost shift invariant" in the early seventies (even before the existence of such ultrafilters was established).

REMARK. The reader is invited to check that if p is an idempotent ultrafilter, then for any $a \in \mathbb{N}$, $a\mathbb{N} \in p$. This, in particular, implies that such p cannot be a principal ultrafilter. (This can also be deduced from the fact that $(\mathbb{N}, +)$ has no idempotents.)

3.3. *The analogy between idempotent ultrafilters and measure preserving systems. A new glimpse at Hindman's theorem*

Each idempotent ultrafilter $p \in \beta\mathbb{N}$ induces a "measure preserving dynamical system" with the phase space $\mathbb{N}$, σ-algebra $\mathcal{P}(\mathbb{N})$, measure p, and "time" being the "p-preserving" $\mathbb{N}$-action induced by the shift. The two peculiarities about such a measure preserving system are that the phase space is countable and that the "invariant measure" is only finitely additive and is preserved by our action not for all, but for almost all instances of "time." Notice that the "Poincaré recurrence theorem" trivially holds: If $A \in p$ then, since there are p-many n for which $(A - n) \in p$, one has, for any such n, $A \cap (A - n) \in p$.

As we saw in the Introduction, it is this defining property of idempotent ultrafilters (arranged there as Proposition 1.14) which is all that one needs for the proof of Hindman's theorem.

The following result gives the "ultrafilter explanation" of Theorem 1.16 in the Introduction. We shall also need it in the proof of Theorem 3.5 below.

THEOREM 3.4. *For any sequence $(x_i)_{i\in\mathbb{N}}$ in $\mathbb{N}$ there is an idempotent $p \in \beta\mathbb{N}$ such that $FS((x_i)_{i\in\mathbb{N}}) \in p$.*

SKETCH OF THE PROOF. Let $\Gamma = \bigcap_{n=1}^{\infty} \overline{FS((x_i)_{i=n}^{\infty})}$. (The closures are taken in the natural topology of $\beta\mathbb{N}$.) Clearly, Γ is compact and nonempty. It is not hard to show that Γ is a subsemigroup of $(\beta\mathbb{N}, +)$. Being a compact left-topological semigroup, Γ has an idempotent. If $p \in \Gamma$ is an idempotent, then $\overline{\Gamma} = \Gamma \ni p$ which, in particular, implies $FS((x_i)_{i=1}^{\infty}) \in p$. □

The space $\beta\mathbb{N}$ has also another natural semigroup structure, namely, the one inherited from the multiplicative semigroup $(\mathbb{N}, \cdot)$, and is a left topological compact semigroup with respect to this structure too. In particular, there are (many) multiplicative idempotents, namely ultrafilters q with the property

$$A \in q \Leftrightarrow \{n \in \mathbb{N}\colon\ A/n \in q\} \in q$$

(where $A/n := \{m \in \mathbb{N}\colon\ mn \in A\}$). By complete analogy with the proof of (the additive version of) Hindman's theorem, one can show that any member of a multiplicative idempotent contains a multiplicative IP set, namely a set of finite products of the form

$$FP(y_n)_{n=1}^{\infty} = \left\{\prod_{i\in\alpha} y_i\colon\ \alpha \subset \mathbb{N},\ 1 \leqslant |\alpha| < \infty\right\}.$$

It follows that for any finite partition $\mathbb{N} = \bigcup_{i=1}^{r} C_i$ there are $i, j \in \{1, 2, \ldots, r\}$ such that C_i contains an additive IP set and C_j contains a multiplicative IP set. The following theorem due to Hindman shows that one can always have $i = j$.

THEOREM 3.5 [84]. *For any finite partition $\mathbb{N} = \bigcup_{i=1}^{r} C_i$, there exists $i \in \{1, 2, \ldots, r\}$ and sequences $(x_n)_{n=1}^{\infty}$ and $(y_n)_{n=1}^{\infty}$ in $\mathbb{N}$ such that*

$$FS\big((x_n)_{n=1}^{\infty}\big) \cup FP\big((y_n)_{n=1}^{\infty}\big) \subseteq C_i.$$

PROOF. Let Γ be the closure in $\beta\mathbb{N}$ of the set of additive idempotents. We claim that $p \in \Gamma$ if and only if every p-large set A contains an additive IP set. Indeed, if $A \in p \in \Gamma$, then $\overline{A}$ is a (clopen) neighborhood of p. It follows that there exists $q \in \overline{A}$ with $q + q = q$. Then $A \in q$ and by Hindman's theorem, A contains an IP set. Conversely, if $\overline{A}$ is a basic neighborhood of p and for some $(x_n)_{n=1}^{\infty}$, $FS((x_n)_{n=1}^{\infty}) \subseteq A$, then by Theorem 3.4, there exists an idempotent q with $FS((x_n)_{n=1}^{\infty}) \in q$, which implies $A \in q$, and hence $p \in \Gamma$.

We will show now that Γ is a right ideal in $(\beta\mathbb{N}, \cdot)$. Let $p \in \Gamma$, $q \in \beta\mathbb{N}$, and let $A \in p \cdot q$. Then $\{x \in \mathbb{N}\colon\ Ax^{-1} \in p\} \in q$ and, in particular, $\{x \in \mathbb{N}\colon\ Ax^{-1} \in p\}$ is non-empty. Let x be such that $Ax^{-1} \in p$. Since $p \in \Gamma$, there exists a sequence $(y_n)_{n=1}^{\infty}$ with $FS((y_n)_{n=1}^{\infty}) \subseteq Ax^{-1}$, which implies $FS((xy_n)_{n=1}^{\infty}) \subseteq A$ and so $p \cdot q \in \Gamma$. We see that Γ is a compact subsemigroup in $(\beta\mathbb{N}, \cdot)$ and hence contains a multiplicative idempotent. To finish the proof, let $\bigcup_{i=1}^{r} C_i = \mathbb{N}$ and let $p \in \Gamma$ satisfy $p \cdot p = p$. Let $i \in \{1, 2, \ldots, r\}$ be such that $C_i \in p$. Then, since $p \in \Gamma$, C_i contains an additive IP set. Also, since p is a multiplicative idempotent, C_i contains a multiplicative IP set. We are done. □

REMARKS.

(1) For an elementary proof of Theorem 3.5, see [22].
(2) Theorem 3.13 below shows that for any finite partition $\bigcup_{i=1}^{r} C_i = \mathbb{N}$ one of the C_i has interesting additional properties. In particular, one of the C_i can be shown to contain in addition to an additive and a multiplicative IP sets, also arbitrarily long arithmetic and arbitrarily long geometric progressions.

3.4. *Minimal idempotents*

Seeing how much mileage one can get by sheer analogy between idempotent ultrafilters and measure preserving systems, it would be natural to inquire (in a hope that this can lead to interesting new results) whether there is a class of idempotents which could be likened to a minimal topological system (with an invariant measure).

To answer this question, let us extend the shift operation $\sigma : n \to n+1, n \in \mathbb{N}$, from $\mathbb{N}$ to $\beta\mathbb{N}$, by the rule $q \to q+1$ (where 1 denotes the principal ultrafilter of sets containing the integer 1), and consider the topological dynamical system $(\beta\mathbb{N}, \sigma)$.

It is customary to refer to a subset I of a semigroup $(S, +)$ as a right (respectively, left) ideal if $I + S \subseteq I$ (respectively, $S + I \subseteq I$). The following theorem establishes the connection between minimal subsystems of $(\beta\mathbb{N}, \sigma)$ and minimal right ideals in $(\beta\mathbb{N}, +)$.

THEOREM 3.6. *The minimal closed invariant subsets of the dynamical system* $(\beta\mathbb{N}, \sigma)$ *are precisely the minimal right ideals of* $(\beta\mathbb{N}, +)$.

PROOF. We first observe that closed σ-invariant sets in $\beta\mathbb{N}$ coincide with right ideals. Indeed if I is a right ideal, i.e. satisfies $I + \beta\mathbb{N} \subseteq I$, then for any $p \in I$ one has $p + 1 \in I + \beta\mathbb{N} \subseteq I$, so that I is σ-invariant. On the other hand, if S is a closed σ-invariant set in $\beta\mathbb{N}$ and $p \in S$, then $p + \beta\mathbb{N} = p + \overline{\mathbb{N}} = \overline{p + \mathbb{N}} \subseteq \overline{S} = S$, which implies $S + \beta\mathbb{N} \subseteq S$.

Now the theorem follows from a simple general fact that any minimal right ideal in a compact left-topological semigroup $(G, \cdot)$ is closed. Indeed, if R is a right ideal in $(G, \cdot)$ and $x \in R$, then xG is compact as the continuous image of G and is an ideal. Hence the minimal ideal containing x is compact as well. (The fact that R contains a minimal ideal follows by an application of Zorn's lemma to the nonempty family $\{I\colon\ I$ is a closed right ideal of G and $I \subseteq R\}$.) □

Our next step is to observe that any minimal right ideal in $(\beta\mathbb{N}, +)$, being a compact left-topological semigroup, contains, by Theorem 3.3, an idempotent.

DEFINITION 3.7. An idempotent p in $(\beta\mathbb{N}, +)$ is called *minimal* if p belongs to a minimal right ideal.

THEOREM 3.8. *Any minimal subsystem of* $(\beta\mathbb{N}, \sigma)$ *is of the form* $(p + \beta\mathbb{N}, \sigma)$ *where* p *is a minimal idempotent in* $(\beta\mathbb{N}, +)$.

PROOF. It is obvious that, for any $p \in (\beta\mathbb{N}, +)$, $p + \beta\mathbb{N}$ is a right ideal. To see that any minimal right ideal is of this form, take any $q \in R$ and observe that $q + \beta\mathbb{N} \subseteq R + \beta\mathbb{N} \subseteq R$. Since R is minimal, we get $q + \beta\mathbb{N} = R$. In particular, one can take q to be an idempotent. □

We shall need the following definition in order to formulate some immediate corollaries of Theorem 3.8.

DEFINITION 3.9. A set $A \subseteq \mathbb{N}$ is piecewise syndetic if it can be represented as an intersection of a syndetic set with an infinite union of intervals $[a_n, b_n]$, where $b_n - a_n \to \infty$.

REMARK. It is not hard to see that $A \subseteq \mathbb{N}$ is piecewise syndetic if and only if there exists a finite set $F \subset \mathbb{N}$ such that the family

$$\left\{\bigcup_{t\in F}(A-t)-n\colon\ n\in\mathbb{N}\right\}$$

has the finite intersection property. While this description of piecewise syndeticity looks somewhat forbidding, it has the advantage of making sense in any semigroup. As we shall see in the proof of Corollary 3.10 below, it is this form of the definition of piecewise syndeticity which is much easier to check when dealing with minimal idempotents.

COROLLARY 3.10. *Let p be a minimal idempotent in $(\beta\mathbb{N}, +)$.*

(i) *For any $A \in p$ the set $B = \{n\colon\ (A - n) \in p\}$ is syndetic.*

(ii) *Any $A \in p$ is piecewise syndetic.*

PROOF. Statement (i) follows immediately from the fact that $(p + \beta\mathbb{N}, \sigma)$ is a minimal system. Indeed, note that the assumption $A \in p$ just means that $p \in \overline{A}$, i.e. $\overline{A}$ is a (clopen) neighborhood of p. Now, in a minimal dynamical system every point x is *uniformly recurrent*, i.e. visits any of its neighborhoods V along a syndetic set. This implies that the set $\{n\colon\ p + n \in \overline{A}\} = \{n\colon\ A \in p + n\} = \{n\colon\ A - n \in p\}$ is syndetic.

(ii) Since the set $B = \{n\colon\ A - n \in p\}$ is syndetic, the union of finitely many shifts of B covers $\mathbb{N}$, i.e. for some finite set $F \subset \mathbb{N}$ one has $\bigcup_{t\in F}(B - t) = \mathbb{N}$. So, for any $n \in \mathbb{N}$ there exists $t \in F$ such that $n \in B - t$, or $n + t \in B$. By the definition of B this implies $(A - (n + t)) \in p$. It follows that for any n the set $\bigcup_{t\in F}(A - t) - n$ belongs to p, and consequently, the family $\{\bigcup_{t\in F}(A - t) - n\colon\ n \in \mathbb{N}\}$ has the finite intersection property. By the remark above, this is equivalent to piecewise syndeticity of A. □

REMARK. It follows from part (ii) of Corollary 3.10 that for any finite partition $\mathbb{N} = \bigcup_{i=1}^{r} C_i$, one of the C_i is piecewise syndetic, and moreover for any finite partition of a piecewise syndetic set, one of the cells of the partition is again piecewise syndetic. One can show that (with the appropriately arranged definition of piecewise syndeticity) this result holds for any infinite semigroup. (In the case of the semigroup $(\mathbb{N}, +)$, this fact can be proved in an elementary fashion, and is apparently originally due to T. Brown [42].)

3.5. *Ultrafilter proof of van der Waerden's theorem*

Note that it follows from the definition above that if A is a syndetic set in $\mathbb{N}$, then, for some finite set $F \subset \mathbb{N}$, the set $\bigcup_{t\in F}(A - t)$ contains arbitrarily long intervals. It follows now from Theorem 1.18 that any piecewise syndetic set contains arbitrarily long arithmetic progressions. (Since any piecewise syndetic set has positive upper Banach density, this fact also follows from Szemerédi's theorem, but this would be an overkill.)

On the other hand, it is clear that since for any minimal idempotent $p \in \beta\mathbb{N}$ and any finite partition $\mathbb{N} = \bigcup_{i=1}^{r} C_i$, one of the C_i belongs to p, van der Waerden's theorem follows from the following result. The proof below is a slight modification of the proof in [16]. (Cf. also [63].)

THEOREM 3.11. *Let $p \in (\beta\mathbb{N}, +)$ be a minimal idempotent and let $A \in p$. Then A contains arbitrarily long arithmetic progressions.*

PROOF. Fix $k \in \mathbb{N}$ and let $G = (\beta\mathbb{N})^k$. Clearly, G is a compact left topological semigroup with respect to the product topology and coordinatewise addition. Let

$$E_0 = \big\{(a, a+d, \ldots, a+(k-1)d)\colon\ a \in \mathbb{N},\ d \in \mathbb{N} \cup \{0\}\big\},$$
$$I_0 = \big\{(a, a+d, \ldots, a+(k-1)d)\colon\ a, d \in \mathbb{N}\big\}.$$

Clearly, E_0 is a semigroup in $\mathbb{N}^k$ and I_0 is an ideal of E_0. Let $E = \mathrm{cl}_G\, E_0$ and $I = \mathrm{cl}_G\, I_0$ be, respectively, the closures of E_0 and I_0 in G. It follows by an easy argument, which we leave to the reader, that E is a compact subsemigroup of G and I is a two-sided ideal of E. Let now $p \in (\beta\mathbb{N}, +)$ be a minimal idempotent and let $\tilde{p} = (p, p, \ldots, p) \in G$. We claim that $\tilde{p} \in I$ and that this implies that each member of p contains a length k arithmetic progression. Indeed, assume that $\tilde{p} \in I$ and let $A \in p$. Then $\overline{A} \times \cdots \times \overline{A} = (\overline{A})^k$ is a neighborhood of $\tilde{p}$. Hence $\tilde{p} \in (\overline{A})^k \cap \mathrm{cl}_G\, I_0 = \mathrm{cl}_G(A^k \cap I_0)$, which implies $A^k \cap I_0 \neq \emptyset$. It follows that for some $a, d \in \mathbb{N}$ $(a, a+d, \ldots, a+(k-1)d) \in A^k$ which finally implies $\{a, a+d, \ldots, a+(k-1)d\} \subset A$.

So it remains to show that $\tilde{p} \in I$. We check first that $\tilde{p} \in E$. Let $A_1, A_2, \ldots, A_k \in p$. Then $\overline{A}_1 \times \overline{A}_2 \times \cdots \times \overline{A}_k \ni \tilde{p}$. If $a \in \bigcap_{i=1}^k A_i$ then $(a, a, \ldots, a) \in (\overline{A}_1 \times \overline{A}_2 \times \cdots \times \overline{A}_k) \cap E_0$ which implies $\tilde{p} \in E$.

Now, since p is a minimal idempotent, there is a minimal right ideal R of $(\beta\mathbb{N}, +)$ such that $p \in R$. Since $\tilde{p} \in E,\ \tilde{p} + E$ is a right ideal of E and there is a minimal right ideal $\widetilde{R}$ of E such that $\widetilde{R} \subseteq \tilde{p} + E$. Let $\tilde{q} = (q_1, q_2, \ldots, q_k)$ be an idempotent in $\widetilde{R}$. Then $\tilde{q} \in \tilde{p} + E$ and for some $\tilde{s} = (s_1, s_2, \ldots, s_k)$ in E we get $\tilde{q} = \tilde{p} + \tilde{s}$. We shall show now that $\tilde{p} = \tilde{q} + \tilde{p}$. Indeed, from $\tilde{q} = \tilde{p} + \tilde{s}$ we get, for each $i = 1, 2, \ldots, k,\ q_i = p + s_i$. This implies $q_i \in R$ and since R is minimal, $q_i + \beta\mathbb{N} = R$. Hence $p \in q_i + \beta\mathbb{N}$. Let, for each $i = 1, 2, \ldots, k,\ t_i \in \beta\mathbb{N}$ be such that $p = q_i + t_i$. Then $q_i + p = q_i + q_i + t_i = q_i + t_i = p$ and so we obtained $\tilde{p} = \tilde{q} + \tilde{p}$.

To finish the proof, we observe that $\tilde{p} = \tilde{q} + \tilde{p}$ implies $\tilde{p} \in \tilde{q} + E = \widetilde{R}$ which, in its turn, implies $\tilde{p} \in I$ (since, as it is not hard to see, any minimal right ideal is contained in a two-sided ideal). We are done. □

3.6. *Central sets*

DEFINITION 3.12. A set $A \subseteq \mathbb{N}$ is called additively (respectively, multiplicatively) central if there is a minimal idempotent $p \in (\beta\mathbb{N}, +)$ (respectively, $p \in (\beta\mathbb{N}, \cdot)$), such that $A \in p$.

As theorems above indicate, central sets are an ideal object for Ramsey-theoretical applications. For example, central sets in $(\mathbb{N}, +)$ not only are large (i.e. piecewise syndetic) but also are combinatorially rich and, in particular, contain IP sets and arbitrarily long arithmetic progressions. Similarly, the multiplicative central sets in $(\mathbb{N}, \cdot)$ (namely, the members of minimal idempotents in $(\beta\mathbb{N}, \cdot)$) are multiplicatively piecewise syndetic, contain finite products sets (i.e. the multiplicative IP sets), arbitrarily long geometric progressions, etc.

The following theorem obtained in collaboration with N. Hindman may be viewed as an enhancement of Theorem 3.5 above.

THEOREM 3.13 [19, p. 312]. *For any finite partition* $\mathbb{N} = \bigcup_{i=1}^{r} C_i$, *one of* C_i *is both additively and multiplicatively central.*

SKETCH OF THE PROOF. Let $M = \mathrm{cl}\{p\colon\ p$ is a minimal idempotent in $(\beta\mathbb{N}, +)\}$. Then one can show that M is a right ideal in $(\beta\mathbb{N}, \cdot)$ (see [19, Theorem 5.4, p. 311]). Let $R \subseteq M$ be a minimal right ideal and pick an idempotent $q = q \cdot q$ in R. Let $i \in \{1, 2, \ldots, r\}$ be such that $C_i \in q$. Since q is a minimal idempotent in $(\beta\mathbb{N}, \cdot)$, C_i is central in $(\mathbb{N}, \cdot)$. Since $C_i \in q$ and $q \in M$, there is some minimal idempotent p in $(\beta\mathbb{N}, +)$ with $C_i \in p$. Hence C_i is also central in $(\mathbb{N}, +)$. □

The following theorem supplies a useful family of examples of additively and multiplicatively central sets in $\mathbb{N}$.

THEOREM 3.14 [22, Lemma 3.3]. *For any sequence* $(a_n)_{n=1}^{\infty}$ *and an increasing sequence* $(b_n)_{n=1}^{\infty}$ *in* $\mathbb{N}$, $\bigcup_{n=1}^{\infty}\{a_n, a_n + 1, a_n + 2, \ldots, a_n + b_n\}$ *is additively central and* $\bigcup_{n=1}^{\infty}\{a_n \cdot 1, a_n \cdot 2, \ldots, a_n \cdot b_n\}$ *is multiplicatively central.*

The original definition of central sets in $(\mathbb{N}, +)$, due to H. Furstenberg, was made in the language of topological dynamics. Before introducing Furstenberg's definition of centrality, we want first to recall some relevant dynamical notions.

Given a compact metric space (X, d), a continuous map $T : X \to X$ and not necessarily distinct points $x_1, x_2 \in X$, one says that x_1, x_2 are *proximal*, if for some sequence $n_k \to \infty$ one has $d(T^{n_k}x_1, T^{n_k}x_2) \to 0$.

A point which is proximal only to itself is called *distal*. In case all the points of X are distal T is called a distal transformation and (X, T) is called a distal system.

Recall that a point x in a dynamical system (X, T) is called *uniformly recurrent* if for any neighborhood V of x the set $\{n\colon\ T^n x \in V\}$ is syndetic. Since in a minimal system any point is uniformly recurrent and since any compact topological system has a minimal subsystem, any topological system has a uniformly recurrent point. A stronger statement, due to J. Auslander [3] and R. Ellis [52] says that in a dynamical system on a compact metric space, any point is proximal to a uniformly recurrent point. (Note that this, in particular, implies that any distal point is uniformly recurrent.)

We are now ready to formulate Furstenberg's original definition of central sets in $(\mathbb{N}, +)$. For the proof of the equivalence of this definition to Definition 3.12 above, see Theorem 3.22.

DEFINITION 3.15. (See [58, p. 161].) A subset $S \subseteq \mathbb{N}$ is a *central* set if there exists a system (X, T), a point $x \in X$, a uniformly recurrent point y proximal to x, and a neighborhood U_y of y such that $S = \{n\colon\ T^n x \in U_y\}$.

In order to prove the equivalence of the two definitions of centrality, we need to introduce first the notion of convergence along ultrafilters. As we shall see, this notion allows one to better understand distality, proximality, and recurrence in topological dynamical systems. We would like to point out that some proofs involving ultrafilters are similar to known proofs involving the so-called Ellis enveloping semigroup. This is not surprising

in view of the fact that the Ellis semigroup is a particular type of compactification and, as such, is in many respects similar to the universal object, the Stone–Čech compactification. In particular, it allows one to much more easily deal with combinatorial applications of topological dynamics.

Given an ultrafilter $p \in \beta\mathbb{N}$ and a sequence $(x_n)_{n\in\mathbb{N}}$ in a topological space X, one writes $p\text{-}\lim_{n\in\mathbb{N}} x_n = y$ if, for every neighborhood U of y, one has $\{n\colon x_n \in U\} \in p$. It is easy to see that if X is a compact Hausdorff space, then $p\text{-}\lim_{n\in\mathbb{N}} x_n$ exists and is unique for any sequence $(x_n)_{n\in\mathbb{N}}$ in X.

THEOREM 3.16. *Let X be a compact Hausdorff space and let $p, q \in \beta\mathbb{N}$. Then for any sequence $(x_n)_{n\in\mathbb{N}}$ in X one has*

$$(q+p)\text{-}\lim_{r\in\mathbb{N}} x_r = p\text{-}\lim_{t\in\mathbb{N}}\, q\text{-}\lim_{s\in\mathbb{N}} x_{s+t}. \tag{3.1}$$

In particular, if p is an idempotent, and $q = p$, one has

$$p\text{-}\lim_{r\in\mathbb{N}} x_r = p\text{-}\lim_{t\in\mathbb{N}}\, p\text{-}\lim_{s\in\mathbb{N}} x_{s+t}.$$

PROOF. Let $x = (q+p)\text{-}\lim_{r\in\mathbb{N}} x_r$. Given a neighborhood U of x we have $\{r\colon x_r \in U\} \in q+p$. Recalling that a set $A \subseteq \mathbb{N}$ is a member of ultrafilter $q+p$ if and only if $\{n \in \mathbb{N}\colon (A-n) \in q\} \in p$, we get

$$\big\{t\colon \big(\{s\colon x_s \in U\} - t\big) \in q\big\} = \big\{t\colon \{s\colon x_{s+t} \in U\} \in q\big\} \in p.$$

This means that, for p-many t, $q\text{-}\lim_{s\in\mathbb{N}} x_{s+t} \in U$ and we are done. □

PROPOSITION 3.17. *Let (X,T) be a topological system and let $x \in X$ be an arbitrary point. Given an idempotent ultrafilter $p \in \beta\mathbb{N}$, let $p\text{-}\lim_{n\in\mathbb{N}} T^n x = y$. Then $p\text{-}\lim_{n\in\mathbb{N}} T^n y = y$. If x is a distal point (i.e. x is proximal only to itself) then $p\text{-}\lim_{n\in\mathbb{N}} T^n x = x$.*

PROOF. Applying Theorem 3.16 (and the fact that $p+p=p$), we have

$$p\text{-}\lim_{n\in\mathbb{N}} T^n y = p\text{-}\lim_{n\in\mathbb{N}} T^n\ p\text{-}\lim_{m\in\mathbb{N}} T^m x = p\text{-}\lim_{n\in\mathbb{N}}\ p\text{-}\lim_{m\in\mathbb{N}} T^{m+n} x = p\text{-}\lim_{n\in\mathbb{N}} T^n x = y.$$

If x is a distal point, then the relations $p\text{-}\lim_{n\in\mathbb{N}} T^n x = y = p\text{-}\lim_{n\in\mathbb{N}} T^n y$ clearly imply $x = y$ and we are done. □

REMARK. Note that Proposition 3.17 implies that a continuous distal self-map T of a compact metric space is onto. It follows that T is invertible and T^{-1} is also distal.

Let R be a minimal right ideal in $\beta\mathbb{N}$. By Theorem 3.8 above, (R,σ), where $\sigma : p \to p+1$, is a minimal (nonmetrizable) system. Given a topological system (X,T) and a point

$x \in X$, let $\varphi : R \to X$ be defined by $\varphi(p) = p\text{-}\lim_{n\in\mathbb{N}} T^n x$. Observe that if the set $Y \subseteq X$ is defined by $Y = \{p\text{-}\lim_{n\in\mathbb{N}} T^n x\colon\ p \in R\}$, then the following diagram is commutative:

$$\begin{array}{ccc} R & \xrightarrow{\sigma} & R \\ \varphi\downarrow & & \downarrow\varphi \\ Y & \xrightarrow{T} & Y \end{array}$$

It follows that (Y, T) is a minimal system. We will use this observation in the proof of the following result.

PROPOSITION 3.18. *If (X, T) is a minimal system then for any $x \in X$ and any minimal right ideal R in $\beta\mathbb{N}$ there exists a minimal idempotent $p \in R$ such that $p\text{-}\lim T^n x = x$.*

PROOF. By the observation above, $X = \{p\text{-}\lim_{n\in\mathbb{N}} T^n x,\ \ p \in R\}$. It follows that the set $\Gamma = \{p \in R\colon\ p\text{-}\lim_{n\in R} T^n x = x\}$ is nonempty and closed. We claim that Γ is a semigroup. Indeed, if $p, q \in \Gamma$, one has:

$$(p+q)\text{-}\lim_{n\in\mathbb{N}} T^n x = q\text{-}\lim_{n\in\mathbb{N}} T^n\ p\text{-}\lim_{m\in\mathbb{N}} T^m x = x.$$

By Theorem 3.3, Γ contains an idempotent which has to be minimal since it belongs to R. We are done. □

We shall need the following simple fact in the proofs below. The proof is immediate and is left as an exercise for the reader.

THEOREM 3.19. *Let (X, T) be a topological system, R a minimal right ideal in $\beta\mathbb{N}$, and let $x \in X$ be a point in X. The following are equivalent*:

(i) *x is uniformly recurrent*;

(ii) *there exists a minimal idempotent $p \in R$ such that $p\text{-}\lim_{n\in\mathbb{N}} T^n x = x$.*

It follows from Proposition 3.17 that for any topological system (X, T), any $x \in X$, and any idempotent ultrafilter p, the points x and $y = p\text{-}\lim_{n\in\mathbb{N}} T^n x$ are proximal. (If (X, T) is a distal system then $y = x$.) The following theorem gives a partial converse of Proposition 3.17.

THEOREM 3.20. *If (X, T) is a topological system and x_1, x_2 are proximal, not necessarily distinct points, and if x_2 is uniformly recurrent, then there exists a minimal idempotent $p \in \beta\mathbb{N}$ such that $p\text{-}\lim_{n\in\mathbb{N}} T^n x_1 = x_2$.*

PROOF. Let $I = \{p \in \beta\mathbb{N}\colon\ p\text{-}\lim_{n\in\mathbb{N}} T^n x_1 = p\text{-}\lim_{n\in\mathbb{N}} T^n x_2\}$. It is not hard to see that I is a nonempty closed subset of $\beta\mathbb{N}$. One immediately checks that I is a right ideal. Let R be a minimal right ideal in I. Since x_2 is uniformly recurrent, its orbital closure is a minimal system. By Proposition 3.18, there exists a minimal idempotent $p \in R$ such that $p\text{-}\lim T^n x_2 = x_2$. Then $p\text{-}\lim_{n\in\mathbb{N}} T^n x_1 = p\text{-}\lim_{n\in\mathbb{N}} T^n x_2 = x_2$ and we are done. □

One can give a similar proof to the following classical result due to J. Auslander [3] and R. Ellis [52].

THEOREM 3.21. *Let (X, T) be a topological system. For any $x \in X$ there exists a uniformly recurrent point y in the orbital closure $\overline{\{T^n x\}}_{n\in N}$, such that x is proximal to y.* (In fact, we will prove that for any minimal right ideal $R \subset \beta\mathbb{N}$ there exists a minimal idempotent $p \in R$ such that $p\text{-}\lim_{n\in\mathbb{N}} T^n x = y$.)

PROOF. Let R be a minimal ideal in $\beta\mathbb{N}$ and let p be a (minimal) idempotent in R. Let $y = p\text{-}\lim_{n\in\mathbb{N}} T^n x$. Clearly, y belongs to the orbital closure of x. By Proposition 3.17, the points x and y are proximal. By Theorem 3.19, y is uniformly recurrent. We are done. □

We are in position now to establish the equivalence of two notions of central that were discussed above. (In [70, Proposition 4.6] Glasner anticipated this result by showing that, if S is a countable Abelian group, then a subset of S is central as defined above if and only if it satisfies conditions similar to Furstenberg's dynamical definition of "central".)

THEOREM 3.22. *The following properties of a set $A \subseteq \mathbb{N}$ are equivalent*:

(i) (*Cf.* [58, *Definition* 8.3]) *There exists a topological system (X, T), and a pair of* (*not necessarily distinct*) *points $x, y \in X$ where y is uniformly recurrent and proximal to x, such that for some neighborhood U of y one has*:

$$A = \{n \in \mathbb{N}\colon\ T^n x \in U\}.$$

(ii) (*See Definition* 3.12 *above, see also* [19, *Definition* 3.1]) *There exists a minimal idempotent $p \in (\beta\mathbb{N}, +)$ such that $A \in p$.*

PROOF. (i) $\Rightarrow$ (ii) By Theorem 3.20, there exists a minimal idempotent p, such that $p\text{-}\lim_{n\in\mathbb{N}} T^n x = y$. This implies that for any neighborhood U of y the set $\{n \in \mathbb{N}\colon T^n x \in U\}$ belongs to p.

(ii) $\Rightarrow$ (i) The idea of the following proof is due to B. Weiss. Let A be a member of a minimal idempotent $p \in \beta\mathbb{N}$. Let $X = \{0, 1\}^{\mathbb{Z}}$, the space of bilateral 0–1 sequences. Endow X with the standard metric, which turns it into a compact space:

$$d(\omega_1, \omega_2) = \inf\left\{\frac{1}{n+1}\colon\ \omega_1(i) = \omega_2(i) \text{ for } |i| < n\right\}.$$

Let $T : X \to X$ be the shift operator: $T(\omega)(n) = \omega(n+1)$. Then T is a homeomorphism of X and (X, T) is a topological dynamical system. Viewing A as a subset of $\mathbb{Z}$, let $x = 1_A \in X$. Finally, let $y = p\text{-}\lim_{n\in\mathbb{N}} T^n x$. By Proposition 3.17, x and y are proximal. Also, since p is minimal, y is, by Theorem 3.19, a uniformly recurrent point. We claim that $y(\mathbf{0}) = 1$. Indeed, define $U = \{z \in X\colon\ z(\mathbf{0}) = y(\mathbf{0})\}$, and note that, since $y = p\text{-}\lim_{n\in\mathbb{N}} T^n x$ and $A \in p$, one can find $n \in A$ such that $T^n x \in U$. But since $x = 1_A$, $(T^n x)(\mathbf{0}) = 1$. But then, given $n \in \mathbb{Z}$, we have: $T^n x \in U \Leftrightarrow (T^n x)(\mathbf{0}) = 1 \Leftrightarrow x(n) = 1 \Leftrightarrow n \in A$. It follows that $A = \{n \in \mathbb{Z}\colon\ T^n x \in U\}$ and we are done. □

Let (X, T) be a topological system. In [58], a point $x \in X$ is called IP* recurrent if for any neighborhood U of x, the set $\{n \in \mathbb{N}\colon\ T^n x \in U\}$ is an IP* set. It is easy to see that a point x is IP* recurrent if and only if for any idempotent $p \in \beta\mathbb{N}$, one has $p\text{-}\lim_{n\in\mathbb{N}} T^n x = x$. Note that the property of a point x being IP* recurrent is much stronger than that of uniform recurrence (which, by Theorem 3.19, is equivalent to the fact that for *some* minimal idempotent p, one has $p\text{-}\lim_{n\in\mathbb{N}} T^n x = x$.) While, in a minimal system, every point is uniformly recurrent, there are minimal systems having no IP* recurrent points. For example, any minimal topologically weakly mixing system has this property. (See [58, Theorem 9.12].) The following theorem shows that distal points (and no others) are IP* recurrent.

THEOREM 3.23. *Let (X, T) be a dynamical system and $x \in X$. The following are equivalent*:

(i) *x is a distal point*;

(ii) *x is IP* recurrent*.

PROOF. (i) $\Rightarrow$ (ii) By Proposition 3.17, for any idempotent p, the points x and $p\text{-}\lim_{n\in\mathbb{N}} T^n x$ are proximal. Since x is distal, this may happen only if $x = p\text{-}\lim T^n x$. But this means that x is an IP* recurrent point.

(ii) $\Rightarrow$ (i) If x is not distal, then there exists $y \neq x$, such that x and y are proximal. But then, by Theorem 3.20, there exists an idempotent p such that $p\text{-}\lim T^n x = y$. Since $y \neq x$, this contradicts (ii). □

3.7. *Diophantine applications*

We shall conclude this section with some Diophantine applications of distal minimal systems. The results which we are going to describe can be viewed as enhancements of classical theorems due to Kronecker, Hardy and Littlewood, and Weyl, and will be based on the following characterization of distal systems. A set $E \subset \mathbb{N}$ is called IP^*_+ if it is a translation of an IP* set.

THEOREM 3.24. *Assume that (X, T) is a minimal system. Then it is distal if and only if for any $x \in X$ and any open set $U \subseteq X$ the set $\{n\colon\ T^n x \in U\}$ is IP^*_+.*

PROOF. Assume that (X, T) is distal. By minimality, there exists $n_0 \in \mathbb{N}$ such that $T^{n_0} x \in U$. By Theorem 3.23, the set $\{n\colon\ T^n(T^{n_0} x) \in U\}$ is IP* which, of course, implies that the set $\{n\colon\ T^n x \in U\}$ is IP^*_+.

Assume now that for any x_1, x_2 and a neighborhood U of x_2 the set $\{n\colon\ T^n x_1 \in U\}$ is IP^*_+. We will find it convenient to call an IP^*_+ set $A \subseteq \mathbb{N}$ *proper* if A is not IP* (i.e. A is a nontrivial shift of an IP* set and, moreover, this shifted IP* set is not IP*). If T were not distal, then for some distinct points x_1, x_2 and idempotents p, q one would have: $p\text{-}\lim_{n\in\mathbb{N}} T^n x_1 = x_2$, $q\text{-}\lim_{n\in\mathbb{N}} T^n x_2 = x_1$ and also $p\text{-}\lim_{n\in\mathbb{N}} T^n x_2 = x_2$, $q\text{-}\lim_{n\in\mathbb{N}} T^n x_1 = x_1$ (see Theorem 3.20 and Proposition 3.17). Let U be a small enough neighborhood of x_2. Then, since $p\text{-}\lim_{n\in\mathbb{N}} T^n x_1 = x_2$, the set $S = \{n\colon\ T^n x_1 \in U\}$ is a

member of p, and hence cannot be a proper IP^*_+ set. But, since $q\text{-}\lim_{n\in\mathbb{N}} T^n x_1 = x_1$, the set S cannot be an improper IP^*_+ set (that is, an IP^* set) either: if U is small enough, $S \notin q$. So T has to be distal. We are done. □

The following theorem was obtained by Hardy and Littlewood in [78] and may be viewed as a polynomial extension of a similar "linear" theorem due to Kronecker [95].

THEOREM 3.25. *If the numbers $1, \alpha_1, \ldots, \alpha_k$ are linearly independent over $\mathbb{Q}$, then for any $d \in \mathbb{N}$ and any kd intervals $I_{lj} \subset [0,1]$, $l = 1, \ldots, d$; $j = 1, \ldots, k$, the set*

$$\Gamma_{dk} = \left\{n \in \mathbb{N}\colon\ n^l \alpha_j \bmod 1 \in I_{lj},\ l = 1, \ldots, d;\ j = 1, \ldots, k\right\}$$

is infinite.

In 1916, H. Weyl [136] introduced the notion of uniform distribution and obtained many strong results extending and enhancing the earlier work of Kronecker, Hardy and Littlewood and others on Diophantine approximations. Perhaps the most famous result obtained in [136] was the theorem on uniform distribution of the sequence $p(n) \pmod 1$, $n = 1, 2, \ldots$, where $p(n)$ is a real polynomial having at least one irrational coefficient other than the constant term. This theorem also admits a nice ergodic proof, via the study of a class of affine transformations of the torus, due to Furstenberg [56].

In connection to the Hardy–Littlewood theorem, Weyl was able to show in [136] that the set Γ_{dk} has positive density equal to the product of the lengths of I_{ij}. This also can be shown by using the dynamical approach of Furstenberg. In the following theorem, we show that the affine transformations of the kind treated by Furstenberg in [56] can also be utilized to prove the following strengthening of the Hardy–Littlewood theorem.

THEOREM 3.26. *Under the assumptions and notation of Theorem* 3.25, *the set Γ_{dk} is IP^*_+.*

PROOF. To make the formulas more transparent we shall put $d = 3$. It will be clear that the same proof gives the general case.

We start with the easily checkable claim that if $T_\alpha : \mathbb{T}^3 \to \mathbb{T}^3$ is defined by $T_\alpha(x, y, z) = (x+\alpha, y+2x+\alpha, z+3x+3y+\alpha)$ then $T^n_\alpha(0,0,0) = (n\alpha, n^2\alpha, n^3\alpha)$. This transformation T is distal (easy) and minimal. The last assertion can actually be derived from the case $k = 1$ of Hardy–Littlewood theorem above, but also can be proved directly. (For example, this fact is a special case of Lemma 1.25, p. 36 in [58].) Our next claim is that if the numbers $1, \alpha_1, \alpha_2, \ldots, \alpha_k$ are linearly independent over Q, then the product map $T = T_{\alpha_1} \times \cdots \times T_{\alpha_k}$ (acting on $\mathbb{T}^{3k}$) is distal and minimal as well. (The distality is obvious, and the minimality follows, again, from an appropriately modified Lemma 1.25 in [58].) By minimality of T, the orbit of zero in $\mathbb{T}^{3k}$ is dense, and this, together with Theorem 3.24, gives the desired result. □

We conclude this section by formulating a general result which may be proved by refining the techniques used above.

THEOREM 3.27. *If real polynomials $p_1(t), p_2(t), \dots, p_k(t)$ have the property that for any nonzero vector $(h_1, h_2, \dots, h_k) \in \mathbb{Z}^k$ the linear combination $\sum_{i=1}^{k} h_i p_i(t)$ is a polynomial with at least one irrational coefficient other than the constant term then for any k subintervals $I_j \subset [0,1]$, $j = 1, \dots, k$, the set*

$$\{n \in \mathbb{N}:\ p_j(n) \bmod 1 \in I_j,\ j = 1, \dots, k\}$$

is IP^*_+.

4. Multiple recurrence

4.1. *Introduction*

One of the common features of the topological multiple recurrence results which were discussed in the previous sections is that they have streamlined, and often relatively short, proofs. In particular, in proving these theorems, one does not have to analyze and distinguish between various types of dynamical behavior which the topological system may possess. In other words, the proofs evolve without taking into account the possibly intricate structure of the system. The situation with measure-theoretical multiple recurrence is, at least at present, quite different. All of the known proofs of dynamical theorems such as the ergodic Szemerédi theorem and other more recent and stronger multiple recurrence results, which will be discussed in this section, are complicated by the fact that systems with different types of dynamical behavior require different types of arguments. Yet, these proofs have a certain (and, in the opinion of the author, quite beautiful) structure which, in some "big" sense, is the same in different proofs.

Our plan for this section is as follows. In Subsection 4.1, we shall analyze the proof of Furstenberg's ergodic Szemerédi theorem, and, in particular, provide complete proofs of some important special cases.

In Subsection 4.2, we will give an overview of (the proofs of) the major multiple recurrence results (as well as their density counterparts) which have appeared since the publication of Furstenberg's groundbreaking paper [57]. An attempt will be made to emphasize the common features of these proofs and to highlight the subtle points. The flow of the discussion in Subsection 4.2 will eventually lead us to some quite recent results and natural open problems.

4.2. *Furstenberg's ergodic Szemerédi theorem*

This subsection is devoted to the thorough discussion of the proof of Furstenberg's ergodic Szemerédi theorem, which corresponds to the case $T_i = T^i$ in Theorem 1.24 formulated in the Introduction.

4.2.1. *Statement of the theorem*

THEOREM 4.1 [57]. *For any probability measure preserving system $(X, \mathcal{B}, \mu, T)$, any $A \in \mathcal{B}$ with $\mu(A) > 0$ and any $k \in \mathbb{N}$, there exists $n \in \mathbb{N}$ such that $\mu(A \cap T^{-n}A \cap \cdots \cap T^{-kn}A) > 0$.*

Actually, we do not know how to prove Theorem 4.1 without proving, at least superficially, a little bit more. (The situation here is analogous to what we encountered when discussing and proving the Sárközy–Furstenberg theorem—see Theorem 1.31 and the subsequent remarks.)

Here is the version of Theorem 4.1 which we will find convenient to work with.

THEOREM 4.2. *For any probability measure preserving system* $(X, \mathcal{B}, \mu, T)$, *any* $A \in \mathcal{B}$ *with* $\mu(A) > 0$, *and any* $k \in \mathbb{N}$, *one has*:

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu(A \cap T^{-n}A \cap \cdots \cap T^{-kn}A) > 0.$$

REMARK 4.3. As a matter of fact, the result proved by Furstenberg in [57] establishes that

$$\liminf_{N-M\to\infty} \frac{1}{N-M} \sum_{n=M}^{N-1} \mu(A \cap T^{-n}A \cap \cdots \cap T^{-kn}A) > 0.$$

This implies (via the Furstenberg correspondence principle) not only that any set of positive upper density in $\mathbb{N}$ contains arithmetic progressions but that the differences of these progressions form a syndetic set. This fact, in turn, follows from a much stronger IP Szemerédi theorem proved by Furstenberg and Katznelson in [63]. (See Theorems 4.31 and 4.32 below.) One of the reasons we have chosen to deal with the formulation as in Theorem 4.2 is that it has a simpler proof which nevertheless will allow us to stress the main ideas and will naturally serve as the basis for a discussion of possible extensions.

There are a few assumptions that we may make without loss of generality.

First, we can assume that the measure μ is nonatomic. (This follows from the fact that the atoms of μ generate an invariant sub-σ-algebra, and Theorems 4.1 and 4.2 are trivially satisfied in the case of atomic measure spaces.)

Second, we can assume that the space $(X, \mathcal{B}, \mu)$ is Lebesgue, i.e. is isomorphic to the unit interval with Lebesgue measure. Indeed, given the set $A \in \mathcal{B}$, we can pass, if needed, to a T-invariant separable sub-σ-algebra of $\mathcal{B}$ with respect to which all of the functions $f_n = 1_A(T^n x)$ and their finite products are measurable. By Caratheódory's theorem (see [123, Chapter 15, Theorem 4]) any separable atomless measure algebra $(X, \mathcal{B}, \mu)$ with $\mu(X) = 1$ is isomorphic to the measure algebra $\mathcal{L}$ induced by the Lebesgue measure on the unit interval. This isomorphism carries T into a Lebesgue-measure preserving isomorphism of $\mathcal{L}$, which by the classical theorem due to von Neumann (see [123, Chapter 15, Theorem 20]) admits realization as a point mapping.

Finally, we can assume that the measure preserving systems that we are dealing with are invertible. Indeed, assuming the invertibility of the measure preserving transformations occurring in the formulations of multiple recurrence theorems such as Theorem 1.24 or Theorem 4.2, not only makes the proofs more convenient, but also is sufficient for combinatorial applications. On the other hand, it is not hard to show that in the case of measure

preserving actions of commutative semigroups with cancellation, the general case follows from the invertible one. See, for example, [58, Chapter 7, Section 4].

These remarks apply also to the other multiple recurrence results formulated below, and we will tacitly keep the above assumptions throughout the rest of this survey.

We could also assume that the measure preserving transformation T in the formulation of Theorems 4.1 and 4.2 is ergodic. Despite the fact that this would make some of the arguments somewhat simpler, we have chosen not to do so since, in more general situations such as, say, the multidimensional ergodic Szemerédi theorem (Theorem 1.24 above), one can assume only that the action of the multidimensional group is ergodic, which does not help things too much. We will however allow ourselves to assume ergodicity of T in dealing with a particular case of Theorem 4.2, namely the case $k = 2$, where, as we shall see below, one can thereby get a short proof via a special argument.

4.2.2. *Some special cases.* To get a better insight we begin by discussing some pertinent special cases.

Theorem 4.2 is clearly trivial if T is periodic, i.e. if for some m, $T^m = Id$. The next case, in order of complexity, is that of T being almost periodic, say a translation by an irrational α on the unit circle. Let $\|x\|$ denote the distance from a real number x to the nearest integer. If α is irrational, then, as it is easy to see, for any $\varepsilon > 0$, the set $\{n \in \mathbb{Z}\colon \|n\alpha\| < \varepsilon\}$ is syndetic. (It is actually IP*.) Hence for a syndetic set of n, the operator T^n is ε-close to the identity operator (in the strong topology on the space of operators). It follows that, in this case, for any $\varepsilon > 0$ the set $\{n\colon |\mu(A \cap T^{-n}A \cap \cdots \cap T^{-kn}A) - \mu(A)| < \varepsilon\}$ is syndetic, which is clearly more than enough for our purposes.

A slightly more general class of measure preserving systems, for which a similar argument works, is the class of so-called *compact systems* (another term: systems with discrete spectrum). These are defined by the requirement that any $f \in L^2(X, \mathcal{B}, \mu)$ is *compact*, i.e. the closure of the orbit $\{T^n f\}_{n\in\mathbb{Z}}$ in L^2 is compact. To see that this is indeed only a slightly more general situation, note that one can show (see [77, Theorem 4]) that if $(X, \mathcal{B}, \mu, T)$ is a compact ergodic system, then it is conjugate to a translation on a compact Abelian group. Now, if $(X, \mathcal{B}, \mu, T)$ is a compact system and $A \in \mathcal{B}$ with $\mu(A) > 0$, then, as before, the set $\{n\colon \|T^n f - f\| < \varepsilon\}$ is syndetic and we see that in this case Theorem 4.2 holds for the same reason as in the case of the irrational translation.

Let us assume now that the system $(X, \mathcal{B}, \mu, T)$ is such that no nonconstant function $f \in L^2$ is compact. In particular, this means that the unitary operator induced on L^2 by T (and which, by the customary abuse of notation, we will often be denoting also by T) has no nontrivial eigenfunctions. Measure preserving systems with this property were introduced, under the name *dynamical systems with continuous spectra*, in [92] and form one of the most important classes of measure preserving systems. Today, such systems are called *weakly mixing systems*. We refer the reader to [18] and [1, Section 3.6], for an overview of mixing properties of measure preserving systems. As we shall see below, Theorem 4.2 can be verified for weakly mixing transformations with relative ease. Before showing this, we want to summarize various equivalent forms of weak mixing in the following theorem. For the proofs see [92] (where the stress is placed on measure preserving $\mathbb{R}$-actions), [86], or any of the more modern texts such as [76,135], or [113]. Note that in most books, either

(i) or (ii) below is taken as the "official" definition of weak mixing, whereas the original definition in [92] corresponds to condition (vi).

THEOREM 4.4. *Let T be an invertible measure preserving transformation of a probability measure space $(X, \mathcal{B}, \mu)$. Let U_T denote the operator defined on measurable functions by $(U_T f)(x) = f(Tx)$. The following conditions are equivalent*:

(i) *For any $A, B \in \mathcal{B}$*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \left|\mu(A \cap T^{-n} B) - \mu(A)\mu(B)\right| = 0;$$

(ii) *For any $A, B \in \mathcal{B}$ there is a set $P \subset \mathbb{N}$ of density zero such that*

$$\lim_{n\to\infty,\, n\notin P} \mu(A \cap T^{-n} B) = \mu(A)\mu(B);$$

(iii) *$T \times T$ is ergodic on the Cartesian square of $(X, \mathcal{B}, \mu)$;*

(iv) *For any ergodic probability measure preserving system $(Y, \mathcal{D}, \nu, S)$ the transformation $T \times S$ is ergodic on $X \times Y$;*

(v) *If f is a measurable function such that for some $\lambda \in \mathbb{C}$, $U_T f = \lambda f$ a.e., then $f =$ const a.e.;*

(vi) *For $f \in L^2(X, \mathcal{B}, \mu)$ with $\int f = 0$ consider the representation of the positive definite sequence $\langle U_T^n f, f\rangle, n \in \mathbb{Z}$, as a Fourier transform of a measure ν on $\mathbb{T}$:*

$$\left\langle U_T^n f, f\right\rangle = \int_{\mathbb{T}} e^{2\pi i n x}\, d\nu, \quad n \in \mathbb{Z}$$

(*this representation is guaranteed by Herglotz theorem, see* [80]). *Then ν has no atoms.*

As we shall see below, it is the relativized version of weak mixing, that is, the notion of *weak mixing relative to a factor*, that plays an important role in the analysis of the structure of an arbitrary dynamical system and which is behind the proof of Theorem 4.2. First, let us verify the validity of Theorem 4.2 for weakly mixing systems.

THEOREM 4.5. *If $(X, \mathcal{B}, \mu, T)$ is a weakly mixing system, then for any $k \in \mathbb{N}$ and any $f_i \in L^\infty(X, \mathcal{B}, \mu)$, $i = 1, 2, \ldots, k$, one has*:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} T^n f_1 T^{2n} f_2 \cdots T^{kn} f_k = \int f_1\, d\mu \int f_2\, d\mu \cdots \int f_k\, d\mu$$

in the L^2-norm.

Theorem 4.5 implies that for any $f_i \in L^\infty$, $i = 0, 1, \ldots, k$, one has

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \int f_0 T^n f_1 \cdots T^{kn} f_k = \int f_0 \, d\mu \int f_1 \, d\mu \cdots \int f_k \, d\mu.$$

Putting $f_i = 1_A$, $i = 0, 1, \ldots, k$, gives us

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu(A \cap T^{-n} A \cap \cdots \cap T^{-kn} A) = \big(\mu(A)\big)^{k+1}.$$

As a matter of fact, Theorem 4.5 implies that for some set $E \subset \mathbb{N}$ having zero density, one has

$$\lim_{\substack{n\to\infty \\ n\notin E}} \mu(A \cap T^{-n} A \cap \cdots \cap T^{-kn} A) = \mu(A)^{k+1}. \tag{4.1}$$

To see this, note first that, since T is weakly mixing, it follows from Theorem 4.4, (iii) and (iv), that not only $T \times T$ is ergodic, but also $(T \times T) \times T = T \times T \times T$ and $(T \times T \times T) \times T = (T \times T) \times (T \times T)$ are ergodic (on X^3 and X^4 respectively). But then $T \times T$ is weakly mixing. Applying Theorem 4.5 to $T \times T$ and performing routine manipulations one gets, for any $f_i \in L^\infty$, $i = 0, 1, \ldots, k$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \left(\int f_0 T^n f_1 \cdots T^{kn} f_k \, d\mu - \int f_0 \, d\mu \int f_1 \, d\mu \cdots \int f_k \, d\mu \right)^2 = 0,$$

which implies (4.1).

In the proof of Theorem 4.5, we shall utilize the following version of the van der Corput trick (cf. Theorem 1.32). For the proof see, for example, [25, p. 445].

THEOREM 4.6. *Let $(u_n)_{n\in\mathbb{N}}$ be a bounded sequence in a Hilbert space $\mathcal{H}$. If for every $h \in \mathbb{N}$ it is the case that $\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \langle u_{n+h}, u_n \rangle$ exists and if*

$$\lim_{H\to\infty} \frac{1}{H} \sum_{h=1}^{H} \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \langle u_{n+h}, u_n \rangle = 0,$$

then $\lim_{N\to\infty} \| \frac{1}{N} \sum_{n=1}^{N} u_n \| = 0$.

PROOF OF THEOREM 4.5. Since any weakly mixing system is ergodic, the claim of the theorem trivially holds for $k = 1$. To see how the induction works, consider the case $k = 2$. Since the case when one of f_1, f_2 is constant brings us back to $k = 1$, we can assume, in view of the identity $f = (f - \int f) + \int f$, that $\int f_1 \, d\mu = 0$. Let now $u_n = T^n f_1 T^{2n} f_2$.

By the ergodicity of T, we have:

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\langle u_{n+h},u_n\rangle &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\int T^{n+h}f_1T^{2n+2h}f_2T^nf_1T^{2n}f_2\,d\mu\\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\int T^hf_1T^{n+2h}f_2f_1T^nf_2\,d\mu\\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\int \bigl(f_1T^hf_1\bigr)T^n\bigl(f_2T^{2h}f_2\bigr)\,d\mu\\ &= \int f_1T^hf_1\,d\mu\int f_2T^{2h}f_2\,d\mu.\end{aligned}$$

We remark now that if T is weakly mixing, then T^2 is also weakly mixing and hence $T\times T^2$ is ergodic (on the product space $(X\times X,\mathcal{B}\times\mathcal{B},\mu\times\mu)$). Writing $f_1\otimes f_2$ for $f_1(x)f_2(y)$ and using the ergodicity of $T\times T^2$ we have:

$$\begin{aligned}&\lim_{H\to\infty}\frac{1}{H}\sum_{h=1}^{H}\int f_1T^hf_1\,d\mu\int f_2T^{2h}f_2\,d\mu\\ &\quad=\lim_{H\to\infty}\frac{1}{H}\sum_{h=1}^{H}\int (f_1\otimes f_2)\bigl(T\times T^2\bigr)^h(f_1\otimes f_2)\,d(\mu\times\mu)\\ &\quad=\left(\int f_1\otimes f_2\,d(\mu\times\mu)\right)^2=\left(\int f_1\,d\mu\right)^2\left(\int f_2\,d\mu\right)^2=0.\end{aligned}$$

The result now follows from Theorem 4.6. Note now that the same argument (in which one uses the ergodicity of $T\times T^2\times\cdots\times T^k$) works for general k. We are done. □

REMARK 4.7.

(1) It is not hard to show that if the system $(X,\mathcal{B},\mu,T)$ is such that for any $f_1,f_2\in L^\infty(X,\mathcal{B},\mu)$, one has $\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}T^nf_1T^{2n}f_2=\int f_1\,d\mu\int f_2\,d\mu$ in the L^2-norm, then T is weakly mixing.

(2) By using a modification of Theorem 4.6, which pertains to the uniform Cesàro averages $\frac{1}{N-M}\sum_{n=M}^{N-1}x_n$ (see, for example, Remark 2.2 in [25]), one can show that in Theorem 4.5 one actually has

$$\lim_{N-M\to\infty}\frac{1}{N-M}\sum_{n=M}^{N-1}T^nf_1T^{2n}f_2\cdots T^{kn}f_k=\int f_1\,d\mu\int f_2\,d\mu\cdots\int f_k\,d\mu,$$

which implies

$$\lim_{N-M\to\infty} \frac{1}{N-M} \sum_{n=M}^{N-1} \mu(A \cap T^{-n}A \cap \cdots \cap T^{-kn}A) = \mu(A)^{k+1}.$$

(3) The reader is invited to check that the proof above actually gives the following more general result, first proved in [6].

THEOREM 4.8. *Assume that, for $k \geqslant 2$, $T_1, T_2, \ldots, T_k$ are commuting measure preserving transformations on a probability space $(X, \mathcal{B}, \mu)$. Then the following are equivalent*:

(i) *For any $f_1, f_2, \ldots, f_k \in L^\infty(X, \mathcal{B}, \mu)$ one has*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} T_1^n f_1 T_2^n f_2 \cdots T_k^n f_k = \int f_1 \, d\mu \int f_2 \, d\mu \cdots \int f_k \, d\mu \quad \text{in } L^2.$$

(ii) *For any $i \neq j$, $T_i T_j^{-1}$ is ergodic on X and $T_1 \times T_2 \times \cdots \times T_k$ is ergodic on X^k.*

The two special cases of Theorem 4.2, which we verified above, correspond on a spectral level to two complementary classes of unitary operators, namely those having discrete spectrum and continuous spectrum. While these two cases are much too special to allow us to conclude the proof of Theorem 4.2 for general k, they are sufficient for $k = 2$ (which constitutes the first nontrivial case of Theorem 4.2). These two special cases are also important in that they indicate a possible line of attack which we will discuss after first completing the proof for $k = 2$.

4.2.3. *The case of $k = 2$*

PROOF OF THEOREM 4.2 FOR $k = 2$. Assume first that T is ergodic, and consider the following splitting of $L^2(X, \mathcal{B}, \mu) = \mathcal{H}$. $\mathcal{H} = \mathcal{H}_c \oplus \mathcal{H}_{wm}$, where the T-invariant subspaces $\mathcal{H}_c$ and $\mathcal{H}_{wm}$ are defined as follows:

$$\begin{aligned} \mathcal{H}_c &= \overline{\text{Span}\{f \in \mathcal{H}\text{: there exists } \lambda \in \mathbb{C} \text{ with } Tf = \lambda f\}} \\ &= \big\{f \in \mathcal{H}\text{: the orbit } (T^n f)_{n\in\mathbb{Z}} \text{ is precompact in norm topology}\big\}, \end{aligned}$$

$$\mathcal{H}_{wm} = \mathcal{H}_c^\perp = \left\{ f \in \mathcal{H}\text{: } \forall g \in \mathcal{H},\ \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \big|\langle T^n f, g\rangle\big| = 0 \right\}.$$

(We remark in passing that this splitting is valid for any unitary operator, and moreover, can be defined in such a way that it makes sense for any group of unitary operators.)

Writing f for 1_A let $f = f_c + f_{wm}$ where $f_c \in \mathcal{H}_c$, $f_{wm} \in \mathcal{H}_{wm}$. Note that $f_c \geqslant 0$ and $\int f_c \, d\mu = \mu(A)$, while $\int f_{wm} \, d\mu = 0$. The nonnegativity of f_c follows from an argument similar to the one used in the proof of Theorem 1.31 to establish the nonnegativity of the projection of 1_A on the space $\mathcal{H}_{\text{rat}}$. Note also that applying this argument again to $1 - f_c$ gives us that $0 \leqslant f_c \leqslant 1$. Since $f_{wm} = 1_A - f_c$, it follows also that f_{wm} satisfies $|f_{wm}| \leqslant 1$.

Using the decomposition $f = f_c + f_{wm}$, we have

$$\frac{1}{N}\sum_{n=0}^{N-1} T^n f T^{2n} f = \frac{1}{N}\sum_{n=0}^{N-1} T^n f_c T^{2n} f_c + \sum_{n=0}^{N-1} T^n f_c T^{2n} f_{wm} + \frac{1}{N}\sum_{n=0}^{N-1} T^n f_{wm} T^{2n} f_c + \frac{1}{N}\sum_{n=0}^{N-1} T^n f_{wm} T^{2n} f_{wm}.$$

We claim that the last three expressions (in which f_{wm} occurs) have zero limit in L^2 as $N \to \infty$. To see this the reader is invited to reexamine the proof of Theorem 4.5 and to observe that it was actually shown there that if $\varphi \in \mathcal{H}_{wm}$, then for any $\psi \in \mathcal{H} = L^2(X, \mathcal{B}, \mu)$, one has (assuming that at least one of φ, ψ is bounded):

$$\lim_{N\to\infty} \left\| \frac{1}{N}\sum_{n=0}^{N-1} T^n \varphi T^{2n} \psi \right\| = \lim_{N\to\infty} \left\| \frac{1}{N}\sum_{n=0}^{N-1} T^n \psi T^{2n} \varphi \right\| = 0.$$

So, we see that in L^2,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} T^n f T^{2n} f = \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} T^n f_c T^{2n} f_c,$$

if the latter limit exists. But the existence of this limit clearly follows from the fact that this is certainly the case when one substitutes for f_c a finite linear combination of eigenfunctions of T and that eigenfunctions span the space $\mathcal{H}_c$. So we have that $\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} T^n f T^{2n} f$ also exists in L^2, and hence

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \int f T^n f T^{2n} f \, d\mu = \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \int 1_A T^n 1_A T^{2n} 1_A \, d\mu = \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \mu\big(A \cap T^{-n} A \cap T^{-2n} A\big)$$

exists as well.

It remains to establish the positivity of the limit in question. Note that, for bounded $g_1, g_2 \in \mathcal{H}_c$, one has $g_1 \cdot g_2 \in \mathcal{H}_c$, and hence f_{wm} is orthogonal to $T^n f_c T^{2n} f_c$. We have:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \mu\big(A \cap T^{-n} A \cap T^{-2n} A\big) = \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \int f T^n f_c T^{2n} f_c \, d\mu$$

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \int (f_c + f_{wm}) T^n f_c T^{2n} f_c \, d\mu$$

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \int f_c T^n f_c T^{2n} f_c \, d\mu.$$

Note now that since f_c is a compact function, the set

$$\{n \in \mathbb{Z}\colon \|T^n f_c - f_c\| < \varepsilon\}$$

is syndetic for every $\varepsilon > 0$, and hence the set

$$S_\varepsilon = \left\{ n \in \mathbb{Z}\colon \left| \int f_c T^n f_c T^{2n} f_c \, d\mu - \int f_c^3 \, d\mu \right| < \varepsilon \right\}$$

is also syndetic. Note also that $\int f_c^3 \, d\mu \geqslant (\int f_c)^3 = (\mu(A))^3$.

Therefore, if ε is small enough, we shall have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu\big(A \cap T^{-n} A \cap T^{-2n} A\big)$$

$$\geqslant \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n \in S_\varepsilon \cap [0, N-1]} \mu\big(A \cap T^{-n} A \cap T^{-2n} A\big) > 0.$$

To finish the proof of this special case one uses the ergodic decomposition. It is not hard to see that in the nonergodic case, both the convergence and the positivity of the limit hold as well. We omit the details. □

As a bonus, we have obtained the fact that for any measure preserving system $(X, \mathcal{B}, \mu, T)$ and any $f, g \in L^\infty$, $\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} T^n f T^{2n} g$ exists in L^2 and equals $\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} T^n f_c T^{2n} g_c$, where f_c, g_c denote the orthogonal projections of f, g on $\mathcal{H}_c$. One, naturally, would like to know whether, in general, one has the convergence of the expressions of the form $\frac{1}{N} \sum_{n=0}^{N-1} T^n f_1 T^{2n} f_2 \cdots T^{kn} f_k$, where $f_i \in L^\infty(X, \mathcal{B}, \mu)$, $i = 1, 2, \ldots, k$. For $k = 3$ the positive answer to this question was provided in [47] for totally ergodic T and in full generality in [66,138] and [87]. The recalcitrant problem of establishing the convergence for general k was solved only recently, in the remarkable work of B. Host and B. Kra [88] and T. Ziegler [139]. See Section 5 below for a discussion of various convergence results which are suggested by combinatorial applications of ergodic theory. See also Appendices A and B written by A. Leibman and A. Quas and M. Wierdl which deal with convergence issues. Note, however, that while the study of convergence is more fundamental from the point of view of ergodic theory, it is (multiple) recurrence, i.e. the positivity of the expressions like $\mu(A \cap T^{-n} A \cap T^{-2n} A \cap \cdots \cap T^{-kn} A)$, which is needed for combinatorial and number-theoretic applications. (Nonetheless, convergence

results may, in some cases, provide the shortest path to establishing recurrence. This point is certainly supported by the proof of Theorem 1.31 and the above discussion of the $k = 2$ case of Theorem 4.2. See also Theorem 5.21(i) below.)

4.2.4. *The structure of ergodic systems.* We return now to our discussion of Theorem 4.5. It turns out that for $k > 2$, the Hilbertian splitting utilized above for $k = 2$ is no longer sufficient, and in order to establish multiple recurrence one has to undertake a deeper study of the structure of general measure preserving systems. In order to describe the main points of Furstenberg's approach, we will review first some general facts. For more information and missing details, the reader is encouraged to consult [57,58] and [61]. Our presentation below follows mainly [61] where a simplified proof of Theorem 4.2 is presented. The only significant point of departure from [61] is in the treatment of compact extensions (see Definition 4.15 below), where we will use a "soft" argument based on van der Waerden's theorem. One of the reasons for this choice is that coloring theorems seem to be indispensable in proving more sophisticated multiple recurrence results and we want to use this opportunity to acquaint the reader with this technique.

Given two probability measure spaces $(X, \mathcal{B}, \mu)$ and $(Y, \mathcal{D}, \nu)$ and a map $\pi : X \to Y$ such that $\pi^{-1}(\mathcal{D}) \subset \mathcal{B}$ and $\pi\mu = \nu$, we say that $(X, \mathcal{B}, \mu)$ is an *extension* of $(Y, \mathcal{D}, \nu)$, and that $(Y, \mathcal{D}, \nu)$ is a *factor* of $(X, \mathcal{B}, \mu)$.

Under mild conditions on the regularity of the space $(X, \mathcal{B}, \mu)$ (which are usually satisfied in the case of Lebesgue spaces—our standing assumption), one can associate with the factor $(Y, \mathcal{D}, \nu)$ a family of measures $\{\mu_y\}_{y\in Y}$ on $(X, \mathcal{B})$ with the following properties:

(i) For each $f \in L^1(X, \mathcal{B}, \mu)$ one has $f \in L^1(X, \mathcal{B}, \mu_y)$ for a.e. $y \in Y$.

(ii) The function $g(y) = \int f\, d\mu_y$ belongs to $L^1(Y, \mathcal{D}, \nu)$ and

$$\int \left(\int f(x)\, d\mu_y(x) \right) d\nu = \int f(x)\, d\mu(x).$$

(iii) If f is measurable with respect to $\pi^{-1}(\mathcal{D})$, then

$$\int f\, d\mu_{\pi(x)} = f(x) \quad \text{a.e.}$$

Using the family $\{\mu_y\}_{y\in Y}$, we will write $\mu = \int \mu_y\, d\nu(y)$ (which means that, for any $A \in \mathcal{B}$, $\mu(A) = \int \mu_y(A)\, d\nu(y)$) and refer to this decomposition as the *disintegration* of μ with respect to the factor $(Y, \mathcal{D}, \nu)$.

For any $1 \leqslant p \leqslant \infty$ one can define the *conditional expectation* operator $E(\cdot|Y)$ from $L^p(X, \mathcal{B}, \mu)$ to $L^p(Y, \mathcal{D}, \nu)$ by the formula

$$E(f|Y)(y) = \int f\, d\mu_y, \quad f \in L^2(X, \mathcal{B}, \mu).$$

Clearly, for $f \geqslant 0$, one has $E(f|Y) \geqslant 0$ and $E(1|Y) = 1$. Also, by property (ii) above, one has $\int f\, d\mu = \int E(f|Y)\, d\nu$.

Note that, given the measure space $(X, \mathcal{B}, \mu)$, there is a natural 1-1 correspondence between its factors and sub-σ-algebras of $\mathcal{B}$. This correspondence allows one to identify the space $L^2(Y, \mathcal{D}, \nu)$ with a closed subspace of $L^2(X, \mathcal{B}, \mu)$ which is of the form $L^2(X, \mathcal{B}_1, \mu)$, where $\mathcal{B}_1 = \pi^{-1}(\mathcal{D})$. This, in turn, leads to a convenient interpretation of conditional expectation operator as the orthogonal projection $L^2(X, \mathcal{B}, \mu) \to L^2(X, \mathcal{B}_1, \mu) \cong L^2(Y, \mathcal{D}, \nu)$.

For any $f \in L^\infty(Y, \mathcal{D}, \nu)$ (viewed as a bounded function in $L^2(X, \mathcal{B}, \mu)$ which is measurable with respect to $\mathcal{B}_1$), one has $E(gf|Y) = fE(g|Y)$. For more details on conditional expectation operators see [58, Chapter 5, Section 3] or [36, Section 34].

Suppose now that $(X_1, \mathcal{B}_1, \mu_1)$ and $(X_2, \mathcal{B}_2, \mu_2)$ are extensions of $(Y, \mathcal{D}, \nu)$ and $\pi_1 : X_1 \to Y$, $\pi_2 : X_2 \to Y$ are the corresponding measure preserving mappings. One can form the *fibre product* space $(X, \mathcal{B}, \mu)$, where

$$X = X_1 \times_Y X_2 = \big\{(x_1, x_2) \in X_1 \times X_2 \colon \pi_1(x_1) = \pi_2(x_2)\big\},$$

$\mathcal{B}$ is the restriction of $\mathcal{B}_1 \times \mathcal{B}_2$ to X, and μ is defined via the disintegrations $\{\mu_y^{(1)}\}_{y\in Y}$, $\{\mu_y^{(2)}\}_{y\in Y}$ by the formula

$$\mu(A) = (\mu_1 \times_Y \mu_2)(A) = \int \big(\mu_y^{(1)} \times \mu_y^{(2)}\big)(A)\, d\nu(y).$$

The notions of extension, factor, and fibre product are naturally extended to measure preserving systems. Given two probability measure preserving systems $\boldsymbol{X} = (X, \mathcal{B}, \mu, T)$ and $\boldsymbol{Y} = (Y, \mathcal{D}, \nu, S)$, one says that $\boldsymbol{X}$ is an extension of $\boldsymbol{Y}$, and $\boldsymbol{Y}$ a factor of $\boldsymbol{X}$, if the corresponding map $\pi : X \to Y$ is not only measure preserving but also satisfies $S\pi(x) = \pi T(x)$ for a.e. $x \in X$. We have now the following formulas:

(iii) For almost every $y \in Y$, $T\mu_y = \mu_{Sy}$, meaning $\mu_y(T^{-1}A) = \mu_{Sy}(A)$ for any $A \in \mathcal{B}$.

(iv) For any $f \in L^2(X, \mathcal{B}, \mu)$, $SE(f|Y) = E(Tf|Y)$.

When the system $\boldsymbol{X} = (X, \mathcal{B}, \mu, T)$ is not ergodic, it has a natural nontrivial factor $\boldsymbol{X}_{\text{inv}} = (X, \mathcal{B}_{\text{inv}}, \mu, T)$, where $\mathcal{B}_{\text{inv}}$ is the σ-algebra of T-invariant sets in $\mathcal{B}$. It is not hard to see that the disintegration of μ corresponding to this factor is nothing but the classical ergodic decomposition of μ (which was treated first in [133]). Another natural example of a factor, which we have, implicitly, encountered already, is associated with the space $\mathcal{H}_c$ of compact functions. Indeed, one has the following theorem.

THEOREM 4.9. (Cf. [93, Theorem 2.2].) *Let $(X, \mathcal{B}, \mu, T)$ be a probability measure preserving system and let $\mathcal{B}_c$ be the smallest σ-algebra in $\mathcal{B}$ with respect to which the elements of $\mathcal{H}_c$ are measurable. Then $\mathcal{B}_c$ is T-invariant and $\mathcal{H}_c \cong L^2(X, \mathcal{B}_c, \mu)$.*

Clearly, $(X, \mathcal{B}_c, \mu, T)$ is a maximal compact factor. As we have already mentioned above, if T is ergodic, the system $(X, \mathcal{B}_c, \mu, T)$ is conjugate to a translation on a compact Abelian group (see [77, Theorem 4]). In this case, $(X, \mathcal{B}_c, \mu, T)$ is often called the (maximal) Kronecker factor.

We can formulate now a criterion in terms of factors for a system to be weakly mixing. (Note that this is just a new way of expressing a familiar concept.)

THEOREM 4.10. *A probability measure preserving system is weakly mixing if and only if it has no nontrivial compact factors*.

In order to prove Theorem 4.2, one has to study the relativized notions of weak mixing and compactness with respect to a factor.

To define a weak mixing extension, one needs the notion of a relative product of measure preserving systems. Let $\boldsymbol{X}_1 = (X_1, \mathcal{B}_1, \mu_1, T_1)$ and $\boldsymbol{X}_2 = (X_2, \mathcal{B}_2, \mu_2, T_2)$ be extensions of $\boldsymbol{Y} = (Y, \mathcal{D}, \nu, S)$. We claim that the measure $\mu_1 \times_Y \mu_2$ is $(T_1 \times T_2)$-invariant, that is, for any measurable $A \subseteq X_1 \times_Y X_2$ one has $\mu_1 \times_Y \mu_2((T_1 \times T_2)^{-1}A) = \mu_1 \times_Y \mu_2(A)$. One needs only to verify this for the sets of the form $A = A_1 \times A_2$ where $A_i \in \mathcal{B}_i$. By definition of $\mu_1 \times_Y \mu_2$ we have

$$\begin{aligned}
&\mu_1 \times_Y \mu_2\big((T_1 \times T_2)^{-1}(A_1 \times A_2)\big) \\
&\quad = \int \mu_y^{(1)} \times \mu_y^{(2)}\big(T_1^{-1}A_1 \times T_2^{-1}A_2\big)\, d\nu(y) \\
&\quad = \int T_1\mu_y^{(1)} \times T_2\mu_y^{(2)}(A_1 \times A_2)\, d\nu(y) = \int \mu_{Sy}^{(1)} \times \mu_{Sy}^{(2)}(A_1 \times A_2)\, d\nu(y) \\
&\quad = \int \mu_y^{(1)} \times \mu_y^{(2)}(A_1 \times A_2)\, dS\nu(y) = \int \mu_y^{(1)} \times \mu_y^{(2)}(A_1 \times A_2)\, d\nu(y) \\
&\quad = \mu_1 \times_Y \mu_2(A_1 \times A_2),
\end{aligned}$$

and so $\boldsymbol{X}_1 \times_{\boldsymbol{Y}} \boldsymbol{X}_2 = (X_1 \times X_2, \mathcal{B}_1 \times \mathcal{B}_2, \mu_1 \times_Y \mu_2, T_1 \times T_2)$ is a measure preserving system, which is called the relative product of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (with respect to $\boldsymbol{Y}$).

DEFINITION 4.11. The system $\boldsymbol{X} = (X, \mathcal{B}, \mu, T)$ is an *ergodic extension* of $\boldsymbol{Y} = (Y, \mathcal{D}, \nu, S)$ if the only T-invariant sets in $\mathcal{B}$ are preimages of the invariant sets in $\mathcal{D}$. The system $\boldsymbol{X}$ is a *weakly mixing extension* of $\boldsymbol{Y}$ if $\boldsymbol{X} \times_{\boldsymbol{Y}} \boldsymbol{X}$ is an ergodic extension of $\boldsymbol{Y}$.

One can show that most properties of the "absolute" weak mixing (and in particular, items (i) through (iv) in Theorem 4.4) extend, with obvious modifications, to statements about relative weak mixing. For example, one has the following fact.

THEOREM 4.12. (Cf. [58, Proposition 6.2].) *A measure preserving system $(X, \mathcal{B}, \mu, T)$ is a weak mixing extension of $(Y, \mathcal{D}, \nu, S)$ if and only if for any $A_1, A_2 \in \mathcal{B}$ one has*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \int \big(\mu_y(A_1 \cap T^{-n}A_2) - \mu_y(A_1)\mu_y(T^{-n}A_2)\big)^2\, d\nu(y) = 0.$$

Moreover, by using Theorem 4.6, one can obtain (by an argument analogous to the one used in the proof of Theorem 4.5) the following result.

THEOREM 4.13. *If $(X, \mathcal{B}, \mu, T)$ is a weak mixing extension of $(Y, \mathcal{D}, \nu, S)$, then for any $A_0, A_1, \ldots, A_k \in \mathcal{B}$ one has*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \int \big(\mu_y\big(A_0 \cap T^{-n} A_1 \cap T^{-2n} A_2 \cdots \cap T^{-kn} A_k\big) - \mu_y(A_0)\mu_y(T^{-n} A_1) \cdots \mu_y(T^{-kn} A_k)\big)^2 \, d\nu(y) = 0.$$

4.2.5. *Multiple recurrence in the general case.* Let us say, following [61], that a system $\boldsymbol{X} = (X, \mathcal{B}, \mu, T)$ has the SZ property, or that T is SZ, if Theorem 4.2 holds for $\boldsymbol{X}$. (For example, as we have already seen above, compact and weakly mixing systems do have the SZ property.)

We have now the following corollary of Theorem 4.13:

THEOREM 4.14 [61, Theorem 8.4]. *If $(X, \mathcal{B}, \mu, T)$ is weakly mixing extension of $(Y, \mathcal{D}, \nu, S)$ and the transformation S is SZ, then $(X, \mathcal{B}, \mu, T)$ has the SZ property.*

PROOF. Let $A \in \mathcal{B}$ with $\mu(A) > 0$ and denote $f = 1_A$. Note that if $a > 0$ is small enough then the set $A_1 = \{y\colon\ E(1_A|Y)(y) \geqslant a\}$ satisfies $\nu(A_1) > 0$. It follows now from Theorem 4.13 (and the formula $E(f|Y)(y) = \int f \, d\mu_y$) that, since $E(1_A|Y) \geqslant a \cdot 1_{A_1}$,

$$\frac{1}{N} \sum_{n=0}^{N-1} \mu(A \cap T^{-n} A \cap \cdots \cap T^{-kn} A) > \frac{1}{2} a^{k+1} \cdot \frac{1}{N} \sum_{n=0}^{N-1} \nu(A_1 \cap S^{-n} A_1 \cap \cdots \cap S^{-kn} A_1)$$

for all large enough N. The result now follows from the assumption that $(Y, \mathcal{D}, \nu, S)$ has the SZ property. □

We will define now relatively compact extensions and show that an analogue of Theorem 4.14 holds. Unlike Theorem 4.14, which is a more or less straightforward extension of Theorem 4.5, the argument needed for the treatment of compact extensions (the mere definition of which is, in our opinion, much less trivial than that of weak mixing extensions) is perhaps the most subtle part of the proof of Theorem 4.2.

DEFINITION 4.15. Let $\boldsymbol{X} = (X, \mathcal{B}, \mu, T)$ be an extension of $\boldsymbol{Y} = (Y, \mathcal{D}, \nu, S)$. Call a function $f \in L^2(X, \mathcal{B}, \mu)$ *almost periodic*, or an AP-function, *relative* to $\boldsymbol{Y}$ if for any $\varepsilon > 0$ there exist $r \in \mathbb{N}$ and functions $g_1, g_2, \ldots, g_r \in L^2(X, \mathcal{B}, \mu)$ such that, for every $n \in \mathbb{Z}$, $\min_{1\leqslant s\leqslant r} \|T^n f - g_s\|_{L^2(\mu_y)} < \varepsilon$ for almost every $y \in Y$. We say that $\boldsymbol{X}$ is a *compact extension* of $\boldsymbol{Y}$ if AP-functions are dense in $L^2(X, \mathcal{B}, \mu)$.

Clearly any compact system (i.e. a system for which the subspace $\mathcal{H}_c$ coincides with $L^2(X, \mathcal{B}, \mu)$) is a compact extension of the trivial one-point system. Note that in this case every element of $L^2(X, \mathcal{B}, \mu)$ is an AP-function.

A less trivial example and, in a sense, a typical one is given by so-called isometric extensions. Let $\boldsymbol{Y} = (Y, \mathcal{D}, \nu, S)$ be an arbitrary system and let Z be a compact metric space equipped with a probability measure η on $\mathcal{B}_Z$, the (completion of the) σ-algebra of Borel sets in Z. Suppose that G is a compact group of isometries of Z and define for some measurable family $\sigma(y)$ of elements of G a transformation T on $X = Y \times Z$ by

$$T(y, z) = \big(Sy, \sigma(y)z\big).$$

One can verify that the system $\boldsymbol{X} = (Y \times Z, \mathcal{D} \times \mathcal{B}_Z, \nu \times \eta, T)$ is a compact extension of $\boldsymbol{Y}$. Perhaps the shortest path to this verification is to consider the case of Z a sphere and G its group of rotations, using the fact that finite-dimensional spaces of spherical functions are invariant under rotations, and these are dense in $L^2(Z)$, and to observe that a similar argument works for general isometric extensions. Note that for nontrivial isometric extensions it is no longer true that every L^2 function is AP relatively to the given factor. Even functions of the simple form $f(y)e^{2\pi i z}$ (where Z is the circle) will not be AP unless f is bounded.

The importance of weakly mixing and compact extensions lies with the fundamental fact, established by Furstenberg in the course of his proof of Szemerédi's theorem, that any system $\boldsymbol{X} = (X, \mathcal{B}, \mu, T)$ appears in a chain (possibly transfinite), $\boldsymbol{X} \to \cdots \to \boldsymbol{X}_{\alpha+1} \to \boldsymbol{X}_\alpha \to \cdots \to \boldsymbol{X}_1 \to \boldsymbol{X}_0$, in which the individual links $\boldsymbol{X}_{\alpha+1} \to \boldsymbol{X}_\alpha$ are either compact or weakly mixing extensions. (As a matter of fact, one can take all of the extensions, with the possible exception of the last link $\boldsymbol{X} = \boldsymbol{X}_{\eta+1} \to \boldsymbol{X}_\eta$, to be compact.)

The topological predecessor of this ergodic-theoretical structure theorem is a similar structure theorem, also due to Furstenberg, for distal systems, which states that any distal system can be seen as a tower of *isometric* extensions. See [56] for details. The structure theory of distal systems works for general locally compact group actions, which hints that measure theoretical structure theory can also be established in this generality. This was done in independent work of Zimmer. (See [140,141].)

Returning to the discussion of the proof of Theorem 4.2, we are in position now to describe the general scheme of the proof. As we shall show in detail below, if $\boldsymbol{X}$ is a compact extension of $\boldsymbol{Y}$ and $\boldsymbol{Y}$ has the SZ property, then $\boldsymbol{X}$ also does. Now, one can show by a routine argument that any totally ordered by inclusion family of factors of a system $(X, \mathcal{B}, \mu, T)$ which have the SZ property has a maximal element. (See [61, Proposition 7.1].) But then it follows from the structure theorem cited above that this maximal factor has to be $(X, \mathcal{B}, \mu, T)$ itself, which gives Theorem 4.2. A somewhat shorter (or, rather, less involved) path, which avoids the full strength of the structure theorem, is via the following proposition.

THEOREM 4.16 [61, Theorem 5.10]. *If $\boldsymbol{X} = (X, \mathcal{B}, \mu, T)$ is an extension of $\boldsymbol{Y} = (Y, \mathcal{D}, \nu, S)$ which is not relatively weak mixing, then there exists a strictly intermediate factor $\boldsymbol{X}^*$ between $\boldsymbol{Y}$ and $\boldsymbol{X}$ such that $\boldsymbol{X}^*$ is a compact extension of $\boldsymbol{Y}$.*

Either way, all that is needed now to bring the proof of Theorem 4.2 to conclusion is the following result.

THEOREM 4.17. *If $\boldsymbol{X} = (X, \mathcal{B}, \mu, T)$ is a compact extension of $\boldsymbol{Y} = (Y, \mathcal{D}, \nu, S)$ and $\boldsymbol{Y}$ has the SZ property, then $\boldsymbol{X}$ also does.*

PROOF. We shall utilize van der Waerden's theorem on arithmetic progressions. To make the ideas clear (and to stress the relevance of van der Waerden's theorem), let us first go back to the "absolute" case and show how the proof works when $(X, \mathcal{B}, \mu, T)$ is a compact system. Let $A \in \mathcal{B}$ with $\mu(A) > 0$, let $f = 1_A$ and let, for a given $\varepsilon > 0$, $g_1, g_2, \ldots, g_r$ be elements of the compact set $K = \overline{\{T^n f\}_{n\in\mathbb{Z}}}$ such that for any $n \in \mathbb{Z}$ there is $j = j(n)$ in $\{1, 2, \ldots, r\}$ satisfying $\|T^n f - g_{j(n)}\| < \varepsilon$. This naturally defines an r-coloring $\mathbb{Z} = \bigcup_{i=1}^r C_i$, and by van der Waerden's theorem, there exists $j \in \{1, 2, \ldots, r\}$ such that for some $m \in \mathbb{Z}$ and $n \in \mathbb{N}$ one has $\|T^{m+in} f - g_j\| < \varepsilon, i = 0, 1, \ldots, k$, which implies $\operatorname{diam}\{T^m f, T^{m+n} f, \ldots, T^{m+kn} f\} < 2\varepsilon$ (in $L^2(X, \mathcal{B}, \mu)$). Note that the set of possible n with this property has positive lower density (and, in fact, is syndetic and even IP*—see the remark following Corollary 2.5). Since T is an isometry, we have $\operatorname{diam}\{f, T^n f, \ldots, T^{kn} f\} < 2\varepsilon$, and hence by choosing ε small enough, we see that for a "large" set of n, $\int f T^n f \cdots T^{kn} f \, d\mu = \mu(A \cap T^{-n} A \cap \cdots \cap T^{-kn} A)$ is arbitrarily close to $\mu(A)$. This certainly implies that

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu(A \cap T^{-n} A \cap \cdots \cap T^{-kn}) > 0.$$

The scheme of usage of van der Waerden's theorem in the case of relatively compact extensions is similar but a little bit more sophisticated. Before embarking on the proof, let us make some convenient reductions. First, note that deleting from a given set $A \in \mathcal{B}$ with $\mu(A) > 0$ portions for which $\mu_y(A) \leqslant \frac{1}{2}\mu(A)$ removes less than half of the measure from A, and hence we can assume without loss of generality that there exists a set $A_1 \in \mathcal{D}$ with $\nu(A_1) \geqslant \frac{1}{2}\mu(A)$ and such that, for $y \in A_1$, $\mu_y(A) \geqslant \frac{1}{2}\mu(A)$ and for $y \notin A_1$, $\mu_y(A) = 0$. Second, one can show that, by removing additional arbitrarily small portions from A, one can assume that $f = 1_A$ is compact relative to $\boldsymbol{Y}$. (See for the details [61, p. 548] or [58, Theorem 6.13].)

Fix a small enough $\varepsilon > 0$ and functions $g_1, g_2, \ldots, g_r$ (one of which is assumed to be 0) such that for any $n \in \mathbb{Z}$,

$$\min_{1\leqslant s\leqslant r} \left\| T^n f - g_s \right\|_y < \varepsilon \quad \text{for a.e. } y \in Y.$$

Let N be such that for any r-coloring of $\{1, 2, \ldots, N\}$ one has a monochromatic progression of length $k + 1$, and assume that for the set $A_1 \in \mathcal{D}$ described above and some $c_1 > 0$, the set $R_N = \{n \in \mathbb{N}\colon\ \nu(A_1 \cap S^{-n} A_1 \cap \cdots \cap S^{-nN} A_1) > c_1\}$ is of positive lower density. We shall show that there exist constants $c_2 > 0$ and $M \in \mathbb{N}$ such that for any $n \in R_N$ there exists $d \in \{1, 2, \ldots, M\}$ with $\mu(A \cap T^{-dn} A \cap \cdots \cap T^{-k(dn)} A) > c_2$. This, clearly, will imply that $\boldsymbol{X}$ has the SZ property. Note that for every $y \in A_1$ and

$n \in R_N$ one has $S^{in}y \in A_1, i = 0, 1, \ldots, N$. Now, for each $y \in A_1$ and each $n \in R_N$, the inequalities $\min_{1 \leqslant s \leqslant r} \|T^{in} f - g_s\|_y < \varepsilon, i = 1, 2, \ldots, N$, define an r-coloring of $\{1, 2, \ldots, N\}$. By van der Waerden's theorem, there exists a monochromatic arithmetic progression $\{i, i + d, \ldots, i + kd\} \subset \{1, 2, \ldots, N\}$ which implies that, for some $g_{s(y)} = g$, $\|T^{(i+jd)n} f - g\|_y < \varepsilon$ for $j = 0, 1, \ldots, k$. This, in turn, implies that $\|T^{jdn} f - g\|_{S^{in}y} < \varepsilon$. It remains now to choose a progression $\{i, i + d, \ldots, i + kd\}$ which occurs for a set A_2 of y of measure at least $\frac{\nu(A_1)}{P}$, where P is the total number of possibilities for the choice of a $(k+1)$-element progression from $\{1, 2, \ldots, n\}$. Note that, since $A_2 \subset A_1$, for each $y \in A_2$, one has $\mu_{S^{in}y}(A) \geqslant \frac{1}{2}\mu(A)$. This implies that (if ε is small enough)

$$\mu_{S^{in}y}(A \cap T^{-dn}A \cap \cdots \cap T^{-k(dn)}A) > \mu_{S^{in}y}(A) - (k+1)\varepsilon > \frac{1}{3}\mu(A).$$

Integrating over the set A_2, we get

$$\mu(A \cap T^{-dn}A \cap \cdots \cap T^{-k(dn)}A) \geqslant \frac{1}{3}\mu(A)\nu(A_2) \geqslant \frac{1}{3P}\mu(A)\nu(A_1) = c_2.$$

We are done. □

4.3. *An overview of multiple recurrence theorems*

Furstenberg's proof of the ergodic Szemerédi theorem was the starting point of a new area: Ergodic Ramsey Theory. In this subsection we shall discuss various multiple recurrence results which are dynamical versions of corresponding Ramsey-theoretical density statements, and which have, so far, no conventional proof. While the proofs of these results are rather involved (which is manifested, in particular, by the length of papers such as [62, 64,98,31,28]), they have a conspicuous commonality of the main structural features. One of our intentions in the following discussion is to stress the structural analogies between various proofs, while paying attention to new ideas whose introduction is necessary in the course of establishing new, stronger, and more refined results.

For a warm-up, let us start the discussion with the density version of Theorem 2.6.

Let V_F be a countably infinite vector space over a finite field F. As in Section 2, let us identify V_F with the direct sum F_∞ of countably many copies of F:

$$F_\infty = \{g = (a_1, a_2, \ldots)\colon a_i \in F \text{ and all but finitely many } a_i = 0\} = \bigcup_{n=1}^{\infty} F_n,$$

where $F_n = \{g = (a_1, a_2, \ldots), a_i \in F, a_i = 0 \text{ for } i > n\}$.

We shall say that a set $E \subset V_F \cong F_\infty$ *has positive upper density* if $\bar{d}_{F_\infty}(E) = \limsup_{N\to\infty} \frac{|E \cap F_n|}{|F_n|} > 0$.

THEOREM 4.18. *Any set of positive upper density in the vector space V_F contains arbitrarily large affine subspaces.*

Note that since F is a finite field, saying "arbitrarily large" in the formulation above is tantamount to saying "of arbitrarily large dimension."

We now formulate an ergodic-theoretical theorem which is analogous to Theorem 4.2 and which implies Theorem 4.18.

THEOREM 4.19. *For any measure preserving action $(T_g)_{g\in F_\infty}$ on a probability measure space $(X, \mathcal{B}, \mu)$ and for any $A \in \mathcal{B}$ with $\mu(A) > 0$, one has*

$$\liminf_{n\to\infty} \frac{1}{|F_n|} \sum_{g\in F_n} \mu\left(\bigcap_{c\in F} T_{cg} A\right) > 0.$$

For the derivation of Theorem 4.18 from Theorem 4.19, one can use the following version of Furstenberg's correspondence principle. Both the version below and the result stated above as Theorem 1.25 are special cases of Theorem 5.8, which will be proved in the next section.

THEOREM 4.20. *For any set $E \subset F_\infty$ with $\bar{d}_{F_\infty}(E) > 0$, there exists a probability measure preserving system $(X, \mathcal{B}, \mu, (T_g)_{g\in F_\infty})$ and a set $A \in \mathcal{B}$ with $\mu(A) = \bar{d}_{F_\infty}(E)$ such that for all $k \in \mathbb{N}$ and any $g_1, g_2, \ldots, g_k \in F_\infty$ one has*

$$\bar{d}_{F_\infty}(E \cap E - g_1 \cap \cdots \cap E - g_k) \geqslant \mu(A \cap T_{g_1} A \cap \cdots \cap T_{g_k} A).$$

Noting that the set $\{cg\}_{c\in F}$ forms a one-dimensional subspace of $V_F \cong F_\infty$, we see that Theorem 4.19 immediately implies, via Furstenberg's correspondence principle, that for some $g \neq \mathbf{0}$, $\bar{d}_{F_\infty}(\bigcap_{c\in F}(E - cg)) > 0$ and hence for any $x \in \bigcap_{c\in F}(E - cg)$, the one-dimensional affine space $\{x + cg\}_{c\in F}$ is contained in E.

To obtain the full strength of Theorem 4.18, one can use the following "iterational" trick (which is very similar to that utilized in the proof of Theorem 1.12). Namely, use Theorem 4.19 to find $g_1 = (b_1, b_2, \ldots, b_k, 0, 0, \ldots) \neq \mathbf{0}$ with the property that the set $A_1 = \bigcap_{c\in F} T_{cg_1} A$ has positive measure. Apply now Theorem 4.19 to the restriction of the action $(T_g)_{g\in F_\infty}$ to the subgroup $G_1 \subset F_\infty$ which is defined by

$$G_1 = \{g = (a_1, a_2, \ldots) \in F_\infty \colon a_1, a_2, \ldots, a_k = 0\}.$$

In other words, the supports of elements from G_1 are disjoint from the support of our $g_1 = (b_1, b_2, \ldots, b_k, 0, 0, \ldots)$. Note also that G_1 is isomorphic to the direct sum of countably many copies of F, and hence is isomorphic to F_∞. Find now $g_2 \in G_1$ with the property that $A_2 = \bigcap_{c\in F} T_{cg_2} A_1$ has positive measure, and continue in this fashion. After m steps of this iterational procedure, we will have found elements $g_1, g_2, \ldots, g_m$ such that the set $A_m = \bigcap_{c_1,c_2,\ldots,c_m\in F} T_{c_1g_1+c_2g_2+\cdots+c_mg_m} A$ has positive measure. It follows now from Theorem 4.20 that $\bar{d}_{F_\infty}(\bigcap_{c_1,c_2,\ldots,c_m\in F} E - (c_1g_1 + \cdots + c_mg_m)) > 0$, and this clearly implies that E contains (many) affine m-dimensional subspaces.

Let us now comment briefly on the proof of Theorem 4.19. It is not hard to check that Theorem 4.19 holds in the two "extremal" cases, namely the case when the action $(T_g)_{g\in F}$

is compact (which, as before, means that for any $f \in L^2(X, \mathcal{B}, \mu)$ the orbit $\{T_g f\}_{g \in F_\infty}$ is precompact), and the weakly mixing case (which can be defined, for example, by postulating the absence of compact functions). Moreover, the proof in each of these two cases is very similar to the analogous case of Theorem 4.2. Perhaps a few remarks are in order to clarify the situation with weak mixing for actions of $(T_g)_{g \in F_\infty}$. First, one can check that, like in the case of $\mathbb{Z}$-actions, weak mixing for F_∞-actions can be characterized in a variety of ways, all parallel to those occurring in the formulation of Theorem 4.4. (This remark actually applies to—properly defined—weak mixing actions of any countable or even locally compact group. See, for example, [33,12,18].)

Second, one can check that an analogue of Theorem 4.6 also holds for more general groups (here the right generality is that of amenable groups; see more discussion in the next section).

Now, one can define, in complete analogy to Definitions 4.11 and 4.15 the notions of relative weak mixing and relative compactness. The analogues of Theorems 4.12, 4.14, and 4.16 can also be established in a more or less similar fashion. So to finish the proof, one has to show that the multiple recurrence property lifts to compact extensions. As the perspicacious reader has probably guessed by now, one can use here the natural analogue of van der Waerden's theorem, namely Theorem 1.27.

4.3.1. *Furstenberg–Katznelson's multidimensional Szemerédi theorem.* Let us discuss now the multidimensional Szemerédi theorem or, rather, its measure-theoretical twin, Theorem 1.24. Here is the version which actually was proved by Furstenberg and Katznelson in [60].

THEOREM 4.21. *For any commuting measure preserving transformations $T_1, T_2, \dots, T_k$ of a probability space $(X, \mathcal{B}, \mu)$ and for any $A \in \mathcal{B}$ with $\mu(A) > 0$ one has*

$$\liminf_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu\big(A \cap T_1^{-n} A \cap \cdots \cap T_k^{-n} A\big) > 0.$$

The main new difficulty which one faces when dealing with k general commuting transformations is that they generate a $\mathbb{Z}^k$-action, which may have different dynamical properties along the sub-actions of different subgroups. In other words, while Theorem 4.2 was about the joint behavior of k commuting transformations of a special form, namely $T, T^2, \dots, T^k$, in Theorem 4.21 we have to study k commuting transformations which are in, so to say, general position. This complicates the underlying structure theory, which has to be "tuned up" to reflect the more complicated situation when different operators in the group generated by $T_1, \dots, T_k$ have different dynamical properties. What saves the day is Theorem 4.24 below, which is at the core of Furstenberg and Katznelson's proof of Theorem 4.21.

We need first to introduce some pertinent definitions. While these definitions make sense for any measure preserving group actions (and are given below a general formulation for future reference), the reader should remember that in the discussion of the proof of Theorem 4.19, the group G which occurs in the next two definitions is meant to stand for $\mathbb{Z}^k$ (and hence the subgroups of G are themselves isomorphic to $\mathbb{Z}^l$ for some $0 \leqslant l \leqslant k$).

DEFINITION 4.22. (*Cf. Definition* 6.3 *in* [58].) An extension $(X, \mathcal{B}, \mu, (T_g)_{g\in G})$ of $(Y, \mathcal{D}, \nu, (S_g)_{g\in G})$ is a weakly mixing extension if for every $g_0 \in G$, $g_0 \neq e$, the system $(X, \mathcal{B}, \mu, T_{g_0})$ is a weakly mixing extension of $(Y, \mathcal{D}, \nu, S_{g_0})$ (in the sense of Definition 4.11).

DEFINITION 4.23. (*Cf. Definition* 6.5 *in* [58].) Assume that $(X, \mathcal{B}, \mu, (T_g)_{g\in G})$ is an extension of $(Y, \mathcal{D}, \nu, (S_g)_{g\in G})$. This extension is called *primitive* if G is a direct product of two subgroups, $G_c \times G_{wm}$, so that $(X, \mathcal{B}, \mu, (T_g)_{g\in G_c})$ is a compact extension of $(Y, \mathcal{D}, \nu, (S_g)_{g\in G_c})$ and $(X, \mathcal{B}, \mu, (T_g)_{g\in G_{wm}})$ is a weakly mixing extension of $(Y, \mathcal{D}, \nu, (S_g)_{g\in G_{wm}})$.

REMARK. We did not explicitly define the notion of compact extension for this more general situation because it is verbatim the same as Definition 4.15. (One just has to replace "for every $n \in \mathbb{Z}$" by "for every $g \in G$".) This should be juxtaposed with Definition 4.22 which, while still coinciding with Definition 4.11 when $G = \mathbb{Z}$, has the emphasis not on the weak mixing behavior of the group action $(T_g)_{g\in G}$, but on the behavior of $\mathbb{Z}$-actions generated by elements $g \in G$, $g \neq e$.

We are now ready to formulate the theorem which provides the main ingredient in the pertinent structure theory. The reader should keep in mind that in the theorem below G stands for $\mathbb{Z}^k$.

THEOREM 4.24. *If* $\boldsymbol{X} = (X, \mathcal{B}, \mu, (T_g)_{g\in G})$ *is an extension of* $\boldsymbol{Y} = (Y, \mathcal{D}, \nu, (S_g)_{g\in G})$, *then there is an intermediate factor* $\boldsymbol{Z}$ *such that* $\boldsymbol{Z}$ *is a primitive extension of* $\boldsymbol{Y}$.

As was the case with Theorem 4.2, one can show that there is always a maximal factor for which Theorem 4.19 is valid. So, in view of Theorem 4.24, it remains only to make sure that the multiple recurrence property in question lifts to primitive extensions. This can be achieved by an argument which puts together the ideas behind the proofs of Theorems 4.14 and 4.17. The fact that primitive extensions utilize the appropriate splitting of $\mathbb{Z}^k$ plays a crucial role. In dealing with the compact part of this splitting, one uses this time the multidimensional van der Waerden theorem. For full details, see [60] and [58, Chapter 7].

4.3.2. *Polynomial Szemerédi theorem.* Note that the coloring theorems used in the proofs of Theorems 4.2, 4.19, and 4.21 are all corollaries of the IP van der Waerden theorem (which was discussed in detail in Section 2). This suggests that there exists perhaps a more general theorem which bears the same relation to the IP van der Waerden theorem (Theorem 2.2 above) as, say, Theorem 4.1 to the (one-dimensional) van der Waerden theorem, and has Theorems 4.2, 4.19, and 4.21 as corollaries. Such a result, which is called the ergodic IP Szemerédi theorem, was established by Furstenberg and Katznelson in [62] and will be briefly discussed below. But before turning our attention to the Furstenberg–Katznelson IP Szemerédi theorem, we want to discuss the polynomial extension of Szemerédi's theorem obtained in [23]. While the paper [23], which appeared in 1996, is more recent than the 1985 paper [62], the structure theory which is utilized in [23] is the same

as that needed for the proof of Theorem 4.21, whereas in [62] the authors deal with IP systems and develop the nontrivial and complicated IP version of structure theory.

Here then is the formulation of the polynomial Szemerédi theorem.

THEOREM 4.25 [23, Theorem B′]. *Let $r, l \in \mathbb{N}$ and let $P : \mathbb{Z}^r \to \mathbb{Z}^l$ be a polynomial mapping satisfying $p(0) = 0$. For any $S \subseteq \mathbb{Z}^l$ with $d^*(S) > 0$ and any finite set $F \subset \mathbb{Z}^r$, there is $n \in \mathbb{N}$ and $u \in \mathbb{Z}^l$ such that $u + P(nF) \subset S$.*

In order to formulate an ergodic result which would imply Theorem 4.25, let us first reformulate Theorem 4.25 in coordinate form.

THEOREM 4.26 [23, Theorem B]. *For $l \in \mathbb{N}$, let $S \subseteq \mathbb{Z}^l$ satisfy $d^*(S) > 0$. Let $p_{1,1}(n), \ldots, p_{1,t}(n), p_{2,1}(n), \ldots, p_{2,t}(n), \ldots, p_{k,1}(n), \ldots, p_{k,t}(n)$ be polynomials with rational coefficients taking integer values on the integers and satisfying $p_{i,j}(0) = 0$, $i = 1, \ldots, k$, $j = 1, \ldots, t$. Then, for any $v_1, \ldots, v_t \in \mathbb{Z}^l$, there exist $n \in \mathbb{N}$ and $v \in \mathbb{Z}^l$ such that $v + \sum_{j=1}^{t} p_{i,j}(n) v_j \in S$ for each $i \in \{1, 2, \ldots, k\}$.*

To see that Theorem 4.25 implies Theorem 4.26, take $k = r$ and apply Theorem 4.25 to the polynomial mapping $P : \mathbb{Z}^r \to \mathbb{Z}^l$ defined by

$$P(n_1, n_2, \ldots, n_r) = \sum_{j=1}^{t} \sum_{i=1}^{r} p_{i,j}(n_i) v_j$$

and the finite set $F = \{(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, 0, \ldots, 1)\} \subset \mathbb{Z}^r$.

To see that Theorem 4.25 follows from Theorem 4.26, let $P : \mathbb{Z}^r \to \mathbb{Z}^l$ be a polynomial mapping satisfying $P(0) = 0$ and let $F = \{w_1, \ldots, w_k\}$ be an arbitrary finite set in $\mathbb{Z}^r$. Letting $t = l$ in Theorem 4.26, define polynomials $p_{i,j}(n)$ by

$$p_{i,j}(n) = P(nw_i)_j, \quad n \in \mathbb{N},\ i = 1, 2, \ldots, k,\ j = 1, 2, \ldots, l.$$

Let $v_1, v_2, \ldots, v_l$ denote the unit vectors from the standard basis in $\mathbb{Z}^l$. Then, by Theorem 4.26 one has, for some $n \in \mathbb{N}$ and $u \in \mathbb{Z}^l$,

$$u + P(nw_i) = u + \sum_{j=1}^{t} P(nw_i)_j v_j \in S, \quad i = 1, 2, \ldots, k,$$

which is the same as $u + P(nF) \subset S$.

We formulate now an ergodic theoretic result which implies (via Furstenberg's correspondence principle) Theorems 4.25 and 4.26, and which may be viewed as a measure preserving analogue of the topological polynomial van der Waerden theorem, Theorem 2.9 above. (See also Theorem 4.43 below.)

THEOREM 4.27 [23, Theorem A]. *Let, for some $t, k \in \mathbb{N}$, $p_{1,1}(n), \ldots, p_{1,t}(n), p_{2,1}(n), \ldots, p_{2,t}(n), \ldots, p_{k,1}(n), \ldots, p_{k,t}(n)$ be polynomials with rational coefficients taking integer values on the integers and satisfying $p_{i,j}(0) = 0$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, t$. Then,*

for any probability space $(X, \mathcal{B}, \mu)$, commuting invertible measure preserving transformations $T_1, T_2, \ldots, T_t$ of X and any $A \in \mathcal{B}$ with $\mu(A) > 0$, one has

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu\left(A \cap \prod_{j=1}^{t} T_j^{-p_{i,j}(n)} A \cap \prod_{j=1}^{t} T_j^{-p_{2,j}(n)} A \cap \cdots \cap \prod_{j=1}^{t} T_j^{-p_{n,j}(n)} A \right) > 0.$$

To get a feeling for how general Theorem 4.27 is (though this feeling, on a combinatorial level, should be provided by the formulation of Theorem 4.25), let us note that as a special case one has, for example, the following refinement of Theorem 4.21.

THEOREM 4.28. *For any commuting invertible measure preserving transformations $T_1, \ldots, T_k$ of a probability space $(X, \mathcal{B}, \mu)$, any polynomials $p_1(n), \ldots, p_k(n)$ which have rational coefficients, take integer values on the integers, and satisfy $p_i(0) = 0$, $i = 1, 2, \ldots, k$, any $A \in \mathcal{B}$ with $\mu(A) > 0$, one has*

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu\big(A \cap T_1^{-p_1(n)} A \cap T_2^{-p_2(n)} A \cap \cdots \cap T_k^{-p_k(n)} A\big) > 0.$$

The proof of Theorem 4.27 in [23] can be described as a "polynomialization" of the proof of Theorem 4.21. To ease the discussion, let us put in Theorem 4.28 $T_i = T$ and consider the expression

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu\big(A \cap T^{-p_1(n)} A \cap T^{-p_2(n)} A \cap \cdots \cap T^{-p_k(n)} A\big).$$

Assume first that $f = 1_A$ is a compact function. In this special case the positivity of the lim inf above easily follows from the following simple fact.

LEMMA 4.29. *Suppose that $p_1(n), p_2(n), \ldots, p_k(n)$ are polynomials with rational coefficients which take integer values on integers and satisfy $p_i(0) = 0$, $i = 1, 2, \ldots, k$. Let T be an isometry of a compact metric space (X, ρ). Then for any $\varepsilon > 0$ there exists a point $x \in X$ such that the set $\bigcap_{i=1}^{k} \{n\colon\ \rho(T^{p_i(n)} x, x) < \varepsilon\}$ is syndetic.*

To prove Lemma 4.29, one can, for example, invoke the fact that if T is an isometry then the dynamical system (X, T) is *semisimple*, i.e. is a disjoint union of minimal systems. Now, it is not hard to show that if the topological system (X, T), where T is an isometry, is minimal, then it is topologically isomorphic (conjugate) to a minimal translation on a compact Abelian group. The desired result then can be deduced from Weyl's result on polynomial Diophantine approximation. Alternatively, one can observe that Lemma 4.29 is a corollary of the polynomial van der Waerden theorem (Theorem 2.9 above).

If $(X, \mathcal{B}, \mu, T)$ is weakly mixing, then the result in question also holds, due to the following refinement of Theorem 4.5.

THEOREM 4.30 [8]. *Let $(X, \mathcal{B}, \mu, T)$ be an invertible weakly mixing system. Assume that the polynomials $p_j(n)$, $j = 1, 2, \ldots, k$, take integer values on integers, have degree greater than or equal to one, and satisfy the condition $p_i(n) - p_j(n) \not\equiv \text{const}$, $i \neq j$. Then for any $f_i \in L^\infty(X, \mathcal{B}, \mu)$ one has*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} T^{p_1(n)} f_1 T^{p_2(n)} f_2 \cdots T^{p_k(n)} f_k = \int f_1 \, d\mu \int f_2 \, d\mu \cdots \int f_k \, d\mu$$

in L^2-norm.

Putting $f_i = 1_A$, $i = 1, 2, \ldots, k$, multiplying by 1_A and integrating gives us

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu\big(A \cap T^{-p_1(n)} A \cap T^{-p_2(n)} A \cap \cdots \cap T^{-p_k(n)} A\big) = \mu(A)^{k+1}.$$

The proof of Theorem 4.30 is achieved by an inductive procedure, based on Theorem 4.6, which is sometimes called PET-induction. See [8] for details.

The proof of Theorem 4.27 in its full generality takes, of course, some more work, but with the help of the polynomial van der Waerden theorem and the appropriately general form of Theorem 4.30, one is able to push the statement through the primitive extensions. See [23] for the details.

4.3.3. *IP Szemerédi theorem.* We pass now to the discussion of the IP Szemerédi theorem which was obtained by Furstenberg and Katznelson in [62]. The reader is encouraged to review the definition of an IP system introduced before the formulation of the IP van der Waerden theorem (Theorem 2.2) and to juxtapose the formulations of Theorem 2.2 and the following statement.

THEOREM 4.31. (See [62, Theorem A].) *Let $(X, \mathcal{B}, \mu)$ be a probability space and G an Abelian group of measure-preserving transformations of X. For any $k \in \mathbb{N}$, any IP systems $\{T_\alpha^{(1)}\}_{\alpha\in\mathcal{F}}, \{T_\alpha^{(2)}\}_{\alpha\in\mathcal{F}}, \ldots, \{T_\alpha^{(k)}\}_{\alpha\in\mathcal{F}}$ in G, and any $A \in \mathcal{B}$ with $\mu(A) > 0$ there exists $\alpha \in \mathcal{F}$ such that*

$$\mu\big(A \cap T_\alpha^{(1)} A \cap T_\alpha^{(2)} A \cap \cdots \cap T_\alpha^{(k)} A\big) > 0.$$

The proof of the IP Szemerédi theorem is achieved via a sophisticated structure theory which could be viewed as an IP variation on the theme of primitive extensions discussed above. Curiously enough, it is not the IP van der Waerden theorem, but the more powerful Hales–Jewett theorem which has to be used when dealing with the IP version of compact extensions. We will give more details on the proof of Theorem 4.31 below, but first we want to discuss some of its corollaries.

Note that, since the notion of an IP set is a generalization of a (semi)group, the notion of an IP system of commuting invertible measure preserving transformations generalizes the

notion of a measure preserving action of a countable Abelian group. It follows that Theorems 4.2, 4.19, and 4.21 are immediate corollaries of Theorem 4.31. It also follows, via an appropriate version of Furstenberg's correspondence principle, that, on a combinatorial level, Theorem 4.21, the IP Szemerédi theorem, implies the multidimensional version of Szemerédi's theorem (Theorem 1.23 above) as well as Theorem 4.18. However, the IP Szemerédi theorem gives more! For example, it follows from it that the sets of configurations, always to be found in sets of positive density in $\mathbb{Z}^k$ or F_∞, are abundant in the sense that the set of parameters of these configurations form IP* sets. (See the discussion at the beginning of Section 2.)

One can derive these IP* versions of combinatorial results from the following corollary of Theorem 4.31. The IP and IP* sets in an Abelian group are defined in complete (and obvious) analogy to the definitions in Sections 1 and 2 which were geared towards $\mathbb{N}$.

THEOREM 4.32. *Let $(X, \mathcal{B}, \mu)$ be a probability space, and let G be a countable Abelian group. For any k commuting measure preserving actions $(T_g^{(1)})_{g\in G}$, $(T_g^{(2)})_{g\in G}, \ldots, (T_g^{(k)})_{g\in G}$ of G on $(X, \mathcal{B}, \mu)$ and any $A \in \mathcal{B}$ with $\mu(A) > 0$, the set*

$$\left\{g \in G\colon \mu\left(A \cap T_g^{(1)}A \cap T_g^{(2)}A \cap \cdots \cap T_g^{(k)}A\right) > 0\right\}$$

is an IP* set in G.

Note that since any IP* set in $\mathbb{N}$ is obviously syndetic, Theorem 4.32 implies, for example, the following fact.

COROLLARY 4.33. *For any commuting transformations $T_1, T_2, \ldots, T_k$ of a probability space $(X, \mathcal{B}, \mu)$ and any $A \in \mathcal{B}$ with $\mu(A) > 0$, the set $\{n \in \mathbb{N}\colon \mu(A \cap T_1^{-n}A \cap T_2^{-n}A \cap \cdots \cap T_k^{-n}A) > 0\}$ is syndetic.*

Note that the conclusion of Corollary 4.33 would follow from Theorem 4.21 if one would be able to replace in its formulation the statement involving the regular Cesàro averages:

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu\left(A \cap T_1^{-n}A \cap \cdots \cap T_k^{-n}A\right) > 0$$

by a similar, but stronger, statement, involving "uniform" averages:

$$\liminf_{N-M\to\infty} \frac{1}{N-M} \sum_{n=M}^{N-1} \mu\left(A \cap T_1^{-n}A \cap \cdots \cap T_k^{-n}A\right) > 0.$$

It is perhaps instructive to pinpoint the exact place in the proof of Theorem 4.21 (or its earlier version, Theorem 4.2) that does not work for the uniform Cesàro averages. This analysis will also allow the reader to get a better feeling for why one is forced to use the Hales–Jewett theorem. Careful examination of the proof reveals that it is actually only the case of compact extensions which causes the trouble.

Let us briefly review the main ingredients of the proof of Theorem 4.2. First, if the system $(X, \mathcal{B}, \mu, T)$ is compact, we saw that the set $\{n \in \mathbb{Z}\colon |\mu(A \cap T^{-n}A \cap \cdots \cap T^{-kn}A) - \mu(A)| < \varepsilon\}$ is syndetic. Second, as was mentioned in Remark 4.7, in the case when $(X, \mathcal{B}, \mu, T)$ is weakly mixing, one has, for any $A \in \mathcal{B}$ with $\mu(A) > 0$,

$$\lim_{N-M\to\infty} \frac{1}{N-M} \sum_{n=M}^{N-1} \mu(A \cap T^{-n}A \cap \cdots T^{-kn}A) = \big(\mu(A)\big)^{k+1},$$

and hence, in this case, the set

$$\big\{n \in \mathbb{Z}\colon \big|\mu(A \cap T^{-n}A \cap \cdots \cap T^{-kn}A) - \big(\mu(A)\big)^{k+1}\big| < \varepsilon\big\}$$

is syndetic.

One can check that the case of relative weak mixing also works for uniform averages. However, it is the case of relatively compact extensions where the syndeticity property is lost in the passage to the extension. Indeed, in the proof of Theorem 4.17 we show that if $\boldsymbol{X} = (X, \mathcal{B}, \mu, T)$ is a relatively compact extension of $\boldsymbol{Y} = (Y, \mathcal{D}, \nu, S)$ and $A \in \mathcal{B}$ with $\mu(A) > 0$, then there is a set $A_1 \in \mathcal{D}$ with $\nu(A_1) > 0$ and a number $M \in \mathbb{N}$ such that for any n which is good for multiple recurrence of A, in $\boldsymbol{Y}$, there is a multiple dn with $d \leqslant M$, which is good for multiple recurrence of A in $\boldsymbol{X}$. So even if one would know in advance that the set $R_{A_1} = \{n_1, n_2, \ldots\}$ of multiple returns of A_1 is a syndetic set, the set of multiples $R_A = \{d_1 n_1, d_2 n_2, \ldots\}$ while being still of positive upper density (due to the fact that $d_i \leqslant M$ for all i), is no longer guaranteed to be syndetic, as it is not hard to see on some trivial examples.

A possible solution of this problem is to use a more powerful coloring theorem instead of van der Waerden's. It turns out that the Hales–Jewett theorem (see Theorems 1.26 and 1.28) which, as we saw in Section 2 (see Proposition 2.7) is very close to van der Waerden's, is strong enough to supply the missing link needed to assure that the syndeticity can be pushed through the transfinite induction. This added strength allows one to get the better result:

$$\liminf_{N-M\to\infty} \frac{1}{N-M} \sum_{n=M}^{N-1} \mu\big(A \cap T_1^{-n}A \cap \cdots \cap T_k^{-n}A\big) > 0.$$

(See [104, Section 5.2], for a presentation of the syndetic version of Theorem 4.2 via the Hales–Jewett theorem.)

We give now more details on the proof of the IP Szemerédi theorem. First, let us introduce, following [58] and [62], some pertinent terminology.

Let us recall that any sequence indexed by the set of nonempty subsets of $\mathbb{N}$ is called an $\mathcal{F}$-sequence. In particular, IP sets and IP systems that we have dealt with in earlier sections are examples of $\mathcal{F}$-sequences. As before, we will be writing, for $\alpha, \beta \in \mathcal{F}$, $\alpha < \beta$ (or $\beta > \alpha$) if $\max\alpha < \min\beta$. Assume that a collection of sets $\alpha_i \in \mathcal{F}$, $i = 1, 2, \ldots$, has the property $\alpha_i < \alpha_{i+1}$ for all $i \in \mathbb{N}$. The set $\mathcal{F}^{(1)} = \{\bigcup_{i\in\beta} \alpha_i\colon \beta \in \mathcal{F}\}$ is called an IP ring. Observe that $\mathcal{F}$ can be viewed as an IP set in the commutative semigroup $(\mathbb{N}, \cup)$, generated

by the singletons $\{i\}$, $i \in \mathbb{N}$. By the same token, the IP ring $\mathcal{F}^{(1)}$ can be viewed as an IP set in $(\mathbb{N}, \cup)$ which is generated by the "atoms" α_i, $i \in \mathbb{N}$, and hence has the same structure as $\mathcal{F}$. More formally, let us define a mapping $\varphi : \mathcal{F} \to \mathcal{F}^{(1)}$ by $\varphi(\beta) = \bigcup_{i \in \beta} \alpha_i$. Clearly, φ is bijective and "structure preserving." It follows that any sequence indexed by the elements of an IP ring may itself be viewed as an $\mathcal{F}$-sequence.

The following is a version of Hindman's theorem, which will be needed below. The reader should have no problem establishing its equivalence to Theorem 1.10.

THEOREM 4.34. *For any finite partition $\mathcal{F} = \bigcup_{i=1}^{r} C_i$, one of C_i contains an IP ring.*

DEFINITION 4.35. Assume that $(x_\alpha)_{\alpha \in \mathcal{F}}$ is an $\mathcal{F}$-sequence in a topological space X. Let $x \in X$ and let $\mathcal{F}^{(1)}$ be an IP ring. We shall write IP-$\lim_{\alpha \in \mathcal{F}^{(1)}} x_\alpha = x$ if for any neighborhood U of x there exists $\alpha_0 = \alpha_0(U)$ such that, for any $\alpha \in \mathcal{F}^{(1)}$ with $\alpha > \alpha_0$, one has $x_\alpha \in U$.

One has now the following IP version of the classical Bolzano–Weierstrass theorem.

THEOREM 4.36. (Cf. [58, Theorem 8.14] and [62, Theorem 1.3].) *If $(x_\alpha)_{\alpha \in \mathcal{F}}$ is an $\mathcal{F}$-sequence in a compact metric space X, then there exist an IP ring $\mathcal{F}^{(1)}$ and $x_0 \in X$ such that the $\mathcal{F}$-sequence $(x_\alpha)_{\alpha \in \mathcal{F}^{(1)}}$ has an IP-limit in X:*

$$\operatorname*{IP\text{-}lim}_{\alpha \in \mathcal{F}^{(1)}} x_\alpha = x_0.$$

SKETCH OF THE PROOF. The proof goes along the lines of the classical "dichotomic" proof of the Bolzano–Weierstrass theorem, in which one replaces the pigeonhole principle by the (much more powerful) Hindman's theorem. For given $\varepsilon > 0$, let $(B_i)_{i=1}^{r}$ be a finite family of open balls of radius $\varepsilon/2$ which covers the compact space X. By Theorem 4.34 one can extract an IP ring $\mathcal{F}^{(1)}$ so that the $\mathcal{F}$-sequence $(x_\alpha)_{\alpha \in \mathcal{F}^{(1)}}$ has all of its elements within distance less than ε of one another. The proof is concluded by the diagonal procedure. □

REMARK. The notions and properties of IP convergence are very similar to those of the convergence along an idempotent ultrafilter, which was introduced and discussed in Section 3. One could advance this analogy even further by introducing $\beta\mathcal{F}$, the Stone–Čech compactification of $\mathcal{F}$. We have preferred to stick to IP convergence for two reasons. First, this allows us to follow more closely the work of Furstenberg and Katznelson in [62]. Second, IP-limits seem, at least as of now, to be a more convenient tool for dealing with the polynomial extensions of the IP Szemerédi theorem. (See Theorems 4.40 and 4.43 below.)

The following result is an IP analogue of Proposition 3.17 above. For the (short) proof see [58, Lemma 8.15], or [62, p. 124].

THEOREM 4.37. *Let $\{T_\alpha\}_{\alpha \in \mathcal{F}}$ be an IP system of continuous transformations of a metric space X. Assume that, for some $x, y \in X$,* IP-$\lim_{\alpha \in \mathcal{F}} T_\alpha x = y$. *Then* IP-$\lim_{\alpha \in \mathcal{F}} T_\alpha y = y$.

Assume now that $(U_\alpha)_{\alpha\in\mathcal{F}}$ is an IP system generated by commuting unitary operators acting on a separable Hilbert space $\mathcal{H}$. By using the fact that the closed ball $B_R = \{x \in \mathcal{H}\colon \|x\| \leqslant R\}$ is a compact metrizable space in the weak topology, one has the following result.

THEOREM 4.38 [62, Theorem 1.7]. *If $\{U_\alpha\}_{\alpha\in\mathcal{F}}$ is an IP system of commuting unitary operators on a Hilbert space $\mathcal{H}$, then there is an IP ring $\mathcal{F}^{(1)}$ such that the IP subsystem $\{U_\alpha\}_{\alpha\in\mathcal{F}^{(1)}}$ converges weakly. Moreover, if one has* IP-$\lim_{\alpha\in\mathcal{F}} U_\alpha = P$ *weakly, then P is an orthogonal projection.*

SKETCH OF THE PROOF. Since, clearly, $\|P\| \leqslant 1$, one needs only to show that $P^2 = P$. But this follows from Theorem 4.37. □

The projection P occurring in the above theorem is an orthogonal projection on the space of *rigid* elements, i.e. elements R, satisfying $U_\alpha f \to f$. Note also that, by a classical exercise, $U_\alpha f \to f$ weakly if and only if $U_\alpha f \to f$ strongly. Assuming that IP-$\lim_{\alpha\in\mathcal{F}} U_\alpha = P$ weakly, we have now the following decomposition of $\mathcal{H}$:

$$\mathcal{H} = \mathcal{H}_r \oplus \mathcal{H}_m, \quad \text{where}$$

$$\mathcal{H}_r = \Big\{ f \in \mathcal{H}\colon \operatorname*{IP\text{-}lim}_{\alpha\in\mathcal{F}} U_\alpha f = f \Big\},$$

$$\mathcal{H}_m = \Big\{ f \in \mathcal{H}\colon \operatorname*{IP\text{-}lim}_{\alpha\in\mathcal{F}} U_\alpha f = 0 \text{ weakly} \Big\}.$$

The reader should view this splitting as the IP analogue of the splitting $\mathcal{H} = \mathcal{H}_c \oplus \mathcal{H}_{wm}$ which was utilized in the proof of Theorem 4.2. This analogy is the starting point of the long list of facts about IP systems of commuting measure preserving transformations which parallel the familiar results pertaining to the structure theory of measure preserving systems and multiple recurrence. For example, when $\mathcal{H} = L^2(X, \mathcal{B}, \mu)$ and the operators U_α are induced by measure preserving transformations T_α on the probability measure space, the space $\mathcal{H}_r$ of rigid functions can be represented, in complete analogy to Theorem 4.6, as $L^2(X, \mathcal{B}_1, \mu)$, where the σ-algebra $\mathcal{B}_1$ consists of sets A for which the indicator function 1_A is rigid. One can go even further and define the notions of relatively rigid and relatively mixing extensions. There is also an IP analogue of the van der Corput trick. (See, for example, Lemma 5.3 in [62].) To handle relatively rigid (or relatively compact, as they are called in [62]) extensions, one uses the Hales–Jewett theorem. Finally (and mainly due to the fact that one deals with a finitely generated group of IP systems), one also has an analogue of primitive extensions and a theorem analogous to Theorem 4.24. (See Theorem 7.10 in [62].)

While many details of the corresponding results demand much work and have to be worked out with care, it is shown in [62] that all this can be glued together to obtain the proof of the IP Szemerédi theorem.

4.3.4. *IP versions of polynomial theorems.* Being encouraged by the IP Szemerédi theorem, one can ask whether the polynomial Szemerédi theorem (Theorem 4.27) also admits

an IP version. This question is already not trivial in the case of single recurrence, and we address it in this context first.

For an arbitrary invertible probability measure preserving system $(X, \mathcal{B}, \mu, T)$, a set $A \in \mathcal{B}$ with $\mu(A) > 0$, and a polynomial $p(n)$ which takes integer values on the integers and satisfies $p(0) = 0$, consider the set

$$R_A = \big\{n\colon\ \mu(A \cap T^{p(n)}A) > 0\big\}.$$

As we have shown in the course of the proof of Theorem 1.31, one has $\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu(A \cap T^{p(n)}A) > 0$, which clearly implies that R_A has positive upper density. Moreover, by using a modification of the van der Corput trick (Theorem 1.32) which deals with limits of the form $\lim_{N-M\to\infty} \frac{1}{N-M} \sum_{n=M}^{N-1} x_n$, one can show that $\lim_{N-M\to\infty} \frac{1}{N-M} \sum_{n=M}^{N-1} \mu(A \cap T^{p(n)}A) > 0$, which implies that, in fact, the set R_A is syndetic.

In order to obtain an IP version of Theorem 1.31, which would guarantee that the set R_A is an IP* set, one has to switch from Cesàro limits to IP limits. The following theorem, which is a special case of a more general result proved in [17], not only implies that R_A is indeed an IP* set, but actually shows that for any $\varepsilon > 0$, the set of returns with large intersections, $\{n\colon\ \mu(A \cap T^{p(n)}A) > \mu(A)^2 - \varepsilon\}$ is also IP*.

THEOREM 4.39. (See [9, Theorem 3.11].) *Assume that $p(t) \in \mathbb{Q}[t]$ satisfies $p(\mathbb{Z}) \subseteq \mathbb{Z}$ and $p(0) = 0$. Then for any invertible probability measure preserving system $(X, \mathcal{B}, \mu, T)$, any $A \in \mathcal{B}$ with $\mu(A) > 0$, and any IP set $(n_\alpha)_{\alpha\in\mathcal{F}} \subset \mathbb{N}$, there exists an IP-ring $\mathcal{F}^{(1)} \subset \mathcal{F}$ such that*

$$\operatorname*{IP\text{-}lim}_{\alpha\in\mathcal{F}^{(1)}} \mu\big(A \cap T^{p(n_\alpha)}A\big) \geqslant \mu(A)^2.$$

The crucial role in the proof of Theorem 4.39 is played by the fact (obtained with the help of an IP version of van der Corput's trick) that there is an IP ring such that (denoting by U_T the unitary operator on $L^2(X, \mathcal{B}, \mu)$ which is induced by T) one has

$$\operatorname*{IP\text{-}lim}_{\alpha\in\mathcal{F}^{(1)}} U_T^{p(n_\alpha)} = P \quad \text{weakly},$$

where P is an orthogonal projection. In particular, it is the fact that P is an orthogonal projection which enables one to get large intersections along the sequence $(p(n_\alpha))_{\alpha\in\mathcal{F}^{(1)}}$. Here is the proof:

$$\begin{aligned}
&\operatorname*{IP\text{-}lim}_{\alpha\in\mathcal{F}^{(1)}} \mu\big(A \cap T^{p(n_\alpha)}A\big)\\
&\quad= \operatorname*{IP\text{-}lim}_{\alpha\in\mathcal{F}^{(1)}} \big\langle U_T^{p(n_\alpha)} 1_A, 1_A\big\rangle\\
&\quad= \langle P1_A, 1_A\rangle = \langle P1_A, P1_A\rangle\langle 1, 1\rangle \geqslant \langle P1_A, 1\rangle^2 = \langle 1_A, 1\rangle^2 = \mu(A)^2.
\end{aligned}$$

While the proof of Theorem 4.39 is, in many respects, just an IP analogue of the proof of Theorem 1.31 above, there is one important distinction which we want to mention here. As we saw in the proof of Theorem 1.31, the splitting $\mathcal{H} = \mathcal{H}_{\text{rat}} \oplus \mathcal{H}_{\text{tot.erg.}}$ works *for all* $p(n) \in \mathbb{Z}[n]$. A novel feature encountered in the proof of the IP analogue of Theorem 1.31 is that the splitting of $\mathcal{H} = L^2(X, \mathcal{B}, \mu)$ which enables one to distinguish between different kinds of asymptotic behavior of $U_T^{p(n)}$ along an IP set $(n_\alpha)_{\alpha\in\mathcal{F}}$ may depend on the polynomial $p(n)$.

However, a much more important novelty which is encountered when one deals with IP analogues of polynomial recurrence theorems is that one has now a bigger family of functions, namely the IP polynomials which form the IP analogue of the conventional polynomials, and for which the IP versions of familiar theorems make sense. Examples of IP polynomial recurrence results in topological dynamics were given in Section 2 (see, for example, Theorems 2.9 and 2.12). The results which we are going to formulate now can be characterized as polynomial IP extensions of Theorems 1.31 and 4.27, and involve a natural subclass of IP polynomials which can be obtained in the following way.

Let $q(t_1, \dots, t_k) \in \mathbb{Z}[t_1, \dots, t_k]$ and let $(n_\alpha^{(i)})_{\alpha\in\mathcal{F}}$, $i = 1, 2, \dots, k$, be IP sets. Then $q(\alpha) = q(n_\alpha^{(1)}, n_\alpha^{(2)}, \dots, n_\alpha^{(k)})$ is an example of an IP polynomial. For example, if $\deg q(t_1, \dots, t_k) = 2$, then $q(\alpha)$ will typically look like

$$g(\alpha) = \sum_{i=1}^{s} n_\alpha^{(i)} m_\alpha^{(i)} + \sum_{i=1}^{r} k_\alpha^{(i)}.$$

The following result, obtained in [17], extends Theorem 4.39 to the case of several commuting transformations and to the family of IP polynomials described above.

THEOREM 4.40 [17, Corollary 2.1]. *Suppose that $(X, \mathcal{B}, \mu)$ is a probability space and that $\{T_1, T_2, \dots, T_t\}$ is a collection of commuting invertible measure preserving transformations of X. Suppose that $(n_\alpha^{(i)})_{\alpha\in\mathcal{F}} \subset \mathbb{N}$ are IP sets, $i = 1, 2, \dots, k$, and that $p_j(x_1, \dots, x_k) \in \mathbb{Z}[x_1, \dots, x_k]$ satisfy $p_j(0, 0, \dots, 0) = 0$ for $j = 1, 2, \dots, t$. Then for any measurable $A \subseteq X$, there exists an IP-ring $\mathcal{F}^{(1)} \subset \mathcal{F}$ such that*

$$\operatorname*{IP-lim}_{\alpha\in\mathcal{F}^{(1)}} \mu\left(A \cap \prod_{i=1}^{t} T_i^{p_i(n_\alpha^{(1)}, n_\alpha^{(2)}, \dots, n_\alpha^{(k)})} A\right) \geqslant \mu(A)^2.$$

Theorem 4.40 is obtained in [17] as a corollary of the following general fact about families of unitary operators, which can be viewed as a polynomial variation of Theorem 4.38. Note that the IP-ring $\mathcal{F}^{(1)}$ which occurs in the formulation, always exists due to the compactness of the weak topology.

THEOREM 4.41 [17, Theorem 1.8]. *Suppose that $\mathcal{H}$ is a Hilbert space, $(U_i)_{i=1}^t$ is a commuting family of unitary operators on $\mathcal{H}$, $(p_i(x_1, \dots, x_k))_{i=1}^t \subset \mathbb{Z}[x_1, \dots, x_k]$ satisfy*

$p_i(0,0,\dots,0)=0$ for $1\leqslant i\leqslant t$, and that $(n_\alpha^{(i)})_{\alpha\in\mathcal{F}}$ are IP sets for $1\leqslant j\leqslant t$. Suppose that $\mathcal{F}^{(1)}$ is an IP-ring such that for each $f\in\mathcal{H}$,

$$\operatorname*{IP-lim}_{\alpha\in\mathcal{F}^{(1)}}\left(\prod_{i=1}^{t}U_i^{p_i(n_\alpha^{(1)},\dots,n_\alpha^{(k)})}\right)f=P_{(p_1,\dots,p_t)}f$$

exists in the weak topology. Then $P_{(p_1,\dots,p_t)}$ is an orthogonal projection. Projections of this type commute, that is, if also $(q_i(x_1,\dots,x_k))_{i=1}^{t}\subset\mathbb{Z}[x_1,\dots,x_k]$ satisfy $q_i(0,0,\dots,0)=0$ for $1\leqslant i\leqslant t$, then

$$P_{(p_1,\dots,p_t)}P_{(q_1,\dots,q_t)}=P_{(q_1,\dots,q_t)}P_{(p_1,\dots,p_t)}.$$

An interesting feature of the proof of Theorem 4.40 is the usage of the following extension of Hindman's theorem, due independently to K. Milliken and A. Taylor. (See [105] and [129].)

THEOREM 4.42 [105,129]. *Suppose that $\mathcal{F}^{(1)}$ is an IP-ring, $l,r\in\mathbb{N}$, and*

$$\left\{(\alpha_1,\dots,\alpha_l)\in\left(\mathcal{F}^{(1)}\right)^l\colon\ \alpha_1<\alpha_2<\cdots<\alpha_l\right\}=\bigcup_{i=1}^{r}C_i.$$

Then there exists j, $1\leqslant j\leqslant r$, and an IP-ring $\mathcal{F}^{(2)}\subset\mathcal{F}^{(1)}$ such that

$$\left\{(\alpha_1,\dots,\alpha_l)\in\left(\mathcal{F}^{(2)}\right)^l\colon\ \alpha_1<\alpha_2<\cdots<\alpha_l\right\}\subset C_j.$$

The next natural step is to (try to) extend Theorem 4.40 to a multiple recurrence result. The following theorem obtained (as a corollary of a more general result) in [32], which we will call the IP polynomial Szemerédi theorem, is an IP extension of Theorem 4.27. (Cf. Theorem 2.9 above.)

THEOREM 4.43 [32, Theorem 0.9]. *Suppose we are given r commuting invertible measure preserving transformations $T_1,\dots,T_r$ of a probability space $(X,\mathcal{B},\mu)$. Let $k,t\in\mathbb{N}$ and suppose that $p_{i,j}(n_1,\dots,n_k)\in\mathbb{Q}[n_1,\dots,n_k]$ satisfy $p_{i,j}(\mathbb{Z}^k)\subseteq\mathbb{Z}$ and $p_{i,j}(0,0,\dots,0)=0$ for $1\leqslant i\leqslant r$, $1\leqslant j\leqslant t$. Then for every $A\in\mathcal{B}$ with $\mu(A)>0$, the set*

$$R_A=\left\{(n_1,\dots,n_k)\in\mathbb{Z}^k\colon\ \mu\left(\bigcap_{j=1}^{t}\left(\prod_{i=1}^{r}T_i^{p_{i,j}(n_1,\dots,n_k)}\right)A\right)>0\right\}$$

is an IP* set in $\mathbb{Z}^k$.

We collect some of the corollaries of Theorem 4.43 in the following list.

(i) Already for $k = 1$, Theorem 4.43 gives a refinement of the polynomial Szemerédi theorem (Theorem 4.27). Indeed, it says that the set

$$\Big\{n \in \mathbb{Z}\colon \mu\big(A \cap T_1^{p_{11}(n)} T_2^{p_{12}(n)} \cdots T_r^{p_{1r}(n)} A \cap \cdots \cap T_1^{p_{t1}(n)} T_2^{p_{t2}(n)} \cdots T_r^{p_{tr}(n)} A\big) > 0\Big\}$$

is IP*, hence syndetic, hence of positive lower density.

(ii) Theorem 4.43 also enlarges the family of configurations which can always be found in sets of positive upper Banach density in $\mathbb{Z}^n$. For example, using Furstenberg's correspondence principle, one obtains the following fact, in which the reader will recognize the density version of Theorem 2.12.

THEOREM 4.44. *Let $P : \mathbb{Z}^r \to \mathbb{Z}^l$, $r, l \in \mathbb{N}$, be a polynomial mapping satisfying $P(0) = 0$, and let $F \subset \mathbb{Z}^r$ be a finite set. Then for any set $E \subset \mathbb{Z}^l$ with $d^*(E) > 0$ and any IP sets $(n_\alpha^{(i)})_{\alpha\in\mathcal{F}}$, $i = 1, \ldots, r$, there exist $u \in \mathbb{Z}^l$ and $\alpha \in \mathcal{F}$ such that*

$$\big\{u + P\big(n_\alpha^{(1)} x_1, n_\alpha^{(2)} x_2, \ldots, n_\alpha^{(r)} x_r\big)\colon (x_1, \ldots, x_r) \in F\big\} \subset S.$$

See [17] for additional applications, both to combinatorics and to ergodic theory.

The proof of Theorem 4.43 that is given in [32] is quite cumbersome (partly due to the fact that in order to push the statement through the transfinite induction over the factors with "manageable" behavior, one has to formulate and prove an even more general result). In a way, it is a *polynomialization* of the proof of the IP Szemerédi theorem in [62]. Not being able to go through the details of the proof here, we would like to mention the two combinatorial facts which play a decisive role in the proof. One of them is the Milliken–Taylor theorem, formulated above as Theorem 4.42. The other one is the polynomial Hales–Jewett theorem, Theorem 2.11.

4.3.5. *The density version of Hales–Jewett theorem.* We are going to discuss now the density versions (and their ergodic counterparts) of three more partition theorems which we encountered in Sections 1 and 2.

We start with Theorem 1.26, the Hales–Jewett theorem. As we saw in Section 2, some major corollaries of the Hales–Jewett theorem, such as the multidimensional van der Waerden theorem and the so-called geometric Ramsey theorem, Theorem 1.27, follow from the IP van der Waerden theorem. The streamlined measure theoretical extension of the IP van der Waerden theorem, the IP Szemerédi theorem (Theorem 4.31) allows one to get the density versions of these corollaries. The IP Szemerédi theorem is, however, still not general enough to give the density version of the Hales–Jewett theorem. This density version, which we will refer to below as dHJ, was established by Furstenberg and Katznelson in [64]. Here is one of a few equivalent formulations of dHJ.

THEOREM 4.45 [64, Theorem E]. *There is a function $R(\varepsilon, k)$, defined for all $\varepsilon > 0$ and $k \in \mathbb{N}$, so that if A is a set with k elements, $W_N(A)$ consists of words in A with length N, and if $N \geqslant R(\varepsilon, k)$, then any subset $S \subset W_N(A)$ with $|S| \geqslant \varepsilon k^N$ contains a combinatorial line.*

In order to formulate the ergodic counterpart of Theorem 4.45 which was proved in [64], we shall need the following definition.

DEFINITION 4.46. (*See* [64, *Definition* 2.7].) Let $W(k)$ denote the free semigroup over the k-element alphabet $\{1, 2, \dots, k\}$. Given k sequences $\{T_n^{(1)}\}_{n=1}^{\infty}$, $\{T_n^{(2)}\}_{n=1}^{\infty}$, $\dots$, $\{T_n^{(k)}\}_{n=1}^{\infty}$ of invertible measure preserving transformations of a probability space $(X, \mathcal{B}, \mu)$, define, for each $w = (w(1), w(2), \dots, w(k)) \in W(k)$,

$$T(w) = T_1^{w(1)} T_2^{w(2)} \cdots T_k^{w(k)}.$$

The family $(T(w), w \in W(k))$ is called a $W(k)$-system.

Here is now the ergodic formulation of dHJ.

THEOREM 4.47. (See [64, Proposition 27].) *Let $\{T(w), w \in W(k)\}$ be a $W(k)$-system of invertible measure preserving transformations of a probability space $(X, \mathcal{B}, \mu)$. For any $A \in \mathcal{B}$ with $\mu(A) > 0$, there exists a combinatorial line $(l(t))_{t\in\{1,2,\dots,k\}}$ in $W(k)$ such that*

$$\mu\big(T\big(l(1)\big)^{-1}A \cap T\big(l(2)\big)^{-1}A \cap \cdots \cap T\big(l(k)\big)^{-1}A\big) > 0.$$

The proof of Theorem 4.47, while following the general scheme of the other proofs discussed above, is significantly more involved, mainly due to the fact that the transformations forming the $W(k)$-system need not commute. (As a matter of fact, in the case where the $W(k)$-system is formed by commutative transformations, the situation is reduced to the IP Szemerédi theorem.) Despite the absence of commutativity, the proof of Theorem 4.47 has a strong IP flavor. In particular, the authors use the IP version of the van der Corput trick, Theorem 4.37, and a (noncommutative) version of Theorem 4.38. Much more importantly, the authors are using an *infinitary* combinatorial result which is a simultaneous extension of the Hindman, Milliken–Taylor, and Hales–Jewett theorems. This combinatorial fact was also obtained by Carlson. (See [63,43] and [14].)

Before moving on with our discussion, we would like to stress that while Theorem 4.47 deals with an action of a free finitely generated semigroup, namely $W(k)$, it is a result about rather special configurations in $W(k)$.

4.3.6. *Multiple recurrence for nilpotent and solvable groups.* Another multiple recurrence theorem involving a noncommutative group is Leibman's nil-Szemerédi theorem obtained in [98], which is a density version of his nil-van der Waerden theorem (Theorem 2.15), and, at the same time, is an extension of Theorem 4.27.

THEOREM 4.48. (Cf. [98, Theorem NM].) *Let $k, t, r \in \mathbb{N}$. Assume that G is a nilpotent group of measure preserving transformations of a probability measure space $(X, \mathcal{B}, \mu)$. Let*

$p_{i,j}(n_1,\dots,n_k) \in \mathbb{Z}[n_1,\dots,n_k]$ with $p_{i,j}(\mathbb{Z}^k) \subseteq \mathbb{Z}$ and $p_{i,j}(0,0,\dots,0) = 0$, $1 \leqslant i \leqslant r$, $1 \leqslant j \leqslant t$. Then for every $A \in \mathcal{B}$ with $\mu(A) > 0$ and any $T_1, T_2, \dots T_r \in G$, the set

$$\left\{(n_1,\dots,n_k) \in \mathbb{Z}^k\colon \mu\left(\bigcap_{j=1}^{t}\left(\prod_{i=1}^{r} T_i^{p_{i,j}(n_1,\dots,n_k)}\right)A\right) > 0\right\}$$

is a syndetic set in $\mathbb{Z}^k$.

In his proof, Leibman builds a nilpotent version of primitive extensions similar to, but more sophisticated (due to the noncommutativity) than that which was introduced in [60]. We will describe it now. Let Y be a measure space and let $\{X_i\}_{i\in I}$ be a system of measure spaces of the form $X_i = Y \times F_i$, $i \in I$; then the measure space $X = Y \times \prod_{i\in I} F_i$ is called *a relatively direct product of X_i, $i \in I$, over Y*.

DEFINITION 4.49. (*Cf.* [98, *Definition* 11.10].) Let G be a finitely generated nilpotent group. An extension $\boldsymbol{X} = (X, \mathcal{B}, \mu, (T_g)_{g\in G})$ of a system $\boldsymbol{Y} = (Y, \mathcal{D}, \nu, (S_g)_{g\in G})$ is *primitive* if X is (isomorphic to) the relatively direct product over Y of a system $\{X_i\}_{i\in I}$ of measure spaces so that

(i) the transformations T_g, $g \in G$, on X permute the spaces X_i in the product: for any $g \in G$ and $i \in I$ one has $T_g(X_i) = X_j$ for some $j \in I$;
(ii) if $T = T_g$ preserves X_i, i.e. $T_g(X_i) = X_i$, then the action of T on X_i is either compact relative to Y or weak mixing relative to Y.

Modulo this definition, the structure theorem for measure preserving actions of a finitely generated nilpotent group is the same as the structure theorem for $\mathbb{Z}^k$-actions, Theorem 4.24 above.

THEOREM 4.50 [98, Theorem 11.11]. *If G is a finitely generated nilpotent group and $\boldsymbol{X} = (X, \mathcal{B}, \mu, (T_g)_{g\in G})$ is an extension of $\boldsymbol{Y} = (Y, \mathcal{D}, \nu, (S_g)_{g\in G})$, then there is an intermediate factor $\boldsymbol{Z}$ such that $\boldsymbol{Z}$ is a primitive extension of $\boldsymbol{Y}$.*

It is worth mentioning that the structure similar to that appearing in the case of measure preserving actions of nilpotent groups can already be observed on the unitary level.

THEOREM 4.51 [99, Theorem N]. *Let $\{T_g\}$ be a unitary action of a finitely generated nilpotent group G on a Hilbert space $\mathcal{H}$. Then $\mathcal{H}$ is representable as the direct sum of a system $\{\mathcal{L}_i\}_{i\in I}$ of closed pairwise orthogonal subspaces so that*:

(i) *the operators T_g, $g \in G$, permute the subspaces $\mathcal{L}_i$: for any $g \in G$ and $i \in I$ one has $T_g(\mathcal{L}_i) = \mathcal{L}_j$ for some $j \in I$,*
(ii) *if $T = T_g$ preserves $\mathcal{L}_i$, i.e. if $T(\mathcal{L}_i) = \mathcal{L}_i$, then either T is scalar on $\mathcal{L}_i$ or T is weakly mixing on $\mathcal{L}_i$.*

Another interesting feature of Leibman's proof of Theorem 4.48 is that in order to lift the recurrence property in question to relatively compact extensions, a coloring theorem is employed which is close in spirit to Theorem 2.17.

Leibman's nil-Szemerédi theorem naturally leads to the question whether the assumptions can be further relaxed and whether, in particular, an analogue of Theorem 1.24 holds true if the measure preserving transformations $T_1, T_2, \dots, T_k$ generate a solvable group. Note that any finitely generated solvable group is either of exponential growth or is virtually nilpotent, i.e. contains a nilpotent group of finite index. (See, for example, [121].) Since Theorem 4.48 easily extends to virtually nilpotent groups, the question boils down to solvable groups of exponential growth. The following result, proved in [27], shows, in a strong way, that for solvable groups of exponential growth the answer to the above question is NO.

THEOREM 4.52 [27, Theorem 1.1(A)]. *Assume that G is a finitely generated solvable group of exponential growth. There exist a measure preserving action $(T_g)_{g\in G}$ of G on a probability measure space $(X, \mathcal{B}, \mu)$, elements $g, h \in G$, and a set $A \in \mathcal{B}$ with $\mu(A) > 0$ such that $T_{g^n} A \cap T_{h^n} A = \emptyset$ for all $n \neq 0$.*

4.3.7. *Density version of polynomial Hales–Jewett theorem—a conjecture.* We conclude this section by formulating a conjecture about a density version of the polynomial Hales–Jewett theorem which, if true, extends both the partition polynomial Hales–Jewett theorem (Theorem 2.11) and the density version of the "linear" Hales–Jewett theorem (Theorem 4.45). For $q, d, N \in \mathbb{N}$, let $\mathcal{M}_{q,d,N}$ be the set of q-tuples of subsets of $\{1, 2, \dots, N\}^d$:

$$\mathcal{M}_{q,d,N} = \big\{(\alpha_1, \alpha_2, \dots, \alpha_q)\colon \alpha_i \subset \{1, 2, \dots, N\}^d, i = 1, 2, \dots, q\big\}.$$

CONJECTURE 4.53. *For any $q, d \in \mathbb{N}$ and $\varepsilon > 0$, there exists $C = C(q, d, \varepsilon)$ such that if $N > C$ and a set $S \subset \mathcal{M}_{q,d,n}$ satisfies $\frac{|S|}{|\mathcal{M}_{q,d,N}|} > \varepsilon$ then S contains a "simplex" of the form*:

$$\big\{(\alpha_1, \alpha_2, \dots, \alpha_q),\ (\alpha_1 \cup \gamma^d, \alpha_2, \dots, \alpha_q),\ (\alpha_1, \alpha_2 \cup \gamma^d, \dots, \alpha_q),\ \dots,$$
$$(\alpha_1, \alpha_2, \dots, \alpha_q \cup \gamma^d)\big\},$$

where $\gamma \subset \mathbb{N}$ is a nonempty set and $\alpha_i \cap \gamma^d = \emptyset$ for all $i = 1, 2, \dots, q$.

5. Actions of amenable groups

5.1. *Generalities*

One of the most striking theorems in mathematics, known as the Hausdorff–Banach–Tarski paradox (see [79] and [5]), claims that given any two bounded sets A and B in $\mathbb{R}^n$, $n \geqslant 3$, each having nonempty interior, one can partition A into finitely many disjoint parts and rearrange them by rigid motions of $\mathbb{R}^n$ to form B. What makes this fact even more striking is that (as was shown by Banach in [4]) the analogous result does not take place in $\mathbb{R}$ or $\mathbb{R}^2$.

It was von Neumann who, in his fundamental work [132], showed that the phenomenon of "paradoxicality" is related not so much to the structure of the space $\mathbb{R}^n$, but rather

to the group of transformations which is used to rearrange the elements of the partition. In particular, von Neumann introduced and studied in [132] a class of groups which he called "messbar" (measurable) and which do not allow for "paradoxical decompositions." These groups are called nowadays amenable (see [111, p. 137], for the origin of the term amenable) and are known to have connections to many mathematical areas, including probability theory, geometry, theory of dynamical systems and representation theory. As we shall see in this section, countable amenable semigroups provide also a natural framework for Furstenberg's correspondence principle. (See [74,114,112] for a comprehensive treatment of different aspects of amenability in the general framework of locally compact groups. See also [134] for a thorough and accessible discussion of amenability for discrete groups with the stress on connections to the Hausdorff–Banach–Tarski paradox.)

DEFINITION 5.1. Let G be a discrete semigroup. For $x \in G$ and $A \subset G$, let $x^{-1}A = \{y \in G\colon\ xy \in A\}$ and $Ax^{-1} = \{y \in G\colon\ yx \in A\}$. A semigroup G is called *left-amenable* (correspondingly, *right-amenable*) if there exists a finitely additive probability measure on the power set $\mathcal{P}(G)$ satisfying $\mu(A) = \mu(x^{-1}A)$ (correspondingly, $\mu(A) = \mu(Ax^{-1})$ for all $A \in \mathcal{P}(G)$ and $x \in G$). We say that G is *amenable* if it is both left- and right-amenable.

It is easy to see (cf. [134, p. 147]) that a semigroup G is amenable if and only if there exists an *invariant mean* on the space $B(G)$ of real-valued bounded functions on G, that is, a positive linear functional $L : B(G) \to \mathbb{R}$ satisfying

(i) $L(1_G) = 1$,

(ii) $L(f_g) = L({}_g f) = L(f)$ for all $f \in B(G)$ and $g \in G$, where $f_g(t) := f(tg)$ and ${}_g f(t) := f(gt)$.

The existence of an invariant mean is only one item from a long list of equivalent properties, (see, for example, [134, Theorem 10.11]), some of which, such as the characterization of amenability given in the next theorem, are far from being obvious and, moreover, are valid for groups (or special classes of semigroups) only. One of the advantages of dealing with groups is that for groups, the notions of left and right amenability coincide. (For an easy proof of this fact see, for example, [81, Theorem 17.11].)

We will find the following characterization of amenability for discrete groups, which was established by Følner in [54], to be especially useful. (See also [107] for a simplified proof.)

THEOREM 5.2. *A countable group G is amenable if and only if it has a left Følner sequence, namely a sequence of finite sets $F_n \subset G$, $n \in \mathbb{N}$, with $|F_n| \to \infty$ and such that $\frac{|F_n \cap gF_n|}{|F_n|} \to 1$ for all $g \in G$.*

REMARK 5.3.

(1) A right Følner sequence is defined (in an obvious way) as a sequence of finite sets $F_n \subset G$, $n \in \mathbb{N}$, for which $|F_n| \to \infty$ and $\frac{|F_n \cap F_n g|}{|F_n|} \to 1$ for all $g \in G$. While in noncommutative groups not every left Følner sequence is necessarily a right Følner sequence and vice versa, it is not hard to show that the existence of a sequence of either type in a semigroup implies the corresponding one-sided version of amenability. As was mentioned above, if G is a group, this is actually enough to get two-sided

amenability. The hard part of Theorem 5.2 is establishing the existence of a Følner sequence.

(2) Theorem 5.2 is also valid for semigroups possessing the cancellation law. See [107] for details.

(3) Theorem 5.2 can also be extended to general locally compact groups. See [74] for details.

It is not known how to construct Følner sequence in a general amenable group defined, say, by a finite set of generators and relations. On the other hand, in many concrete, especially Abelian, situations, one has no problem finding a Følner sequence. For example, it is easy to see that the sets F_n which occurred in the proof of Theorem 2.6, form a Følner sequence in F_∞. The reader should also have no problem verifying that d-dimensional parallelepipeds $\Pi_n = [a_n^{(1)}, b_n^{(1)}] \times [a_n^{(2)}, b_n^{(2)}] \times \cdots \times [a_n^{(d)}, b_n^{(d)}]$, where $\min_{1 \leqslant i \leqslant d} |a_i - b_i| \to \infty$ as $n \to \infty$, form a Følner sequence in $\mathbb{Z}^d$. Let us indicate how one can construct a Følner sequence in the cancellative Abelian semigroup $(\mathbb{N}, \cdot)$.

Let $(a_n)_{n \in \mathbb{N}}$ be an arbitrary sequence in $\mathbb{N}$ and let

$$F_n = \left\{ a_n p_1^{i_1} p_2^{i_2} \cdots p_n^{i_n} \colon 0 \leqslant i_j \leqslant k_{j,n},\ j = 1, 2, \ldots, n \right\},$$

where $k_{j,n}$ is a doubly indexed sequence of positive integers such that, for every j, $k_{j,n} \to \infty$ as $n \to \infty$, and $\{p_n\}$ is the sequence of primes taken in arbitrary order. It is not hard to check that $\{F_n\}_{n \in \mathbb{N}}$ is a Følner sequence in $(\mathbb{N}, \cdot)$.

The following theorem summarizes some general facts about amenable (semi)groups which were established already in [132]. (For accessible proofs, see [74, Chapter 1] and [134, Theorem 10.4].)

THEOREM 5.4.

(i) *Any Abelian semigroup is amenable.*

(ii) *Homomorphic images and subgroups of amenable groups are amenable.*

(iii) *If N is a normal subgroup of an amenable group G, then G/N is amenable.*

(iv) *If N is a normal subgroup of a group G, and if both N and G/N are amenable, then G is amenable.*

(v) *If a group G is a union of a family of amenable subgroups $\{H_\alpha\}_{\alpha \in I}$ so that for any H_α, H_β there exists H_γ with $H_\gamma \supset H_\alpha \cup H_\beta$, then G is amenable.*

It follows that the class of amenable groups is quite rich. In particular, it contains all solvable groups, since they can be obtained from Abelian groups by successive extensions with the help of Abelian groups. It follows also that a group is amenable if and only if all of its finitely generated subgroups are amenable. This, in turn, implies that all locally finite groups (i.e. the groups in which every finite subset generates a finite subgroup) are amenable.

On the other hand, the group $F_2 = \langle a, b \rangle$ (the free group on two generators) and hence any group containing it as a subgroup, is not amenable.

To see that F_2 is not amenable, one can argue as follows. Let $F_2 = A^+ \cup A^- \cup B^+ \cup B^- \cup \{e\}$, where e is the unit of F_2 (the "empty" word) and the sets A^+, A^-, B^+, and B^-

consist of the reduced words starting with a, a^{-1}, b, and b^{-1} respectively. Assume that μ is a finitely additive probability measure on $\mathcal{P}(F_2)$ satisfying $\mu(A) = \mu(gA)$ for any $A \in \mathcal{P}(F_2)$ and $g \in F_2$. Clearly, $\mu(\{e\}) = 0$. (If $\mu(\{e\}) = c > 0$, then, by translation invariance of μ, for any $n \in \mathbb{N}$, $\mu(\{b^n\}) = c$ and the set $\{e, b, b^2, \ldots, b^N\}$, where $N \geqslant \frac{1}{c}$, would have measure bigger than one.) Assume that $\mu(A^+) = c > 0$. (The same proof will work for any other set of our partition which has positive measure.) Let $A_n = b^n A$, $n \geqslant 0$. Clearly, the sets A_n are disjoint and, by translation invariance, have the same measure $c > 0$. It follows that the set $\bigcup_{n=0}^{N} A_n$, where, as before, $N \geqslant \frac{1}{c}$, has measure bigger than one, which gives a contradiction. (Note that the simple argument used here is similar to that utilized in the proof of the abstract version of the Poincaré recurrence theorem, Theorem 1.4.)

It follows now that groups such as $SL(n, \mathbb{Z})$ with $n \geqslant 2$ or $SO(3, \mathbb{R})$ (with the discrete topology) are not amenable, since one can show that they contain a subgroup isomorphic to F_2. As was observed by von Neumann in [132], it is the latter fact that is behind the Hausdorff–Banach–Tarski paradox. See [134] for a reader-friendly explanation of this fact. On the other hand, not every nonamenable group has to contain a subgroup isomorphic to F_2. Moreover, nonamenable groups can even be periodic. (See [109,110] and [112, p. 182].)

As is well known to aficionados, many classical notions and results pertaining to 1-parameter group actions extend naturally to amenable groups. Here is, for example, a version of von Neumann's ergodic theorem for actions of countable amenable groups.

THEOREM 5.5. *Let G be a countable amenable group. Assume that $(U_g)_{g\in G}$ is an antirepresentation of G as a group of unitary operators acting on a Hilbert space $\mathcal{H}$ (i.e. $U_{g_1}U_{g_2} = U_{g_2 g_1}$ for all $g_1, g_2 \in G$). Let P be the orthogonal projection on the space $\mathcal{H}_{\text{inv}} = \{f \in \mathcal{H}\colon\ U_g f = f\ \forall g \in G\}$. Then for any left Følner sequence $(F_n)_{n\in\mathbb{N}}$ in G, one has*

$$\lim_{n\to\infty} \left\| \frac{1}{|F_n|} \sum_{g\in F_n} U_g f - Pf \right\| = 0.$$

SKETCH OF THE PROOF. It is not hard to check that, in complete analogy to $\mathbb{Z}$-actions, the orthogonal complement of $\mathcal{H}_{\text{inv}}$ in $\mathcal{H}$, which we will denote by $\mathcal{H}_{\text{erg}}$, coincides with the space $\overline{\text{Span}}\{f - U_g f\colon\ f \in \mathcal{H}, g \in G\}$. So it remains to verify that on $\mathcal{H}_{\text{erg}}$, the limit in question is zero. It is enough to check this for elements of the form $f - U_{g_0} f$. We have:

$$\begin{aligned}
&\left\| \frac{1}{|F_n|} \sum_{g\in F_n} U_g (f - U_{g_0} f) \right\| \\
&\quad = \left\| \frac{1}{|F_n|} \sum_{g\in F_n} U_g f - \frac{1}{|F_n|} \sum_{g\in F_n} U_{g_0 g} f \right\| \\
&\quad = \left\| \frac{1}{|F_n|} \sum_{g\in F_n} U_g f - \frac{1}{|F_n|} \sum_{g\in g_0 F_n} U_g f \right\| \leqslant \frac{|F_n \Delta g_0 F_n|}{|F_n|} \|f\|.
\end{aligned}$$

Since, by the definition of a left Følner sequence, $\frac{|F_n \Delta g_0 F_n|}{|F_n|} \underset{n\to\infty}{\longrightarrow} 0$, we are done. □

Recall that a measure preserving action $(T_g)_{g\in G}$ of a group G on a probability space $(X, \mathcal{B}, \mu)$ is *ergodic* if any set $A \in \mathcal{B}$ which satisfies $\mu(T_g A \Delta A) = 0$ for all $g \in G$ has either measure zero or measure one. The reader should have no problem in verifying the following corollary of Theorem 5.5.

THEOREM 5.6. *Assume that $(T_g)_{g\in G}$ is an ergodic measure preserving action of a countable amenable group G. Then for any (left or right) Følner sequence $(F_n)_{n\in\mathbb{N}}$ of G, and any $A_1, A_2 \in \mathcal{B}$, one has*

$$\frac{1}{|F_n|}\sum_{g\in F_n}\mu(A_1 \cap T_g A_2) \underset{n\to\infty}{\longrightarrow} \mu(A_1)\mu(A_2).$$

Here is another useful result, whose proof can be transferred almost verbatim from the proof of the classical Bogoliouboff–Kryloff theorem. (See, for example, [135, Theorem 6.9 and Corollary 6.9.1].)

THEOREM 5.7. *Let $(T_g)_{g\in G}$ be an action of an amenable group G by homeomorphisms of a compact metric space X. Then there is a probability measure on the Borel σ-algebra $\mathcal{B}(X)$ such that for any $A \in \mathcal{B}(X)$ and any $g \in G$, one has $\mu(A) = \mu(T_g A)$.*

REMARK. Unlike the von Neumann ergodic theorem, the pointwise theorem for actions of amenable groups is a much harder and more delicate result, which was proved in the right generality (that is, for any locally compact amenable group and for functions in L^1) only recently, in a remarkable paper of E. Lindenstrauss [103]. For a comprehensive survey of pointwise ergodic theorems for general group actions see [2].

We are now going to discuss Ramsey-theoretical aspects of amenable groups.

Given a countable amenable group G and, say, a left Følner sequence $\{F_n\}_{n\in\mathbb{N}}$ in G, one can define the upper density with respect to $\{F_n\}_{n\in\mathbb{N}}$ by $\bar{d}_{\{F_n\}}(E) = \limsup_{n\to\infty}\frac{|E\cap F_n|}{|F_n|}$, $E \subset G$. Note that it immediately follows from the definition of a left Følner sequence that for all $g \in G$ and $E \subset G$, one has $\bar{d}_{\{F_n\}}(gE) = \bar{d}_{\{F_n\}}(E)$. By analogy with some known results about sets of positive density in Abelian or nilpotent groups which were discussed in previous sections, one can expect that large sets in G, i.e. sets having positive upper density with respect to some Følner sequence, will contain some nontrivial configurations. The results which we will formulate below support this point of view and lead to a general conjecture, which will be formulated at the end of this section.

5.2. *Correspondence principle for countable amenable groups*

We start by formulating and proving a version of Furstenberg's correspondence principle for countable amenable groups.

THEOREM 5.8. (See [11, Theorem 6.4.17].) *Let G be a countable amenable group and assume that a set $E \subset G$ has positive upper density with respect to some left Følner sequence $\{F_n\}_{n\in\mathbb{N}}$: $\bar{d}_{\{F_n\}}(E) = \limsup_{n\to\infty} \frac{|E\cap F_n|}{|F_n|} > 0$. Then there exists a probability measure preserving system $(X, \mathcal{B}, \mu, (T_g)_{g\in G})$ and a set $A \in \mathcal{B}$ with $\mu(A) = \bar{d}_{\{F_n\}}(E)$ such that for any $k \in \mathbb{N}$ and $g_1, \ldots, g_k \in G$, one has*

$$\bar{d}_{\{F_n\}}\big(E \cap g_1^{-1}E \cap \cdots \cap g_k^{-1}E\big) \geqslant \mu\big(A \cap T_{g_1}^{-1}A \cap \cdots T_{g_k}^{-1}A\big).$$

PROOF. We show first that there exists a left-invariant mean L on the space $B(G)$ of bounded real-valued functions on G such that

(i) $L(1_E) = \bar{d}_{\{F_n\}}(E)$,

(ii) for any $k \in \mathbb{N}$ and any $g_1, \ldots, g_k \in G$, one has

$$\bar{d}_{\{F_n\}}\big(E \cap g_1^{-1}E \cap \cdots \cap g_k^{-1}E\big) \geqslant L(1_E \cdot 1_{g_1^{-1}E} \cdot \cdots \cdot 1_{g_k^{-1}E}).$$

Let $\mathcal{S}$ be the (countable) family of subsets of G of the form $\bigcap_{j=1}^k g_j^{-1}E$, where $k \in \mathbb{N}$ and $g_j \in G$, $j = 1, 2, \ldots, k$. By using the diagonal procedure, we can pass to a subsequence $\{F_{n_i}\}_{i=1}^\infty$ of our Følner sequence such that for our set E we have $\bar{d}_{\{F_n\}}(E) = \lim_{i\to\infty} \frac{|E\cap F_{n_i}|}{|F_{n_i}|}$ and for any $S \in \mathcal{S}$ the limit $L(S) = \lim_{i\to\infty} \frac{|S\cap F_{n_i}|}{|F_{n_i}|} = \lim_{i\to\infty} \frac{1}{|F_{n_i}|} \sum_{g\in F_{n_i}} 1_S(g)$ exists. Observe that for a typical set $S = \bigcap_{j=1}^k g_j^{-1}E \in \mathcal{S}$ this will give

$$\begin{aligned}
\bar{d}_{\{F_n\}}\left(\bigcap_{j=1}^k g_j^{-1}E\right) &= \limsup_{n\to\infty} \frac{\big|\big(\bigcap_{j=1}^k g_j^{-1}E\big) \cap F_n\big|}{|F_n|} \\
&\geqslant \lim_{i\to\infty} \frac{\big|\big(\bigcap_{j=1}^k g_j^{-1}E\big) \cap F_{n_i}\big|}{|F_{n_i}|} \\
&= L\left(\bigcap_{j=1}^k 1_{g_j^{-1}E}\right).
\end{aligned}$$

Extending by linearity, we will get a positive linear functional L on the subspace $V \subset B(G)$ of finite linear combinations of characteristic functions of sets in $\mathcal{S}$. Note that it follows from the definition of a left Følner sequence that this functional L on V is left-invariant, i.e. for any $f \in V$ and $g \in G$, one has $L(f) = L({}_g f)$, where as before ${}_g f(t) = f(gt)$.

To extend L from V to $B(G)$, define the Minkowski functional $P(f)$ by $P(f) = \limsup_{i\to\infty} \frac{1}{|F_{n_i}|} \sum_{g\in F_{n_i}} f(g)$. Clearly, for any $f_1, f_2 \in B(G)$, one has $P(f_1 + f_2) \leqslant P(f_1) + P(f_2)$, and for any nonnegative t, $P(tf) = tP(f)$. Note also that, on V, $L(f) = P(f)$. By the Hahn–Banach theorem, there is an extension of L (which we will denote by L as well) to $B(G)$ satisfying $L(f) \leqslant P(f)$ for all $f \in B(G)$. Clearly, L is a left-invariant mean satisfying conditions (i) and (ii) above.

Finally, in preparation for the next stage of the proof, let us note that L can be naturally extended to the space $B_{\mathbb{C}}(G)$ of complex valued bounded functions. We shall continue to denote this extension by L.

Let now $f(h) = 1_E(h)$ be the characteristic function of E and let $\mathcal{A}$ be the uniformly closed and closed under conjugation functional algebra generated by the function f and all of the functions of the form ${}_g f$, where $g \in G$. Then $\mathcal{A}$ is a separable (G is countable and linear combinations with rational coefficients are dense in $\mathcal{A}$) commutative C*-algebra with respect to the sup norm. By the Gelfand representation theorem, $\mathcal{A}$ is isomorphic to an algebra of the form $C(X)$, where X is a compact metric space. The linear functional L, which we constructed above, induces a positive linear functional $\tilde{L}$ on $C(X)$. By the Riesz representation theorem, there exists a regular measure μ on the Borel σ-algebra $\mathcal{B}$ of X such that for any $\varphi \in \mathcal{A}$,

$$L(\varphi) = \tilde{L}(\tilde{\varphi}) = \int_X \tilde{\varphi}\, d\mu,$$

where $\tilde{\varphi}$ denotes the image of φ in $C(X)$. Notice that since the Gelfand transform, establishing the isomorphism between $\mathcal{A}$ and $C(X)$, preserves the algebraic operations, and since the characteristic functions of sets are the only idempotents in $C(X)$, it follows that the image $\tilde{f}$ of our $f(h) = 1_E(h)$ is the characteristic function of some set $A \subset X$: $\tilde{f}(x) = 1_A(x)$. This gives

$$\bar{d}_{\{F_n\}}(E) = L(1_E) = \tilde{L}(1_A) = \int_X 1_A\, d\mu = \mu(A).$$

Notice also that the translation operators $\varphi(h) \to \varphi(gh)$, $\varphi \in \mathcal{A}$, $g \in G$, form an anti-action of G on $\mathcal{A}$, which induces an anti-action $(T_g)_{g\in G}$ on $C(X)$ defined for any $\varphi \in \mathcal{A}$ by $(T_g)\tilde{\varphi} = {}_g\tilde{\varphi}$. The transformations T_g, $g \in G$ are C*-isomorphisms of $C(X)$ (since they are induced by C*-isomorphisms $\varphi \to {}_g\varphi$ of $\mathcal{A}$). Now, it is known that algebra isomorphisms of $C(X)$ are induced by homeomorphisms of X, which we, by a slight abuse of notation, will also be denoting by T_g, $g \in G$. These homeomorphisms $T_g : X \to X$ form an action of G on X and preserve the measure μ. To see this, let $C \in \mathcal{B}$ and let $\varphi \in \mathcal{A}$ be the preimage of 1_C (so that $\tilde{\varphi} = 1_C$). For an arbitrary $g \in G$ we have:

$$\begin{aligned}\mu(C) &= \int_X 1_C(x)\, d\mu(x) = \tilde{L}(\tilde{\varphi}) = L(\varphi) = L({}_g\varphi) = \tilde{L}({}_g\tilde{\varphi}) \\ &= \tilde{L}\big(\tilde{\varphi}(T_g x)\big) = \int_X 1_C(T_g x)\, d\mu(x) = \int 1_{T_g^{-1}C}(x)\, d\mu(x) \\ &= \mu\big(T_g^{-1}C\big).\end{aligned}$$

Notice also that since $L(\mathbf{1}) = 1$, $\mu(X) = \tilde{L}(1_X) = 1$. It follows that $(X, \mathcal{B}, \mu, (T_g)_{g\in G})$ is a probability measure preserving system. (As a bonus, we have that, in this representation, the measure preserving transformations T_g are homeomorphisms of a compact metric

space.) We finally have, for $f = 1_E$, $g_0 = e$, and any $g_1, \dots, g_k \in G$:

$$\bar{d}_{\{F_n\}}\left(\bigcap_{j=0}^{k} g_j^{-1}E\right) \geqslant L\left(\prod_{j=0}^{k} g_j f\right) = \tilde{L}\left(\prod_{j=0}^{k} g_j \tilde{f}\right)$$
$$= \tilde{L}\left(\prod_{j=0}^{k}\left((T_{g_j})f\right)\right) = \int_X \prod_{j=0}^{k} 1_{T_{g_j}^{-1}A} = \mu\left(\bigcap_{j=0}^{k} T_{g_j}^{-1}A\right).$$

We are done. □

REMARK 5.9. It is not hard to modify the proof above to make Theorem 5.8 valid for any countable amenable semigroup possessing a left Følner sequence. As a matter of fact, Furstenberg's correspondence principle can be extended to general countable amenable semigroups if, instead of using Følner sequences, one defines a set $E \subset G$ to be large if for some left-invariant mean L on $B(G)$, one has $L(1_E) > 0$. (See [31, Theorem 2.1].) The proof in [31] is different also in that it avoids the usage of the Gelfand transform. See also Remark 6.4.21 in [11], which describes an approach to the proof of Theorem 5.8 which does not make use of C*-algebras.

The following useful lemma will be used repeatedly in the sequel. (Note that while, in view of the pending applications, it is arranged in the amenable set-up, the lemma is actually completely general and has, in principle, very little to do with amenability.)

LEMMA 5.10. (Cf. [7, Theorem 1.1].) *Let $\{F_n\}_{n\in\mathbb{N}}$ be a (left or right) Følner sequence in a countable amenable semigroup G, let $(X, \mathcal{B}, \mu)$ be a probability space, and let, for every $g \in G$, there be given $A_g \in \mathcal{B}$ with $\mu(A_g) \geqslant a > 0$. Then there exists a set $S \subset G$ with $d_{\{F_n\}}(S) \geqslant A$, such that for any finite set $F \subset S$ one has $\mu(\bigcap_{g\in F} A_g) > 0$.*

PROOF. For any finite set $F \subset G$, let $A_F = \bigcap_{g\in F} A_g$. Deleting, if needed, a set of measure zero from $\bigcup_{g\in G} A_g$, we may and will assume that if $A_F \neq 0$ then $\mu(A_F) > 0$. Let now

$$f_n(x) = \frac{1}{|F_n|}\sum_{g\in F_n} 1_{A_g}(x).$$

Note that $0 \leqslant f_n(x) \leqslant 1$ for all x and that $\int f_n\, d\mu \geqslant a > 0$ for all $n \in \mathbb{N}$. Let $f(x) = \limsup_{n\to\infty} f_n(x)$. By Fatou's lemma, we have

$$\int_X f\, d\mu = \int_X \limsup_{n\to\infty} f\, d\mu \geqslant \limsup_{n\to\infty} \int_X f_n\, d\mu \geqslant a.$$

Thus $\int_X f\, d\mu \geqslant a$ and, since $\mu(X) = 1$, there exists $x_0 \in X$ such that

$$\limsup_{n\to\infty} f_n(x_0) = f(x_0) \geqslant a.$$

It follows that there is a sequence $n_i \to \infty$ such that

$$f_{n_i}(x_0) = \frac{1}{|F_{n_i}|} \sum_{g \in F_{n_i}} 1_{A_g}(x) \underset{i \to \infty}{\longrightarrow} f(x_0) \geqslant a. \tag{5.1}$$

Let $P = \{g \in G\colon\ x_0 \in A_g\}$. It follows from (5.1) that $\bar{d}_{\{F_n\}}(P) \geqslant a$, and, since $x_0 \in A_g$ for all $g \in P$, we have that $\mu(A_F) > 0$ for every finite nonempty $F \subset P$. □

Applying Theorem 5.8 (more precisely, the version of Theorem 5.8 for semigroups possessing a Følner sequence), we immediately obtain the following result.

COROLLARY 5.11. *Let $\{F_n\}_{n\in\mathbb{N}}$ be a left Følner sequence in an amenable semigroup G, and let $E \subset G$ satisfy $\bar{d}_{\{F_n\}}(E) = c > 0$. Then there exists a set $P \subset G$ with $\bar{d}_{\{F_n\}}(P) \geqslant c$ such that for any $g_1, \ldots, g_k \in P$, $\bar{d}_{\{F_n\}}(\bigcap_{i=1}^{k} g_i^{-1} E) > 0$.*

In order to formulate another application of Lemma 5.10, we need to introduce first the following definition.

DEFINITION 5.12. Let G be a countable semigroup. A set $R \subset G$ is called a set of *measurable recurrence* if for any measure preserving action $(T_g)_{g\in G}$ on a probability space $(X, \mathcal{B}, \mu)$ and any $A \in \mathcal{B}$ with $\mu(A) > 0$, there exists $g \in R$, $g \neq e$, such that $\mu(A \cap T_g^{-1} A) > 0$.

Different semigroups have all kinds of peculiar sets of recurrence. For example, it follows from Theorem 1.31 that for any polynomial $p(n) \in \mathbb{Z}[n]$ with $p(0) = 0$, the set $\{p(n)\colon\ n \in \mathbb{Z}\}$ is a set of measurable recurrence for $\mathbb{Z}$-actions. Moreover, Theorem 4.39 tells us that for any IP set $(n_\alpha)_{\alpha\in\mathcal{F}} \subset \mathbb{N}$, the set $\{p(n_\alpha)\colon\ \alpha \in \mathcal{F}\}$ is a set of measurable recurrence. More generally, one can show (see [17]) that, for any $p_1(n), \ldots, p_k(n) \in \mathbb{Z}[n]$ satisfying $p_i(0) = 0$, $i = 1, \ldots, k$, and any IP sets $(n_\alpha^{(1)})_{\alpha\in\mathcal{F}}, \ldots, (n_\alpha^{(k)})_{\alpha\in\mathcal{F}}$, the set

$$\bigl\{p_1\bigl(n_\alpha^{(1)}\bigr), \ldots, p_k\bigl(n_\alpha^{(k)}\bigr)\colon\ \alpha \in \mathcal{F}\bigr\} \subset \mathbb{Z}^k$$

is a set of measurable recurrence. Sets of the form $\{1 + \frac{1}{k}\colon\ k \in \mathbb{N}\}$ can be shown to be sets of measurable recurrence for the multiplicative group of positive rationals. This list can be continued indefinitely.

The following theorem shows that, for countable amenable semigroups possessing a Følner sequence, the notion of a set of measurable recurrence coincides with the notion of "density recurrence":

THEOREM 5.13. *Let S be an amenable semigroup having a left Følner sequence. Then $R \subset S$ is a set of measurable recurrence if and only if for any left Følner sequence $\{F_n\}_{n\in\mathbb{N}}$ in G and any $E \subset G$ with $\bar{d}_{\{F_n\}}(E) > 0$ there exists $g \in R$, $g \neq e$, such that $E \cap g^{-1} E \neq \emptyset$.*

PROOF. In one direction, the claim of the theorem immediately follows from Furstenberg's correspondence principle. So, it remains to show that if for any left Følner sequence

$\{F_n\}_{n\in\mathbb{N}}$ and $E \subset G$ with $\bar{d}_{\{F_n\}}(E) > 0$ there exists $g \in R$, $g \neq e$, such that $E \cap g^{-1}E \neq \emptyset$, then R is a set of measurable recurrence. Let $(T_g)_{g\in G}$ be a measure preserving action on a probability space $(X, \mathcal{B}, \mu)$ and let $A \in \mathcal{B}$ with $\mu(A) > 0$. It follows from (the proof of) Lemma 5.10 that we may assume that, if $A \cap T_g^{-1}A \neq \emptyset$, then $\mu(A \cap T_g^{-1}A) > 0$, and that there exist a set $P \subset G$ with $\bar{d}_{\{F_n\}}(P) \geqslant \mu(A)$, and a point $x \in X$ such that, for any $g \in P$, one has $T_g x \in A$.

By our assumptions, there exists $g \in R$, $g \neq e$, such that $P \cap g^{-1}P \neq \emptyset$. Letting $h \in P \cap g^{-1}P$, we have $h, gh \in P$. It follows that $T_h x \in A$ and $T_{gh}x = T_g(T_h x) \in A$. This implies that, simultaneously, $T_h x \in A$ and $T_h x \in T_g^{-1}A$, which gives $A \cap T_g^{-1}A \neq \emptyset$ and, hence, $\mu(A \cap T_g^{-1}A) > 0$. We are done. □

The following version of Theorem 5.13 is valid for any amenable semigroup. (See [31, Theorem 2.2].)

THEOREM 5.14. *Suppose that S is a countable left amenable semigroup. Then $R \subset S$ is a set of measurable recurrence if and only if for every left-invariant mean L and every $E \subset S$ with $L(1_E) > 0$, one has $E \cap g^{-1}E \neq \emptyset$ for some $g \in R$, $g \neq e$.*

REMARK 5.15. One can show that both Furstenberg's correspondence principle and Theorem 5.13 fail for $\mathbb{R}$-actions. Indeed, it is proved in [15] that

(i) For any $\alpha > 0$, $\{n^\alpha\colon n \in \mathbb{N}\}$ is a set of measurable recurrence for (continuous) measure preserving $\mathbb{R}$-actions.

(ii) For all but countably many $\alpha > 1$, one can find a measurable set $E \subset R$ such that

$$d(E) = \lim_{t\to\infty} \frac{m(E \cap [0, t])}{t} = \frac{1}{2}$$

and $E \cap (E - n^\alpha) = \emptyset$ for all $n \in \mathbb{N}$.

5.3. *Applications to multiplicatively large sets*

DEFINITION 5.16. A set $E \subseteq \mathbb{N}$ is called multiplicatively large if for some Følner sequence $\{F_n\}_{n\in\mathbb{N}}$ in $(\mathbb{N}, \cdot)$, one has $\bar{d}_{\{F_n\}}(E) > 0$.

We shall use now Lemma 5.10 to obtain some new results about multiplicatively large sets.

We start by remarking that the notions of largeness for sets in $\mathbb{N}$, which are based on additive and multiplicative structures, are different. For example, the set O of odd natural numbers has (additive!) density $\frac{1}{2}$ with respect to any Følner sequence in $(\mathbb{N}, +)$. On the other hand, it is not hard to see that the set O will have zero density along any Følner sequence in $(\mathbb{N}, \cdot)$. In the other direction, consider, for example, a Følner sequence $\{a_n F_n\}_{n\in\mathbb{N}}$ in $(\mathbb{N}, \cdot)$, which is defined as follows. Let

$$F_n = \left\{p_1^{i_1} p_2^{i_2} \cdots p_n^{i_n}, 0 \leqslant i_j \leqslant n, 1 \leqslant j \leqslant n\right\},$$

where p_i, $i = 1, 2, \ldots$, are primes in arbitrary order, and let the integers a_n satisfy $a_n > |F_n|$, $n \in \mathbb{N}$. Let now $S = \bigcup_{n=1}^{\infty} a_n F_n$. It is easy to see that S has zero additive density with respect to any Følner sequence in $(\mathbb{N}, +)$. At the same time, S has multiplicative density one with respect to the Følner sequence $\{a_n F_n\}_{n\in\mathbb{N}}$.

As may be expected by mere analogy with additively large sets, multiplicatively large sets always contain (many) geometric progressions. (This can be derived, for example, with the help of the IP Szemerédi theorem, see Section 4.2.) It turns out, however, that multiplicatively large sets also contain arbitrarily long arithmetic progressions and some other, somewhat unexpected, configurations.

THEOREM 5.17. (See [13, Theorem 3.2].) *Any multiplicatively large set $E \subseteq \mathbb{N}$ contains arbitrarily long arithmetic progressions.*

PROOF. Invoking Furstenberg's correspondence principle, let $(X, \mathcal{B}, \mu, (T_n)_{n\in\mathbb{N}})$ be the corresponding measure preserving system (where $(T_n)_{n\in\mathbb{N}}$ is a measure preserving action of $(\mathbb{N}, \cdot)$), and let $A \in \mathcal{B}$ be the set of positive measure corresponding to E. Let $A_n = T_n^{-1}A$. Clearly, $\mu(A) = \mu(A_n)$ for all $n \in \mathbb{N}$. By Lemma 5.10, there exists an additively large set S with the property that for any finite $F \subset S$, one has $\mu(\bigcap_{n\in F} T_n^{-1}A) > 0$. Using Szemerédi's theorem, we get, for arbitrary $k \in \mathbb{N}$, an arithmetic progression $P_k = \{n + id\colon i = 0, 1, \ldots, k-1\} \subset S$ such that

$$\mu\left(\bigcap_{n\in P_k} T_n^{-1}A\right) > 0.$$

Applying again Furstenberg's correspondence principle, we see that the set $\bigcap_{n\in P_k} E/n$ is multiplicatively large and, in particular, nonempty. This implies that, for some $n \in \mathbb{N}$, $E \supset mP_k$. □

By working a little bit harder, one can show that any multiplicatively large set contains *geoarithmetic* progressions, namely configurations of the form $\{bq^j(a+id)\colon 0 \leqslant i, j \leqslant n\}$. (See [13, Theorem 3.11].) The following result describes yet another type of geoarithmetic configurations, which can always be found in multiplicatively large sets. (See [13] for more results on, and a discussion of, the combinatorial richness of multiplicatively large sets in $\mathbb{N}$.)

THEOREM 5.18. (See [13, Theorem 3.15].) *Let $E \subset \mathbb{N}$ be a multiplicatively large set. For any $k \in \mathbb{N}$, there exist $a, b, d \in \mathbb{N}$ such that*

$$\left\{b(a+id)^j\colon 0 \leqslant i, j \leqslant k\right\} \subset E.$$

5.4. *Multiple recurrence for amenable groups*

We shall address now the question about possible amenable extensions of the multiple recurrence results discussed in Section 4. While it is not clear at all how to even formulate

an amenable generalization of the one-dimensional Szemerédi theorem (either ergodic or combinatorial), it is, curiously enough, not too hard to guess what should be an amenable version of the multidimensional Szemerédi theorem.

CONJECTURE 5.19. *Let G be a countable amenable group with a Følner sequence $\{F_n\}_{n\in\mathbb{N}}$. Let $(T_g^{(1)})_{g\in G},\dots,(T_g^{(k)})_{g\in G}$ be k pairwise commuting measure preserving actions of G on a probability measure space $(X,\mathcal{B},\mu)$. ("Pairwise commuting" means here that for any $1\leqslant i\neq j\leqslant k$ and any $g,h\in G$, one has $T_g^{(i)}T_h^{(j)}=T_h^{(j)}T_g^{(i)}$.) Then for any $A\in\mathcal{B}$ with $\mu(A)>0$ one has*:

$$\lim_{n\to\infty}\frac{1}{|F_n|}\sum_{g\in F_n}\mu\big(A\cap T_g^{(1)}A\cap T_g^{(1)}T_g^{(2)}A\cap\cdots\cap T_g^{(1)}T_g^{(2)}\cdots T_g^{(k)}A\big)>0.$$

REMARK 5.20. The "triangular" expressions

$$A\cap T_g^{(1)}A\cap T_g^{(1)}T_g^{(2)}A\cap\cdots\cap T_g^{(1)}T_g^{(2)}\cdots T_g^{(k)}A$$

appearing in the formulation above, seem to be the "right" configurations to consider. See the discussion and the counterexamples in [20].

The following theorem lists the known instances of the validity of Conjecture 5.19.

THEOREM 5.21. *Conjecture* 5.19 *holds true in the following situations*:

(i) [30] *For $k=2$.*

(ii) [34] *For general k but under the additional assumption that each of the following actions is ergodic*:

$$\begin{aligned}
&\big(T_g^{(k)}\big)_{g\in G},\\
&\big(T_g^{(k-1)}\otimes T_g^{(k-1)}T_g^{(k)}\big)_{g\in G},\\
&\qquad\vdots\\
&\big(T_g^{(2)}\otimes T_g^{(2)}T_g^{(3)}\otimes\cdots\otimes T_g^{(2)}T_g^{(3)}\cdots T_g^{(k)}\big)_{g\in G},\\
&\big(T_g^{(1)}\otimes T_g^{(1)}T_g^{(2)}\otimes\cdots\otimes T_g^{(1)}T_g^{(2)}\cdots T_g^{(k)}\big)_{g\in G}.
\end{aligned}$$

(In this case, the limit in question equals $(\mu(A))^{k+1}$.)

While the case $k=2$ corresponds to the intersection of three sets only, it allows one to derive some interesting combinatorial corollaries, some of which are brought together in the following theorem.

THEOREM 5.22.

(i) [30, *Theorem* 6.1] *Suppose G is a countable amenable group and that $E \subset G \times G$ has positive upper density with respect to a left Følner sequence $\{F_n\}_{n\in\mathbb{N}}$ for $G \times G$. Then the set*

$$\big\{g \in G\colon \text{ there exists } (a,b) \in G \times G \text{ such that } \big\{(a,b),(ga,b),(ga,gb)\big\} \subset E\big\}$$

is a synetic set in G.

(ii) [30, *Corollary* 7.2] *Suppose that G is a countable amenable group, $r \in \mathbb{N}$, $G \times G \times G = \bigcup_{i=1}^{r} C_i$. Then the set*

$$\big\{g \in G\colon \text{ there exist } i, 1 \leqslant i \leqslant r, \text{ and } (a,b,c) \in G \times G \times G \text{ such that } \big\{(a,b,c),(ga,b,c),(ga,gb,c),(ga,gb,gc)\big\} \subset C_i\big\}$$

is a syndetic set in G.

(iii) [31, *Theorem* 3.4] *Suppose that G is a countable amenable group and that $G = \bigcup_{i=1}^{r} C_i$ is a finite partition. Let $A = \{g \in G\colon [G : C(g)] < \infty\}$, where $C(g)$ is the centralizer of g. If $[G : A] = \infty$, then there exist $x, y \in G$ and i, $1 \leqslant i \leqslant r$, with $xy \neq yx$ and such that $\{x, y, xy, yx\} \subset C_i$.*

We conclude this section by fulfilling the promise made in Section 1:

PROOF OF THEOREM 1.6. Let $A = \{a_1, a_2, \ldots\} \subset \Gamma$ be an infinite set. Denoting by $[A]^2$ the set of two-element subsets of A, let us define a finite coloring of $[A]^2$ by assigning to each $\{a_i, a_j\} \in [A]^2$, $i < j$, the coset $(a_i - a_j)\Gamma$. (This coloring is finite since Γ is of finite index in F^*.) We will apply now the infinite version of Ramsey's theorem (see [72, Theorem 5, p. 16]), which says that for any finite coloring of the set of two-element subsets $[P]^2$ of an infinite set P there exists an infinite subset $P_1 \subset P$ such that the set of two-element subsets of P_1, $[P_1]^2$, is monochromatic. It follows that there is $c \in F^*$ and an infinite set $B = \{a_{n_1}, a_{n_2}, \ldots\} \subset A$ such that each member of $[B]^2$ has the same color, $c\Gamma$. This, in turn, says that $a_{n_i} - a_{n_j} \in c\Gamma$ for all $i < j$. Writing $b_i = c^{-1}a_{n_i}$, we see that Γ itself contains an infinite difference set $\{b_i - b_j\}_{i<j}$.

Using now the amenability of the Abelian group F, let μ be a finitely additive translation-invariant probability measure on $\mathcal{P}(F)$. Since there are only finitely many disjoint cosets of Γ in F^* and since, clearly, $\mu(\{0\}) = 0$, one of the cosets, call it $c\Gamma$, has to satisfy $\mu(c\Gamma) > 0$. Let $x \in F^*$ be an arbitrary element and consider the sets $c\Gamma + xb_i$, $i \in \mathbb{N}$. Applying the familiar by now reasoning, we see that for some $i < j$, $\mu((c\Gamma + xb_i) \cap (c\Gamma + xb_j)) > 0$. This implies that $c\Gamma \cap (c\Gamma - x(b_i - b_j)) \neq \emptyset$, which, in turn, gives us $x(b_i - b_j) \in c\Gamma - c\Gamma$. Since $b_i - b_j \in \Gamma$, we get $x \in c\Gamma - c\Gamma$, and, since x was arbitrary, it gives us $F = c\Gamma - c\Gamma$, and, after the cancellation, $F = \Gamma - \Gamma$. We are done. □

REMARK 5.23. The methods used in the above proof can be applied to more general (and not necessarily commutative) rings. Some of the generalizations are given in [35].

By invoking stronger combinatorial theorems, such as the IP-Szemerédi theorem, one can actually show that if F is an infinite field and Γ is a multiplicative subgroup of finite index in F^*, then for any finite set $S \subset F$ there is $\gamma \in \Gamma$ such that $\gamma + S = \{\gamma + x\colon\ x \in S\} \subset \Gamma$. This, in turn, implies that there exists a finitely additive translation-invariant probability measure μ on F such that $\mu(\Gamma) = 1$.

6. Issues of convergence

This relatively short section is devoted to the discussion of ergodic theorems which are related to combinatorial and number-theoretic applications of ergodic theory.

Various convergence results and conjectures that we have already encountered in the previous sections typically emerged as a means of establishing various recurrence results. Yet, from a purely ergodic-theoretical point of view, these results are of significant interest on their own. While, in order to obtain combinatorial corollaries, one is perfectly satisfied with establishing the positivity of a lim inf of Cesàro averages (see, for example, Theorem 4.2), the ideology and tradition of ergodic theory immediately leads to questions whether the limit of a pertinent Cesàro sum exists in norm or almost everywhere.

These questions usually lead to the development of new strong analytic techniques which, in turn, not only provide deeper knowledge about the structure of dynamical systems, but also enhance our understanding of the mutually perpetuating connections between ergodic theory, combinatorics, and number theory.

Consider, for example, Theorem 1.31. As we have seen in Section 1, a convenient way of showing that, for any measure preserving system $(X, \mathcal{B}, \mu, T)$, any $A \in \mathcal{B}$ with $\mu(A) > 0$, and any polynomial $p(n) \in \mathbb{Z}[n]$ satisfying $p(0) = 0$, one has $\mu(A \cap T^{p(n)}A) > 0$, is to consider the averages

$$\frac{1}{N}\sum_{n=0}^{N-1} \mu(A \cap T^{p(n)}A)$$

and to show that the limit of these averages is positive. This implies that the set

$$\left\{n \in \mathbb{N}\colon\ \mu(A \cap T^{p(n)}A) > 0\right\} \tag{6.1}$$

has positive upper density, which, in turn, implies that the equation $x - y = p(n)$ has "many" integer solutions (x, y, n) with $x, y \in E$, $n \in \mathbb{N}$. At this point the interests of combinatorial number theory and conventional ergodic theory part. While the Cesàro averages are of little help if one wants to undertake the more refined study of the set (6.1) (see Theorem 4.39 and the discussion preceding it), it is the focus of the classical ergodic theory on the equidistribution of orbits, which makes the following question interesting.

QUESTION 6.1. Given an invertible probability measure preserving system $(X, \mathcal{B}, \mu, T)$, a polynomial $p(n) \in \mathbb{Z}[n]$ and a function $f \in L^p(X, \mathcal{B}, \mu)$, where $p \geqslant 1$, is it true that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^{p(n)}x) \tag{6.2}$$

exists almost everywhere?

Note that while the norm convergence of the averages (6.2) is not hard to establish (we did it at the end of Section 1 for $p = 2$, which almost immediately implies the norm convergence in any L^p space for $p \geqslant 1$), the pointwise convergence is quite a bit harder. It was J. Bourgain who developed in the late eighties a powerful technique which allowed him to answer Question 6.1 in the affirmative first for $p = 2$ [39] and soon after for any $p > 1$ [40]. The case $p = 1$ is still open and is perhaps one of the central open problems in that branch of ergodic theory which deals with almost everywhere convergence.

For an excellent survey of Bourgain's methods and a thorough discussion of various positive and negative results on pointwise ergodic theorems, the reader is referred to [122]. See also Appendix B, where A. Quas and M. Wierdl present a reader-friendly simplified proof of Bourgain's theorem on a.e. convergence along the set of squares (i.e. for $p(n) = n^2$) for functions in the L^2 space.

In view of the multiple recurrence results discussed in Section 4, the following question naturally suggests itself.

QUESTION 6.2. Let $T_1, T_2, \ldots, T_k$ be invertible measure preserving transformations which act on a probability space $(X, \mathcal{B}, \mu)$ and generate a nilpotent group. Is it true that for any polynomials $p_i(n) \in \mathbb{Z}[n]$ and $f_i \in L^\infty(X, \mathcal{B}, \mu)$, $i = 1, 2, \ldots, k$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f_1\big(T_1^{p_1(n)}x\big) f_2\big(T_2^{p_2(n)}x\big) \cdots f_k\big(T_k^{p_k(n)}x\big) \tag{6.3}$$

exists in the L^2 norm? Almost everywhere?

We are going to describe the status of current knowledge in the following brief comments.

The only known result on almost everywhere convergence for $k > 1$ is due, again, to Bourgain, who showed in [41] that for $k = 2$, $p_1(n) = an$, $p_2(n) = bn$, $a, b \in \mathbb{Z}$, the limit in (6.3) exists a.e. for any f_1, $f_2 \in L^\infty$. (It is not hard to show that this implies also the a.e. result for any f_1, $f_2 \in L^2$.)

Assume now that $T_1 = T_2 = \cdots = T_k = T$. As was already mentioned in Section 4, the convergence of the averages

$$\frac{1}{N} \sum_{n=0}^{N-1} f_1(T^n x) f_2(T^{2n} x) \cdots f_k(T^{kn} x) \tag{6.4}$$

in L^2 norm was only recently established by Host and Kra [88] and, independently, Ziegler [139]. Appendix A, written by A. Leibman, gives, among other things, a glimpse into the structure of the proofs contained in the impressive papers [88] and [139]. (While the main result proved in Appendix A is somewhat special and deals with the so-called *characteristic factors* for the averages $\frac{1}{N}\sum_{n=0}^{N-1} f_1(T^{a_1 n}x) f_2(T^{a_2 n}x)\cdots f_k(T^{a_k n}x)$, the apparatus and techniques utilized there should provide the reader with a better understanding of the methods involved in the study of the averages (6.4).)

The following result, obtained very recently by A. Leibman [101], shows that the Host–Kra and Ziegler theorems can be extended to polynomial expressions.

THEOREM 6.3 [101]. *For $T_1 = T_2 = \cdots = T_k = T$, the averages* (6.3) *converge in L^2.*

It was shown in [66] that if T is totally ergodic (i.e. T^n is ergodic for any $n \neq 0$), then for any $f, g \in L^\infty(X, \mathcal{B}, \mu)$, one has

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f_1(T^n x) f_2\big(T^{n^2} x\big) = \int f_1 \, d\mu \int f_2 \, d\mu$$

in the L^2 norm.

The following theorem, proved in [55], gives a nice generalization of this fact.

THEOREM 6.4 [55]. *Assume that $(X, \mathcal{B}, \mu, T)$ is an invertible totally ergodic system. Then for any rationally independent polynomials $p_1(n), p_2(n), \ldots, p_k(n) \in \mathbb{Z}[n]$ and any $f_i \in L^\infty(X, \mathcal{B}, \mu)$, $i = 1, 2, \ldots, k$, one has*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f_1\big(T^{p_1(n)} x\big) f_2\big(T^{p_2(n)} x\big) \cdots f_k\big(T^{p_k(n)} x\big) = \int f_1 \, d\mu \int f_2 \, d\mu \cdots \int f_k \, d\mu.$$

The L^2-convergence of the averages $\frac{1}{N}\sum_{n=0}^{N-1} f_1(T_1^n x) f_2(T_2^n x)$ for commuting T_1, T_2 (where $f_1, f_2 \in L^\infty(X, \mathcal{B}, \mu)$) was established in [46]. The following result, obtained in [25], provides a nilpotent extension of this fact.

THEOREM 6.5 [25]. *Let T_1, T_2 be measure preserving transformations of a probability space $(X, \mathcal{B}, \mu)$ generating a nilpotent group. Then for any $f_1, f_2 \in L^\infty(X, \mathcal{B}, \mu)$,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f_1\big(T_1^n x\big) f_2\big(T_2^n x\big) \tag{6.5}$$

exists in the L^2 norm.

REMARK 6.6. Similarly to the situation with recurrence (see Theorem 4.52), one can show that if T_1, T_2 generate a solvable group of exponential growth, then the averages (6.5) do not always converge. See Theorem 1.1(B) in [27].

Due to our specific interest in ergodic theorems related to the material surveyed in the previous sections, we have focused here only on rather special (but important) convergence issues. For more information on pointwise convergence, the reader is referred to [94] and [68] as well as to the survey [122] mentioned above and the article of A. Nevo [2], in this volume.

Acknowledgement

The author would like to thank Ronnie Pavlov and Sasha Leibman for their assistance in preparing this survey, and Neil Hindman, Emmanuel Lesigne, and Randall McCutcheon for helpful comments on the preliminary version of this survey. Special thanks go also to the editors of this volume, Anatole Katok and Boris Hasselblatt for their infinite patience.

Appendix A. Host–Kra and Ziegler factors and convergence of multiple ergodic averages, by A. Leibman

A.1. *Multiple ergodic averages*

The *nonconventional*, or *multiple* ergodic averages

$$\frac{1}{N}\sum_{n=1}^{N} T^n f_1 \cdot \dots \cdot T^{kn} f_k, \tag{A.1}$$

where T is a measure preserving transformation of a probability measure space X and $f_1, \dots, f_k$ are (bounded) measurable functions on X, were introduced by H. Furstenberg in his ergodic-theoretical proof of Szemerédi's theorem [57]. In order to prove Szemerédi's theorem, it was sufficient to show that, in the case $f_1 = \dots = f_k \geqslant 0, \not\equiv 0$, the lim inf of the averages (A.1) is nonzero, and Furstenberg had confined himself to proving this fact. The question whether the limit of the multiple ergodic averages exists in L^1-sense was an open problem for more than twenty years, until it was answered positively by Host and Kra [88] and, independently, by Ziegler [139]. The way of solving this problem was suggested in [57]: one has to determine a factor Z of X which is *characteristic* for the averages (A.1), which means that the limiting behavior of (A.1) only depends on the conditional expectation of f_i with respect to Z:

$$\left\| \frac{1}{N}\sum_{n=1}^{N} \bigl(T^n f_1 \cdot \dots \cdot T^{kn} f_k - T^n E(f_1|Z) \cdot \dots \cdot T^{kn} E(f_k|Z)\bigr) \right\|_{L^1(X)} \underset{N\to\infty}{\longrightarrow} 0$$

for any $f_1, \dots, f_k \in L^\infty(X)$, or equivalently, that $\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} T^n f_1 \cdot \dots \cdot T^{kn} f_k = 0$ whenever one of $E(f_i|Z)$, $i = 1, \dots, k$, is equal to 0. Once a characteristic factor Z has been found, the problem is restricted to the system (Z, T); one therefore succeeds if he/she manages to show that every system (X, T) possesses a characteristic factor with a relatively simple structure, so that the convergence of averages (A.1) can be easily established for it. For example, under the assumption that T is ergodic, one can show that the Kronecker factor K of X is characteristic for the two-term averages $\frac{1}{N} \sum_{n=1}^{N} T^n f_1 \cdot T^{2n} f_2$ (see [57]; see also Section 4.2 of this survey). Since K has a structure of a compact Abelian group on which T acts as a translation, it is not hard to see that the averages above converge for $f_1, f_2 \in L^\infty(K)$.

A k-step nilsystem is a pair (N, T) where N is a compact homogeneous space of a k-step nilpotent group G and T is a translation of N defined by an element of G. When G is a nilpotent Lie group, N is called *a k-step nilmanifold*; if G is an inverse limit of nilpotent Lie groups, N is called *a k-step pro-nilmanifold*. After Conze and Lesigne had shown [46–48] that the characteristic factor for the three-term multiple ergodic averages is a two-step nilsystem, it was natural to conjecture that the characteristic factor for the averages (A.1) with arbitrary k is a $(k-1)$-step nilsystem. Host and Kra, and, independently, Ziegler have confirmed this conjecture by constructing such factors.

Ziegler's factors $Y_{k-1}(X, T)$, $k = 2, 3, \dots$, are characteristic for the averages of the form

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} T^{a_1 n} f_1 \cdot \dots \cdot T^{a_k n} f_k \tag{A.2}$$

for any $a_1, \dots, a_k \in \mathbb{Z}$. Ziegler's construction is a (very complicated) extension of that of Conze and Lesigne: she obtains the factor $Y_k(X, T)$ as a product of $Y_{k-1}(X, T)$ and a compact Abelian group H so that T acts as a skew-product transformation on $Y_k(X, T) = Y_{k-1}(X, T) \times H$, $T(y, h) = (Ty, h + \rho(y))$, with ρ satisfying certain conditions that allow one to impose on $Y_k(X, T)$ the structure of a k-step pro-nilmanifold with T being a translation on it. She also shows that $Y_{k-1}(X, T)$ is the minimal factor of X which is characteristic for all averages of the form (A.2), and the maximal factor of X having the structure of a $(k-1)$-step pro-nilmanifold.

Host and Kra used another, very elegant construction. They first describe the characteristic factor for the (numerical) averages of the form

$$\lim_{N_k\to\infty} \frac{1}{N_k} \sum_{n_k=1}^{N_k} \cdots \lim_{N_1\to\infty} \frac{1}{N_1} \sum_{n_1=1}^{N_1} \int_X \prod_{\varepsilon_1,\dots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_1 n_1 + \dots + \varepsilon_k n_k} f_{\varepsilon_1,\dots,\varepsilon_k}. \tag{A.3}$$

(These averages are not introduced in [88] explicitly, but can be clearly observed in the very construction of the Host–Kra factors; see Proposition A.9 below.) While the expression (A.3) looks forbidding, it is quite natural. (For instance, when $k = 2$ it is just $\lim_{N_2\to\infty} \frac{1}{N_2} \sum_{n_2=1}^{N_2} \lim_{N_1\to\infty} \frac{1}{N_1} \sum_{n_1=1}^{N_1} \int_X f_{0,0} \cdot T^{n_1} f_{1,0} \cdot T^{n_2} f_{0,1} \cdot T^{n_1+n_2} f_{1,1}$.) The corresponding characteristic factor, which will be denoted by $Z_{k-1}(X, T)$, can be easily constructed inductively (we will describe this construction below), and Host and Kra

prove that, for each k, the factor $Z_{k-1}(X,T)$ possesses a structure of a $(k-1)$-step pro-nilmanifold. The averages (A.3) turn out to be "universal": successive applications of the van der Corput lemma (see [8]; see also Theorems 1.32 and 4.6 of the main text) allow one to majorize by the averages (A.3), with suitable $k = k(l)$, all averages of the form

$$\lim_{N\to\infty} \frac{1}{|\Phi_N|} \sum_{u\in\Phi_N} T^{\varphi_1(u)} f_1 \cdot \ldots \cdot T^{\varphi_l(u)} f_l, \tag{A.4}$$

where $\varphi_1, \ldots, \varphi_l$ are linear functions $\mathbb{Z}^d \to \mathbb{Z}$ and $\{\Phi_N\}_{N=1}^{\infty}$ is any Følner sequence in $\mathbb{Z}^d$. (The averages (A.3) also majorize the "polynomial" averages of the form $\lim_{N\to\infty} \frac{1}{|\Phi_N|} \sum_{u\in\Phi_N} T^{p_1(u)} f_1 \cdot \ldots \cdot T^{p_l(u)} f_l$, where $p_1, \ldots, p_l$ are polynomials $\mathbb{Z}^d \to \mathbb{Z}$; see [89] and [101].) It follows that the factors $Z_k(X,T)$, with $k = k(l)$, are characteristic for the averages (A.4). In particular, it is shown in [88] that $Z_{k-1}(X,T)$ is characteristic for the averages (A.1), and in [89] that $Z_k(X,T)$ is characteristic for the averages of the form (A.2).

In this note we first describe the Host–Kra construction. We then show that the Host–Kra factors associated with a nontrivial power T^l of a transformation T are the same as the factors associated with T itself. (In [89] this was done for the case of a totally ergodic T only; we give a different proof of this fact.) Next, we prove that, actually, for $k \geqslant 2$ already $Z_{k-1}(X,T)$ is characteristic for the averages (A.4) and, in particular, (A.2). (The existence of the limit (A.4) will now follow from two facts: (i) $(Z_{k-1}(X,T), T)$ is isomorphic to a nil-system on a pro-nilmanifold, and (ii) the averages (A.4) converge for such a nilsystem; for a (quite nontrivial) proof of the first fact see [88], for a proof of the second fact see [102] and [100].) As a corollary, we obtain that the Host–Kra factors $Z_{k-1}(X,T)$ coincide with the corresponding Ziegler factors $Y_{k-1}(X,T)$. Indeed, being a $(k-1)$-step pro-nilmanifold, $Z_{k-1}(X,T)$ is a factor of $Y_{k-1}(X,T)$; on the other hand, since $Y_{k-1}(X,T)$ is the minimal characteristic factor for the averages (A.2), it is a factor of $Z_{k-1}(X,T)$.

A.2. *Construction of Host–Kra factors*

We will set up some terminology and notation. We will assume that the measure spaces we deal with are regular, that is, are metric spaces endowed with a probability Borel measure. (Any separable measure preserving system has a regular model; see, for example, [58, Chapter 5].) Let $\pi : X \to V$ be a measurable mapping from a measure space $(X, \mathcal{B}, \mu)$ to a measure space $(V, \mathcal{D}, \nu)$. If π is measure preserving, that is, $\mu(\pi^{-1}(A)) = \nu(A)$ for all $A \in \mathcal{D}$, V is called *a factor* of X. (Note that here V is a factor of a space, not of a dynamical system.) We will denote by X_v the fiber $\pi^{-1}(v)$, $v \in V$. Let $f \in L^1(X)$; then $\mu_f(D) = \int_{\pi^{-1}(D)} f\, d\mu$ is a (signed) measure on V absolutely continuous with respect to ν. By the Radon–Nikodym theorem $d\mu_f / d\nu$ is an integrable function on V; it is denoted by $E(f|V)$ and is called the *conditional expectation* of f with respect to V. The fibers X_v, $v \in V$, may be given a structure of measure spaces with probability measures μ_v, $v \in V$, such that $\int_{X_v} f\, d\mu_v = E(f|V)(v)$ for all $f \in L^1(X)$. (See [58, Chapter 5].) We will refer to the partition $X = \bigcup_{v\in V} X_v$ as to *the decomposition of X with respect to V*.

If $(V, \mathcal{D}, \nu)$ is a factor of $(X, \mathcal{B}, \mu)$, with $\pi : X \to V$ being the factorization mapping, then $\pi^{-1}(\mathcal{D})$ is a sub-σ-algebra of $\mathcal{B}$, which we will identify with $\mathcal{D}$. Conversely, with any sub-σ-algebra $\mathcal{D}$ of $\mathcal{B}$ a factor V of X is associated; roughly speaking, V is the partition of X induced by $\mathcal{D}$. (One can construct V in the following way. Choose a countable system $\{D_n\}_{n\in\mathbb{N}}$ generating $\mathcal{D}$. (Such a system may not, actually, exist; we then take a countable system that generates the subsets from $\mathcal{D}$ up to measure zero.) For each n, let $D_n^0 = D_n$ and $D_n^1 = X \setminus D_n$. Put $V = \{0,1\}^{\mathbb{N}}$, and for each $v = (e_1, e_2, \ldots) \in V$ put $X_v = \bigcap_{n=1}^{\infty} D_n^{e_n}$. This defines a mapping $X \to V$, $X_v \mapsto \{v\}$; a measure on V is inherited from X.)

Let $(V, \mathcal{D}, \nu)$ be a factor of $(X, \mathcal{B}, \mu)$ and $\pi : X \to V$ be the factorization mapping. *The relative square* $X \times_V X$ is the subspace $\{(x_1, x_2)\colon\ \pi(x_1) = \pi(x_2)\} = \bigcup_{v\in V} X_v \times X_v$ of $X \times X$, with the measure $\mu \times_V \mu = \int_V \mu_v \times \mu_v \, d\nu$ thereon. A mapping $X \times_V X \to V$ is naturally defined by $(x_1, x_2) \mapsto \pi(x_1)(= \pi(x_2))$, and turns V into a factor of $X \times_V X$ with fibers $(X \times_V X)_v = X_v \times X_v$, $v \in V$, so that $\bigcup_{v\in V} X_v \times X_v$ is the decomposition of $X \times_V X$ with respect to V. (Note that, if we start with a σ-subalgebra $\mathcal{D}$ of $\mathcal{B}$, the space representing the corresponding factor V is not defined canonically, and $X \times_V X$ is only defined up to measure zero. Usually, no new underlying space is introduced for V, and V is simply taken to be the nonregular measure space $(X, \mathcal{D}, \mu)$. The relative square $X \times_V X$ is then defined as X^2 with the measure given by $(\mu \times_V \mu)(A \times B) = \int_V \mu_v(A)\mu_v(B)\, d\nu$, $A, B \in \mathcal{B}$. We use the "set-theoretical" approach to make the geometric picture more transparent. This however leads to some delicate problems related to the fact that our constructions are only defined up to measure zero. The reader is referred to [58, Chapter 5] for a detailed treatment of measure-theoretical issues.) To simplify notation, starting from this moment we will not designate measures; for each space appearing below it will be clear from the context what measure is assumed thereon.

Now let T be a measure preserving transformation of X. We will denote by $\mathcal{I}(X, T)$ the σ-algebra of T-invariant measurable subsets of X and by $I(X, T)$ the factor of X associated with $\mathcal{I}(X, T)$. The decomposition $X = \bigcup_{v\in I(X,T)} X_v$ of X with respect to $I(X, T)$ is then the ergodic decomposition of X. To simplify notation, we will write $X \times_T X$ for $X \times_{I(X,T)} X$.

The Host–Kra factors of X with respect to T are constructed in the following way. One puts $X_T^{[0]} = X$, $T^{[0]} = T$, and when $X_T^{[k]}$ and $T^{[k]}$ have been defined for certain k, let $X_T^{[k+1]} = X_T^{[k]} \times_{T^{[k]}} X_T^{[k]}$ and let $T^{[k+1]}$ be the restriction of $T^{[k]} \times T^{[k]}$ on $X_T^{[k+1]}$. For any $k = 0, 1, \ldots$, $X_T^{[k]}$ is a measurable subspace of X^{2^k}; let $\mathcal{Z}_k(X, T)$ be the minimal σ-algebra on X such that $\mathcal{I}(X_T^{[k]}, T^{[k]}) \subseteq \mathcal{Z}_k(X, T)^{\otimes 2^k}$. *The* kth *Host–Kra factor* $Z_k(X, T)$ *of* X *with respect to* T is the factor of X associated with $\mathcal{Z}_k(X, T)$.

Assume that $(V, \mathcal{D})$ is a factor of X such that the fibers X_v, $v \in V$, are T-invariant, $T(X_v) = X_v$. Since "life is independent" in distinct fibers X_v, we have:

LEMMA A.1. *For any* k, *the spaces* $X_T^{[k]}$ *and* $I(X_T^{[k]}, T^{[k]})$ *decompose with respect to* V *to, respectively,* $\bigcup_{v\in V}(X_v)_T^{[k]}$ *and* $\bigcup_{v\in V} I((X_v)_T^{[k]}, T^{[k]})$.

PROOF. Let $X = \bigcup_{\alpha\in I(X,T)} X_\alpha$ be the decomposition of X with respect to $I(X, T)$, that is, the ergodic decomposition of X. Elements of the σ-algebra $\mathcal{D} \subseteq \mathcal{B}$ are preserved by T, thus $\mathcal{D} \subseteq \mathcal{I}(X, T)$, and V is a factor of $I(X, T)$. Let $I(X, T) = \bigcup_{v\in V} I_v$ be the

decomposition of $I(X,T)$ with respect to V. For (almost) every $v \in V$ the decomposition $X_v = \bigcup_{\alpha \in I_v} X_\alpha$ is the ergodic decomposition of X_v, and thus $I(X_v,T) = I_v$. Hence, $\bigcup_{v\in V} I(X_v,T)$ is the decomposition of $I(X,T)$ with respect to V. Now,

$$X_T^{[1]} = \bigcup_{\alpha\in I(X,T)} X_\alpha \times X_\alpha = \bigcup_{v\in V}\bigcup_{\alpha\in I_{X_v,T}} X_\alpha \times X_\alpha = \bigcup_{v\in V} (X_v)_T^{[1]}.$$

(To be accurate, we also have to check that the measure on $X_T^{[1]}$ agrees with this decomposition. It does:

$$\mu \times_{I(X,T)} \mu = \int_{I(X,T)} \mu_\alpha \times \mu_\alpha \, d\alpha = \int_V \int_{I_v} \mu_\alpha \times \mu_\alpha \, d\alpha \, dv = \int_V \mu_v \, dv.)$$

We then proceed by induction on k. □

Let $\bigcup_{i\in V} X_i$ be a finite measurable partition of X. The finite set V can then be considered as a factor of X (with measure defined by $\nu(\{i\}) = \mu(X_i)$, and the fibers X_i having measures $\mu_i = \mu/\mu(X_i)$, $i \in V$). For this case, Lemma A.1 says that $X_T^{[k]}$ and $I(X_T^{[k]}, T^{[k]})$ partition to, respectively, $\bigcup_{i\in V}(X_i)_T^{[k]}$ and $\bigcup_{i\in V} I((X_i)_T^{[k]}, T^{[k]})$. It follows that $\mathcal{I}(X_T^{[k]}, T^{[k]}) = \prod_{i\in V} \mathcal{I}((X_i)_T^{[k]}, T^{[k]})$, and that $\mathcal{Z}_k(X,T) = \prod_{i\in V} \mathcal{Z}_k(X_i,T)$ and $Z_k(X,T) = \bigcup_{i\in V} Z_k(X_i,T)$.

A.3. *Host–Kra factors for T^l*

Our first goal is to investigate the Host–Kra factors associated with T^l, $l \neq 0$.

THEOREM A.2. *For any $l \neq 0$ and $k \geqslant 1$ the kth Host–Kra factor $Z_k(X, T^l)$ of X with respect to T^l coincides with the kth Host–Kra factor $Z_k(X,T)$ of X with respect to T.*

PROOF. We fix a nonzero integer l. It follows from Lemma A.1 that it suffices to prove Theorem A.2 for an ergodic T only. We first assume that T^l is also ergodic. Given a measure preserving transformation S of a measure space Y, let us denote by $\mathcal{E}_\lambda(Y,S)$ the eigenspace of S in $L^1(Y)$ corresponding to the eigenvalue λ, $\mathcal{E}_\lambda(Y,S) = \{f \in L^1(Y)\colon Sf = \lambda f\}$. In particular, $\mathcal{E}_1(Y,S)$ is the space of S-invariant integrable functions on Y, which we will denote by $\mathcal{L}(Y,S)$. □

LEMMA A.3. (Cf. [89].) *Let S be a measure preserving transformation of a measure space Y. If S^l is ergodic, then $I(Y\times Y, S^l \times S^l) = I(Y\times Y, S\times S)$.*

PROOF. S^l is ergodic means that $\mathcal{E}_\lambda(Y,S) = \{0\}$ for all $\lambda \neq 1$ with $\lambda^l = 1$. We have $\mathcal{L}(Y\times Y, (S\times S)^l) \subseteq \operatorname{span}\{\mathcal{E}_\lambda(Y\times Y, S\times S)\colon \lambda^l = 1\}$. For any $\lambda \in \mathbb{C}$, $|\lambda| = 1$, the space $\mathcal{E}_\lambda(Y\times Y, S\times S)$ is spanned by the functions of the form $f \otimes g$ where $f \in \mathcal{E}_{\lambda_1}(Y,S)$ and $g \in \mathcal{E}_{\lambda_2}(Y,S)$ with $\lambda_1\lambda_2 = \lambda$. For such a function, $fg \in \mathcal{E}_\lambda(Y,S)$. If $\lambda \neq 1$ and

$\lambda^l = 1$, we have $fg = 0$; since S is ergodic, $|f| = \text{const}$ and $|g| = \text{const}$, so either $f = 0$ or $g = 0$. Thus, for any $\lambda \neq 1$ with $\lambda^l = 1$ we have $\mathcal{E}_\lambda(Y \times Y, S \times S) = \{0\}$. Hence, $\mathcal{L}(Y \times Y, S^l \times S^l) \subseteq \mathcal{E}_1(Y \times Y, S \times S) = \mathcal{L}(Y \times Y, S \times S)$. With the evident opposite inclusion $\mathcal{L}(Y \times Y, S \times S) \subseteq \mathcal{L}(Y \times Y, S^l \times S^l)$ this implies $\mathcal{I}(Y \times Y, S^l \times S^l) = \mathcal{I}(Y \times Y, S \times S)$. □

LEMMA A.4. (Cf. [89].) *Let T be a measure preserving transformation of a measure space X. If T^l is ergodic then $X^{[k]}_{T^l} = X^{[k]}_T$ and $I(X^{[k]}_{T^l}, (T^l)^{[k]}) = I(X^{[k]}_T, T^{[k]})$ for all $k \geqslant 0$.*

PROOF. For $k = 0$ the statement is trivial. Assume by induction that, for some $k \geqslant 0$, $Y = X^{[k]}_{T^l} = X^{[k]}_T$ and $I = I(Y, (T^l)^{[k]}) = I(Y, T^{[k]})$. Then $X^{[k+1]}_{T^l} = X^{[k+1]}_T = Y \times_I Y$. Let $Y = \bigcup_{\alpha \in I} Y_\alpha$ be the decomposition of Y with respect to I and for each $\alpha \in I$ let $S_\alpha = T^{[k]}|_{Y_\alpha}$. By the induction assumption S^l_α is ergodic on Y_α for every $\alpha \in I$, thus by Lemma A.1 and Lemma A.3 applied to the systems (Y_α, S_α),

$$
\begin{aligned}
&I\big(Y \times_I Y, (T^l)^{[k]} \times (T^l)^{[k]}\big) \\
&\quad = \bigcup_{\alpha \in I} I\big(Y_\alpha \times Y_\alpha, S^l_\alpha \times S^l_\alpha\big) \\
&\quad = \bigcup_{\alpha \in I} I(Y_\alpha \times Y_\alpha, S_\alpha \times S_\alpha) = I\big(Y \times_I Y, T^{[k]} \times T^{[k]}\big).
\end{aligned}
$$
□

It follows that $Z_k(X, T^l) = Z_k(X, T)$ for all $k \geqslant 0$, which proves Theorem A.2 in the case T^l is ergodic.

Now assume that T is ergodic whereas T^l is not. We may assume that l is a prime integer. In this case X is partitioned, up to a subset of measure 0, to measurable subsets $X_0, \ldots, X_{l-1}$ such that $T(X_i) = X_{i+1}$ for all $i \in \mathbb{Z}_l$. (We identify $\{0, \ldots, l-1\}$ with $\mathbb{Z}_l = \mathbb{Z}/(l\mathbb{Z})$ in order to have $(l-1)+1 = 0$.)

LEMMA A.5. *Let X be a disjoint union of measure spaces $X_0, \ldots, X_{l-1}$ and let T be an invertible measure preserving transformation of X such that $T(X_i) = X_{i+1}$, $i \in \mathbb{Z}_l$. Then $X_0, \ldots, X_{l-1} \in \mathcal{Z}_1(X, T)$.*

PROOF. We may assume that T is ergodic; otherwise we pass to the ergodic components of X with respect to T. Then $X^{[1]}_T = X^2$ and $T^{[1]} = T \times T$. The "diagonal" $W = X^2_0 \cup \cdots \cup X^2_{l-1} \subseteq X^{[1]}_T$ is $T^{[1]}$-invariant and therefore W is $\mathcal{Z}_1(X, T) \otimes \mathcal{Z}_1(X, T)$-measurable. By Fubini's theorem the "fibers" $X_0, \ldots, X_{l-1}$ of W are $\mathcal{Z}_1(X, T)$-measurable. □

LEMMA A.6. *Let Y be a disjoint union of measure spaces $Y_0, \ldots, Y_{l-1}$ and let S be an invertible measure preserving transformation of Y such that $S(Y_i) = Y_{i+1}$, $i \in \mathbb{Z}_l$. Then $Y \times_S Y$ is partitioned to $\bigcup_{i,j \in \mathbb{Z}_l} Y_{i,j}$ where $Y_{i,i} = Y_i \times_{S^l} Y_i$ for all $i \in \mathbb{Z}_l$, and for all $i, j, s, t \in \mathbb{Z}_l$, $(S^s \times S^t)|_{Y_{i,j}}$ is an isomorphism between $Y_{i,j}$ and $Y_{i+s,j+t}$. In particular, $(S \times S)(Y_{i,j}) = Y_{i+1,j+1}$ for all i, j, thus the subsets $V_i = \bigcup_{j \in \mathbb{Z}_l} Y_{j,j+i}$, $i \in \mathbb{Z}_l$, are $S \times S$-invariant and partition $Y \times_S Y$, and $\mathrm{Id}_{Y_0} \times S^i$ is an isomorphism between V_0 and V_i.*

PROOF. We first determine $I(Y,S)$. Let A be a measurable S-invariant subset of Y. Let $A_i = A \cap Y_i$, $i \in \mathbb{Z}_l$. Then A_0 is S^l-invariant, and $A_i = S^i(A_0)$ for $i \in \mathbb{Z}_l$. So, the mapping $A \mapsto A \cap Y_0$ is an isomorphism between $\mathcal{I}(Y,S)$ and $\mathcal{I}(Y_0,S^l)$, which induces an isomorphism between $I(Y,S)$ and $I(Y_0,S^l)$.

Let $Y_0 = \bigcup_{\alpha\in I} Y_{0,\alpha}$ be the decomposition of Y_0 with respect to $I = I(Y_0,S^l)$. For every $\alpha \in I$ and $i \in \mathbb{Z}_l \setminus \{0\}$ define $Y_{i,\alpha} = S^i(Y_{0,\alpha})$ and $Y_\alpha = \bigcup_{i\in\mathbb{Z}_l} Y_{i,\alpha}$. Then $Y = \bigcup_{\alpha\in I} Y_\alpha$ is the decomposition of Y with respect to I. We have

$$Y_S^{[1]} = \bigcup_{\alpha\in I} Y_\alpha \times_S Y_\alpha = \bigcup_{\alpha\in I}\bigcup_{i,j\in\mathbb{Z}_l} Y_{i,\alpha}\times Y_{j,\alpha} = \bigcup_{i,j\in\mathbb{Z}_l}\bigcup_{\alpha\in I} Y_{i,\alpha}\times Y_{j,\alpha} = \bigcup_{i,j\in\mathbb{Z}_l} Y_{i,j},$$

where $Y_{i,j} = \bigcup_{\alpha\in I} Y_{i,\alpha}\times Y_{j,\alpha}$. In particular, $Y_{i,i} = \bigcup_{\alpha\in I} Y_{i,\alpha}\times Y_{i,\alpha} = Y_i \times_{S^l} Y_i$ for all $i \in \mathbb{Z}_l$. □

LEMMA A.7. *Let X be a disjoint union of measure spaces $X_0,\dots,X_{l-1}$ and let T be an invertible measure preserving transformation of X such that $T(X_i) = X_{i+1}$, $i \in \mathbb{Z}_l$. Then for any $k \geqslant 0$, $X_T^{[k]}$ can be partitioned, $X_T^{[k]} = \bigcup_{j=1}^{l^k} W_j$, into $T^{[k]}$-invariant measurable subsets $W_1,\dots,W_{l^k}$, such that $W_1 = \bigcup_{i\in\mathbb{Z}_l}(X_i)_{T^l}^{[k]}$ with $T^{[k]}((X_i)_{T^l}^{[k]}) = (X_{i+1})_{T^l}^{[k]}$ for each i, and for each $j = 2,\dots,l^k$ there exists an isomorphism $\tau_j : W_1 \to W_j$, which in each coordinate is given by a power of T (that is, if $\pi_n : X^{[k]} \to X$, $n = 1,\dots,2^k$, are the projection mappings, for each n there exists $m \in \mathbb{Z}$ such that $\pi_n \circ \tau_j = T^m \circ \pi_n|_{W_1}$).*

PROOF. We use induction on k; for $k = 0$ the statement is trivial. Assume that it holds for some $k \geqslant 0$. Then by Lemma A.1, $X_T^{[k+1]} = \bigcup_{j=1}^{l^k} W_j \times_{T^{[k]}} W_j$. The isomorphisms τ_j between W_1 and W_j, commuting with $T^{[k]}$, induce isomorphisms $\tau_j \times \tau_j$ between $W_1 \times_{T^{[k]}} W_1$ and $W_j \times_{T^{[k]}} W_j$, $j = 1,\dots,l^k$, and $\tau_j \times \tau_j$ acts on coordinates as powers of T if τ_j does. Thus, we may focus on $W_1 \times_{T^{[k]}} W_1$ only.

By Lemma A.6 applied to $W_1 = \bigcup_{i\in\mathbb{Z}_l}(X_i)_{T^l}^{[k]}$ and $T^{[k]}|_{W_1}$, $W_1 \times_{T^{[k]}} W_1$ is partitioned into $T^{[k]} \times T^{[k]} = T^{[k+1]}$-invariant subsets $V_0,\dots,V_{l-1}$ such that

$$V_0 = \bigcup_{i\in\mathbb{Z}_l} (X_i)_{T^l}^{[k]} \times_{(T^{[k]})^l} (X_i)_{T^l}^{[k]} = \bigcup_{i\in\mathbb{Z}_l} (X_i)_{T^l}^{[k+1]}$$

and $V_1,\dots,V_{l-1}$ are isomorphic to V_0 by isomorphisms whose projections on the factors $(X_i)_{T^l}^{[k]}$ coincide with some powers of $T^{[k]}$. □

END OF THE PROOF OF THEOREM A.2. Assume that T is ergodic on X, l is a prime integer and T^l is not ergodic on X. Let $k \geqslant 1$. Ignoring a subset of measure 0 in X, partition X to measurable subsets $X_0,\dots,X_{l-1}$ such that, for each i, $T(X_i) = X_{i+1}$. Let $k \geqslant 1$ and let $W_1,\dots,W_{l^k}$ be as in Lemma A.7. Since $X_0,\dots,X_{l-1}$ are T^l-invariant, by Lemma A.1 we have $\mathcal{I}(X^{[k]},(T^l)^{[k]}) = \prod_{i\in\mathbb{Z}_l}\mathcal{I}(X_i^{[k]},(T^l)^{[k]})$ and $\mathcal{Z}_k(X,T^l) = \prod_{i\in\mathbb{Z}_l}\mathcal{Z}_k(X_i,T^l)$. Any $T^{[k]}$-invariant measurable subset A of $W_1 = \bigcup_{i\in\mathbb{Z}_l}(X_i)_{T^l}^{[k]}$ has form $A = \bigcup_{i\in\mathbb{Z}_l} A_i$ where $A_i \in \mathcal{I}(X_i,(T^l)^{[k]})$ and $T^{[k]}(A_i) = A_{i+1}$, $i \in \mathbb{Z}_l$. Thus, $\mathcal{I}(W_1,T^{[k]}) \subseteq \mathcal{I}(X^{[k]},(T^l)^{[k]}) \subseteq$

$\mathcal{Z}_k(X,T^l)^{\otimes 2^k}$. Since $\mathcal{Z}_k(X,T^l)$ is T-invariant and $W_n=\tau_n(W_1)$ where τ_n is an isomorphism acting on each coordinate as a power of T, $\mathcal{I}(W_n,T^{[k]})\subseteq\mathcal{Z}_k(X,T^l)^{\otimes 2^k}$ for any n. Hence, $\mathcal{Z}_k(X,T)\subseteq\mathcal{Z}_k(X,T^l)$.

We will now show that for any $i\in\mathbb{Z}_l$ and any $B\in\mathcal{I}(X_i^{[k]},(T^l)^{[k]})$ one has $B\in\mathcal{Z}_k(X,T)^{\otimes 2^k}$; this will imply that $\mathcal{Z}_k(X,T^l)\subseteq\mathcal{Z}_k(X,T)$. Put $A_j=(T^{[k]})^{j-i}(B)$, $j\in\mathbb{Z}_l$, and $A=\bigcup_{j\in\mathbb{Z}_l}A_j$. Then $A\in\mathcal{I}(W_1,T^{[k]})\subseteq\mathcal{Z}_k(X,T)^{\otimes 2^k}$. By Lemma A.5, $X_i\in\mathcal{Z}_1(X,T)\subseteq\mathcal{Z}_k(X,T)$, thus $(X_i)^{[k]}_{T^l}\in\mathcal{Z}_k(X,T)^{\otimes 2^k}$, and therefore $B=A_i=A\cap(X_i)^{[k]}_{T^l}\in\mathcal{Z}_k(X,T)^{\otimes 2^k}$. □

A.4. *Characteristic factors for multiple averages*

We now pass to our second result:

THEOREM A.8. *For any $k\geqslant 2$, any $d\in\mathbb{N}$, any linear functions $\varphi_1,\dots,\varphi_k:\mathbb{Z}^d\to\mathbb{Z}$ and any Følner sequence $\{\Phi_N\}_{N=1}^{\infty}$ in $\mathbb{Z}^d$, $Z_{k-1}(X,T)$ is a characteristic factor for the averages $\frac{1}{|\Phi_N|}\sum_{u\in\Phi_N}T^{\varphi_1(u)}f_1\cdot\dots\cdot T^{\varphi_k(u)}f_k$ in $L^1(X)$, that is,*

$$\lim_{N\to\infty}\Big\|\frac{1}{|\Phi_N|}\sum_{u\in\Phi_N}\big(T^{\varphi_1(u)}f_1\cdot\dots\cdot T^{\varphi_k(u)}f_k - T^{\varphi_1(u)}E\big(f_1|Z_{k-1}(X,T)\big)\cdot\dots\cdot T^{\varphi_k(u)}E\big(f_k|Z_{k-1}(X,T)\big)\big)\Big\|_{L^1(X)}=0 \tag{A.5}$$

for any $f_1,\dots,f_k\in L^\infty(X)$.

In order to prove Theorem A.8 we will first show that $Z_{k-1}(X,T)$ is a characteristic factor for averages of a very special form. Let us bring more facts from [88]. Starting from this moment, we will only be considering real-valued functions on X. Given $f_0,f_1\in L^\infty(X)$, by the ergodic theorem we have

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\int_X f_0\cdot T^n f_1=\int_{I(X,T)}E\big(f_0|I(X,T)\big)\cdot E\big(f_1|I(X,T)\big)=\int_{X^{[1]}_T}f_0\otimes f_1.$$

Applying this twice we get, for $f_{0,0},f_{0,1},f_{1,0},f_{1,1}\in L^\infty(X)$,

$$\lim_{N_2\to\infty}\frac{1}{N_2}\sum_{n_2=1}^{N_2}\lim_{N_1\to\infty}\frac{1}{N_1}\sum_{n_1=1}^{N_2}\int_X f_{0,0}\cdot T^{n_1}f_{1,0}\cdot T^{n_2}f_{0,1}\cdot T^{n_1+n_2}f_{1,1}$$

$$= \lim_{N_2\to\infty} \frac{1}{N_2} \sum_{n_2=1}^{N_2} \lim_{N_1\to\infty} \frac{1}{N_1} \sum_{n_1=1}^{N_2} \int_X \bigl(f_{0,0} \cdot T^{n_2} f_{0,1}\bigr) \cdot T^{n_1} \bigl(f_{1,0} \cdot T^{n_2} f_{1,1}\bigr)$$

$$= \lim_{N_2\to\infty} \frac{1}{N_2} \sum_{n_2=1}^{N_2} \int_{X^{[1]}} (f_{0,0} \otimes f_{1,0}) \cdot T^{n_2} (f_{0,1} \otimes f_{1,1})$$

$$= \int_{X^{[2]}} (f_{0,0} \otimes f_{1,0}) \otimes (f_{0,1} \otimes f_{1,1}).$$

By induction, for any k and any collection $f_{\varepsilon_1,\dots,\varepsilon_k} \in L^\infty(X)$, $\varepsilon_1, \dots, \varepsilon_k \in \{0,1\}$,

$$\lim_{N_k\to\infty} \frac{1}{N_k} \sum_{n_k=1}^{N_k} \cdots \lim_{N_1\to\infty} \frac{1}{N_1} \sum_{n_1=1}^{N_1} \int_X \prod_{\varepsilon_1,\dots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_1 n_1 + \cdots + \varepsilon_k n_k} f_{\varepsilon_1,\dots,\varepsilon_k}$$

$$= \int_{X^{[k]}} \bigotimes_{\varepsilon_1,\dots,\varepsilon_k\in\{0,1\}} f_{\varepsilon_1,\dots,\varepsilon_k}$$

(where the tensor product is taken in a certain order, which we do not specify here).

For $k \in \mathbb{N}$ and $f \in L^\infty(X)$ the seminorm $|||f|||_{T,k}$ associated with T is defined by $|||f|||_{T,k} = (\int_{X_T^{[k]}} f^{\otimes 2^k})^{1/2^k}$. Equivalently,

$$|||f|||_{T,k}^{2^k} = \lim_{N_k\to\infty} \frac{1}{N_k} \sum_{n_k=1}^{N_k} \cdots \lim_{N_1\to\infty} \frac{1}{N_1} \sum_{n_1=1}^{N_1} \int_X \prod_{\varepsilon_1,\dots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_1 n_1 + \cdots + \varepsilon_k n_k} f.$$

It is proved in [88] that for any $f_1, \dots, f_{2^k} \in L^\infty(X)$ one has

$$\left| \int_{X_T^{[k]}} \bigotimes_{j=1}^{2^k} f_j \right| \leqslant \prod_{j=1}^{2^k} |||f_j|||_{T,k}.$$

For any $k \in \mathbb{N}$ and $f \in L^\infty(X)$ we have

$$|||f|||_{T,k}^{2^k} = \int_{X_T^{[k]}} f^{\otimes 2^k} = \int_{I(X_T^{[k-1]}, T^{[k-1]})} E\bigl(f^{\otimes 2^{k-1}} | I\bigl(X_T^{[k-1]}, T^{[k-1]}\bigr)\bigr)^2.$$

Since $\mathcal{I}(X_T^{[k-1]}, T^{[k-1]}) \subseteq \mathcal{Z}_{k-1}(X,T)^{\otimes 2^{k-1}}$, one has $|||f|||_{T,k} = 0$ whenever $E(f|Z_{k-1}(X,T)) = 0$.

PROPOSITION A.9. *For any $k \geqslant 2$, nonzero integers $l_1, \dots, l_k$ and a collection $f_{\varepsilon_1,\dots,\varepsilon_k} \in L^\infty(X)$, $\varepsilon_1, \dots, \varepsilon_k \in \{0,1\}$, if $E(f_{\varepsilon_1,\dots,\varepsilon_k} | Z_{k-1}(X,T)) = 0$ for some $\varepsilon_1, \dots, \varepsilon_k$ then*

$$\lim_{N_k\to\infty} \frac{1}{N_k} \sum_{n_k=1}^{N_k} \cdots \lim_{N_1\to\infty} \frac{1}{N_1} \sum_{n_1=1}^{N_1} \int_X \prod_{\varepsilon_1,\dots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_1 l_1 n_1 + \cdots + \varepsilon_k l_k n_k} f_{\varepsilon_1,\dots,\varepsilon_k} = 0.$$

PROOF. Let l be a common multiple of $l_1, \dots, l_k$. Since, by Theorem A.2, $Z_{k-1}(X, T^l) = Z_{k-1}(X, T)$, $E(f_{\varepsilon_1,\dots,\varepsilon_k}|Z_{k-1}(X,T)) = 0$ implies $\|\!|\!| f_{\varepsilon_1,\dots,\varepsilon_k} |\!|\!|_{T^l,k} = 0$.

Let $r_i = l/l_i$, $i = 1, \dots, k$. We have

$$\begin{aligned}
&\lim_{N_k\to\infty} \frac{1}{N_k} \sum_{n_k=1}^{N_k} \cdots \lim_{N_1\to\infty} \frac{1}{N_1} \sum_{n_1=1}^{N_1} \int_X \prod_{\varepsilon_1,\dots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_1 l_1 n_1+\cdots+\varepsilon_k l_k n_k} f_{\varepsilon_1,\dots,\varepsilon_k} \\
&\quad = \frac{1}{r_1\cdots r_k} \sum_{m_k=0}^{r_k-1} \cdots \sum_{m_1=0}^{r_1-1} \lim_{N_k\to\infty} \frac{1}{N_k} \sum_{n_k=1}^{N_k} \cdots \lim_{N_1\to\infty} \frac{1}{N_1} \sum_{n_1=1}^{N_1} \\
&\qquad \int_X \prod_{\varepsilon_1,\dots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_1 l n_1+\cdots+\varepsilon_k l n_k} \bigl(T^{\varepsilon_1 l_1 m_1+\cdots+\varepsilon_k l_k m_k} f_{\varepsilon_1,\dots,\varepsilon_k}\bigr) \\
&\quad = \frac{1}{r_1\cdots r_k} \sum_{m_k=0}^{r_k-1} \cdots \sum_{m_1=0}^{r_1-1} \int_{X^{[k]}_{T^l}} \bigotimes_{\varepsilon_1,\dots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_1 l_1 m_1+\cdots+\varepsilon_k l_k m_k} f_{\varepsilon_1,\dots,\varepsilon_k}.
\end{aligned}$$

And for any $m_{\varepsilon_1,\dots,\varepsilon_k} \in \mathbb{Z}$, $\varepsilon_1, \dots, \varepsilon_k \in \{0,1\}$,

$$\begin{aligned}
&\left| \int_{X^{[k]}_{T^l}} \bigotimes_{\varepsilon_1,\dots,\varepsilon_k\in\{0,1\}} T^{m_{\varepsilon_1,\dots,\varepsilon_k}} f_{\varepsilon_1,\dots,\varepsilon_k} \right| \\
&\quad \leqslant \prod_{\varepsilon_1,\dots,\varepsilon_k\in\{0,1\}} \bigl\|\!\bigl| T^{m_{\varepsilon_1,\dots,\varepsilon_k}} f_{\varepsilon_1,\dots,\varepsilon_k} \bigr|\!\bigr\|_{T^l,k} \\
&\quad = \prod_{\varepsilon_1,\dots,\varepsilon_k\in\{0,1\}} \bigl\|\!\bigl| f_{\varepsilon_1,\dots,\varepsilon_k} \bigr|\!\bigr\|_{T^l,k} = 0.
\end{aligned}$$

□

Let $\varphi : \mathbb{Z}^d \to \mathbb{Z}$ be a nonzero linear function, that is, a function of the form $\varphi(n_1, \dots, n_d) = a_1 n_1 + \cdots + a_d n_d$ with $a_1, \dots, a_d \in \mathbb{Z}$ not all zero. Then for any measure preserving system (Y, S), any $f \in L^1(Y)$ and any Følner sequence $\{\Phi_N\}_{N=1}^\infty$ in $\mathbb{Z}^d$ one has $\lim_{N\to\infty} \frac{1}{|\Phi_N|} \sum_{u\in\Phi_N} S^{\varphi(u)} f = E(f|I(Y, S^l)) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^N S^{ln} f$, where $l = \gcd(a_1, \dots, a_d)$. Applying this fact k times, we come to the following generalization of Proposition A.9:

PROPOSITION A.10. *For any $k \geqslant 2$, positive integers $d_i \in \mathbb{N}$, nonzero linear functions $\varphi_i : \mathbb{Z}^{d_i} \to \mathbb{Z}$, Følner sequences $\{\Phi_{i,N}\}_{N=1}^\infty$ in $\mathbb{Z}^{d_i}$, $i = 1, \dots, k$, and a collection of functions $f_{\varepsilon_1,\dots,\varepsilon_k} \in L^\infty(X)$, $\varepsilon_1, \dots, \varepsilon_k \in \{0,1\}$, if $E(f_{\varepsilon_1,\dots,\varepsilon_k}|Z_{k-1}(X,T)) = 0$ for some $\varepsilon_1, \dots, \varepsilon_k$ then*

$$\begin{aligned}
&\lim_{N_k\to\infty} \frac{1}{|\Phi_{k,N_k}|} \sum_{u_k\in\Phi_{k,N_k}} \cdots \lim_{N_1\to\infty} \frac{1}{|\Phi_{1,N_1}|} \sum_{u_1\in\Phi_{1,N_1}} \\
&\qquad \int_X \prod_{\varepsilon_1,\dots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_1\varphi_1(u_1)+\cdots+\varepsilon_k\varphi_k(u_k)} f_{\varepsilon_1,\dots,\varepsilon_k} = 0.
\end{aligned}$$

The proof of Theorem A.8 will be based on the following lemma:

LEMMA A.11. *For any linear functions* $\varphi_1, \dots, \varphi_k : \mathbb{Z}^d \to \mathbb{Z}$ *and any* $f_1, \dots, f_k \in L^\infty(X)$,

$$\begin{aligned}
&\limsup_{N\to\infty} \left\| \frac{1}{|\Phi_N|} \sum_{u\in\Phi_N} T^{\varphi_1(u)} f_1 \cdot \dots \cdot T^{\varphi_k(u)} f_k \right\|_{L^2(X)} \\
&\leqslant \Bigg(\lim_{N_1\to\infty} \frac{1}{|\Phi_{N_1}|^2} \sum_{(v_1,w_1)\in\Phi_{N_1}^2} \lim_{N_k\to\infty} \frac{1}{|\Phi_{N_k}|^2} \\
&\qquad \sum_{(v_k,w_k)\in\Phi_{N_k}^2} \cdots \lim_{N_2\to\infty} \frac{1}{|\Phi_{N_2}|^2} \sum_{(v_2,w_2)\in\Phi_{N_2}^2} \int_X \\
&\qquad \prod_{\varepsilon_1,\varepsilon_2,\dots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_1\varphi_1(v_1-w_1)+\varepsilon_2(\varphi_1-\varphi_2)(v_2-w_2)+\dots+\varepsilon_k(\varphi_1-\varphi_k)(v_k-w_k)} f_1 \Bigg)^{1/2^k} \\
&\quad \cdot \prod_{i=2}^{k} \|f_i\|_{L^\infty(X)}.
\end{aligned}$$

PROOF. Let $\{\Phi_N\}_{N=1}^\infty$ be a Følner sequence in $\mathbb{Z}^d$. We will use the van der Corput lemma in the following form: if $\{f_u\}_{u\in\mathbb{Z}^d}$ is a bounded family of elements of a Hilbert space, then

$$\begin{aligned}
&\limsup_{N\to\infty} \left\| \frac{1}{|\Phi_N|} \sum_{u\in\Phi_N} f_u \right\|^2 \\
&\leqslant \limsup_{N_1\to\infty} \frac{1}{|\Phi_{N_1}|^2} \sum_{v,w\in\Phi_{N_1}} \limsup_{N\to\infty} \frac{1}{|\Phi_N|} \sum_{u\in\Phi_N} \langle f_u, f_{u+v-w} \rangle.
\end{aligned}$$

We may assume that $|f_2|, \dots, |f_k| \leqslant 1$. By the van der Corput lemma we have:

$$\begin{aligned}
&\limsup_{N\to\infty} \left\| \frac{1}{|\Phi_N|} \sum_{u\in\Phi_N} T^{\varphi_1(u)} f_1 \cdot \dots \cdot T^{\varphi_k(u)} f_k \right\|_{L^2(X)}^2 \\
&\leqslant \limsup_{N_1\to\infty} \frac{1}{|\Phi_{N_1}|^2} \sum_{v,w\in\Phi_{N_1}} \limsup_{N\to\infty} \frac{1}{|\Phi_N|} \sum_{u\in\Phi_N} \int_X T^{\varphi_1(u)} f_1 \cdot \dots \cdot T^{\varphi_k(u)} f_k \\
&\quad \cdot T^{\varphi_1(u+v-w)} f_1 \cdot \dots \cdot T^{\varphi_k(u+v-w)} f_k \\
&= \limsup_{N_1\to\infty} \frac{1}{|\Phi_{N_1}|^2} \sum_{v,w\in\Phi_{N_1}} \limsup_{N\to\infty} \frac{1}{|\Phi_N|} \sum_{u\in\Phi_N} \int_X T^{\varphi_1(u)} \big(f_1 \cdot T^{\varphi_1(v-w)} f_1 \big) \cdot \dots \\
&\quad \cdot T^{\varphi_k(u)} \big(f_k \cdot T^{\varphi_k(v-w)} f_k \big)
\end{aligned}$$

$$
\begin{aligned}
&= \limsup_{N_1\to\infty} \frac{1}{|\Phi_{N_1}|^2} \sum_{v,w\in\Phi_{N_1}} \limsup_{N\to\infty} \frac{1}{|\Phi_N|} \\
&\quad \sum_{u\in\Phi_N} \int_X T^{\varphi_1(u)-\varphi_k(u)}(f_1\cdot T^{\varphi_1(v-w)}f_1)\cdot\ldots \\
&\quad \cdot T^{\varphi_{k-1}(u)-\varphi_k(u)}\big(f_{k-1}\cdot T^{\varphi_{k-1}(v-w)}f_{k-1}\big)\cdot\big(f_k\cdot T^{\varphi_k(v-w)}f_k\big) \\
&= \limsup_{N_1\to\infty} \frac{1}{|\Phi_{N_1}|^2} \\
&\quad \sum_{v,w\in\Phi_{N_1}} \limsup_{N\to\infty} \int_X \bigg(\frac{1}{|\Phi_N|}\sum_{u\in\Phi_N} T^{(\varphi_1-\varphi_k)(u)}\big(f_1\cdot T^{\varphi_1(v-w)}f_1\big)\cdot\ldots \\
&\quad \cdot T^{(\varphi_{k-1}-\varphi_k)(u)}\big(f_{k-1}\cdot T^{\varphi_{k-1}(v-w)}f_{k-1}\big)\bigg)\cdot\big(f_k\cdot T^{\varphi_k(v-w)}f_k\big) \\
&\leqslant \limsup_{N_1\to\infty} \frac{1}{|\Phi_{N_1}|^2} \\
&\quad \sum_{(v,w)\in\Phi_{N_1}^2} \limsup_{N\to\infty} \bigg\|\frac{1}{|\Phi_N|}\sum_{u\in\Phi_N} T^{(\varphi_1-\varphi_k)(u)}\big(f_1\cdot T^{\varphi_1(v-w)}f_1\big)\cdot\ldots \\
&\quad \cdot T^{(\varphi_{k-1}-\varphi_k)(u)}\big(f_{k-1}\cdot T^{\varphi_{k-1}(v-w)}f_{k-1}\big)\bigg\|_{L^2(X)}.
\end{aligned}
$$

By the induction hypothesis, applied to the linear functions $\varphi_i-\varphi_k:\mathbb{Z}^d\to\mathbb{Z}$ and to the functions $f_i\cdot T^{\varphi_i(v-w)}f_i\in L^\infty(X)$, $i=1,\ldots,k-1$, for any $(v,w)\in\mathbb{Z}^{2d}$ we have

$$
\begin{aligned}
&\limsup_{N\to\infty} \bigg\|\frac{1}{|\Phi_N|}\sum_{u\in\Phi_N} T^{(\varphi_1-\varphi_k)(u)}\big(f_1\cdot T^{\varphi_1(v-w)}f_1\big)\cdot\ldots \\
&\quad \cdot T^{(\varphi_{k-1}-\varphi_k)(u)}\big(f_{k-1}\cdot T^{\varphi_{k-1}(v-w)}f_{k-1}\big)\bigg\|_{L^2(X)} \\
&\leqslant \bigg(\lim_{N_k\to\infty}\frac{1}{|\Phi_{N_k}|^2}\sum_{(v_k,w_k)\in\Phi_{N_k}^2}\cdots\lim_{N_2\to\infty}\frac{1}{|\Phi_{N_2}|^2}\sum_{(v_2,w_2)\in\Phi_{N_2}^2}\int_X \\
&\quad \prod_{\varepsilon_2,\ldots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_2(\varphi_1-\varphi_2)(v_2-w_2)+\cdots+\varepsilon_k(\varphi_1-\varphi_k)(v_k-w_k)}\big(f_1\cdot T^{\varphi_1(v-w)}f_1\big)\bigg)^{\frac{1}{2^{k-1}}} \\
&= \bigg(\lim_{N_k\to\infty}\frac{1}{|\Phi_{N_k}|^2}\sum_{(v_k,w_k)\in\Phi_{N_k}^2}\cdots\lim_{N_2\to\infty}\frac{1}{|\Phi_{N_2}|^2}\sum_{(v_2,w_2)\in\Phi_{N_2}^2}\int_X \\
&\quad \prod_{\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_1\varphi_1(v-w)+\varepsilon_2(\varphi_1-\varphi_2)(v_2-w_2)+\cdots+\varepsilon_k(\varphi_1-\varphi_k)(v_k-w_k)}f_1\bigg)^{\frac{1}{2^{k-1}}}.
\end{aligned}
$$

Thus,

$$\begin{aligned}
&\limsup_{N\to\infty}\left\|\frac{1}{|\Phi_N|}\sum_{u\in\Phi_N} T^{\varphi_1(u)}f_1\cdot\ldots\cdot T^{\varphi_k(u)}f_k\right\|_{L^2(X)}\\
&\quad\leqslant\Bigg(\limsup_{N_1\to\infty}\frac{1}{|\Phi_{N_1}|^2}\sum_{(v,w)\in\Phi_{N_1}^2}\Bigg(\lim_{N_k\to\infty}\frac{1}{|\Phi_{N_k}|^2}\sum_{(v_k,w_k)\in\Phi_{N_k}^2}\cdots\\
&\qquad\lim_{N_2\to\infty}\frac{1}{|\Phi_{N_2}|^2}\sum_{(v_2,w_2)\in\Phi_{N_2}^2}\int_X\\
&\qquad\prod_{\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_1\varphi_1(v-w)+\varepsilon_2(\varphi_1-\varphi_2)(v_2-w_2)+\cdots+\varepsilon_k(\varphi_1-\varphi_k)(v_k-w_k)}f_1\Bigg)^{\frac{1}{2^{k-1}}}\Bigg)^{\frac{1}{2}}\\
&\quad\leqslant\Bigg(\lim_{N_1\to\infty}\frac{1}{|\Phi_{N_1}|^2}\sum_{(v,w)\in\Phi_{N_1}^2}\lim_{N_k\to\infty}\frac{1}{|\Phi_{N_k}|^2}\sum_{(v_k,w_k)\in\Phi_{N_k}^2}\cdots\\
&\qquad\lim_{N_2\to\infty}\frac{1}{|\Phi_{N_2}|^2}\sum_{(v_2,w_2)\in\Phi_{N_2}^2}\int_X\\
&\qquad\prod_{\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_k\in\{0,1\}} T^{\varepsilon_1\varphi_1(v-w)+\varepsilon_2(\varphi_1-\varphi_2)(v_2-w_2)+\cdots+\varepsilon_k(\varphi_1-\varphi_k)(v_k-w_k)}f_1\Bigg)^{\frac{1}{2^k}}.\ \square
\end{aligned}$$

PROOF OF THEOREM A.8. Because of the multilinearity of (A.5), it suffices to show that $\lim_{N\to\infty}\frac{1}{|\Phi_N|}\sum_{u\in\Phi_N} T^{\varphi_1(u)}f_1\cdot\ldots\cdot T^{\varphi_k(u)}f_k=0$ in $L^1(X)$ whenever $E(f_1|Z_{k-1}(X,T))=0$. We may assume that the functions $\varphi_1,\ldots,\varphi_k$ are all nonzero and distinct. Then, combining Lemma A.11 and Proposition A.10, applied to the nonzero linear functions $\varphi_1(v-w)$, $(\varphi_1-\varphi_2)(v-w),\ldots,(\varphi_1-\varphi_k)(v-w)$ on $\mathbb{Z}^{2d}$ and the Følner sequence $\{\Phi_N^2\}_{N=1}^{\infty}$ in $\mathbb{Z}^{2d}$, we get $\lim_{N\to\infty}\frac{1}{|\Phi_N|}\sum_{u\in\Phi_N} T^{\varphi_1(u)}f_1\cdot\ldots\cdot T^{\varphi_k(u)}f_k=0$ in $L^2(X)$ and so, in $L^1(X)$. $\square$

Acknowledgement

I thank V. Bergelson and E. Lesigne for valuable comments on the draft of this note.

Appendix B. Ergodic averages along the squares, by A. Quas and M. Wierdl

B.1. *Enunciation of the result*

In this note we want to present a proof of the almost everywhere convergence of the ergodic averages along the sequence of squares.

THEOREM B.1. *Let τ be a measurable, measure preserving transformation of the σ-finite measure space (X, Σ, μ).*

Then, for $f \in L^2$, the averages

$$S_t f(x) = \frac{1}{t} \sum_{n \leqslant t} f\left(\tau^{n^2} x\right)$$

converge for almost every $x \in X$.

The theorem is due to J. Bourgain. To keep our presentation as continuous as possible, we present historical remarks, and cite references in the last section, Section B.6.

B.2. *Subsequence lemma*

The main idea of the proof is to analyze the Fourier transform $\widehat{S}_t(\alpha) = 1/t \sum_{n \leqslant t} e^{2\pi i n^2 \alpha}$ of the averages. This analysis permits us to replace the averages S_t by other operators that are easier to handle. The replaceability of the sequence (S_t) by another sequence (A_t) means that we have an inequality of the form

$$\int \sum_t |S_t f - A_t f|^2 < c \int |f|^2. \tag{B.1}$$

Now, if somehow we prove that the sequence $(A_t f(x))$ converges for a.e. x, then the above inequality implies, since its left-hand side is finite for $f \in L^2$, that the sequence $(S_t f(x))$ converges a.e. as well.

Well, we will not be able to prove an inequality of the type (B.1) exactly. In the real inequality, we will be able to have an inequality where the t runs through a lacunary sequence. But this is quite all right since it is enough to prove the a.e. convergence of the $(S_t f)$ along a lacunary sequence:

LEMMA B.2. *For $\sigma > 1$ denote*

$$I = I_\sigma = \{t \mid t = \sigma^n \text{ for some positive integer } n\}.$$

Suppose that for each fixed $\sigma > 1$, the sequence $(S_t f)_{t \in I}$ converges a.e.

Then the full (S_t) sequence converges a.e.

PROOF. We can assume that the function f is nonnegative. For a given t, choose k so that $\sigma^k \leqslant t < \sigma^{k+1}$. We can then estimate as

$$S_t f(x) \leqslant \frac{1}{\sigma^k} \sum_{n \leqslant \sigma^{k+1}} f\left(\tau^{n^2} x\right) = \sigma \cdot S_{\sigma^{k+1}} f(x),$$

and similarly, we have $\sigma^{-1} \cdot S_{\sigma^k} f(x) \leqslant S_t f(x)$. This means that

$$\sigma^{-1} \cdot \lim_k S_{\sigma^k} f(x) \leqslant \liminf_t S_t f(x) \leqslant \limsup_t S_t f(x) \leqslant \sigma \cdot \lim_k S_{\sigma^k} f(x).$$

Choosing now $\sigma_p = 2^{2^{-p}}$, we get that $\lim_k S_{\sigma_p^k} f(x)$ is independent of p for a.e. x, and, by the above estimates, it is equal to $\lim_t S_t f(x)$. □

For the rest of the proof, we fix $\sigma > 1$, and unless we say otherwise, we always assume that $t \in I = I_\sigma$.

DEFINITION B.3. If two sequences (A_t) and (B_t) of $L^2 \to L^2$ operators satisfy

$$\int \sum_{t \in I_\sigma} |A_t f - B_t f|^2 < c \int |f|^2; \quad f \in L^2,$$

then we say that (A_t) and (B_t) are *equivalent*.

B.3. *Oscillation and an instructive example*

One standard way of proving a.e. convergence for the usual ergodic averages $1/t \sum_{n \leqslant t} f(\tau^n x)$ is to first prove a maximal inequality, and then note that there is a natural dense class for which a.e. convergence holds.

Unfortunately, the second part of this scheme does not work for the averages along the squares, since there is no known class of functions for which it would be easy to prove a.e. convergence of the averages.

Instead, for the squares, we will prove a so called *oscillation inequality*: for any $t(1) < t(2) < \cdots$ with $t(k) \in I$, there is a constant c so that we have

$$\int \sum_k \sup_{t(k) < t < t(k+1)} |S_t f - S_{t(k+1)} f|^2 \leqslant c \int f^2. \tag{B.2}$$

We leave it to the reader to verify why an oscillation inequality implies a.e. convergence of the sequence $(S_t f)$. We also leave it to the reader to verify that if two operator sequences (A_t) and (B_t) are equivalent and (A_t) satisfies an oscillation inequality, then so does (B_t).

An important remark is that by the so called *transference principle* of Calderón, it is enough to prove the inequality in (B.2) on the integers $\mathbb{Z}$ which we consider equipped with the counting measure and the right shift. In this case, we have $S_t f(x) = 1/t \sum_{n \leqslant t} f(x + n^2)$.

To see how Fourier analysis can help in proving an oscillation inequality, let us look at a simpler example first: the case of the usual ergodic averages $U_t f(x) = 1/t \sum_{n \leqslant t} f(x + n)$ (by the transference principle, we only need to prove the oscillation inequality on the integers).

Let us assume that we already know the maximal inequality

$$\int_{\mathbb{Z}} \sup_{t\in I_\sigma} |U_t f|^2 \leqslant c \cdot \int_{\mathbb{Z}} |f|^2.$$

For the Fourier transform $\widehat{U}_t(\alpha) = 1/t \sum_{n\leqslant t} e^{2\pi i n\alpha}$, $\alpha \in (-1/2, 1/2)$, we easily obtain the estimates

$$\left|\widehat{U}_t(\alpha) - 1\right| \leqslant c \cdot t \cdot |\alpha|; \tag{B.3}$$

$$\left|\widehat{U}_t(\alpha)\right| \leqslant \frac{c}{t \cdot |\alpha|}. \tag{B.4}$$

The first estimate is effective (nontrivial) when $|\alpha| < 1/t$ and it says that $\widehat{U}_t(\alpha)$ is close to 1. The second estimate is effective when $|\alpha| > 1/t$, and it says that then $|\widehat{U}_t(\alpha)|$ is small. In other words, the estimates in (B.3) and (B.4) say that the function $\mathbb{1}_{(-1/t,1/t)}(\alpha)$ captures the "essence" of $\widehat{U}_t(\alpha)$. How? Let us define the operator A_t via its Fourier transform as $\hat{A}_t(\alpha) = \mathbb{1}_{(-1/t,1/t)}(\alpha)$. The great advantage of the (A_t) is that it is a monotone sequence of projections. We'll see in a minute how this can help. First we claim that the sequences (U_t) and (A_t) are equivalent. To prove this claim, start by observing that

$$\widehat{U_t f}(\alpha) = \widehat{U}_t(\alpha) \cdot \hat{f}(\alpha); \qquad \widehat{A_t f}(\alpha) = \widehat{A}_t(\alpha) \cdot \hat{f}(\alpha),$$

and then estimate, using Parseval's formula, as

$$\begin{aligned}
\int_{\mathbb{Z}} \sum_{t\in I} |A_t f - U_t f|^2 &= \int_{-1/2}^{1/2} \sum_{t\in I} \left|\widehat{A}_t(\alpha) - \widehat{U}_t(\alpha)\right|^2 \cdot \left|\hat{f}(\alpha)\right|^2 d\alpha \\
&\leqslant \int_{-1/2}^{1/2} \left|\hat{f}(\alpha)\right|^2 d\alpha \cdot \sup_\alpha \sum_{t\in I} \left|\widehat{U}_t(\alpha) - \hat{A}_t(\alpha)\right|^2 \\
&= \int_{\mathbb{Z}} f^2 \cdot \sup_\alpha \sum_{t\in I} \left|\widehat{U}_t(\alpha) - \hat{A}_t(\alpha)\right|^2.
\end{aligned}$$

It follows that it is enough to prove the inequality

$$\sup_\alpha \sum_{t\in I} \left|\widehat{U}_t(\alpha) - \hat{A}_t(\alpha)\right|^2 < \infty.$$

To see this, for a fixed α, divide the summation on t into two parts, $t < |\alpha|^{-1}$ and $t > |\alpha|^{-1}$. For the case $t < |\alpha|^{-1}$, use the estimate in (B.3) and in case $t > |\alpha|^{-1}$ use the estimate in (B.4). In both cases, we end up with a geometric progression with quotient $1/\sigma$.

Since (U_t) and (A_t) are equivalent and (U_t) satisfies a maximal inequality, the operators A_t also satisfy a maximal inequality. But then the sequence $(A_t f(x))$ satisfies an oscillation inequality. To see this, first note that if $t(k) \leqslant t \leqslant t(k+1)$ then

$A_t f(x) - A_{t(k+1)} f(x) = A_t(A_{t(k)} f(x) - A_{t(k+1)} f(x))$. It follows, that

$$\int_{\mathbb{Z}} \sup_{t(k)<t<t(k+1)} |A_t f - A_{t(k+1)} f|^2 = \int_{\mathbb{Z}} \sup_t \left|A_t(A_{t(k)} f - A_{t(k+1)} f)\right|^2$$
$$\leqslant c \cdot \int_{\mathbb{Z}} |A_{t(k)} f - A_{t(k+1)} f|^2,$$

since the sequence (A_t) satisfies a maximal inequality. But now the oscillation inequality follows from the inequality

$$\int_{\mathbb{Z}} \sum_k |A_{t(k)} f - A_{t(k+1)} f|^2 \leqslant \int_{\mathbb{Z}} f^2.$$

This inequality, in turn, follows by examining the Fourier transform of the left-hand side.

Now the punchline is that the ergodic averages (U_t) also satisfy the oscillation inequality since (U_t) and (A_t) are equivalent.

Let us summarize the scheme above: the maximal inequality for (U_t) implies a maximal inequality for the (A_t) since the two sequences are equivalent. But the (A_t), being a monotone sequence of projections, satisfy an oscillation inequality. But then, again appealing to the equivalence of the two sequences, the (U_t) satisfies an oscillation inequality.

What we have learned is that if a sequence of operators $(B_t f)$ satisfies a maximal inequality, and it is equivalent to a monotone sequence (A_t) of projections, then $(B_t f)$ satisfies an oscillation inequality.

In the remaining sections we will see that the scheme of proving an oscillation inequality for the averages along the squares (S_t) is similar, and ultimately it will be reduced to proving a maximal inequality for a monotone sequence of projections.

B.4. *Periodic systems and the circle method*

The difference between the usual ergodic averages and the averages along squares is that the squares are not uniformly distributed in residue classes. Indeed, for example no number of the form $3n - 1$ is a square. This property of the squares is captured well in the behavior of the Fourier transform, $\widehat{S}_t(\alpha) = 1/t \sum_{n \leqslant t} e^{2\pi i n^2 \alpha}$: for a typical rational $\alpha = b/q$, $\lim_{t\to\infty} \widehat{S}_t(\alpha)$ is nonzero (while it would be 0 if the squares were uniformly distributed mod q).

We need some estimates on the Fourier transform $\widehat{S}_t(\alpha)$. Since we will often deal with the function $e^{2\pi i\beta}$, we introduce the notation $e(\beta) = e^{2\pi i\beta}$. Also, the estimates for the Fourier transform $\widehat{S}_t(\alpha)$ are simpler if instead of the averages $\frac{1}{t}\sum_{n\leqslant t} \tau^{n^2} f(x)$ we consider the weighted averages $1/t \sum_{n^2\leqslant t}(2n-1)\tau^{n^2} f(x)$. The weight $2n - 1$ is motivated by $n^2 - (n-1)^2 = 2n - 1$. Everything we said about the averages along the squares applies equally well to these new weighted averages. Furthermore, it is an exercise in summation by parts to show that the a.e. convergence of the weighted and nonweighted averages is equivalent.

So from now on, we use the notation

$$S_t f(x) = \frac{1}{t} \sum_{n^2 \leqslant t} (2n-1) \tau^{n^2} f(x).$$

Let $\hat{\Lambda}(\alpha) = \lim_t \widehat{S}_t(\alpha)$. By Weyl's theorem, $\hat{\Lambda}(\alpha) = 0$ for irrational α and for rational $\alpha = b/q$, if b/q is in reduced terms, we have the estimate

$$\left|\hat{\Lambda}(b/q)\right| \leqslant \frac{c}{q^{1/2}}. \tag{B.5}$$

This inequality tells us that while the squares are not uniformly distributed in residue classes mod q, at least they try to be: $\hat{\Lambda}_t(b/q) \to 0$ as $q \to \infty$.

Now the so-called *circle method* of Hardy and Littlewood tells us about the structure of $\widehat{S}_t(\alpha)$. Let us introduce the notations $P(t) = t^{1/3}$, $Q(t) = 2t/P(t) = 2t^{2/3}$. According to the circle method, we have the following estimates

$$\left|\widehat{S}_t(\alpha) - \hat{\Lambda}(b/q) \cdot \widehat{U}_t(\alpha - b/q)\right| \leqslant c \cdot t^{-1/6},$$
$$\text{when } q \leqslant P(t),\ |\alpha - b/q| < 1/Q(t), \tag{B.6}$$
$$\left|\widehat{S}_t(\alpha)\right| < c \cdot t^{-1/6}, \quad \text{otherwise}, \tag{B.7}$$

where recall that U_t denotes the usual ergodic averages so $\widehat{U}_t(\beta) = 1/t \sum_{n \leqslant t} e(n\beta)$. In other words, the estimate above tells us that $\widehat{S}_t(\alpha)$ is close to $\hat{\Lambda}(b/q) \cdot \widehat{U}_t(\alpha - b/q)$ if α is close to a rational point b/q with small denominator, and otherwise $|\widehat{S}_t(\alpha)|$ is small.

Given these estimates, it is easy to see that the sequence (S_t) is equivalent to the sequence (A_t) defined by its Fourier transform as

$$\hat{A}_t(\alpha) = \sum_{\substack{b/q \\ q \leqslant P(t)}} \hat{\Lambda}(b/q) \cdot \widehat{U}_t(\alpha - b/q) \cdot \mathbb{1}_{(-1/Q(t), 1/Q(t))}(\alpha - b/q).$$

It remains to prove an oscillation inequality for the A_t. To do this, first we group those b/q for which q is of similar size:

$$E_p = \left\{b/q \mid 2^p \leqslant q < 2^{p+1}\right\}.$$

By the estimates in (B.5), we have

$$\sup_{b/q \in E_p} \left|\hat{\Lambda}(b/q)\right| \leqslant c \cdot 2^{-p/2}. \tag{B.8}$$

Note also that if $b/q \in E_p$ then the term $\hat{\Lambda}(b/q)$ occurs in the definition of A_t only when $t > 2^{3p}$. Define the operator $A_{p,t}$ by its Fourier transform as

$$\hat{A}_{p,t}(\alpha) = \sum_{b/q \in E_p} \hat{\Lambda}(b/q) \cdot \widehat{U}_t(\alpha - b/q) \cdot \mathbb{1}_{(-1/Q(t), 1/Q(t))}(\alpha - b/q), \quad t > 2^{3p}.$$

Using the triangle inequality for the summation in p, we see that an oscillation inequality for (A_t) would follow from the inequality

$$\int_{\mathbb{Z}} \sum_k \sup_{t(k)<t<t(k+1)} |A_{p,t} f - A_{p,t(k+1)} f|^2 \leqslant c \cdot \frac{p^2}{2^p} \cdot \int_{\mathbb{Z}} f^2. \tag{B.9}$$

We have learned in the previous section, Section B.3, that it is useful to try work with projections. As a step, we introduce the operators $B_{p,t}$ defined via

$$\widehat{B}_{p,t}(\alpha) = \sum_{b/q \in E_p} \hat{\Lambda}(b/q) \cdot \mathbb{1}_{(-1/t,1/t)}(\alpha - b/q), \quad t > 2^{3p}.$$

Note that for each α there is at most one $b/q \in E_p$ so that $\mathbb{1}_{(-1/t,1/t)}(\alpha - b/q) \neq 0$ or $\mathbb{1}_{(-1/Q(t),1/Q(t))}(\alpha - b/q) \neq 0$ for some $t > 2^{3p}$. Fix α and let b/q be the corresponding point of E_p. Hence, using the estimates in (B.3), (B.4), and (B.8), we get

$$\left|\hat{A}_{p,t}(\alpha) - \widehat{B}_{p,t}(\alpha)\right| \leqslant c \cdot 2^{-p/2} \cdot \min\left\{t|\alpha - b/q|, \left(t|\alpha - b/q|\right)^{-1}\right\}; \quad t > 2^{3p}.$$

It follows that we can replace the $(A_{p,t})$ by the $(B_{p,t})$:

$$\int_{\mathbb{Z}} \sum_{t>2^{3p}} |A_{p,t} f - B_{p,t} f|^2 \leqslant c \cdot 2^{-p} \cdot \int_{\mathbb{Z}} f^2.$$

In order to prove the required oscillation inequality for the $B_{p,t}$, we make one more reduction. Namely, we claim that defining $C_{p,t}$ by

$$\widehat{C}_{p,t}(\alpha) = \sum_{b/q \in E_p} \mathbb{1}_{(-1/t,1/t)}(\alpha - b/q), \quad t > 2^{3p}$$

(so $\widehat{C}_{p,t}$ is just $\widehat{B}_{p,t}$ without the multipliers $\hat{\Lambda}(b/q)$), we need to prove

$$\int_{\mathbb{Z}} \sum_k \sup_{t(k)<t<t(k+1)} |C_{p,t} - C_{p,t(k+1)}|^2 \leqslant c \cdot p^2 \cdot \int_{\mathbb{Z}} f^2. \tag{B.10}$$

To see that this is sufficient, define the function g by its Fourier transform as

$$\hat{g}(\alpha) = \sum_{b/q \in E_p} \hat{\Lambda}(b/q) \cdot \mathbb{1}_{(-2^{-3p},2^{-3p})}(\alpha - b/q) \cdot \hat{f}(\alpha).$$

Indeed, then $B_{p,t} f(x) = C_{p,t} g(x)$ and $\int_{\mathbb{Z}} g^2 \leqslant c \cdot 2^{-p} \int_{\mathbb{Z}} f^2$ by (B.8).

Now, the $C_{p,t}$ form a monotone (in t) sequence of projections, and hence they will satisfy the oscillation inequality in (B.10) once they satisfy the maximal inequality

$$\int_{\mathbb{Z}} \sup_{t>2^{3p};\, t\in I} |C_{p,t}|^2 \leqslant c \cdot p^2 \cdot \int_{\mathbb{Z}} f^2. \tag{B.11}$$

To encourage the reader, we emphasize that our only remaining task is to prove the inequality in (B.11) above.

B.5. *The main inequality*

Since the least common multiple of the denominators of rational numbers in the set E_p is not greater than 2^{cp2^p} and the distance between two elements of E_p is at least 2^{-2p}, the estimate in (B.11) follows from the following result

THEOREM B.4. *Let $0 < \delta < 1/2$ and $e(\alpha_1), e(\alpha_2), \ldots, e(\alpha_J)$ be distinct complex Qth roots of unity with $|\alpha_i - \alpha_j| > \delta/2$ for $i \neq j$. We assume that $\delta^{-1} \leqslant Q$. Define for $t \in I$ the projections R_t by*

$$\widehat{R_t}(\alpha) = \sum_{j \leqslant J} \mathbb{1}_{(-1/t,1/t)}(\alpha - \alpha_j).$$

Then we have, with an absolute constant c,

$$\int_{\mathbb{Z}} \sup_{t \geqslant \delta^{-1};\, t \in I} |R_t f|^2 \leqslant c \cdot (\log\log Q)^2 \cdot \int_{\mathbb{Z}} |f|^2.$$

We restrict the range on t to $t \geqslant \delta^{-1}$, because then the sum making up R_t contains pairwise orthogonal elements—as a result of the separation hypothesis $|\alpha_i - \alpha_j| > \delta/2$.

PROOF. Two essentially different techniques will be used to handle the supremum. The first technique will handle the range $\delta^{-1} \leqslant t \leqslant Q^4$, and the other technique will handle the remaining $t > Q^4$ range.

Let us start with proving the inequality

$$\int_{\mathbb{Z}} \sup_{\delta^{-1} \leqslant t \leqslant Q^4;\, t \in I} |R_t f|^2 \leqslant c \cdot (\log\log Q)^2 \cdot \int_{\mathbb{Z}} |f|^2. \tag{B.12}$$

We can assume that Q^4 is a power of σ, say $Q^4 = \sigma^S$, and then the range $\delta^{-1} \leqslant t \leqslant Q^4$ can be rewritten as $c \log \delta^{-1} \leqslant s \leqslant S$, where we take logs to base σ. Introduce the monotone sequence of projections $P_s = R_{\sigma^{S-s}}$, $s \leqslant S - c \log \delta^{-1}$. All follows from

$$\int_{\mathbb{Z}} \sup_{s \leqslant S - c\log\delta^{-1}} |P_s f|^2 \leqslant c \cdot \log^2 S \cdot \int_{\mathbb{Z}} |f|^2.$$

It is clearly enough to show the inequality for dyadic $S - c \log \delta^{-1}$:

$$\int_{\mathbb{Z}} \sup_{s \leqslant 2^M} |P_s f|^2 \leqslant c \cdot M^2 \cdot \int_{\mathbb{Z}} |f|^2.$$

For each integer $m \leqslant M$ consider the sets

$$H_m = \left\{ P_{(d+1)\cdot 2^m} - P_{d\cdot 2^m} \mid d = 0, 1, \ldots, 2^{M-m} - 1 \right\}.$$

If the dyadic expansion of s is $s = \sum_{m \leqslant M} \varepsilon_m \cdot 2^m$, where ε_m is 0 or 1, then for some $X_m \in H_m$, $P_s = \sum_{m \leqslant M} \varepsilon_m \cdot X_m$. It follows that

$$\left| P_s f(x) \right|^2 \leqslant M \cdot \sum_{m \leqslant M} \left| X_m f(x) \right|^2.$$

For each m, we have

$$\left| X_m f(x) \right|^2 \leqslant \sum_{d \leqslant 2^{M-m}} \left| P_{(d+1)\cdot 2^m} f(x) - P_{d\cdot 2^m} f(x) \right|^2,$$

hence

$$\begin{aligned} \int_{\mathbb{Z}} \sup_{s \leqslant 2^M} |P_s f|^2 &\leqslant M \cdot \int_{\mathbb{Z}} \sum_{m \leqslant M} \sum_{d \leqslant 2^{M-m}} \left| P_{(d+1)\cdot 2^m} f(x) - P_{d\cdot 2^m} f(x) \right|^2 \\ &\leqslant M \cdot \sum_{m \leqslant M} \sum_{s \leqslant 2^M} \int_{\mathbb{Z}} |P_{s+1} f - P_s f|^2 \\ &\leqslant M^2 \cdot \int_{\mathbb{Z}} |f|^2, \end{aligned}$$

where we are using for the second inequality the fact that the P_s are a monotone sequence of projections.

Let us now handle the remaining range for t. We want to prove

$$\int_{\mathbb{Z}} \sup_{t > Q^4} |R_t f|^2 \leqslant c \cdot \int_{\mathbb{Z}} |f|^2. \tag{B.13}$$

It seems best if we replace the operators R_t by the operators

$$A_t f(x) = \frac{1}{t} \sum_{n \leqslant t} \sum_{j \leqslant J} e(n\alpha_j) f(x+n).$$

This replacement is possible if we prove the following two inequalities

$$\int_{\mathbb{Z}} \sum_{t > \delta^{-2}} |A_t f - R_t f|^2 \leqslant c \cdot \int_{\mathbb{Z}} |f|^2 \tag{B.14}$$

and

$$\int_{\mathbb{Z}} \sup_{t > Q^4} |A_t f|^2 \leqslant c \cdot \int_{\mathbb{Z}} |f|^2. \tag{B.15}$$

Let us start with proving (B.14). By Parseval's formula, we need to prove

$$\sup_{\alpha} \sum_{t} \left| \hat{A}_t(\alpha) - \widehat{R}_t(\alpha) \right|^2 < \infty.$$

Fix α. Without loss of generality we can assume that of the α_j, the point α_1 is closest to α. Possibly dividing the sum over j into two parts and reindexing them, we assume that $\alpha_1 < \cdots < \alpha_J$. Using the separation hypothesis $|\alpha_i - \alpha_j| > \delta/2$, we have that $|\alpha - \alpha_j| > (j-1)\delta/2$ for $j > 1$.

For $t \leqslant 1/|\alpha - \alpha_1|$ we can thus estimate (recall that $\widehat{U}_t(\beta) = 1/t \sum_{n \leqslant t} e(n\beta)$) as

$$\begin{aligned}
\left| \hat{A}_t(\alpha) - \widehat{R}_t(\alpha) \right| &\leqslant \left| \widehat{U}_t(\alpha - \alpha_1) - 1 \right| + \sum_{2 \leqslant j \leqslant J} \left| \widehat{U}_t(\alpha - \alpha_j) \right| \\
&\leqslant c \cdot \left(t|\alpha - \alpha_1| + \sum_{2 \leqslant j \leqslant J} \frac{1}{t(j-1)\delta} \right) \\
&\leqslant c \cdot \left(t|\alpha - \alpha_1| + \log J/(\delta t) \right) \\
&\leqslant c \cdot \left(t|\alpha - \alpha_1| + \delta^{-2}/t \right),
\end{aligned}$$

where for the second inequality we used (B.3) and (B.4) and for the last estimate we used that $J \leqslant \delta^{-1}$. Summing this estimate over $t \in I$ with $\delta^{-2} \leqslant t \leqslant 1/|\alpha - \alpha_1|$ we get a finite bound independent of α.

For $t > 1/|\alpha - \alpha_1|$, we have

$$\left| \hat{A}_t(\alpha) - \widehat{R}_t(\alpha) \right| \leqslant \sum_{1 \leqslant j \leqslant J} \left| \widehat{U}_t(\alpha - \alpha_j) \right| \leqslant c \cdot \frac{\delta^{-2}}{t},$$

which, upon summing over the full range $\delta^{-2} < t$, again gives a finite bound independent of α.

Let us single out a consequence of inequality (B.14): there is a constant c so that

$$\int_{\mathbb{Z}} |A_t f|^2 \leqslant c \cdot \int_{\mathbb{Z}} |f|^2; \quad t > \delta^{-2}. \tag{B.16}$$

Our only remaining task is to prove inequality (B.15).

For a given t, let q be the largest integer so that $q Q^2 \leqslant t$. Note that $q \geqslant Q^2$ since $t > Q^4$. We can estimate as

$$\begin{aligned}
&\left| \sum_{n \leqslant t} \sum_{j \leqslant J} e(n\alpha_j) f(x+n) \right| \\
&\quad \leqslant \left| \sum_{n \leqslant q Q^2} \sum_{j \leqslant J} e(n\alpha_j) f(x+n) \right| + \left| \sum_{q Q^2 < n \leqslant t} \sum_{j \leqslant J} e(n\alpha_j) f(x+n) \right|.
\end{aligned} \tag{B.17}$$

We estimate the second term on the right trivially as

$$\left| \sum_{qQ^2 < n \leqslant t} \sum_{j \leqslant J} e(n\alpha_j) f(x+n) \right| \leqslant J \cdot \sum_{qQ^2 < n \leqslant (q+1)Q^2} \left| f(x+n) \right|.$$

With this, we have

$$\begin{aligned}
&\sup_{t > Q^4} \left(\frac{1}{t} \left| \sum_{qQ^2 < n \leqslant t} \sum_{j \leqslant J} e(n\alpha_j) f(x+n) \right| \right)^2 \\
&\quad \leqslant \sup_{q \geqslant Q^2} \left(\frac{J}{qQ^2} \cdot \sum_{qQ^2 < n \leqslant (q+1)Q^2} \left| f(x+n) \right| \right)^2 \quad \text{by Cauchy's inequality} \\
&\quad \leqslant \sup_{q \geqslant Q^2} \frac{J^2 \cdot Q^2}{qQ^2} \cdot \frac{\sum_{qQ^2 < n \leqslant (q+1)Q^2} |f(x+n)|^2}{qQ^2} \\
&\quad \leqslant \sum_{q \geqslant Q^2} \frac{J^2}{q^2} \cdot \frac{1}{Q^2} \sum_{qQ^2 < n \leqslant (q+1)Q^2} \left| f(x+n) \right|^2.
\end{aligned}$$

Integrating the last line, we obtain the bound

$$\sum_{q \geqslant Q^2} \frac{J^2}{q^2} \cdot \int_{\mathbb{Z}} |f|^2 \leqslant c \cdot \frac{J^2}{Q^2} \int_{\mathbb{Z}} |f|^2 \leqslant c \cdot \int_{\mathbb{Z}} |f|^2$$

since $J \leqslant Q$.

Let us now handle the first term on the right of (B.17). Since $e(\alpha_j)$ satisfies $e((mQ^2 + h)\alpha_j) = e(h\alpha_j)$ (this is the first and last time we use that the $e(\alpha_j)$ are Qth roots of unity), we can write, defining $Tg(x) = g(x + Q^2)$,

$$\left| \frac{1}{t} \sum_{n \leqslant qQ^2} \sum_{j \leqslant J} e(n\alpha_j) f(x+n) \right| \leqslant \left| \frac{1}{q} \sum_{m \leqslant q} T^m \frac{1}{Q^2} \sum_{h \leqslant Q^2} \sum_{j \leqslant J} e(h\alpha_j) f(x+h) \right|.$$

By the ergodic maximal inequality, applied to T, the ℓ^2 norm of our maximal operator is bounded by the ℓ^2 norm of

$$\frac{1}{Q^2} \sum_{h \leqslant Q^2} \sum_{j \leqslant J} e(h\alpha_j) f(x+h).$$

But the estimate in (B.16) says, the ℓ^2 norm of the above is bounded independently of Q since $Q^2 > \delta^{-2}$ by assumption. □

B.6. *Notes*

More details. More details and references can be found in [122]. In particular, the circle method and the transference principle are described in complete details—though no proof of the main inequality of Bourgain, Theorem B.4, is given. The inequalities (B.6) and (B.7) appear as (4.23) and (4.24) in [122].

Theorem B.1. The result is due to Bourgain [39]. He later extended the result to $f \in L^p$, $p > 1$; cf. [40]. The case $p = 1$ is the most outstanding unsolved problem in this subject.

Idea of proof. The basic structure of the proof is that of Bourgain's [40] but we used ideas from Lacey's paper [96] as well—not to mention some personal communication with M. Lacey.

Other sequences. The sequence of primes is discussed in [137]. But we'd like to emphasize that the L^2 theory of the primes is identical to the case of the squares. The only difference is in the estimates in (B.6) and (B.7).

A characterization of sequences which are good for the pointwise and mean ergodic theorems can be found in [38].

Acknowledgement

The authors thank E. Lesigne for useful comments.

References

Surveys in volume 1A and this volume

[1] B. Hasselblatt and A. Katok, *Principal structures*, Handbook of Dynamical Systems, Vol. 1A, B. Hasselblatt and A. Katok, eds Elsevier, Amsterdam (2002), 1–203.

[2] A. Nevo, *Pointwise ergodic theorems for actions of groups*, Handbook of Dynamical Systems, Vol. 1B, B. Hasselblatt and A. Katok, eds, Elsevier, Amsterdam (2006), 871–980.

Other sources

[3] J. Auslander, *On the proximal relation in topological dynamics*, Proc. Amer. Math. Soc. **11** (1960), 890–895.

[4] S. Banach, *Sur le problème de la mesure*, Fund. Math. **4** (1923), 7–33.

[5] S. Banach and A. Tarski, *Sur la décomposition des ensembles de points en parties respectivement congruentes*, Fund. Math. **6** (1924), 244–277.

[6] D. Berend and V. Bergelson, *Jointly ergodic measure-preserving transformations*, Israel J. Math. **49** (4) (1984), 307–314.

[7] V. Bergelson, *Sets of recurrence of $\mathbb{Z}^m$-actions, and properties of sets of differences in $\mathbb{Z}^m$*, J. London Math. Soc. (2) **31** (1985), 295–304.
[8] V. Bergelson, *Weakly mixing PET*, Ergodic Theory Dynamical Systems **7** (1987), 337–349.
[9] V. Bergelson, *Ergodic Ramsey theory—an update*, Ergodic Theory of $\mathbb{Z}^d$-actions (Warwick, 1993–1994), London Math. Soc. Lecture Note Ser., Vol. 228, Cambridge University Press, Cambridge (1996), 273–296.
[10] V. Bergelson, *The multifarious Poincare recurrence theorem*, Descriptive Set Theory and Dynamical Systems (Marseille-Luminy, 1996), London Math. Soc. Lecture Note Ser., Vol. 277, Cambridge University Press, Cambridge (2000), 31–57.
[11] V. Bergelson, *Ergodic theory and Diophantine problems*, Topics in Symbolic Dynamics and Applications (Temuco, 1997), London Math. Soc. Lecture Note Ser., Vol. 279, Cambridge University Press, Cambridge (2000), 167–205.
[12] V. Bergelson, *Minimal idempotents and ergodic Ramsey theory*, Topics in Dynamics and Ergodic Theory, London Math. Soc. Lecture Note Ser., Vol. 310, Cambridge University Press, Cambridge (2003), 8–39.
[13] V. Bergelson, *Multiplicatively large sets and ergodic Ramsey theory*, Israel J. Math., to appear.
[14] V. Bergelson, A. Blass and N. Hindman, *Partition theorems for spaces of variable words*, Proc. London Math. Soc. (3) **68** (3) (1994), 449–476.
[15] V. Bergelson, M. Boshernitzan and J. Bourgain, *Some results on nonlinear recurrence*, J. Anal. Math. **62** (1994), 29–46.
[16] V. Bergelson, H. Furstenberg, N. Hindman and Y. Katznelson, *An algebraic proof of van der Waerden's theorem*, Enseign. Math. (2) **35** (3–4) (1989), 209–215.
[17] V. Bergelson, H. Furstenberg and R. McCutcheon, *IP-sets and polynomial recurrence*, Ergodic Theory Dynamical Systems **16** (5) (1996), 963–974.
[18] V. Bergelson and A. Gorodnik, *Weakly mixing group actions: A brief survey and an example*, Modern Dynamical Systems and Applications, B. Hasselblatt, M. Brin and Y. Pesin, eds, Cambridge University Press, New York (2004), 3–25.
[19] V. Bergelson and N. Hindman, *Nonmetrizable topological dynamics and Ramsey theory*, Trans. Amer. Math. Soc. **320** (1990), 293–320.
[20] V. Bergelson and N. Hindman, *Some topological semicommutative van der Waerden type theorems and their combinatorial consequences*, J. London Math. Soc. (2) **45** (3) (1992), 385–403.
[21] V. Bergelson and N. Hindman, *Additive and multiplicative Ramsey theorems in N—some elementary results*, Combin. Probab. Comput. **2** (3) (1993), 221–241.
[22] V. Bergelson and N. Hindman, *On* IP^* *sets and central sets*, Combinatorica **14** (3) (1994), 269–277.
[23] V. Bergelson and A. Leibman, *Polynomial extensions of van der Waerden's and Szemerédi's theorems*, J. Amer. Math. Soc. **9** (1996), 725–753.
[24] V. Bergelson and A. Leibman, *Set-polynomials and polynomial extension of the Hales–Jewett theorem*, Ann. of Math. (2) **150** (1) (1999), 33–75.
[25] V. Bergelson and A. Leibman, *A nilpotent Roth theorem*, Invent. Math. **147** (2) (2002), 429–470.
[26] V. Bergelson and A. Leibman, *Topological multiple recurrence for polynomial configurations in nilpotent groups*, Adv. Math. **175** (2) (2003), 271–296.
[27] V. Bergelson and A. Leibman, *Failure of the Roth theorem for solvable groups of exponential growth*, Ergodic Theory Dynamical Systems **24** (2004), 45–53.
[28] V. Bergelson, A. Leibman and R. McCutcheon, *Polynomial Szemerédi theorems for countable modules over integral domains and finite fields*, J. Anal. Math., to appear.
[29] V. Bergelson and R. McCutcheon, *Uniformity in the polynomial Szemerédi theorem*, Ergodic Theory of $\mathbb{Z}^d$-actions (Warwick, 1993–1994), London Math. Soc. Lecture Note Ser., Vol. 228, Cambridge University Press, Cambridge (1996), 273–296.
[30] V. Bergelson, R. McCutcheon and Q. Zhang, *A Roth theorem for amenable groups*, Amer. J. Math. **119** (6) (1997), 1173–1211.
[31] V. Bergelson and R. McCutcheon, *Recurrence for semigroup actions and a non-commutative Schur theorem*, Topological Dynamics and Applications (Minneapolis, MN, 1995), Contemp. Math., Vol. 215, Amer. Math. Soc., Providence, RI (1998), 205–222.
[32] V. Bergelson and R. McCutcheon, *An ergodic IP polynomial Szemerédi theorem*, Mem. Amer. Math. Soc. **146** (695) (2000), viii+106 pp.
[33] V. Bergelson and J. Rosenblatt, *Mixing actions of groups*, Illinois J. Math. **32** (1) (1988), 65–80.

[34] V. Bergelson and J. Rosenblatt, *Joint ergodicity for group actions*, Ergodic Theory Dynamical Systems **8** (3) (1988), 351–364.
[35] V. Bergelson and D. Shapiro, *Multiplicative subgroups of finite index in a ring*, Proc. Amer. Math. Soc. **116** (4) (1992), 885–896.
[36] P. Billingsley, *Probability and Measure*, Wiley, New York (1986).
[37] A. Blaszczyk, S. Plewik and S. Turek, *Topological multidimensional van der Waerden theorem*, Comment. Math. Univ. Carolin. **30** (4) (1989), 783–787.
[38] M. Boshernitzan, G. Kolesnik, A. Quas and M. Wierdl, *Ergodic averaging sequences*, J. Anal. Math., 33 pages, to appear. Available at http://www.csi.hu/mw/hardy_ergav.pdf or http://www.csi.hu/mw/hardy_ergav.dvi
[39] J. Bourgain, *On the maximal ergodic theorem for certain subsets of the integers*, Israel J. Math. **61** (1988), 39–72.
[40] J. Bourgain, *Pointwise ergodic theorems for arithmetic sets (appendix: The return time theorem)*, Publ. Math. I.H.E.S. **69** (1989), 5–45.
[41] J. Bourgain, *Double recurrence and almost sure convergence*, J. Reine Angew. Math. **404** (1990), 140–161.
[42] T.C. Brown, *An interesting combinatorial method in the theory of locally finite semigroups*, Pacific J. Math. **36** (1971), 285–289.
[43] T.J. Carlson, *Some unifying principles in Ramsey theory*, Discrete Math. **68** (1988), 117–169.
[44] P. Civin and B. Yood, *The second conjugate space of a Banach algebra as an algebra*, Pacific J. Math. **11** (1961), 847–870.
[45] W. Comfort and S. Negrepontis, *The Theory of Ultrafilters*, Springer, Berlin (1974).
[46] J.P. Conze and E. Lesigne, *Théorèmes ergodiques pour des mesures diagonales*, Bull. Soc. Math. France **112** (2) (1984) 143–175.
[47] J.P. Conze and E. Lesigne, *Sun un théorème ergodique pour des mesures diagonales*, Probabilités Publ. Inst. Rech. Math. Rennes, 1987-1, Univ. Rennes I, Rennes (1988), 1–31.
[48] J.P. Conze and E. Lesigne, *Sun un théorème ergodique pour des mesures diagonales*, C. R. Acad. Sci. Paris Sér. I Math. **306** (12) (1988), 491–493.
[49] L.E. Dickson, *On the congruence* $x^n + y^n \equiv z^n \pmod{p}$, J. Reine Angew. Math. **135** (1909), 134–141.
[50] L.E. Dickson, *History of the Theory of Numbers, Vol. II (Diophantine Analysis)*, Chelsea, New York (1971).
[51] R. Ellis, *Distal transformation groups*, Pacific J. Math. **8** (1958), 401–405.
[52] R. Ellis, *A semigroup associated with a transformation group*, Trans. Amer. Math. Soc. **94** (1960), 272–281.
[53] P. Erdős and P. Turán, *On some sequences of integers*, J. London Math. Soc. **11** (1936), 261–264.
[54] E. Følner, *On groups with full Banach mean values*, Math. Scand. **3** (1955), 243–254.
[55] N. Frantzikinakis and B. Kra, *Polynomial averages converge to the product of integrals*, Israel J. Math., to appear.
[56] H. Furstenberg, *The structure of distal flows*, Amer. J. Math. **85** (1963), 477–515.
[57] H. Furstenberg, *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Anal. Math. **31** (1977), 204–256.
[58] H. Furstenberg, *Recurrence in Ergodic Theory and Combinatorial Number Theory*, Princeton University Press (1981).
[59] H. Furstenberg, *Poincaré recurrence and number theory*, Bull. Amer. Math. Soc. (N.S.) **5** (1981), 211–234.
[60] H. Furstenberg and Y. Katznelson, *An ergodic Szemerédi theorem for commuting transformations*, J. Anal. Math. **34** (1978), 275–291.
[61] H. Furstenberg, Y. Katznelson and D. Ornstein, *The ergodic theoretical proof of Szemerédi's theorem*, Bull. Amer. Math. Soc. (N.S.) **7** (3) (1982), 527–552.
[62] H. Furstenberg and Y. Katznelson, *An ergodic Szemerédi theorem for IP-systems and combinatorial theory*, J. Anal. Math. **45** (1985), 117–168.
[63] H. Furstenberg and Y. Katznelson, *Idempotents in compact semigroups and Ramsey theory*, Israel J. Math. **68** (1990), 257–270.
[64] H. Furstenberg and Y. Katznelson, *A density version of the Hales–Jewett theorem*, J. Anal. Math. **57** (1991), 64–119.

[65] H. Furstenberg and B. Weiss, *Topological dynamics and combinatorial number theory*, J. Anal. Math. **34** (1978), 61–85.
[66] H. Furstenberg and B. Weiss, *A mean ergodic theorem for* $\frac{1}{N}\sum_{n=1}^{N} f(T^n x)g(T^{n^2}x)$, Convergence in Ergodic Theory and Probability (Columbus, OH, 1993), Ohio State Univ. Math. Res. Inst. Publ., Vol. 5, de Gruyter, Berlin (1996), 193–227.
[67] H. Furstenberg and B. Weiss, *Markov processes and Ramsey theory for trees*, Combinatorics, Probability and Computing **12** (2003), 547–563.
[68] A. Garcia, *Topics in Almost Everywhere Convergence*, Markham, Chicago (1970).
[69] L. Gillman and M. Jerison, *Rings of Continuous Functions*, Springer, New York (1976).
[70] E. Glasner, *Divisibility properties and the Stone–Čech compactification*, Canad. J. Math. **32** (1980), 993–1007.
[71] T. Gowers, *Hypergraph regularity and the multidimensional Szemerédi theorem*, Preprint.
[72] R. Graham, B. Rothschild and J. Spencer, *Ramsey Theory*, Wiley, New York (1980).
[73] B. Green and T. Tao, *The primes contain arbitrarily long arithmetic progressions*. Available at http://arxiv.org/abs/math.NT/0404188.
[74] F. Greenleaf, *Invariant Means on Topological Groups*, van Nostrand, New York (1969).
[75] A.W. Hales and R.I. Jewett, *Regularity and positional games*, Trans. Amer. Math. Soc. **106** (1963), 222–229.
[76] P. Halmos, *Lectures on Ergodic Theory*, Mathematical Society of Japan, Tokyo (1956).
[77] P. Halmos and J. von Neumann, *Operator methods in classical mechanics, II*, Ann. of Math. **43** (1942), 332–350.
[78] G. Hardy and J. Littlewood, *Some problems of Diophantine approximation*, Acta Math. **37** (1914), 155–191.
[79] F. Hausdorff, *Grundzüge der Mengenlehre*, Verlag von Veit, Leipzig (1914). Reprinted by Chelsea, New York (1949).
[80] H. Helson, *Harmonic Analysis*, Addison-Wesley, Reading, MA (1983).
[81] E. Hewitt and K. Ross, *Abstract Harmonic Analysis I*, Springer (1963).
[82] D. Hilbert, *Über die Irreduzibilität ganzer rationaler Funktionen mit ganzzahligen Koeffizienten*, J. Math. **110** (1892), 104–129.
[83] N. Hindman, *Finite sums from sequences within cells of a partition of* $\mathbb{N}$, J. Combin. Theory (Ser. A) **17** (1974), 1–11.
[84] N. Hindman, *Partitions and sums and products of integers*, Trans. Amer. Math. Soc. **247** (1979), 227–245.
[85] N. Hindman and D. Strauss, *Algebra in the Stone–Čech Compactification. Theory and Applications*, de Gruyter Expositions in Mathematics, Vol. 27, Walter de Gruyter, Berlin (1998), xiv+485 pp.
[86] E. Hopf, *Ergodentheorie*, Chelsea, New York (1948).
[87] B. Host and B. Kra, *Convergence of Conze–Lesigne averages*, Ergodic Theory Dynamical Systems **21** (2) (2001), 493–509.
[88] B. Host and B. Kra, *Nonconventional ergodic averages and nilmanifolds*, Ann. of Math., to appear.
[89] B. Host and B. Kra, *Convergence of polynomial ergodic averages*, Israel J. Math., to appear.
[90] T. Kamae and M. Mendès-France, *Van der Corput's difference theorem*, Israel J. Math. **31** (3–4) (1978), 335–342.
[91] A.Y. Khintchine, *Three Pearls of Number Theory*, Graylock Press, Rochester, NY (1952).
[92] B.O. Koopman and J. von Neumann, *Dynamical systems of continuous spectra*, Proc. Nat. Acad. Sci. U.S.A. **18** (1932), 255–263.
[93] U. Krengel, *Weakly wandering vectors and weakly independent partitions*, Trans. Amer. Math. Soc. **164** (1972), 199–226.
[94] U. Krengel, *Ergodic Theorems*, de Gruyter, Berlin (1985).
[95] L. Kronecker, *Die Periodensysteme von Funktionen Reeller Variablen*, Berliner Sitzungsberichte (1884), 1071–1080.
[96] M. Lacey, *On an inequality due to Bourgain*, Illinois J. Math. **41** (1997), 231–236.
[97] A. Leibman, *Multiple recurrence theorem for nilpotent group actions*, Geom. Funct. Anal. **4** (6) (1994), 648–659.
[98] A. Leibman, *Multiple recurrence theorem for measure preserving actions of a nilpotent group*, Geom. Funct. Anal. **8** (1998), 853–931.

[99] A. Leibman, *The structure of unitary actions of finitely generated nilpotent groups*, Ergodic Theory Dynamical Systems **20** (2000), 809–820.
[100] A. Leibman, *Pointwise convergence of ergodic averages for polynomial sequences of translations on a nilmanifold*, Ergodic Theory Dynamical Systems, to appear.
[101] A. Leibman, *Convergence of multiple ergodic averages along polynomials of several variables*, Israel J. Math., to appear.
[102] E. Lesigne, *Sur une nil-variété, les parties minimales assocées à une translation sont uniquement ergodiques*, Ergodic Theory Dynamical Systems **11** (1991), 379–391.
[103] E. Lindenstrauss, *Pointwise theorems for amenable groups*, Invent. Math. **146** (2001), 259–295.
[104] R. McCutcheon, *Elemental Methods in Ergodic Ramsey Theory*, Lecture Notes in Math., Vol. 1722, Springer, Berlin (1999).
[105] K. Milliken, *Ramsey's theorem with sums or unions*, J. Combin. Theory (Ser. A) **18** (1975), 276–290.
[106] B. Nagle, V. Rödl and M. Schacht, *The counting lemma for regular k-uniform hypergraphs*, Random Structures Algorithms, to appear.
[107] I. Namioka, *Følner's condition for amenable semi-groups*, Math. Scand. **15** (1964), 18–28.
[108] K. Numakura, *On bicompact semigroups*, Math. J. Okayama University **1** (1952), 99–108.
[109] A. Olshanskii, *On the question of the existence of an invariant mean on a group*, Uspekhi Mat. Nauk **35** (4) (214) (1980), 199–200.
[110] A. Olshanskii, *An infinite group with subgroups of prime orders*, Izv. Akad. Nauk SSSR Ser. Mat. **44** (2) (1980), 309–321.
[111] D. Ornstein and B. Weiss, *Entropy and isomorphism theorems for actions of amenable groups*, J. Anal. Math. **48** (1987), 1–141.
[112] A.L.T. Paterson, *Amenability*, Mathematical Surveys and Monographs, Vol. 29, Amer. Math. Soc., Providence (1988).
[113] K. Petersen, *Ergodic Theory*, Cambridge University Press, Cambridge (1981).
[114] J.-P. Pier, *Amenable Locally Compact Groups*, Wiley (1984).
[115] H. Poincaré, *Sur le problème des trois corps et les équations de la dynamique*, Acta Math. **13** (1890), 1–270.
[116] R. Rado, *Note on combinatorial analysis*, Proc. London Math. Soc. **48** (1993), 122–160.
[117] F.P. Ramsey, *On a problem of formal logic*, Proc. London Math. Soc. **30** (1930), 264–286.
[118] P. Ribenboim, *13 Lectures on Fermat's Last Theorem*, Springer, New York (1979).
[119] V. Rödl and J. Skokan, *Regularity lemma for k-uniform hypergraphs*, Random Structures Algorithms, to appear.
[120] V. Rödl and J. Skokan, *Applications of the regularity lemma for uniform hypergraphs*, Preprint.
[121] J. Rosenblatt, *Invariant measures and growth conditions*, Trans. Amer. Math. Soc. **193** (1974), 33–53.
[122] J. Rosenblatt and M. Wierdl, *Pointwise ergodic theorems via harmonic analysis*, Ergodic Theory and Its Connection with Harmonic Analysis, K. Petersen and I. Salama, eds, London Math. Soc. Lecture Note Ser., Vol. 205, Cambridge University Press, Cambridge (1995), 3–151.
[123] H.L. Royden, *Real Analysis*, Macmillan, New York (1968).
[124] A. Sárközy, *On difference sets of integers III*, Acta Math. Acad. Sci. Hungar. **31** (1978), 125–149.
[125] W. Schmidt, *Equations over Finite Fields: An Elementary Approach*, Springer, Berlin (1976).
[126] I. Schur, *Über die Kongruenz $x^m + y^m \equiv z^m \pmod p$*, Jahresber. Deutsch. Math.-Verein. **25** (1916), 114–117.
[127] E. Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27** (1975), 199–245.
[128] T. Tao, *A variant of the hypergraph removal lemma*, Preprint.
[129] A. Taylor, *A canonical partition relation for finite subsets of ω*, J. Combin. Theory (Ser. A) **17** (1974), 1–11.
[130] B. van der Waerden, *Beweis einer Baudetschen Vermutung*, Nieuw. Arch. Wisk. **15** (1927), 212–216.
[131] B. van der Waerden, *How the proof of Baudet's conjecture was found*, Studies in Pure Mathematics presented to Richard Rado, L. Mirsky, ed., Academic Press, London (1971), 251–260.
[132] J. von Neumann, *Zur algemeinen Theorie des Masses*, Fund. Math. **13** (1929), 73–116.
[133] J. von Neumann, *Zur Operatorenmethode in der klassischen Mechanik*, Ann. of Math. **33** (1932), 587–642.
[134] S. Wagon, *The Banach–Tarski Paradox*, Cambridge University Press, Cambridge (1985).

[135] P. Walters, *An Introduction to Ergodic Theory*, Springer, New York (1982).
[136] H. Weyl, *Über die Gleichverteilung von Zahlen mod. Eins*, Math. Ann. **77** (1916), 313–352.
[137] M. Wierdl, *Pointwise ergodic theorem along the prime numbers*, Israel J. Math. **64** (1988), 315–336.
[138] Q. Zhang, *On convergence of the averages* $\frac{1}{N}\sum_{n=1}^{N} f_1(R^n x) f_2(S^n x) f_3(T^n x)$, Monatsh. Math. **122** (1996), 275–300.
[139] T. Ziegler, *Universal characteristic factors and Furstenberg averages*, Preprint.
[140] R. Zimmer, *Extensions of ergodic group actions*, Illinois J. Math. **20** (3) (1976), 373–409.
[141] R. Zimmer, *Ergodic actions with generalized discrete spectrum*, Illinois J. Math. **20** (4) (1976), 555–588.

This page intentionally left blank

CHAPTER 13

Pointwise Ergodic Theorems for Actions of Groups

Amos Nevo*

Department of Mathematics, Technion, Haifa, Israel
E-mail: *anevo@tx.technion.ac.il*

Contents

*The author was supported in part by ISF Grant #126-01.

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

1. Introduction

In recent years a number of significant results and developments related to pointwise ergodic theorems for general measure-preserving actions of locally compact second countable (lcsc) groups have been established, including the solution of several long-standing open problems. The exposition that follows aims to survey some of these results and their proofs, and will include, in particular, an exposition of the following results.

(1) A complete solution to the ball averaging problem on Lie, and more generally lcsc, groups of polynomial volume growth. Namely, a proof that for any metric quasi-isometric to a word metric (and in particular, Riemannian metrics on nilpotent Lie groups), the normalized ball averages satisfy the pointwise ergodic theorem in L^1. This brings to a very satisfactory close a long-standing problem in ergodic theory, dating at least to Calderon's 1952 paper on groups satisfying the doubling condition.

(2) In fact, two independent solutions will be described regarding the ball averaging problem in the case of connected Lie groups with polynomial volume growth, but both have the following in common. They resolve, in particular, a long-standing conjecture in the theory of amenable groups, dating at least to F. Greenleaf's 1969 book [58]. The conjecture asserts that the sequence of powers of a neighborhood on an amenable group constitute an asymptotically invariant sequence, namely has the Følner property. This conjecture was disproved for solvable groups with exponential growth, but has been now verified for groups with polynomial growth.

(3) The pointwise ergodic theorem in L^1 for a tempered sequence of asymptotically invariant sets was established recently, improving on the case of L^2 established earlier. This result resolves the long-standing problem of constructing *some* pointwise ergodic sequence in L^1 on an *arbitrary* amenable group. The ideas of the two available proofs will be briefly described.

(4) A new and streamlined account of the classical Dunford–Zygmund method will be described. This account allows the derivation of pointwise ergodic theorems for asymptotically invariant sequence on any lcsc amenable algebraic (or Lie) group over any local field, generalizing the Greenleaf–Emerson theorem. It also allows the construction of pointwise ergodic sequences on any lcsc algebraic (or Lie) group over any local field, generalizing Templeman's theorem for lcsc connected groups.

(5) A general spectral method will be described for the derivation of pointwise ergodic theorems for ball averages on Gelfand pairs. This method will be demonstrated for the ball averages on any lcsc simple algebraic group (over any local field). Pointwise theorems will be demonstrated also for the natural singular spherical averages on some of the Gelfand pairs.

(6) The proof of a pointwise ergodic theorem for actions of the free groups, generalizing Birkhoff's and Wiener's theorems for $\mathbb{Z}$ and $\mathbb{Z}^d$ will be described, using the general spectral method referred to above.

(7) The derivation of pointwise ergodic theorems for actions of simple algebraic groups with an explicit exponentially fast rate of convergence to the ergodic mean will be described. The same result will be described also for certain discrete lattice subgroups.

(8) Some ergodic theorems for semisimple Lie groups of real rank at least two will be described, which are in marked contrast to the results that Euclidean analogs might suggest, a contrast which has its roots in the exponential volume growth on semisimple groups.

Our goal is to elucidate some of the main ideas used in the proof of the pointwise ergodic theorems alluded to above. Our account of the pointwise ergodic theorems for groups with polynomial volume growth will be quite detailed, as these results are very recent and have not appeared before elsewhere. However, in the case of the spectral method, we have specifically attempted to give an account of the proofs which is as elementary as possible and demonstrated using the simplest available examples. The motivation for these choices is that the spectral methods which we employ require considerable background in the structure theory and representation theory of semisimple Lie groups, as well as classical singular integral theory. At this time these methods are not yet part of the standard tool kit in ergodic theory, and consequently it seems appropriate to give an exposition which focuses on the ergodic theorems and explains some of the main ideas in their proof, but requires as little as possible by way of background.

We have also tried to emphasized the pertinent open problems in the theory, many of which are presented along the way.

Ergodic theorems for actions of connected Lie groups, and particularly equidistribution theorems on homogeneous spaces and moduli spaces, have been developed and used in a rapidly expanding array of applications, many of which are presented in the two volumes of the present handbook. Thus it seems reasonable to limit the scope of our discussion in the present exposition and concentrate specifically on pointwise ergodic theorems, which have not been treated elsewhere.

We must note however that even within the more limited scope of pointwise ergodic theorems for general group actions our account has some important omissions. We mention some of these below, and offer as our rationale the fact that there already exist good expositions of these topics in the literature, some of which are referred to below. These omissions includes the analytic theory of homogeneous nilpotent Lie groups, and in particular the extensive theory of convolution operators, harmonic analysis, maximal functions and pointwise convergence theorems for diverse averages on Euclidean spaces, Heisenberg (-type) groups, homogeneous nilpotent groups and harmonic AN-groups (see [141] and [38] for an introduction to some of these topics). They also include the extensive results on equidistribution on homogeneous spaces (see [41] and [138] for surveys, [57] for some new results), as well as the general theory of actions of amenable locally compact second countable (lcsc) groups (see [121], and [134,92,154] for more recent results). Another omission is the mean ergodic theorem for semisimple groups proved in [150], and other ergodic theorems on moduli spaces and their applications, which are described in detail in the present volume.

It is also natural to include in a discussion of maximal inequalities for group actions a discussion of convolution operators, particularly radial averages on general lcsc groups and their homogeneous spaces. This subject, for which the theory is very incomplete receives only very scant mention here, and we refer to [111] for a short survey and many open problems.

2. Averaging along orbits in group actions

2.1. *Averaging operators*

Let G be a locally compact second countable (lcsc) group, X an lcsc space on which G acts (continuously) as a group of homeomorphisms, or more generally, a standard Borel space on which G acts (measurably) by Borel automorphisms. Let m be a G-invariant σ-finite Borel measure on X. The G-action on X gives rise to a representation π of G as a group of isometries of the Banach spaces $L^p(X)$, given by $\pi(g)f(x) = f(g^{-1}x)$, and for $1 \leqslant p < \infty$ the representation $\pi : G \to \mathrm{Iso}(L^p(X))$ of G into the isometry group is strongly continuous.

For any Borel probability measure on G, we can consider the averaging operator given, for every $f \in L^p(X)$, $1 \leqslant p \leqslant \infty$, by $\pi(\mu)f(x) = \int_G f(g^{-1}x)\,d\mu(g) = \int_G \pi(g)f(x)\,d\mu(g)$. The last equation is well-defined, and does indeed determine unambiguously an element of $L^p(X)$, and let us very briefly recall the well-known arguments proving this fact. Fix two Borel measurable functions f_i, $i = 1, 2$, on X, which have finite L^p-norm and are equal almost everywhere, and $h \in L^q(X)$, q the dual exponent. Then by Fubini's theorem for m-almost all $x \in X$ the two functions $g \mapsto f_i(g^{-1}x)h(x)$ are μ-integrable and equal, and so the values $\int_G f_i(g^{-1}x)\,d\mu(g) = \pi(\mu)f_i(x)$ are equal for m-almost all $x \in X$. Hence the latter integral, denoted by $\pi(\mu)f$, uniquely determines a function class (namely up to m-measure zero) for any $f \in L^p(X)$. Furthermore, $\pi(\mu)f$ has finite L^p-norm, and in fact by Hölder's inequality $\|\pi(\mu)f\|_p \leqslant \|f\|_p$. In addition, $\pi(\mu)f$ defined above coincides, as an element of $L^p(X)$, with the Lebesgue integral (w.r.t. μ) of the Banach-space valued measurable function (strongly continuous if $1 \leqslant p < \infty$) given by $g \mapsto \pi(g)f$ from G to $L^p(X)$. Detailed proofs of these well-known facts can be found in [44, Chapter III, §11, Theorem 17, Chapter VIII, §7].

2.1.1. *The regular representation and the action by convolutions.* Consider the case where $X = G$, and the measure $m = m_G$ is left-invariant Haar measure. Then the operators $\lambda(g)f(x) = f(g^{-1}x)$ are isometric in every $L^p(G, m_G)$, and $g \mapsto \lambda(g)$ is the left regular representation. If μ is absolutely continuous with density $d\mu = b(g)\,dm_G(g)$, then the operator $\lambda(\mu)$ is the operator of left convolution by μ:

$$\lambda(\mu)f(x) = \int_G f\big(g^{-1}x\big)b(g)\,dm_G(g) = \mu * f(x).$$

We can similarly consider the right regular representation of G, given by $\rho(g)f(x) = f(xg)$. Of course, in general $\rho(g)$ is *not* an isometric operator in $L^p(G, m_G)$, unless the left Haar measure m_G is also right-invariant, namely unless G is unimodular. Note also that the operator $\rho(\mu)$ is given by $\rho(\mu)f(x) = \int_G f(xg)b(g)\,dm_G(g)$, and is equal to the convolution operator $f * \mu^\vee$ if and only if G is unimodular (here $\mu^\vee(A) = \mu(A^{-1})$). In particular if G is unimodular and the measure μ is symmetric (namely satisfies $\mu^\vee = \mu$), then $\rho(\mu)f = f * \mu$.

2.2. *Ergodic theorems*

We will generally consider a family of probability measures μ_t ($t \in \mathbb{R}_+$) on G, such that $t \mapsto \mu_t$ is w^*-continuous as a map into $M(G) = C_0(G)^*$. Usually each μ_t will have compact support, depending on t.

We will focus our attention on the case of probability-measure preserving actions, namely we assume that $(X, \mathcal{B}, m)$ is probability space, and G is a group of measure-preserving transformations $g : X \to X$, and $m(gA) = m(A)$. Here the probability measures μ_t on G can be regarded as family of averaging operators producing a sampling method along the group orbits $G \cdot x$ of G in X, via $\pi(\mu_t) f(x) = \int_G f(g^{-1}x)\, d\mu_t(g)$.

Ergodicity of the G-action is defined as usual by the condition that every G-invariant integrable function is constant almost everywhere.

Our main goal below will be to establish a pointwise ergodic theorem in L^p for interesting families of averages μ_t on G, in a general ergodic action (X, m). By that we mean establishing the following convergence theorem:

$$\lim_{t\to\infty} \pi(\mu_t) f(x) = \int_X f\, dm.$$

For m-almost every $x \in X$, and in the L^p-norm, for all $f \in L^p(X)$, where $1 \leqslant p < \infty$.

The ergodicity condition above is of course equivalent to the condition that the σ-algebra $\mathcal{I}$ of G-invariant sets is the trivial subalgebra of $\mathcal{B}$, consisting of sets of measure zero or one. In a general, not necessarily ergodic action we can consider the conditional expectation operator $\mathcal{E} : L^1(X, \mathcal{B}) \to L^1(X, \mathcal{I})$, and the sampling error along an orbit takes the form $|\pi(\mu_t) f(x) - \mathcal{E} f(x)|$. We recall that it is a well-known consequence of the standard ergodic decomposition theorem that in order to establish a pointwise ergodic theorem in an arbitrary action, it suffices to establish it for ergodic actions. In more detail, if for every $f \in L^p(X)$ $\lim_{t\to\infty} \pi(\mu_t) f(x) = \int_X f\, dm$ almost everywhere for every probability-preserving ergodic action, then $\lim_{t\to\infty} \pi(\mu_t) f(x) = \mathcal{E} f(x)$ almost everywhere for every probability-preserving action. We shall therefore often assume in what follows that the G-action is ergodic, when convenient.

In general, one would like to allow as many choices of sampling methods μ_t as possible, since different choices play absolutely crucial roles in different applications. A very basic distinction that arises is between the following cases:

(1) μ_t *is absolutely continuous* w.r.t. Haar measure on G. A fundamental question is to understand the case when $\mu_t = \beta_t$ are the normalized averages on a ball of radius t and center e w.r.t. an invariant metric on G. This includes, for example, an invariant Riemannian metric on a connected Lie group, or a word metric on an lcsc group.

(2) μ_t *is a singular measure*, namely non-atomic and supported on a closed subset of G of Haar measure zero. This includes, when G is a Lie group, closed submanifolds of positive codimension in G, for example, $\mu_t = \sigma_t$ the normalized averages on a sphere of t and center e defined using the restriction to the sphere of an invariant Riemannian metric.

(3) μ_t *is a discrete* (*atomic*) *measure*. This includes averages supported on discrete subgroups, for example, when μ_t is supported on a lattice subgroup of G, or averages

supported on integer points in a subvariety, when G is a real algebraic group defined over $\mathbb{Q}$, for example.

We will begin by considering some absolutely continuous measures and some singular measures, and will comment on the important and interesting problem of discrete averages later, in Sections 10.5, 12.3 and 13. It is interesting to note that our discussion below will make it clear that the analysis of singular averages (for example, sphere averages) is a natural and indispensable ingredient in developing the theory of absolutely continuous averages, when the groups in question have exponential volume growth.

We note also that we will devote much of our attention in what follows to the study of averages defined geometrically via an invariant metric on the group, such as ball averages, shell averages and spherical averages. However, the spectral methods described in Sections 9–12 below apply more generally and are not confined to radial averages. For semisimple groups, for example, we will consider in Section 12 also horospherical and many other non-radial averages.

2.3. *Maximal functions*

The maximal function associated with the family of averaging operators μ_t is defined by, for $f \in L^p(X)$:

$$M_\mu^* f(x) = f_\mu^*(x) = \sup_{t \in \mathbb{R}_+} \left|\pi(\mu_t) f(x)\right|.$$

Let us hasten to note that in general it is not a-priori clear that the maximal function is well-defined and measurable for function classes in $L^p(X)$, and indeed, this is not always the case (see Section 2.3.1 below for further discussion).

We recall that a strong maximal inequality is an L^p-norm inequality for the maximal function, of the form $\|M_\mu^* f\|_p \leqslant C_p \|f\|_p$, $\forall f \in L^p(X)$, where $1 < p \leqslant \infty$. A weak-type maximal inequality is an estimate of the distribution function associated with the maximal function. Of particular interest is the weak-type $(1, 1)$ maximal inequality given by:

$$m\left\{x \in X;\, f_\mu^*(x) > \delta\right\} \leqslant \frac{C}{\delta}\|f\|_1, \quad \forall f \in L^1(X).$$

In the case of probability-measure-preserving actions it is natural to consider the maximal function given (for ergodic actions) by the following formula:

$$\tilde{f}_\mu^*(x) = \sup_{t \geqslant 0} \left|\pi(\mu_t) f(x) - \int_X f\, dm\right|.$$

The quantity $\int_X f\, dm$ is called the space average of the function f, and $\pi(\mu_t) f(x)$ are called the time averages along the orbit $G \cdot x$. Therefore, $\tilde{f}_\mu^*(x)$ is a bound for the largest error performed when sampling the values of f along the G-orbit of $x \in X$, using μ_t as the sampling method. The strong L^p-Maximal inequality: $\|\tilde{f}_\mu^*\|_p \leqslant C_p \|f\|_p$ bounds the total

size of the error performed during the sampling process by the size of the function sampled, size being measured here by the L^p-norm. Similarly, the weak-type maximal inequality bounds the size of the set where the sampling error is larger than a fixed positive constant in terms of the size of the L^1-norm of f. In the case of general, not necessarily ergodic actions, the space average $\int_X f\,dm$ must be replaced by the conditional expectation $\mathcal{E}f$ of f w.r.t. the σ-algebra of G-invariant functions. Of course, when (X,m) is a probability space, the operators $\tilde{f}^*_\mu$ and f^*_μ satisfy exactly the same maximal inequalities and so we will continue with the analysis of just one of them, whose choice will be dictated by the problem at hand.

Finally, we recall that $L(\log^k L)(X)$ denotes the subspace of $L^1(X)$ consisting of functions for which $\int_X |f(\log^+ f)^k|\,dm(x)$ is finite.

2.3.1. *Measurability of the maximal function.* The maximal function is *not* necessarily a well-defined measurable function, even for very natural choices of the averaging operators μ_t. Let us first note that if the averages μ_t are all absolutely continuous with respect to Haar measure on G, and $t \mapsto \mu_t$ is continuous w.r.t. the $L^1(G)$ norm, then for every $f \in L^p(X)$, and for m-almost every $x \in X$, $t \mapsto \pi(\mu_t)f(x)$ is a continuous function of t, see, e.g., [44, Chapter VIII, §7, p. 686]. It then follows that f^*_μ is indeed measurable, since the supremum may be taken of the countable set of rational numbers. However, absolute continuity is not a necessary condition for measurability of the maximal function, and we will discuss below many singular averages which give rise to a measurable maximal function.

Indeed absolute continuity is not necessary even for the continuity of $t \mapsto \pi(\mu_t)f(x)$, for almost every $x \in X$. Thus, for example, when $\mu_t = \sigma_t$ are the sphere averages on $\mathbb{R}^n$, $n \geqslant 3$, such a result for say bounded functions with bounded support, is established in [36, II.4], see also [141, Chapter XI, §3.5].

Interestingly, for $\mathbb{R}^n$ (and other connected Lie groups) measurability of the maximal functions seems closely connected to considerations related to curvature, as is indicated by the following fact. Let ∂Q_t denote the family of sets in the group $G = \mathbb{R}^2$ given by the boundaries of squares Q_t centered at the origin, and let q_t denote the usual (uniformly distributed, linear Lebesgue) probability measure on ∂Q_t. Let σ_t be the rotation-invariant probability measure on the circle $S_t = \partial B_t$ of radius t centered at the origin. Then the maximal function f^*_q is *not* always measurable, even when f is the characteristic function of a measurable set of finite measure, but f^*_σ is measurable for such f [14].

Clearly, when the action of G on X is a continuous action on a locally compact second countable (lcsc) space, the maximal function $f^*_\mu(x) = \sup_{t\in\mathbb{R}_+} |\pi(\mu_t)f(x)|$ is certainly measurable provided that $f \in C_c(X)$ and in addition $t \mapsto \int_G f(g^{-1}x)\,d\mu_t$ is continuous for every $f \in C_c(X)$. This condition will be satisfied by all the averages we will discuss below. A maximal inequality for f^*_μ when $f \in C_c(X)$ serves as an *a-priori* inequality, and is used to extend both the measurability of f^*_μ as well as the maximal inequality to the appropriate Lebesgue spaces. We refer to [106, Appendix A] and [114, §2.2] for more information on these arguments.

2.4. *A general recipe for proving pointwise ergodic theorems*

A proof of a pointwise ergodic theorem for a family of bounded operators $\pi(\mu_t) = T_t$ acting in $L^p(X, m)$ for some $1 \leqslant p < \infty$ can be obtained using the following four-step recipe.

(1) *Prove a mean ergodic theorem in* $L^p(X)$ for the averages, namely show that $\lim_{t\to\infty} \|T_t f - \int_X f\, dm\|_p = 0$.
(2) *Find a dense subspace of functions* $V \subset L^p(X)$ *for which pointwise convergence holds*, namely $\lim_{t\to 0} T_t f(x) = \int_X f dm$, for almost all $x \in X$, and for all $f \in V$.
(3) *Establish a strong* L^p*-maximal inequality* for the maximal function $f^*(x) = \sup_{t\geqslant 0} |T_t f(x)|$, where $f \in L^p(X)$, namely $\|f^*\|_p \leqslant C_p \|f\|_p$.
(4) *Use interpolation theory, either real or complex*, to establish a maximal inequality for the action of T_t in $L^s(X)$, $s \neq p$.

We recall that the fact that (2) and (3) taken together imply pointwise convergence almost everywhere of $T_t f(x)$ for *every* $f \in L^p(X)$ is a formulation of the well-known Banach principle (see, e.g., [53]). The identification of the limit is achieved in (1), by the mean ergodic theorem. Obviously, many variations are possible on this basic theme, for example, the use of weak-type maximal inequalities, variational maximal inequalities, establishing the maximal inequality in (3) only on an a-priori dense subspace, as well as applying a wide array of interpolation methods.

Thus the basic problems we shall address below are establishing maximal inequalities for the family μ_t, the corresponding pointwise convergence theorems for a dense subspace, the identification of the limit in the ergodic theorem, and applying interpolation techniques in the Lebesgue spaces $L^p(X)$.

3. Ergodic theorems for commutative groups

3.1. *Flows of* 1*-parameter groups: Birkhoff's theorem*

In order to motivate the discussion below, let us start very concretely by considering one of the most basic example in ergodic theory. This will serve as our point of departure for several later developments.

EXAMPLE 3.1 (*Lines in* $\mathbb{R}^2$ *with an irrational slope*). Let $\ell = sw, s \in \mathbb{R}$, be a line in $\mathbb{R}^2$ with an irrational slope, and let f be a function on the space $X = \mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ (i.e. $\mathbb{Z}^2$-periodic function on $\mathbb{R}^2$). Let $T_s = R_{sw} : X \to X$ denote the transformation given by translation on X, and let m denote the translation-invariant probability measure on X. We note that m is in fact the *unique* probability measure on X invariant under the transformation T_s, $s \neq 0$ a fact which is immediate upon considering the Fourier coefficients of such an invariant measure. The time averages of f along the orbit of ℓ passing through x are defined by $\beta_t f(x) = \frac{1}{2t} \int_{-t}^{t} f(T_s x)\, ds$.

Recall the following classical results for the probability space (X, m) and the transformations T_s:

(1) *Weyl's Equidistribution Theorem.* $\forall f \in C(X), \forall x \in X$:

$$\lim_{t\to\infty} \beta_t f(x) = \int_X f(u)\, dm(u) = \text{space average of } f.$$

(2) *Wiener's Differentiation Theorem.* $\forall f \in L^1(X)$:

$$\lim_{t\to 0} \beta_t f(x) = f(x), \quad \text{for almost all } x \in X.$$

(3) *Birkhoff's Pointwise Ergodic Theorem for flows.* $\forall f \in L^1(X)$:

$$\lim_{t\to\infty} \beta_t f(x) = \int_X f\, dm, \quad \text{for almost all } x \in X.$$

(4) *Birkhoff's Pointwise Ergodic Theorem for invertible transformations.* The averages $\beta_n(\mathbb{Z}) f(x) = \frac{1}{2n+1} \sum_{k=-n}^{n} f(T_1^k(x)$, satisfy conclusion (1) and (3), namely $\frac{1}{2n+1} \sum_{k=-n}^{n} f(T_1^k x)$ have the same convergence properties as their continuous analogs.

Anticipating some later developments let us note that the classical ergodic averages β_t are absolutely continuous measures on the line ℓ, and in fact constitute the normalized averages on a ball of radius t in $\mathbb{R}$, w.r.t. an invariant Riemannian metric (which is unique here up to scalar). $\beta_n(\mathbb{Z})$ are the normalized ball averages w.r.t. the induced invariant metric on the integer lattice $\mathbb{Z}$. This metric is here also a word metric on $\mathbb{Z}$ w.r.t. the set of generators $\{\pm 1\}$.

Let us further note the following:

(1) The equidistribution theorem stated in (1) holds in fact under the sole condition that X be a compact metric space, and T_s a homeomorphism (or a 1-parameter group of homeomorphisms) possessing a *unique* invariant probability measure. The case $X = \mathbb{T}^n$ was originally considered by H. Weyl.

(2) The differentiation theorem stated in (2) holds in fact for any standard measure space (X, m), and any 1-parameter group of measure-preserving transformations. This result is due to N. Wiener [155, Theorem III′], and his proof combines Lebesgue's and Hardy–Littlewood's [68] differentiation theorems on the real line with a principle of local transfer to a general measure-preserving flow (see the discussion in Section 5.4.2 for more details).

(3) The pointwise ergodic theorems stated in (3) and (4) hold in fact for any probability space (X, m), and any invertible measure preserving transformation or 1-parameter group, under the sole condition of ergodicity. This is the content of Birkhoff's theorem [12].

Actions of the real line with an invariant probability measure exist in great abundance. We mention briefly the following examples, focusing mostly on those that will reappear again later.

EXAMPLE 3.2.

(1) Any complete, divergence free vector field on a Riemannian manifold generates an associated volume-preserving flow. The total Riemannian volume if of course finite if the manifold is compact.
(2) In particular, the geodesic flow on a compact Riemannian manifold gives an $\mathbb{R}$-action preserving a finite volume. This includes the geodesic flow on a compact (or more generally, finite volume) locally symmetric space.
(3) The horocycle flow on a compact (or finite volume) surface of constant negative curvature also gives a volume-preserving flow. Similarly, analogous flows can be defined for all locally-symmetric spaces associated with semisimple Lie groups, by considering actions of one-parameter unipotent subgroups.
(4) Any 1-parameter subgroup of the connected Lie group G acting by translations on G/Γ, where Γ is a discrete lattice subgroup of G.

Actions of $\mathbb{Z}$ with invariant measure are just as abundant. Indeed, both $\mathbb{Z}$ and $\mathbb{R}$ have the all-important property that each of their actions by homeomorphisms of a compact metric space possesses at least one invariant ergodic measure, and this fact gives of course rise to a vast collection of examples. We mention here however only one, which will be particularly important in what follows. Let G be any lcsc group, and $\Gamma \subset G$ a lattice. Then any element $g \in G$, as well as any subgroup $H \subset G$ act by measure-preserving transformations on the probability space $(G/\Gamma, m)$. This includes the case where Γ is a lattice in a simple non-compact algebraic group over a locally compact non-discrete field. In the latter case, the Howe–Moore mixing theorem asserts the remarkable fact that the action of every element g is not only ergodic, but in fact mixing, provided only that the powers of g are not confined to a compact subgroup, and that G has no compact factors. We recall that mixing means that the correlations $\langle T^n f, f' \rangle$ converge to zero, if f or f' has zero integral.

We remark that equidistribution for *every* orbit does not hold in many of the examples above, e.g., for the case of geodesic flows on compact surfaces of constant negative curvature. Convergence for *every* orbit fails even if the function is assumed continuous, or even smooth. The restriction to *almost* every starting point is thus essential in the pointwise ergodic theorem.

3.2. *Flows of commutative multi-parameter groups: Wiener's theorem*

Naturally, the next problem to consider is the generalization of the pointwise ergodic theorem from the case of a one-parameter flow to that of several commuting flows, and in particular from ball averages in actions of $\mathbb{R}$ (or $\mathbb{Z}$) to ball averages in actions of $\mathbb{R}^d$ (or $\mathbb{Z}^d$). This problem was solved by N. Wiener in [155], where he introduced several key ideas that came to play an important role in the further development of ergodic theory for groups with polynomial volume growth, and more generally amenable groups. We will focus below on Wiener's covering lemma [155, Lemma C′], and the introduction of a transfer principle [155, Proof of Theorem IV′]. These arguments are geometric by nature and in particular, as already noted by Wiener, they apply equally well to ball averages on $\mathbb{R}^d$ or the lattice $\mathbb{Z}^d$, as well as to many other averages, e.g., on Euclidean cells centered at the origin (see

also Pitt's discussion [126]). Even further, they were generalized by A. Calderon [19] to apply to certain non-commutative groups with polynomial volume growth. In Section 5 we will present these arguments and generalizations thereof, but before delving into the proofs, it may be instructive to consider some examples to which Wiener's (or Pitt's) theorem applies.

EXAMPLE 3.3.

(1) Let $\mathcal{L}_n$ denote the space of unimodular lattices in $\mathbb{R}^n$. Clearly $SL_n(\mathbb{R})$ acts on $\mathcal{L}_n$, and the action is easily seen to be transitive. The stability group of the lattice $\mathbb{Z}^n$ is $SL_n(\mathbb{Z})$, so that we can identify $\mathcal{L}_n$ with the homogeneous space $SL_n(\mathbb{R})/SL_n(\mathbb{Z})$. The latter space carries a finite measure m which is $SL_n(\mathbb{R})$-invariant, and we normalize it to be a probability measure. We fix a given non-trivial representation of $SL_3(\mathbb{R})$ in $SL_n(\mathbb{R})$, and let $A \cong \{a_t b_s \mid (t,s) \in \mathbb{R}^2\}$ denote the two-dimensional vector group of diagonal matrices in $SL_3(\mathbb{R}) \subset SL_n(\mathbb{R})$, $n \geqslant 3$. Then for all $f \in L^1(\mathcal{L}_n)$ and for almost every $x \in \mathcal{L}_n$

$$\lim_{r\to\infty} \frac{1}{\pi r^2} \int_{t^2+s^2\leqslant r^2} f(a_t b_s x)\, dt\, ds = \int_{\mathcal{L}_n} f(x)\, dm.$$

(2) Let $A_1, \ldots, A_d$ be d commuting $n \times n$ matrices with integer entries and determinant $\{\pm 1\}$, and consider their natural action on $\mathbb{T}^n = \mathbb{R}^n/\mathbb{Z}^n$ by group automorphisms. Then, for every $f \in L^1(\mathbb{T}^n)$ and almost every $x \in \mathbb{T}^n$

$$\lim_{r\to\infty} \frac{1}{(2r+1)^d} \sum_{-r\leqslant i_j\leqslant r} f\left(A_1^{i_1} \cdots A_d^{i_d} x\right) = \int_{\mathbb{T}^n} f(x)\, dm$$

provided every integrable function invariant under $A_1, \ldots, A_d$ is constant. This is the case if at least one of the A_i has no roots of unity as eigenvalues, for example.

(3) Consider the compact Abelian group $K = (\mathbb{Z}/2\mathbb{Z})^{\mathbb{Z}^2}$, with Haar measure given by the natural product measure. The $\mathbb{Z}^2$-action by coordinate shifts is an ergodic measure-preserving action by automorphisms of the compact Abelian group K. Denoting by a_1 and b_1 unit translations in the direction of the axes, we have: $\forall f \in L^1(K)$ and for almost every $x \in K$

$$\lim_{n\to\infty} \frac{1}{|B_n \cap \mathbb{Z}^2|} \sum_{(k,m)\in\mathbb{Z}^2;\, k^2+m^2\leqslant n^2} f(a_k b_m x) = \int_K f(x)\, dm.$$

The example above can of course be greatly generalized, by replacing $\mathbb{Z}^2$ be other infinite Abelian groups, replacing $\mathbb{Z}_2$ by other compact groups, and also considering the extensive collection of closed shift-invariant subgroups arising in such actions. This gives rise to a wealth of ergodic measure preserving actions of multi-parameter groups by automorphisms of compact Abelian groups. For a discussion of this topic and its connections to commutative algebra and number theory we refer to [135].

In Section 5 we will considerably expand the scope of the discussion, formulate the ball averaging problem in ergodic theory and describe the complete solution of the ball averaging problem for all lcsc groups with polynomial volume growth. This result constitutes a common generalization of the pointwise and maximal ergodic theorems of Birkhoff, Wiener and Calderon, and brings to a satisfactory close a long line of development in ergodic theory. The result depends on some very recent developments and has not appeared before elsewhere. The proof is based on the arguments of Wiener and Calderon, together with one further ingredient, namely the asymptotic invariance (under translation) of the balls in these groups. This fact was recently established for connected Lie groups of polynomial volume growth by two interesting independent arguments which we will describe. In the general case of lcsc group asymptotic invariance of the balls ultimately depends also on considerations introduced by M. Gromov in his celebrated theorem [63] asserting that groups with polynomial volume growth are virtually nilpotent.

We will have on a number of occasions to use results on the (polynomial or exponential) volume growth of balls in most of the groups that will come under consideration below, and so we therefore now turn to this issue.

4. Invariant metrics, volume growth, and ball averages

4.1. *Growth type of groups*

Let G be a compactly generated locally compact second countable (lcsc) group. If V is a compact set generating G, namely $\bigcup_{n\in\mathbb{N}} V^n = G$, we can define a distance function $|g|_V$ on G via $|g|_V = \min\{n;\ g \in V^n\}$, where $V^0 = \{e\}$ and $V^n = V \cdot V \cdots V$ is the set of n-fold products of elements of V. The distance function is inversion-invariant if and only if the set V is symmetric, namely $V = V^{-1}$, a condition equivalent to the symmetry of the function $d_V(g,h) = |g^{-1}h|_V$. We will assume $V = V^{-1}$ from now on, and then the associated function d_V is a left-G-invariant metric on G, which we will call the word metric determined by V. Since G is lcsc, for some $n \in \mathbb{N}$ V^n contains an open set and then $m_G(V^{n+k}) > 0$ for all $k \in \mathbb{N}$, where m_G is (any) Haar measure. Clearly $V^n V^k = V^{n+k}$, and thus the sequence $\log m_G(V^n)$ is subadditive, and it follows that the limit $\lim_{n\to\infty} \frac{1}{n} \log m_G(V^n) = h_V$ exists. It is straightforward to see that if $h_V > 0$ for some V, then $h_{V'} > 0$ for any other compact generating set V' (and any Haar measure). If $h_V > 0$ G is called a group of exponential volume growth, and if $h_V = 0$ for some, or equivalently, all compact generating sets V, G is called a group of subexponential growth (and it is then necessarily unimodular). In that case we can consider the quantity $\limsup_{n\to\infty} \frac{\log m_G(V^n)}{\log n} = q_V$, and we recall the well known fact that if $q_V < \infty$ for one compact generating set V, then $q_{V'} < \infty$ for any other compact generating set V'. In fact, in this case there exists a unique $0 < q(G) < \infty$ depending only on G, such that the ratio $m_G(V^n)/n^q$ is bounded for $n \in \mathbb{N}$, for any V as above. This follows immediately from the fact that for large enough k, a ball of radius k of one metric contains a ball of radius ck of the second metric, and is contained is a ball of radius Ck of the second metric, where $c \leqslant C$ are fixed positive constants. If $q_V < \infty$ for some compact generating set then G is called a group of polynomial volume growth, and if $h_V = 0$ and $q_V = \infty$ for some compact generating set V, then G is called a group

of intermediate growth. As is well known, this possibility does in fact arise (see [42] for an accessible exposition of such a group, constructed by R. Grigorchuk). However, such groups do not arise as subgroups of connected Lie groups or algebraic groups.

We have used the metrics of the form d_V described above to define the growth type of a group G. There are of course many other left-invariant metrics on a group, but since invariant metrics can be easily rescaled, not all of them will give the same growth type. For example, taking the metric $d(t,s) = \log(1+|t-s|)$ on $G = \mathbb{R}$, we obtain an invariant metric whose balls centered at 0 (denoted B_t) satisfy $m_{\mathbb{R}}(B_t) = \exp(t) - 1$, namely have exponential growth. It therefore natural to define a left-invariant metric d on G to be admissible, if the quantities $h_d = \lim_{t\to\infty} \frac{1}{t} \log m_G(B_t)$ and $q_d = \limsup_{t\to\infty} \frac{\log m_G(B_t)}{\log t}$ exhibit the same behavior as the quantities h_V and q_V defined by the metrics d_V associated with compact generating sets V. In other words, $h_d > 0$ iff $h_V > 0$, and $q_d < \infty$ iff $q_V < \infty$. In the sequel, we will consider only admissible metrics on G.

4.2. *Invariant metrics*

In general, we can let G be an lcsc group, and let $N : G \to \mathbb{R}_+$ be a continuous proper function satisfying $N(gh) \leqslant C(N(g) + N(h))$, namely a quasi-norm on G. Then $d_N(g,h) = N(g^{-1}h)$ defines a left-invariant quasi-metric on G. We can consider the sets $B_t^N = \{g \in G; N(g) \leqslant t\}$ and the normalized measures β_t^N with density $\chi_{B_t^N}/m_G(B_t^N)$. When G is say a connected Lie group and N is sufficiently regular, it is also possible to define for every positive radius t the natural probability measure σ_t^N supported on $\partial B_t^N = S_t^N = \{g \in G; N(g) = t\}$. Our discussion below can be in principle extended to this more general context, but in the interest of simplicity we will restrict ourselves to a discussion of functions of the form $N(g) = d(e,g)$ where d is an invariant admissible metric (so that the constant C in the definition of a quasi-norm is equal to 1). We remark that one interesting natural set of examples consist of homogeneous invariant quasi-norms on homogeneous connected nilpotent Lie groups. Such quasi-norms give rise to a quasi-distance which is not necessarily a metric, but in fact an equivalent homogeneous invariant norm can always be found. We refer, e.g., to [141, Chapter XIII, §5 and §7B] and [69] for more on this topic.

Another important general example of a left-invariant metric is the function $d(g,h) = \log(1 + \|\tau(g)^{-1}\tau(h)\|)$, where $\|\cdot\|$ is a symmetric operator (or even just linear) norm on $M_n(\mathbb{R})$, and $\tau : G \to GL_n(\mathbb{R})$ is a faithful linear representation (and we assume that d is admissible).

It is natural to introduce certain properties of metrics that facilitate the discussion and are relevant to questions of growth and ball averaging. Let us start by defining two notions of equivalence between metrics.

DEFINITION 4.1. Consider two metric space (X,d) and (X',d') and a function $f : X \to X'$.

(1) $Y \subset X$ is called C-dense in X if there exists $C \geqslant 0$ such that every $x \in X$ is at distance at most C from Y.

(2) f is called a *Quasi-isometry*, if f satisfies, for some $B, b > 0$ and all $x, y \in X$

$$\frac{1}{B}d'\big(f(x), f(y)\big) - b \leqslant d(x, y) \leqslant Bd'\big(f(x), f(y)\big) + b$$

and in addition $f(X)$ is C-dense.

(3) f is called a *Coarse-isometry*, if f satisfies, for some $b \geqslant 0$ and all $x, y \in X$

$$d'\big(f(x), f(y)\big) - b \leqslant d(x, y) \leqslant d'\big(f(x), f(y)\big) + b$$

and in addition $f(X)$ is C-dense.

Next, consider the following possibilities regarding approximate analogs to geodesics in a metric space.

DEFINITION 4.2. Let (X, d) be a metric space. The metric d is called

(1) *Discretely coarsely geodesic* [2] if there exists $C \geqslant 0$ such that for any $x, y \in X$ it is possible to find a finite sequence of points $x = x_0, x_1, \ldots, x_n = y$ satisfying $d(x_{i-1}, x_i) \leqslant C$, $1 \leqslant i \leqslant n$, and $d(x, y) = \sum_{i=1}^{n} d(x_{i-1}, x_i)$.

(2) *Asymptotically geodesic* [15] if for every $\varepsilon > 0$ there exists $C_\varepsilon > 0$, such that given $x, y \in X$, it is possible to find a finite sequence of points $x = x_0, x_1, \ldots, x_n = y$ satisfying $d(x_{i-1}, x_i) \leqslant C_\varepsilon$, $1 \leqslant i \leqslant n$, and $d(x, y) \geqslant (1 - \varepsilon) \sum_{i=1}^{n} d(x_{i-1}, x_i)$.

(3) *Monotone* [148] if there exists $C \geqslant 1$ such that for any $x, y \in X$ it is possible to find a finite sequence of points $x = x_0, x_1, \ldots, x_n = y$ satisfying $d(x_{i-1}, x_i) \leqslant C$, and $d(x, x_{i-1}) + 1 \leqslant d(x, x_i)$, $0 < i \leqslant n$.

The precise behavior of the volume growth function $m_G(B_t)$ for balls defined by an admissible metric d on an lcsc group G is a very important characteristic from several perspectives, and is fundamental in consideration related to ergodic theorems, as we shall see below. The volume growth problem has seen some important recent progress, which we will describe in Sections 4.4 and 4.5. But before doing so, let us first formulate the following basic problem in ergodic theory.

4.3. *The ball averaging problem in ergodic theory*

Our discussion so far, including of course the description of the basic pointwise ergodic theorems of Birkhoff and Wiener for Abelian groups, leads naturally to the formulation of the following problems.

Let G be an lcsc group, d an admissible metric on G, B_t the corresponding balls of radius t and center e, and let β_t be normalized ball averages on G. By this we mean that β_t are the absolutely continuous probability measures on G whose density with respect to left Haar measure m_G on G is the characteristic function of B_t, normalized by $m_G(B_t)$.

When G is a connected Lie group, a particularly basic case arises when d is the distance function on G associated to a left-invariant Riemannian metric on G. Note that in this case,

we can also consider the spheres S_t, and the probability measures σ_t supported on S_t, which is given canonically by the volume form arising from the restriction of the Riemannian metric to the spheres. In general, when G is lcsc and the metric assumes integer values, we can consider the sequence σ_t, $t \in \mathbb{N}$, of probability measures defined by the restriction of a left Haar measure to the sphere. Anticipating some of the developments of the succeeding sections, we formulate the following.

(1) *The ball averaging problem*: Establish when β_t satisfies the pointwise ergodic theorem in L^1 (or at least in every L^p, $1 < p < \infty$), namely for every ergodic probability-preserving action

$$\lim_{t\to\infty} \pi(\beta_t) f(x) = \int_X f \, dm, \quad \text{for almost all } x \in X$$

and in the L^p-norm.

(2) *Sphere averaging problem*: Establish when there exists $p_0 < \infty$ such that the sphere averages σ_t satisfy the pointwise ergodic theorem in L^p, $p > p_0$, namely

$$\lim_{t\to\infty} \pi(\sigma_t) f(x) = \int_X f \, dm, \quad \text{for almost all } x \in X$$

and in the L^p-norm.

(3) *Spherical differentiation*: For G a connected Lie group, d a Riemannian metric, establish when the singular spherical differentiation theorem holds for $f \in L^p$, $p > p_0$, namely

$$\lim_{t\to 0} \pi(\sigma_t) f(x) = f(x), \quad \text{for a.e. } x.$$

It is evident that the first two problems raised above are inextricably linked with the study of volume growth for balls and spheres, to which we now turn.

4.4. *Exact volume growth*

4.4.1. *Exact polynomial volume growth.* When G has polynomial volume growth the discussion of Section 4.1 allows us only to conclude that there exists a positive number q, such that for any Haar measure and any admissible metric on G, we have the following estimate for the volume of balls defined by the metric: $m_G(B_t) \leqslant C(G,d)t^q$. When G has exponential volume growth h, we can only conclude that $c_\varepsilon(h-\varepsilon)^t \leqslant m_G(B_t) \leqslant C_\varepsilon(h+\varepsilon)^t$ for every $\varepsilon > 0$. As we shall see below, these estimates are entirely unsatisfactory for many purposes, and so it is natural to introduce the following definitions (see [64, Definition I.2]).

DEFINITION 4.3.

(1) The pair (G,d) is said to have exact $t^q e^{ct}$ volume growth if the balls B_t defined by the left-invariant admissible metric d satisfy

$$\lim_{t\to\infty} \frac{m_G(B_t)}{Ct^q \exp ct} = 1$$

for some non-negative q, c and C (which are necessarily uniquely determined).

(2) The pair (G, d) is said to have strict $t^q e^{ct}$ volume growth if

$$bt^q \exp ct \leqslant m_G(B_t) \leqslant Bt^q \exp ct$$

for some non-negative q and c (which are necessarily uniquely determined) and two positive constants $b \leqslant B$.

Important recent progress has been obtained in establishing strict or exact volume growth in several context. We will describe the developments regarding exact growth in the present section, and regarding strict growth in the following one.

The most obvious examples of groups G and metrics d with exact t^q (namely polynomial) volume growth are:

(1) $G = \mathbb{R}^k$, and d the metric defined by any norm on $\mathbb{R}^k$.
(2) $G = H_n = \mathbb{C}^n \times \mathbb{R}$ the Heisenberg groups, with the metric determined by the homogeneous norm $N(z, t) = (\|z\|^4 + t^2)^{1/4}$.
(3) More generally, G a connected nilpotent homogeneous group, with d the metric derived from a homogeneous norm.

In these examples the homogeneity of the metric on G, namely the fact that $B_t = \alpha_t(B_1)$ (where α_t is the dilation automorphism group) immediately implies that $m_G(B_t) = m_G(B_1)t^q$, where q is the homogeneous dimension. Thus in particular the polynomial volume growth for homogeneous metrics is exact. Riemannian metrics are not homogeneous in general, and here, under one further assumption, the following result was established by P. Pansu.

THEOREM 4.4 [123]. *Let G be a simply connected nilpotent Lie group admitting a lattice subgroup Γ. Then*

(1) *G has exact polynomial volume growth w.r.t. any Riemannian metric on G which is Γ-invariant.*
(2) *Any lattice Γ in G has exact polynomial volume growth with respect to any word metric.*

Pansu's theorem has the following two consequences [123] (see also the discussion in [42, Chapter VII.C]).

COROLLARY 4.5.

(1) *Every discrete group containing a finitely generated nilpotent group of finite index has exact polynomial volume growth with respect to any word metric. Indeed, by a well-known result of Malcev, Γ can be embedded as a lattice in a simply connected nilpotent Lie group.*
(2) *Every discrete group Γ with polynomial volume growth has exact polynomial volume growth with respect to any word metric. Indeed, by Gromov's theorem* [63], *Γ is virtually nilpotent, namely contains a nilpotent subgroup of finite index.*

We note that a sharper form of Pansu's theorem has been established by M. Stoll [137], namely that $|\#B_t - Ct^q| \leqslant Bt^{q-1}$ for 2-step nilpotent groups.

For a simply connected nilpotent Lie group G, the condition that G contains a lattice subgroup is equivalent to the condition that the Lie algebra of G admits a basis with rational structure constants. Thus there exists only a countable set of such groups. So while Pansu's theorem (in combination with Gromov's) imply exact volume growth for every discrete group with polynomial growth, it leaves much to be desired in the case of general connected nilpotent Lie groups (and more generally lcsc groups with polynomial growth). Very recently, E. Breuillard has obtained a solution to this problem in the simply connected nilpotent case, in the following form.

THEOREM 4.6 (Exact polynomial volume growth for simply connected nilpotent Lie groups [15]). *Let G be a simply connected nilpotent Lie group G, let H be a closed cocompact subgroup* (*e.g.*, *$H = G$*), *and let d be an H-invariant Riemannian or word metric on G. Then G has exact polynomial volume growth with respect to d. In fact, the conclusion holds for any locally bounded, proper, asymptotically geodesic metric d.*

The method of proof used in [15] is motivated by [123], and is based on showing that the ratio of the volume of the balls defined by d is asymptotically the same as the volume of the balls defined by an appropriate Carnot–Caratheodory metric on G. For the latter the scaling property of the volume is clear, and thus exactness for the volume of the d-balls follows. Furthermore, exactness is also proved in [15] for metrics on some connected non-nilpotent Lie groups of polynomial volume growth. Here use is made of the construction of the nil-shadow of such a group, and this allows the reduction of the problem to the nilpotent case. In view of the structure theorem for general lcsc groups of polynomial volume growth which will be discussed further in Section 5.5, the methods developed in [15] may well lead to a complete solution of the problem of establishing exact growth on an arbitrary lcsc group of polynomial volume growth.

4.4.2. *Exact exponential volume growth.* Important recent progress on exact growth has recently been obtained also for a class of groups with exponential volume growth, namely semisimple Lie groups. Let us first note that a very natural choice for a metric on semisimple groups is the bi-K-invariant Riemannian metric on the group which is associated with the Killing form (K a maximal compact subgroup). The exact $t^{r-1}\exp 2\|\rho\|t$ volume growth (see Section 10.4 for notation) of balls in this case follows, for example, from the sharper results of [86] on the asymptotic volume of the spheres, but can also be proved more directly.

Another natural family of metrics is given as follows. Fix a linear representation τ of G into $GL_n(\mathbb{R})$, fix any (vector-space) norm $\|\cdot\|$ on $M_n(\mathbb{R})$, and let $d(g,h) = \log(1 + \|\tau(g^{-1})\tau(h)\|)$. d will be symmetric if $\|g^{-1}\| = \|g\|$, and satisfy the triangle inequality if $\|gh\| \leqslant \|g\|\|h\|$, e.g., if $\|\cdot\|$ is a symmetric operator norm of $M_n(\mathbb{R})$. The following general result has been established by A. Gorodnik and B. Weiss.

THEOREM 4.7 (Exact volume growth for semisimple Lie groups [57]). *Let G be a connected semisimple Lie group with finite center, $\tau : G \to GL_n(\mathbb{R})$ a linear representation, and $\|\cdot\|$ any norm on $M_n(\mathbb{R})$. Then the sets given by $G_t = \{g \in G;\ d(e,g) \leqslant t\}$ have exact*

$t^q \exp ct$ volume growth, for some q and $c > 0$ depending on the representation τ and the norm $\|\cdot\|$.

4.5. *Strict volume growth*

4.5.1. *Strict polynomial growth.* A thorough investigation of growth as well as strict growth has been conducted by Y. Guivarc'h [64] (see also [78] and [6]). For connected Lie groups with polynomial volume growth, and so in particular for nilpotent Lie groups, the fundamental result on strict t^q-volume growth for word metrics is as follows.

THEOREM 4.8 (Strict volume growth for nilpotent groups [64, Theorem II.3]). *Let G be any connected Lie group of polynomial growth. Then any word metric, and thus also any metric quasi-isometric to a word metric, and in particular invariant Riemannian metric has strict polynomial volume growth.*

Furthermore, it is noted in [64] that since a finitely generated torsion free nilpotent group can always be embedded in a simply connected Lie group, strict volume growth holds for word metrics on countable nilpotent groups. Thus, using Gromov's theorem [63], strict growth holds for countable discrete groups of polynomial volume growth.

4.5.2. *The volume doubling condition.* One fundamental consequence of strict polynomial volume growth is the volume doubling condition, introduced by Calderon. Since our discussion centers around the ball averaging problem, we formulate it in the following form.

DEFINITION 4.9 (*Calderon's doubling volume condition* [19]). G is said to satisfy the doubling volume condition w.r.t. the invariant admissible metric d if the balls $B_t = \{g;\, d(g,e) \leqslant t\}$ satisfy $m_G(B_{2t}) \leqslant C(G,d) m_G(B_t)$, for all $t > 0$. Here m_G denote some (left or right) Haar measure on G.

REMARK 4.10.

(1) Clearly the volume doubling condition immediately implies that $m_G(B_{2^n}) \leqslant C(G,d)^n m_G(B_1)$, and so it follows that (G,d) has polynomial volume growth and is therefore unimodular.
(2) Clearly when (G,d) satisfies exact polynomial volume growth, the balls satisfy the doubling condition. But in fact, for the doubling condition already strict polynomial volume growth is sufficient. Thus for connected Lie groups of polynomial growth or for discrete nilpotent groups, volume doubling follows from Theorem 4.8.

4.5.3. *Strict exponential volume growth.* For semisimple Lie groups, using the polar coordinates on semisimple group, it can be easily established that any bi-K-invariant Riemannian metric (K a maximal compact subgroup) have strict $t^q \exp ct$ volume growth, with q and $c > 0$ depending on the metric. More generally, a bi-K-invariant metric on G

which restricts to a Weyl-group-invariant norm on the Lie algebra of a (split) Cartan subgroup (see Section 10.4 for the definitions), will be called norm-like. The following recent result of H. Abels and G. Margulis [2] establishes that norm-like metrics are the yardsticks for very general metrics on G, as follows.

THEOREM 4.11 (Word metrics on semisimple groups are coarsely isometric to norm-like metrics [2]). *Let G be a connected semisimple Lie group with finite center, and d_V a word metric on G, associated with a bounded symmetric open set. Then there exists a bi-K-invariant norm-like metric d such that $|d_V - d|$ is bounded on $G \times G$.*

We note that in fact the last result holds more generally for all reductive algebraic groups, and for every left-invariant coarsely geodesic quasi-metric on G satisfying certain natural properness and boundedness condition, and refer to [2] for the details.

COROLLARY 4.12. *The balls associated with a norm-like distance on a semisimple Lie group have strict $t^q \exp ct$ volume growth, and hence also the balls defined by d_V have the same property.*

As to discrete groups with exponential volume growth, let us recall the following result due to M. Coornaert [31]. Let Γ be a word-hyperbolic group, and S a finite symmetric set of generators. Then the spheres, and hence also the balls, have strict $\exp ct$ volume growth, namely $b \exp ct \leqslant |S_t| \leqslant B \exp ct$.

Finally, let us note that establishing exact, and even strict volume growth is an open problem for most other groups and metrics. For some further results in this direction regarding discrete groups we refer to the [42].

4.6. *Balls and asymptotic invariance under translations*

Before formulating the ergodic theorems for balls on groups with polynomial volume growth, consider the following properties of the family of balls in G. These properties obviously hold under the assumption of exact polynomial volume growth, and they are easily seen to be mutually equivalent.

PROPOSITION 4.13. *Let (G, d) have exact polynomial volume growth w.r.t. the balls B_t defined by an admissible invariant metric d on G. Then the following properties hold*

(1) *The volume of the ball is asymptotically stable, namely for every $r > 0$,*

$$\lim_{t\to\infty} \frac{m_G(B_{t+r})}{m_G(B_t)} = 1.$$

(2) *The volume of a shell is asymptotically negligible when compared with the volume of the ball, namely for every $r > 0$,*

$$\lim_{t\to\infty} \frac{m_G(B_{t+r} \setminus B_t)}{m_G(B_t)} = 0.$$

(3) *The ball averages are asymptotically invariant under* (*right*) *convolution, namely for every* $r > 0$,

$$\lim_{t\to\infty} \|\beta_t * \beta_r - \beta_t\|_{L^1(G)} = 0.$$

(4) *The balls are asymptotically uniformly invariant under* (*right*) *translations, namely for every compact set* $Q \subset G$,

$$\lim_{t\to\infty} \frac{m_G(B_t \cdot Q \Delta B_t)}{m_G(B_t)} = 0.$$

(5) *The balls are asymptotically invariant under* (*right*) *translations, namely for every* $g \in G$,

$$\lim_{t\to\infty} \frac{m_G((B_t g)\Delta B_t)}{m_G(B_t)} = 0.$$

This very pleasant property of balls is discussed in F. Greenleaf [58], where the question is raised whether it holds for balls defined by a word metric, for all lcsc groups admitting *any* sequence of asymptotically invariant compact sets of positive finite Haar measure. It is referred to in [58] as the localization conjecture, as it locates a specific asymptotically invariant sequence in the group—namely the powers of a compact neighborhood of the identity. In this generality the conjecture is false, and in fact fails already for the $ax + b$ group, as shown in [100]. In fact, a general result is that for connected exponential solvable Lie groups, no subsequence of the sequence of balls w.r.t. a Riemannian metric (say) can be asymptotically invariant, as shown in [127].

Nevertheless, for some groups of polynomial growth the localization conjecture was very recently given two very interesting independent solutions, each yielding significantly more information that just asymptotic invariance. One solution is due to E. Breuillard, applies to all word metrics on simply connected nilpotent Lie groups (as well as some further Lie groups of polynomial volume growth) and in fact gives the sharper result that the volume growth of balls is exact, and hence they are of course asymptotically invariant. In fact the result applies to more general metrics—see Theorem 4.6.

Another solution is due to R. Tessera [148], applies also to all word metrics on connected Lie groups with polynomial volume growth, and also yields a result sharper than asymptotic invariance. Indeed, the fact that the volume of the shell of width r namely $m_G(B_{t+r} \setminus B_t)$ is asymptotically negligible when compared to the volume of the ball $m_G(B_t)$, can be given a precise quantitative form, as follows.

THEOREM 4.14 (Balls satisfying the doubling condition are asymptotically invariant [148]). *Let G be an lcsc group, and let d be a word metric satisfying the doubling condition. Then there exist positive constants δ and C, such that $m_G(B_{t+1} \setminus B_t) \leqslant C t^{-\delta} m_G(B_t)$ for all $t \in \mathbb{N}$. In particular, the sequence of balls B_t is asymptotically invariant.*

In fact, in [148] a more general result is proved, namely the same estimate is established for every monotone metric on a metric-measure space. We remark that the doubling con-

dition follows from strict growth, a result established by Y. Guivarc'h for connected Lie groups of polynomial volume growth and stated in Theorem 4.8.

For completeness, we indicate here the elegant elementary argument given in [148] for word metrics in the group case, which proves Theorem 4.14.

PROOF. Let $V \subset G$ be a compact symmetric neighborhood of the identity, and d the corresponding left invariant word metric. Let B_n be the sequence of ball centered at the identity, and let $C_{n,n+k} = B_{n+k} \setminus B_n$ be the shell of width k and inner radius n. We are interested in comparing the size of the shell $C_{n-1,n}$ and the ball B_n. To do that, first note that the doubling condition is equivalent to the statement that for some fixed positive constant c independent of n, $|B_n| = |C_{0,n}| \geqslant c|C_{n,2n}| = |B_{2n} \setminus B_n|$.

Thus a natural generalization of the doubling condition from balls to shells would be the estimate $|C_{m-k,m}| \geqslant c|C_{m,m+k}|$, for some fixed constant c and all $k \leqslant m$.

Assuming this estimate for the moment, consider the following disjoint union of shell whose width doubles at each step:

$$D_j = C_{n-1,n} \cup C_{n-2,n-1} \cup C_{n-4,n-2} \cup \cdots \cup C_{n-2^j,n-2^{j-1}} = C_{n-2^j,n},$$

D_{j+1} is obtained by adding one more shell to D_j, and the foregoing estimate (taking $k = 2^j$, $m = n - 2^j$) implies that the last shell added to D_j to produce D_{j+1} satisfies

$$|C_{n-2^{j+1},n-2^j}| \geqslant c|C_{n-2^j,n}| = c|D_j|.$$

Thus $|D_{j+1}| \geqslant (1+c)|D_j|$ and hence $|D_j| \geqslant (1+c)^j |C_{n-1,n}|$. Letting $i = [\log_2 n]$, note that clearly $D_i \subset B_n$ and hence

$$|B_n| \geqslant |D_i| \geqslant (1+c)^i |C_{n-1,n}| \geqslant \frac{1}{2}(1+c)^{\log n} |C_{n-1,n}| = \frac{1}{2} n^{\log_2(1+c)} |C_{n-1,n}|.$$

Thus the size of the shell does indeed satisfy $|C_{n-1,n}| \leqslant Bn^{-\delta}|B_n|$.

Now to obtain the estimate $|C_{m-k,m}| \geqslant c|C_{m,m+k}|$, one uses the doubling condition, as follows. First, note that by the triangle inequality, for $1 \leqslant k \leqslant n/4$

$$B_k \cdot C_{n-2k,n-2k+1} \subset C_{n-4k,n} \quad \text{and} \quad C_{n,n+4k} \subset B_{8k} C_{n-2k,n-2k+1}.$$

Let $x_i, i \in I$ be a maximal k-net in $C_{n-2k,n-2k+1}$, namely such that $B_k(x_i) \cap B_k(x_j) = \emptyset$ if $i \neq j$. By maximality, we have $C_{n-2k,n-2k+1} \subset \bigcup_{i \in I} B_{2k}(x_i)$ and therefore $C_{n,n+4k} \subset \bigcup_{i \in I} B_{10k}(x_i)$. However, the doubling condition clearly implies that $|B_{10k}| \leqslant A|B_k|$, and thus we conclude that

$$|C_{n-4k,n}| \geqslant |C_{n-2k,n-2k+1}| \geqslant \sum_{i \in I} |B_k(x_i)| \geqslant \frac{1}{A} \sum_{i \in I} |B_{10k}(x_i)| \geqslant \frac{1}{A} |C_{n,n+4k}|.$$

The inequality $|C_{m-k,m}| \geqslant c|C_{m,m+k}|$ for all $k \leqslant m$ follows similarly. □

5. Pointwise ergodic theorems for groups of polynomial volume growth

Our main purpose in this section is to give a complete account of the proofs of the following results, which generalize Birkhoff's, Wiener's and Calderon's pointwise ergodic theorems. We start with the following basic result, which is a variation on the classical results developed by Wiener [155], Riesz [131] and Calderon [19], and relies on their arguments.

THEOREM 5.1 (Pointwise ergodic theorem for groups with volume doubling, asymptotically invariant balls). *Let G be locally compact second countable group of polynomial volume growth with respect to the admissible metric d. If the family of balls B_t satisfies the doubling condition and is asymptotically invariant under translations, then the family of ball averages β_t defined by d satisfies the pointwise ergodic theorem in L^p, $1 \leqslant p < \infty$.*

Recall now that according to Theorem 4.6, balls w.r.t. word metrics on (say) simply connected nilpotent Lie groups have exact polynomial growth. Thus they satisfy the doubling condition and are asymptotically invariant under translations. Together with Theorem 5.1 this proves the following result, due to E. Breuillard.

THEOREM 5.2 (Pointwise ergodic theorem for simply connected nilpotent Lie groups [15]).

(1) *Let G be a simply connected nilpotent Lie group. Let d be the distance function derived from a G-invariant Riemannian metric, a homogeneous G-invariant metric, or a word metric, and β_t the corresponding ball averages. Then β_t satisfy the pointwise ergodic theorem in every L^p, $1 \leqslant p < \infty$.*
(2) *The same result holds for any finitely generated nilpotent group, with d any word metric.*

We note that the result for discrete nilpotent groups follows already from Pansu's Theorem (Theorem 4.4).

An alternative proof of a generalization of Theorem 5.2 follows by combining Guivarc'h's results on growth, and Tessera's result on asymptotic invariance. Recall that Theorem 4.14 asserts that the doubling condition implies asymptotic invariance. Theorem 4.8 establishes that connected Lie groups of polynomial volume growth have strict polynomial growth and thus satisfy the doubling condition. Thus together with Theorem 5.1, these results imply pointwise convergence in L^1 of ball averages, for every Lie group of polynomial volume growth.

In fact, utilizing fully the results concerning growth in [64], together with Losert's structure theorem [93] for lcsc groups of polynomial volume growth based on Gromov's theorem [63], it is possible to give a complete solution to the ball averaging problem for all metrics (quasi-isometric to) word metrics on all lcsc groups of polynomial volume growth. We formulate this results as follows, and will outline its proof in Section 5.5.

THEOREM 5.3 (Pointwise ergodic theorem for groups with polynomial volume growth). *For every locally compact second countable group G of polynomial volume growth, the*

family of ball averages β_t defined by any word metric satisfies the pointwise ergodic theorem in L^p, $1 \leqslant p < \infty$. The same holds true for the balls determined by any left-invariant metric quasi-isometric with a word metric.

In the next four subsections of the present section we will give a complete proof of Theorem 5.1, demonstrating the four steps called for by the recipe of Section 2.3. In the fifth we will describe the ingredients needed to complete the proof of Theorem 5.3.

PROOF OF THEOREM 5.1. Given a group of polynomial volume growth with asymptotically invariant balls, the plan of the proof is to follow the four steps outlined in the recipe of Section 2.3, and we start with

5.1. *Step I: The mean ergodic theorem*

For the proof of the mean ergodic Theorem we will use here a standard variant of F. Riesz's [131] classical proof of von Neumann's mean ergodic theorem.

First let us note that $L^2(X)$ is the direct (orthogonal) sum of the subspaces $\mathcal{I}$ consisting of G-invariant vectors, and $\mathcal{K} = \overline{\text{span}}\{(\pi(g) - I)f;\ f \in L^2(X),\ g \in G\}$. Now note that if $h \in \mathcal{K}$ is of the form $h = \pi(g)f - f$, then clearly,

$$\begin{aligned}\left\|\pi(\beta_t)h\right\|_{L^2(X)} &= \left\|\left(\pi(\beta_t)\pi(g) - \pi(\beta_t)\right)f\right\|_{L^2(X)} \\ &\leqslant \|\beta_t * \delta_g - \beta_t\|_{L^1(G)}\|f\|_{L^2(X)}.\end{aligned}$$

But since by Proposition 4.13(4), as $t \to \infty$

$$\frac{1}{m_G(B_t)} m_G(B_t g \Delta B_t) \to 0$$

we have $\lim_{t\to\infty} \pi(\beta_t)h = 0$ in L^2-norm for a dense set of $h \in \mathcal{K}$, hence for all $h \in \mathcal{K}$, by an obvious approximation argument.

Since clearly $\pi(\beta_t)f = f$ for every $f \in \mathcal{I}$, we conclude that for every $f \in L^2(X) = \mathcal{K} \oplus \mathcal{I}$

$$\lim_{t\to\infty} \left\|\pi(\beta_t)f - \mathcal{E}f\right\|_{L^2(X)} = 0,$$

where $\mathcal{E}$ is the projection on the space $\mathcal{I}$ of G-invariant vectors, so that the mean ergodic theorem holds in $L^2(X)$. The mean ergodic theorem also holds in every $L^p(X)$, $1 \leqslant p < \infty$, by standard approximation argument, using the fact that L^∞ is norm-dense in every L^p.

5.2. *Step II: Pointwise convergence on a dense subspace*

Let $1 \leqslant p < \infty$, and consider the space

$$\mathcal{K}' = \text{span}\left\{h = \pi(g)f - f;\ f \in L^\infty(X),\ g \in G\right\}$$

and the space $\mathcal{I}$ of G-invariant functions in $L^\infty(X)$. The sum of these two spaces is dense in $L^p(X)$, as follows from the following two facts. First, every $u \in L^q(X)$ (where $\frac{1}{p}+\frac{1}{q}=1$) which integrates to zero against every function in $\mathcal{K}'$ is a G-invariant function, since L^∞ is norm-dense in every L^p. Second, since u is measurable w.r.t. the σ-algebra of G-invariant functions, if it integrates to zero against the characteristic function of every G-invariant set, then necessarily $u=0$.

Now again as in Riesz's argument of Section 5.2, if $h=\pi(g)f-f$ then for almost every $x \in X$

$$\left|\pi(\beta_t)h(x)\right| = \left|\pi(\beta_t * \delta_g - \beta_t)f(x)\right| \leqslant \frac{2\|f\|_{L^\infty(X)}}{m_G(B_t)} m_G(B_t g \Delta B_t) \to 0.$$

Thus $\pi(\beta_t)f(x) \to \int_X f\,dm$ almost everywhere for every f in the dense subspace $\mathcal{K}' \oplus \mathcal{I}$ of $L^p(X)$.

REMARK 5.4. Anticipating some arguments needed in the sequel, we note that the proof of the mean ergodic theorem is based solely on property (5) in Proposition 4.13 above, namely asymptotic invariance under translation. The same remark applies to the proof of the existence of a dense subspace where pointwise convergence holds.

5.3. *Step III: The maximal inequality for ball averages*

The maximal inequality for ball averages will be established in two stages, in Sections 5.4.1 and 5.4.2. First the maximal inequality for the special action of the ball averages by convolution on the group manifold will be established, using the volume doubling condition which holds in all groups of strict polynomial volume growth. Then a transfer principle will be formulated, which will allow us to deduce the maximal inequality for an arbitrary action from its validity for convolutions.

5.3.1. *Maximal inequality for convolutions: the volume doubling condition.* The method of proof of the weak-type $(1,1)$ maximal inequality for convolutions which we will present originates in Wiener's proof for balls in $\mathbb{R}^d$. It was later observed by Calderon [19] that the proof only depends on the fact that the volume of a Euclidean ball of a given radius is bounded by a fixed multiple of the volume of a ball of half the radius. In [19] this volume doubling condition is introduced for more general families of sets $N_t \subset G$, but here we will continue to focus on the ball averaging problem.

We now turn to the covering lemma, which we reproduce in the original finite form given in [155, Lemma C′].

LEMMA 5.5 (Wiener–Calderon covering argument [155,19]). *Assume d is an admissible metric on a lcsc group G, which satisfies the doubling volume condition. Then every finite family of balls $\{B_{t_i}, i \in I\}$ in G contains a subfamily $\{B_{t_j}, j \in J \subset I\}$ of disjoint balls whose total volume $m_G(\bigcup_{j\in J} B_{t_j})$, is at least $\delta \cdot m_G(\bigcup_{i\in I} B_{t_i})$, where $\delta = \delta(G) > 0$. Thus the total volume of the disjoint subcover is at least a fixed fraction of the total volume of the original family.*

PROOF. The volume doubling condition obviously implies that $m_G(B_t) \geqslant \delta m_G(B_{3t})$. I being finite, choose one of the balls in the family which has maximal radius, and label it B_{t_1}. Consider now the subfamily $\{B_{t_i'};\ i' \in I' \subset I\}$ of all balls intersecting B_{t_1}. The union of all the balls $B_{t_i'}$, $i' \in I'$, is contained in a ball of radius at most three times that of B_{t_1}, with the same center. Therefore keeping B_{t_1} and deleting all other balls intersecting it, we keep at least a fraction δ of the total volume of the subfamily $\{B_{t_i'};\ i' \in I'\}$. We therefore put the index t_1 in J, and apply the same argument again to the family $\{B_{t_i};\ i \in I \setminus I'\}$, which consists only of balls disjoint from B_{t_1}. Proceeding finitely many times, we obtain a disjoint sequence of balls whose total volume occupies at least a fraction δ of the total volume $m_G(\bigcup_{i\in I} B_{t_i})$. □

A variant of the previous argument is the following covering lemma, which can be proved in much the same way—see [97, Chapter IV, §1] for a more general formulation.

LEMMA 5.6 (Vitali covering lemma). *Assume d is an admissible metric on an lcsc group G, whose balls satisfy the doubling volume condition. Given any set A of positive Haar measure, there exists a disjoint sequence of balls B_{t_i}, $i \in \mathbb{N}$, satisfying $m_G(A \setminus \bigcup_{i\in\mathbb{N}} B_{t_i}) = 0$.*

Using Lemma 5.5, the following maximal inequality was proved by Calderon [19] (for more general families of sets), following [155]. It generalizes the Hardy–Littlewood maximal inequality for averages on the real line, as well as Wiener's maximal inequality for balls in Euclidean space.

THEOREM 5.7 (Maximal inequality for convolutions with ball averages satisfying the doubling condition [155,19]). *Assume d is an admissible metric on an lcsc group G, whose balls satisfy the doubling volume condition. Then the family of ball averages β_t satisfies the weak-type $(1,1)$-maximal inequality for convolutions, given by*

$$m_G\left\{g \in G;\ \sup_{0<t<\infty} \left|F * \beta_t(g)\right| > \varepsilon\right\} \leqslant \frac{C(G)}{\varepsilon}\|F\|_{L^1(G)}.$$

PROOF. Since G is unimodular and the balls are symmetric, each of the measures β_t is symmetric. The convolution operators are thus given by (see Section 2.1)

$$F * \beta_t(g) = \frac{1}{m_G(B_t)}\int_{B_t} F(gh)\,dm_G(h) = \frac{1}{m_G(B_t)}\int_{y\in B_t(g)} F(y)\,dm_G(y),$$

where $B_t(g) = gB_t(e)$ is the ball of radius t and center g. We denote as usual $F_\beta^* = \sup_{t>0}|F * \beta_t(g)|$. Since $|F * \beta_t(g)| \leqslant |F| * \beta_t(g)$ we can and will assume that $F \geqslant 0$, without loss of generality. Let $U_\varepsilon = \{g \in G;\ F_\beta^*(g) > \varepsilon\}$, and let $W \subset U_\varepsilon$ be compact. By definition, for each $w \in W$ there is a ball $B_{r_w}(w) = wB_{r_w}(e)$ with center w and radius r_w satisfying

$$m_G\big(B_{r_w}(e)\big) = m_G\big(B_{r_w}(w)\big) < \frac{1}{\varepsilon}\int_{y\in B_{r_w}(w)} F(y)\,dm_G(y).$$

There exists a finite covering of W using the collection of open balls $wB_{r_w}(e)$, and let us denote a finite covering family by $\{B_i;\ i \in I\}$. By Lemma 5.5 we can choose a finite disjoint subfamily $\{B_j;\ j \in J\}$ whose union retains at least a fraction $\delta(G)$ of the measure of the union of the balls B_i, $i \in I$. Combining these two estimates,

$$m_G(W) \leqslant m_G\Big(\bigcup_{i\in I} B_i\Big) \leqslant \frac{1}{\delta(G)} m_G\Big(\bigcup_{j\in J} B_j\Big)$$
$$\leqslant \frac{1}{\delta(G)\varepsilon} \int_{\cup_{j\in J} B_j} F(g)\, dm_G(g) \leqslant \frac{1}{\delta(G)\varepsilon} \|F\|_{L^1(G)}.$$

Taking the supremum over all compact sets contained in U_ε we conclude that the same estimate holds for the measure of the set U_ε and this concludes the proof of the weak-type $(1,1)$ maximal inequality. □

We have thus established the maximal weak-type $(1,1)$ inequality for the family of operators $F \mapsto F * \beta_t$ of right convolutions by the ball averages β_t.

REMARK 5.8. The covering arguments of Lemmas 5.6 and 5.7 described above fail for groups with exponential volume growth. However, the maximal inequality for ball averages acting by convolutions is often true, as we shall see below. Thus for semisimple Lie groups G the action of the ball averages by convolution on the symmetric space G/K satisfies the weak-type $(1,1)$ maximal inequality, as shown in [143]. There does not seem to be an lcsc group for which the weak type $(1,1)$ maximal inequality for admissible ball averages is known to fail, for convolutions or otherwise. We refer to [111] for more information on this subject.

5.3.2. *Maximal inequality for general actions: the transfer principle.* We now come to a basic observation whose origin is in Wiener's proof of the maximal inequality for ball averages in actions of $\mathbb{R}^d$, and was subsequently considerably generalized and expanded (as we shall see below). Wiener recognized that in order to prove the maximal inequality for ball averages in a measure-preserving action of the group on a *general space*, it is sufficient to prove the maximal inequality for the *action of the group on itself by translation*, provided that the balls are asymptotically invariant. This observation was later termed the transfer principle, and the idea underlying it is the following (see [155, proof of Theorem IV′]). Apply the operator β_r to the function f^t, which is f restricted to the subset $B_t \cdot x$ of the G-orbit of x in X. Replacing $\beta_r f(x)$ by $\beta_r f^t(x)$ produces an error which is controlled by the (normalized) difference in volumes between B_{r+t} and B_t. Since $(|B_{r+t} \setminus B_t|)/|B_t| \to 0$, it follows that we may consider in effect (almost) every orbit individually. Thus maximal inequalities for functions on X are reduced to maximal inequalities for *convolution operators* on the group manifold. We will give here first a very simple formulation of the transfer principle for strong maximal inequalities, and defer a more general formulation to Section 6.

Explicitly, let $d(g,h)$ be a left-invariant admissible metric on G, and let $|g| = d(e,g)$. Then clearly $|gh| \leqslant |g| + |h|$. Consider as usual the family β_t of probability measures on G defined by the balls.

THEOREM 5.9 (The transfer principle for ball averages on groups with polynomial volume growth). *Suppose $\rho(\beta_t)$, $0 \leqslant t \leqslant r$, satisfy the strong L^p-maximal inequality*

$$\left\| \sup_{0<t\leqslant r} \rho(\beta_t)F \right\|_p \leqslant C_p \|F\|_p$$

for $F \in L^p(G)$, and some $1 < p < \infty$, where ρ is the right regular representation. Then

(1) *$\pi(\beta_t)$ satisfy the maximal inequality*

$$\left\| \sup_{0<t\leqslant r} \left|\pi(\beta_t)f\right| \right\|_p \leqslant C_p \left(\frac{m_G(B_R)}{m_G(B_{R-r})} \right)^{1/p} \|f\|_p$$

in $L^p(X)$, for any measure preserving action π of G on a σ-finite measure space X. Here $R \geqslant 2r$ is any positive number.

(2) *In particular, if $\lim_{R\to\infty} m_G(B_R)/m_G(B_{R-r}) = 1$, and β_t, $0 < t < \infty$ satisfies the maximal inequality for right convolutions on the group manifold, then β_t satisfies the maximal inequality in any measure-preserving action.*

PROOF. Given $f \in L^p(X)$, fix $x \in X$ and define: $F_x(g) = f(g^{-1}x) = \pi(g)f(x)$ if $|g| \leqslant R$ and $F_x(g) = 0$ otherwise. Clearly, if $|g| \leqslant R-r$, $|h| \leqslant r$, then $F_x(gh) = \pi(g)\pi(h)f(x)$. Now integrate over h w.r.t. the measure β_t, $t \leqslant r$. Then we clearly have $\pi(g)\pi(\beta_t)f(x) = \rho(\beta_t)F_x(g)$, as long as $|g| \leqslant R-r$. Taking the supremum over $0 \leqslant t \leqslant r$ of the pth power, we obtain:

$$\pi(g) \sup_{0<t\leqslant r} \left|\pi(\beta_t)f(x)\right|^p = \sup_{0<t\leqslant r} \left|\rho(\beta_t)F_x(g)\right|^p.$$

Now integrate w.r.t. $g \in B_{R-r}$ (recall that G is unimodular), and use the maximal inequality for $\rho(\beta_t)$, to get:

$$\begin{aligned}\int_{B_{R-r}} \pi(g) \sup_{0<t\leqslant r} \left|\pi(\beta_t)f(x)\right|^p dg &\leqslant \int_G \sup_{0<t\leqslant r} \left|\rho(\beta_t)F_x(g)\right|^p dg \\ &\leqslant C_p^p \int_G \left|F_x(g)\right|^p dg.\end{aligned}$$

Since $F_x(g)$ has support in B_R, the last integral equals $C_p^p \int_{B_R} |F_x(g)|^p\, dg$. Finally, we integrate over X and use the fact that G is measure preserving, to obtain:

$$\int_{B_{R-r}}\int_X \pi(g) \sup_{0<t\leqslant r} \left|\pi(\beta_t)f(x)\right|^p dm\, dg \leqslant C_p^p \int_{B_R}\int_X \left|f\left(g^{-1}x\right)\right|^p dm(x)\, dg.$$

Hence

$$\left\| \sup_{0<t\leqslant r} \left|\pi(\beta_t)f\right| \right\|_p \leqslant C_p \left(\frac{|B_R|}{|B_{R-r}|} \right)^{1/p} \|f\|_p.$$

This concludes the proof of the maximal inequality stated in part (1). The proof of (2) is an immediate consequence. □

REMARK 5.10. A similar argument establishes also the weak-type $(1,1)$ maximal inequality for the ball averages. We have elected to present first the proof of the strong L^p-maximal inequality, for simplicity. The transfer of the weak-type $(1,1)$ maximal inequality will be demonstrated in greater generality in Section 6 below.

5.4. *Step IV: Interpolation arguments*

Thus far, taking Remark 5.10 for granted, we have established the weak-type $(1,1)$ maximal inequality for general measure-preserving actions of G, together with the mean ergodic theorem and pointwise convergence on a dense subspace. According to the recipe of Section 2.3, Theorem 5.1 has therefore been established for $f \in L^1(X)$. To conclude the proof of Theorem 5.1 is an easy matter and it remains only to note the following.

The family β_t consists of Markov operators, and clearly each $\pi(\beta_t)$ has norm bounded by 1 as an operator on $L^1(X)$ and $L^\infty(X)$. It is clear that

$$\sup_{t>0} \left|\pi(\beta_t) f(x)\right| = f_\beta^*(x) \leqslant \|f\|_{L^\infty(X)}.$$

Given the weak-type $(1,1)$-maximal inequality for ball averages in an arbitrary measure-preserving action, by Marcinkiewicz's interpolation theorem, f_β^* satisfies the strong L^p-maximal inequality for $1 < p < \infty$.

This concludes the proof of Theorem 5.1. □

5.5. *Groups of polynomial volume growth: general case*

We now proceed with

PROOF OF THEOREM 5.3. According to Theorem 5.1, to show that the balls of a given admissible metric d on a group G with polynomial volume growth satisfy the pointwise ergodic theorem, it suffices to show that they are asymptotically invariant under translations and volume doubling. By Theorem 4.14, in fact a sufficient condition for asymptotic invariance is that the balls satisfy the volume doubling property. As noted already above, if the balls have strict polynomial volume growth, then they clearly satisfy the doubling volume condition. Furthermore, if the balls for a given metric have strict polynomial growth, then it is clear that the balls with respect to any quasi-isometric metric also have strict polynomial growth, and thus also satisfy the doubling condition. Therefore Theorem 5.3 will be completely proved once we establish the following

PROPOSITION 5.11. *Given any lcsc group G with polynomial volume growth, G has strict polynomial volume growth w.r.t. any metric quasi-isometric to a word metric.*

PROOF. The argument utilizes the definitive results on strict polynomial growth of Lie groups, obtained by Y. Guivarc'h [64], together with a structure theorem for lcsc groups with polynomial volume growth obtained by V. Losert [150], generalizing Gromov's result [63] in the discrete case.

Thus according to [150, Theorem 2] any lcsc group with polynomial volume growth G admits a normal series $C \triangleleft R \triangleleft N \triangleleft G$ with C and G/N compact, R/C a connected solvable Lie group of polynomial volume growth, and N/R a finitely generated discrete nilpotent group.

By [64, Theorem I.4], G has strict growth if G/C does, so we may assume $C = \{e\}$. R is a closed subgroup of G, and thus has polynomial growth by [64, Theorem I.2]. Now by [64, Theorem I.4], since N is normal and co-compact in G, it has a growth function equivalent to that of G, and hence N has polynomial growth, and G has strict growth if N does. So it suffices to show that N has strict growth, and since R is solvable, and N/R nilpotent, it follows that N is a solvable Lie group with connected component R and polynomial volume growth. Thus by [64, Theorem III.5] it follows that N has strict polynomial volume growth. □

We remark that Proposition 5.11, in combination with Theorem 4.14, implies of course the following fact, which we record for completeness.

PROPOSITION 5.12. *Let G be an lcsc group with polynomial volume growth, d any metric quasi-isometric to a word metric, and B_t the corresponding balls. Then*

(1) *The balls B_t are asymptotically invariant under translation.*
(2) *The shells $C_t = B_{t+1} \setminus B_t$ satisfy $m_G(C_t) \leqslant Ct^{-\delta} m_G(B_t)$, for some positive δ and C, for $t \geqslant 1$.*

5.5.1. *Subsequence theorem for groups with subexponential growth.* Let us note the following regarding pointwise ergodic theorems for subsequences of ball averages.

(1) Calderon's original formulation [19] of his pointwise ergodic theorem did not prove or assume strict polynomial volume growth, but instead noted that the doubling condition implies the following property for the volume of the balls. There exists a set $D \subset \mathbb{R}$ of density 1, such that for any $s > 0$, we have $\lim_{t\to\infty} m_G(B_{t\pm s})/m_G(B_t) = 1$. Thus, using the arguments above, it is shown in [19] that the pointwise ergodic theorem holds for $\pi(\beta_t) f(x)$, provided that as $t \to \infty$ it assumes only values from the set D. Thus it was shown in [19] that in every group satisfying the doubling condition, there is a subsequence β_{t_n} satisfying the pointwise ergodic theorem in L^1.
(2) Similarly, subexponential growth (which is equivalent to polynomial volume growth in the connected Lie group case but not in general) implies that a subsequence of the sequence of balls is asymptotically invariant. Hence in particular the mean ergodic theorem is true for the subsequence.
(3) For a discrete subgroup of exponential growth, it follows easily from the definition that no subsequence of balls can be a Følner sequence. We recall that it has been established in [127] that when G is a connected solvable Lie group and has exponen-

tial volume growth, again no subsequence of the sequence of balls is asymptotically invariant.

6. Amenable groups: Følner averages and their applications

6.1. *The transfer principle for amenable groups*

As we shall see presently, inspection of the proof of Theorem 5.9 reveals that the condition which is essential for the transfer principle to hold (for an arbitrary family μ_t of probability measures with compact supports, not just β_t) is the existence of a sequence of sets F_n (given there by the balls B_n) with the following property.

DEFINITION 6.1 (*Følner conditions*). A sequence F_n of compact sets of positive Haar measure in an lcsc group G is called

(1) (right) *Følner sequence* if for every $g \in G$

$$\lim_{n\to\infty} \frac{\eta(F_n g \Delta F_n)}{\eta(F_n)} = 0;$$

(2) (right) *uniform Følner sequence* if for any given compact set $Q \subset G$ we have,

$$\lim_{n\to\infty} \frac{\eta(F_n Q \Delta F_n)}{\eta(F_n)} = 0.$$

It then follows also that for every compact set Q

$$\lim_{n\to\infty} \frac{\eta(F_n Q)}{\eta(F_n)} = 1,$$

where η is right Haar measure on G.

The fact that the existence of a Følner sequence is sufficient for the validity of a transfer principle (generalizing that of Wiener [155]) was proved in various different formulations by a number of authors, starting with Calderon [20], followed by Coifman and Weiss [30], Emerson [46], Herz [72] and Tempelman [147].

The existence of a Følner sequence in an lcsc group G is equivalent to G being amenable, and as is well-known, polynomial volume growth implies amenability, but there are many amenable groups of exponential volume growth. Thus the Følner condition yields a significantly more general transfer principle than Theorem 5.9. We now turn to the formulation and proof of a version of the transfer principle for amenable groups that will be found useful below.

THEOREM 6.2 (The transfer principle for amenable groups). *Let μ_t, $0 < t < \infty$, be probability measures with compact supports on an lcsc group G. Assume that for $t \leqslant R$ we have* $\mathrm{supp}(\mu_t) \subset Q$, *where Q is a compact subset* (*possibly depending on R*). *Suppose η is*

right Haar measure on G, and that μ_t, $0 \leqslant t \leqslant R$, satisfy the following maximal inequality for $F \in L^p(G,\eta)$, where $1 < p < \infty$

$$\left\| \sup_{0<t\leqslant R} \left| \int_G F(gh)\, d\mu_t(h) \right| \right\|_{L^p(G,\eta)} \leqslant C_p \|F\|_{L^p(G,\eta)}.$$

(1) *If A is any compact set of positive measure in G, then $\pi(\mu_t)$ satisfy the strong L^p-maximal inequality*

$$\left\| \sup_{0<t\leqslant R} \left|\pi(\mu_t) f\right| \right\|_p \leqslant C_p \left(\frac{\eta(AQ)}{\eta(A)} \right)^{1/p} \|f\|_p$$

for any measure preserving action π of G on a σ-finite measure space (X,m).

(2) *If the weak-type $(1,1)$-maximal inequality holds for the action by translation, namely for $F \in L^1(G,\eta)$*

$$\eta\left\{ g \in G;\ \sup_{0<t\leqslant R} \left| F(gh)\, d\mu_t(h) \right| > \delta \right\} < \frac{C}{\delta} \|F\|_{L^1(G,\eta)}.$$

Then for any measure-preserving action on (X,m) the following weak-type $(1,1)$-maximal inequality holds (for any A as in (1))

$$m\left\{ x \in X;\ \sup_{0<t\leqslant R} \left|\pi(\mu_t) f(x)\right| > \delta \right\} < \frac{\eta(AQ)}{\eta(A)} \frac{C}{\delta} \|f\|_{L^1(X)}.$$

(3) *When A_n satisfies $\lim_{n\to\infty} \eta(A_n \cdot Q)/\eta(A_n) = 1$, the maximal inequalities for μ_t hold in $L^p(X)$, $1 < p < \infty$ (with the same bound as in $L^p(G)$) namely*

$$\left\| \sup_{0<t\leqslant R} \left|\pi(\mu_t) f\right| \right\|_{L^p(X)} \leqslant C_p \|f\|_{L^p(X)}$$

and

$$m\left\{ x \in X;\ \sup_{0<t\leqslant R} \left|\pi(\mu_t) f(x)\right| > \delta \right\} < \frac{C}{\delta} \|f\|_{L^1(X)}.$$

(4) *Finally, if A_n satisfy the foregoing condition for every compact Q (namely if A_n form a right uniform Følner sequence) and the assumption regarding the maximal inequalities for convolutions is satisfied for $R = \infty$, then so is the conclusion.*

PROOF. (1) Given $f \in L^p(X)$, $1 < p < \infty$, fix $x \in X$ and a compact set $A \subset G$ of positive measure, and define:

$$F_A(g) = f\left(g^{-1}x\right) = \pi(g) f(x) \quad \text{if } g \in A, \quad \text{and} \quad F_A(g) = 0 \quad \text{otherwise.}$$

Clearly, if $k \in A$ and $h \in Q$, then $F_{AQ}(kh) = \pi(k)\pi(h) f(x)$.

By assumption Q contains the support of μ_t, $0 < t \leqslant R$, and we can therefore integrate the last equation over $h \in Q$ w.r.t. the measure μ_t and write, as long as $k \in A$

$$\pi(k)\pi(\mu_t) f(x) = \int_G F_{AQ}(kh)\, d\mu_t(h).$$

Taking the supremum over $0 < t \leqslant R$ of the pth power, we obtain:

$$\pi(k) \sup_{0<t\leqslant R} \left|\pi(\mu_t) f(x)\right|^p = \sup_{0<t\leqslant R} \left| \int_G F_{AQ}(kh)\, d\mu_t(h) \right|^p.$$

Now integrate over $k \in A$ using right-invariant Haar measure η, and extend the integration to all of G on the right-hand side. This yields the obvious inequality

$$\int_A \pi(k) \sup_{0<t\leqslant R} \left|\pi(\mu_t) f(x)\right|^p d\eta(k) \leqslant \int_G \sup_{0\leqslant t\leqslant R} \left| \int_G F_{AQ}(kh)\, d\mu_t(h) \right|^p d\eta(k)$$

using the strong L^p-maximal inequality which we assumed for the right-hand side, together with the fact that (by definition) F_{AQ} is supported in $AQ \subset G$, we obtain that the last integral is bounded by:

$$C_p^p \int_G \left|F_{AQ}(g)\right|^p d\eta(g) = C_p^p \int_{AQ} \left|F_{AQ}(g)\right|^p d\eta(g).$$

Finally, we integrate both sides of the inequality over X, and use Fubini's theorem to obtain:

$$\begin{aligned} &\int_A \int_X \pi(k) \sup_{0<t\leqslant R} \left|\pi(\mu_t) f(x)\right|^p dm(x)\, d\eta(k) \\ &\quad \leqslant C^p \int_{AQ} \int_X \left|f\left(g^{-1}x\right)\right|^p dm(x)\, d\eta(g). \end{aligned}$$

Hence since the G-action is measure preserving:

$$\left\| \sup_{0<t\leqslant R} \left|\pi(\mu_t) f\right| \right\|_{L^p(X)} \leqslant C_p \left(\frac{\eta(AQ)}{\eta(A)} \right)^{1/p} \|f\|_{L^p(X)}.$$

This concludes the proof of part (1) of the theorem.

(2) As to part (2), fix a compact set A, and let us define the set

$$\mathcal{D}(\delta) = \left\{ (k,x) \in A \times X;\ \sup_{0<t\leqslant R} \pi(k) \left|\pi(\mu_t) f(x)\right| > \delta \right\}.$$

The first coordinate sections of $\mathcal{D}(\delta)$ are given, for each $k \in A$, by

$$\mathcal{D}^k(\delta) = \left\{ x \in X;\ \sup_{0<t\leqslant R} \pi(k) \left|\pi(\mu_t) f(x)\right| > \delta \right\}.$$

In particular, the set whose measure we are interested in estimating is

$$\mathcal{D}^e(\delta) = \Big\{ x \in X;\ \sup_{0<t\leqslant R} \big|\pi(\mu_t) f(x)\big| > \delta \Big\}.$$

Note that clearly, $\mathcal{D}^k(\delta) = k(\mathcal{D}^e(\delta))$, and since each $k \in G$ is measure preserving, we have $m(\mathcal{D}^k(\delta)) = m(\mathcal{D}^e(\delta))$ for every $k \in A$.

The second coordinate sections of $\mathcal{D}(\delta)$ are given, for each $x \in X$, by

$$\mathcal{D}_x(\delta) = \left\{ k \in A;\ \sup_{0<t\leqslant R} \left| \int_G \pi(kh) f(x)\, d\mu_t(h) \right| > \delta \right\}.$$

By Fubini's theorem, we have

$$\eta \times m(\mathcal{D}_\delta) = \int_X \eta\big(\mathcal{D}_x(\delta)\big)\, dm(x) = \int_A m\big(\mathcal{D}^k(\delta)\big)\, d\eta(k) = \eta(A) m\big(\mathcal{D}^e(\delta)\big).$$

Now by assumption, the action by translation satisfies the weak-type $(1,1)$-maximal inequality. Keeping the notation introduced in the proof of part (1), we have $|\pi(k)\pi(\mu_t) f(x)| \leqslant \int_G |F|_{AQ}(kh)\, d\mu_t(h)$ if $k \in A$, $h \in Q$. Hence

$$\eta\big(\mathcal{D}_x(\delta)\big) \leqslant \eta\left\{ k \in A;\ \sup_{0<t\leqslant R} \left| \int_G \big|F_{AQ}(kh)\big|\, d\mu_t(h) \right| > \delta \right\}.$$

Combining the two foregoing arguments, we conclude

$$\eta(A) m\big(\mathcal{D}^e(\delta)\big) = \int_X \eta\big(\mathcal{D}_x(\delta)\big)\, dm(x) \leqslant \int_X \frac{C}{\delta} \|F_{AQ}\|_{L^1(G,\eta)}\, dm(x)$$

finally, using the fact that G is measure-preserving and the definition of F_{AQ}, we obtain

$$m\big(\mathcal{D}^e(\delta)\big) = m\Big\{ x \in X;\ \sup_{0<t\leqslant R} \big|\pi(\mu_t) f(x)\big| > \delta \Big\} \leqslant \frac{\eta(AQ)}{\eta(A)} \frac{C}{\delta} \|f\|_{L^1(X)}$$

and the proof of part (2) is complete.

(3) Part (3) follows immediately upon applying the following arguments. If A_n satisfies for every compact set $Q \subset G$

$$\lim_{n\to\infty} \frac{\eta(A_n Q)}{\eta(A_n)} = 1$$

we take the limit as $n \to \infty$ in part (1) and conclude that for each Q,

$$\Big\| \sup_{0<t\leqslant R} \big|\pi(\mu_t) f\big| \Big\|_{L^p(X)} \leqslant C_p \|f\|_{L^p(X)}.$$

(4) Since C_p is fixed and independent of Q, we therefore choose a sequence of compact sets $Q_m \subset Q_{m+1}$ whose union is G. Thus $\mathrm{supp}(\mu_t) \subset Q_m$ for $t \leqslant R_m$, where $R_m \to \infty$. Applying the foregoing result to each set Q_m, a straightforward application of the monotone convergence theorem or Fatou's lemma allows us to conclude that also

$$\left\| f_\mu^* \right\|_{L^p(X)} = \left\| \sup_{0<t<\infty} \left| \pi(\nu_t) f \right| \right\|_{L^p(X)} \leqslant C_p \| f \|_{L^p(X)}.$$

A similar argument proves the corresponding result for the weak-type $(1,1)$ maximal inequality, and for the case $R \to \infty$.

This concludes the proof of the transfer principle of amenable groups. □

REMARK 6.3. The formulation of the transfer principle in Theorem 6.2 is similar to the one given by Tempelman in [145, Chapter 5, §1.4]. It differs somewhat from those of Calderon [19], Emerson [46] and Coifman and Weiss [30]. The latter formulations all consider the transfer of an arbitrary operator T on $L^1_{\mathrm{loc}}(G)$ satisfying the following properties.

(1) T is sublinear,
(2) T commutes with right translations,
(3) T is semilocal, i.e. if $\mathrm{supp}(F) \subset C$ then $\mathrm{supp}(TF) \subset QC$, for some fixed compact set Q depending on T,
(4) T maps the space L^1_{loc} into the space of continuous functions on G.

Under these condition, if the operator T is bounded on $L^p(G)$, the transferred operator is defined and is bounded on $L^p(X)$ with the same bound, for an arbitrary measure-preserving action. The same holds for the maximal function associated with a sequence T_n of such operators. In the context of convolution operators, condition (4) would usually require the absolute continuity of the measure μ_t. However, we would like to emphasize that the transfer principle is valid also for any family μ_t of singular measures on G (many of which will appear below) and does not require absolute continuity.

REMARK 6.4. Let us note further that in the transfer principle formulated in Theorem 6.2 (as compared to Theorem 5.9)

(1) The family of measures μ_t whose maximal inequalities are being transfered is arbitrary, and in particular the measures are not required to be symmetric (or, as already noted, absolutely continuous).
(2) G need not be unimodular, and furthermore the Følner sequence which guarantees the validity of the transfer principle can be arbitrary and no growth conditions (such as the doubling condition) on it are assumed.
(3) A principle of local transfer, namely when all the measures μ_t on G have their support contained in a fixed compact set, holds without any restriction at all on the group G, which need not be amenable in this case.
(4) Similar results can be easily formulated for semigroup actions. Note that then we must consider the *anti-representation* $\pi'(g) f(x) = f(gx)$ (satisfying $\pi'(hg) = \pi'(g)\pi'(h)$), and the operators $\pi'(\mu_t) f(x) = \int_G f(gx)\, d\mu_t$.

6.2. *Generalizations of the doubling condition: regular Følner sequences*

In Section 6.1 it was established that the transfer principle for amenable groups is a direct consequence of the existence of a Følner sequence, and hence is valid for any amenable group. Thus to obtain a maximal inequality for μ_t in a general action of an amenable groups G it suffices to establish a maximal inequality for the action of G by right translation. This problem turns out to be a very difficult one for many of the most natural averages μ_t. So far, we have seen in Section 5 how to prove such a result for balls w.r.t. an admissible metric, provided they satisfy the doubling volume condition, which then implies polynomial volume growth.

Calderon's original doubling volume condition [19] is formulated for an increasing family N_t of compact symmetric neighborhoods of the identity, which generate G, and satisfies $N_t N_s \subset N_{t+s}$ together with $m_G(N_{2t}) \leqslant C m_G(N_t)$. A group possessing such a family is necessarily unimodular, so any Haar measure can be taken. Note that for balls B_t w.r.t. an invariant metric, $B_{2t} = B_t \cdot B_t$, so Calderon's condition can be written as $m_G(B_t^{-1} \cdot B_t) \leqslant C m_G(B_t)$.

Thus, a natural generalization of the doubling condition of [19] (as well as the conditions considered by Pitt [126] and Cotlar [33]), is given by the following condition, introduced by A. Tempelman [146].

DEFINITION 6.5 (*Regular sequences* [146]). A sequence N_k of sets of positive finite measure in an lcsc group G is called regular if

$$m_G\left(N_k^{-1} \cdot N_k\right) \leqslant C m_G(N_k)$$

for some C independent of k, and a left Haar measure m_G on G.

Regular sequences have been utilized to prove the following result, proved in [147,26, 11] for the unimodular case, and in [46] for general amenable groups. (For simplicity of notation, we switch to anti-representations of G here.)

THEOREM 6.6 (Pointwise ergodic theorem for regular Følner sequences [146,26,11,46]). *Assume G is an amenable lcsc group, m_G left Haar measure, and $N_k \subset G$ is an increasing left Følner sequence, with $\bigcup_{k\in\mathbb{N}} N_k = G$, satisfying $m_G(N_k^{-1} N_k) \leqslant C m_G(N_k)$, i.e. a regular sequence. Then*

(1) *The maximal operator* $\sup_{k\in\mathbb{N}} |\frac{1}{m_G(N_k)} \int_{N_k} F(gh)\, dm_G(g)|$ *satisfies the weak-type* $(1,1)$ *and strong* L^p *maximal inequalities for* $F \in L^p(G, m_G)$.
(2) *The operators* $\pi(\eta_k) f(x) = \frac{1}{m_G(N_k)} \int_{g\in N_k} f(gx)\, dm_G(g)$ *satisfy the weak-type* $(1,1)$ *and strong* L^p *maximal inequalities, in every measure-preserving action of* G *on* (X, m).
(3) *The sequence η_k satisfies the pointwise ergodic theorem in L^1, for every probability measure preserving action of G.*

As to the proof of Theorem 6.6, we note the following. Given the transfer principle of Theorem 6.2, to obtain the maximal inequalities stated in part (2) of Theorem 6.6, it suffices

to prove the maximal inequality stated in part (1) for translations. The proof of the latter is similar to that of the corresponding result for ball averages discussed in Section 5, and uses a natural generalization of the covering arguments employed by Wiener and Calderon to the present context. Namely, it is shown that under the condition of regularity, given a finite set of translates $\{N_k g_i;\ k \in K, i \in I\}$ which covers a given compact set F, it is possible to select a subcover consisting of *disjoint* translates, which still covers a fixed fraction of F. The weak-type $(1,1)$ inequality then follows as in Section 5.

To get the full pointwise ergodic theorem stated in Theorem 6.6(3), one needs also, according to the recipe of Section 2.3, a mean ergodic theorem and pointwise convergence on a dense subspace. We thus note the following

PROPOSITION 6.7 (Mean ergodic theorem for Følner sequences). *For every Følner sequence F_n on an amenable lcsc group G, the normalized averages on F_n satisfy the mean ergodic theorem in $L^2(X)$ in every probability preserving action. Furthermore, there exists in every $L^p(X)$, $1 \leqslant p < \infty$, a dense subspace on which pointwise convergence holds.*

PROOF. Looking at the proofs given in Sections 5.2 and 5.3 of the corresponding statements, it is clear that they are valid for every Følner sequence, as already noted there. □

REMARK 6.8.

(1) A practical criterion for the existence of a regular Følner sequence has never been found. In particular it seems unknown which connected amenable Lie groups with exponential volume growth (if any) posses such a sequence. For the latter class a pointwise ergodic theorem will be proved for certain Følner sequences in Section 7, using a different approach.
(2) It was shown in [92] that there exists a discrete amenable group of exponential volume growth (the "lamplighter" group) for which no regular Følner sequence exists.

6.3. *Subsequence theorems: tempered Følner sequences*

The discussion of Section 6.2 does not resolve the problem of the existence of a Følner sequence which satisfies the pointwise ergodic theorem in L^1 in an *arbitrary* lcsc amenable group. This problem was resolved in L^2 using a more general condition, which was introduced by A. Shulman (see [145, Chapter 5]) for this purpose, as follows.

DEFINITION 6.9 (*Tempered sequences*). A sequence of sets of positive finite measure N_k in G is called tempered if for some fixed C

$$m_G\big(\widetilde{N}_n^{-1} \cdot N_{n+1}\big) \leqslant C m_G(N_{n+1}),$$

where $\widetilde{N}_n = \bigcup_{k \leqslant n} N_k$.

We note that this condition is very different from the regularity condition of Section 6.2. Indeed, regularity is primarily a growth condition, and it is often incompatible with the

Følner property, as noted in [60]. Furthermore, regular Følner sequences have never been found and are unlikely to exist in groups with exponential volume growth. On the other hand, temperedness is primarily an invariance condition, implying that the $(n+1)$th set is almost invariant under (inverse) left translations by all the n previously chosen sets. Thus an easy induction argument shows that starting with any Følner sequence, one can choose a tempered subsequence [92, Proposition 1.4].

It was shown by A. Shulman that the averages associated with a tempered Følner sequence satisfy the pointwise ergodic theorem in L^2 (see [145, §5.6]). The complete result in L^1 has been established by E. Lindenstrauss [92], as follows.

THEOREM 6.10 (Pointwise ergodic theorem for tempered Følner sequences [92]). *Let N_k be a tempered left Følner sequence on an amenable lcsc group G. The normalized averages $\eta_k f(x) = \frac{1}{m_G(N_k)} \int_{N_k} f(gx)\, dg$ satisfy the pointwise ergodic theorem, and the weak-type $(1,1)$ maximal inequality in L^1, in every probability-preserving action of G. Thus every amenable lcsc group admits a Følner sequence satisfying the pointwise ergodic theorem in L^1.*

As to the proof of Theorem 6.10, we note that it proceeds as usual by establishing the weak-type $(1,1)$ maximal inequality for the operators η_k acting on $X = G$, and then appeals to the transfer principle of Theorem 6.2. The maximal inequality uses a covering argument in G, where the covering sets are taken from the set of all translates $\mathcal{F} = \{N_n g;\ n \in \mathbb{N},\ g \in G\}$. The proof of the covering argument introduces an important new probabilistic technique to the discussion, as follows. Given a compact set $F \subset G$ to be covered, it is shown that a probability distribution can be introduced on the set of subcollections of $\mathcal{F}$, such that typically (w.r.t. the probability distribution) a random subcollection consists of almost disjoint sets, and these cover most of F, approximately evenly. This allows the construction of a subcover which retains at least a fixed fraction of the measure of the original set F, and which is almost disjoint, and thus the weak-type $(1,1)$-maximal follows in a manner similar to the proof of Theorem 5.9.

We note that another proof of the same result is due to B. Weiss [154]. The proof establishes a weak-type $(1,1)$-maximal inequality for averaging with respect to a tempered Følner sequence, based of a covering argument for the family of translates of the sequence. The covering lemma is based on an interesting direct combinatorial argument utilizing temperedness, and is thus a deterministic one.

REMARK 6.11. The process of refining a given Følner sequence to a tempered subsequence may be rather drastic, namely the resulting subsequence may be very sparse. Thus in [92, Corollary 5.6] it is shown that any tempered Følner sequence on the lamplighter group must satisfy $\lim_{k\to\infty} |N_{k+1}|/|N_k| = \infty$, and in particular be of super-exponential growth. A question raised in [92] is whether there always exists a Følner sequence with exponential growth satisfying the pointwise ergodic theorem in L^1.

7. A non-commutative generalization of Wiener's theorem

The present section is devoted to an exposition of some non-commutative generalizations of Birkhoff's and Wiener's pointwise ergodic theorems. These results are based on a method introduced independently by Dunford [43] and Zygmund [156], and first utilized in their proof of a fundamental result on (dominated) pointwise convergence of averages of product type on the product of several not-necessarily-commuting one-parameter flows. We will demonstrate below that the method can be applied to yield a large collection of pointwise convergence theorems and strong maximal inequalities for families of measures on groups, provided the groups can be represented as a product of more basic flows, for example one-parameter flows. The Dunford–Zygmund method is thus an ideal tool for proving ergodic theorems for connected Lie groups, when they admit global Lie coordinates of the second kind, and as we shall see also for algebraic groups, which admit a variety of decomposition theorems. In particular, the method was used by Emerson and Greenleaf [60] to give a proof of a pointwise ergodic theorem for certain sequences of Følner averages on any connected amenable group. Below we will present the proof of a generalization of the Dunford–Zygmund theorem, as well as a sharper form of the Greenleaf–Emerson theorem. We will also give a proof of a pointwise ergodic theorems for groups with an Iwasawa decomposition, generalizing Tempelman's theorem for connected Lie groups. Finally we will indicate further results based on these methods, which apply for example to general algebraic groups (see [112] for details).

We note however that the averages that the Dunford–Zygmund method apply to are of a very specific form, and in particular they usually bear no resemblance to ball averages w.r.t. an invariant metric on the group.

7.1. *The Dunford–Zygmund method*

Let us start by formulating and proving the Dunford–Zygmund theorem in the most basic special case. As will become clear below, however, this case already demonstrates the main ideas involved.

PROPOSITION 7.1 (The pointwise ergodic theorem for two non-commuting flows [43, 156]). *Let u_t, $t \in \mathbb{R}$, and v_s, $s \in \mathbb{R}$, be two not-necessarily-commuting $\mathbb{R}$-flows, namely representations of $\mathbb{R}$ as measure preserving transformations of a probability space (X, m). Then*

(1) *The strong L^p-maximal inequality, $1 < p \leqslant \infty$, holds for rectangle averages, namely*

$$\left\| \sup_{T,S>0} \left| \frac{1}{4TS} \int_{-T}^{T} \int_{-S}^{S} u_t v_s f \, dt \, ds \right| \right\|_p \leqslant C_p \| f \|_p.$$

(2) *Let $\mathcal{U}$ (resp. $\mathcal{V}$) be the conditional expectation w.r.t. the σ-algebra of sets invariant under every u_t, $t \in \mathbb{R}$ (resp. every v_s, $s \in \mathbb{R}$). Then for every $f \in L^p(X)$, $1 < p < \infty$, and for almost every $x \in X$*

$$\lim_{\min(T,S)\to\infty} \frac{1}{4TS} \int_{-T}^{T}\int_{-S}^{S} u_t v_s f(x)\, dt\, ds = \mathcal{U}\mathcal{V} f(x)$$

and the convergence is also in the L^p-norm.

(3) *Pointwise convergence also holds for $f \in L(\log L)(X)$.*

PROOF. (1) The maximal inequality stated follows from the following simple observations. As usual assume without loss of generality that $f \geqslant 0$, and then clearly

$$\begin{aligned}
&\sup_{0<T\leqslant T_0,\, 0<S\leqslant S_0} \frac{1}{4TS} \int_{-T}^{T}\int_{-S}^{S} u_t v_s f(x)\, dt\, ds \\
&\leqslant \sup_{0<T\leqslant T_0} \frac{1}{2T} \int_{-T}^{T} u_t \left(\sup_{0<S} \frac{1}{2S} \int_{-S}^{S} v_s f(x)\, ds \right) dt \\
&\leqslant M_U^* \left(M_V^* f \right)(x).
\end{aligned}$$

Now the maximal function $M_V^* f$, associated with the averages $1/2S \int_{-S}^{S} v_s\, ds$, has an L^p-norm bound, by the maximal inequality for one-parameter flows, which is a consequence of Theorem 5.7. The same of course holds for the maximal function M_U^* which is associated with the averages $1/2T \int_{-T}^{T} u_t\, dt$. Hence the maximal inequality for averaging over $|t| \leqslant T$, $|s| \leqslant S$ follows. The desired strong maximal inequality then follows from the monotone convergence theorem.

(2) According to the recipe of Section 2.3, given the maximal inequality, pointwise almost everywhere convergence follows provided it can be established on a dense subspace of $L^p(X)$ ($p < \infty$). We again apply the analog of Riesz's argument [131] used in the proof of Theorem 5.3 and define the space $\mathcal{K} = \operatorname{span}\{v_s f - f \mid s \in \mathbb{R},\ f \in L^\infty(X)\}$. The sum of $\mathcal{K}$ and the space of $\{v_s;\ s \in \mathbb{R}\}$-invariant functions is dense in $L^p(X)$, $1 \leqslant p < \infty$. Now if $h = (v_{s_0} f - f) \in \mathcal{K}$, then for $S \geqslant |s_0|$, and every $T > 0$

$$\left| \frac{1}{2T} \int_{-T}^{T} u_t \left(\frac{1}{2S} \int_{-S}^{S} v_s (v_{s_0} f - f)(x)\, ds \right) dt \right| \leqslant \frac{2|s_0| \cdot \|f\|_\infty}{S} \longrightarrow 0$$

as $S \to \infty$. Therefore

$$\lim_{\min(S,T)\to\infty} \frac{1}{2T} \int_{-T}^{T} u_t \left(\frac{1}{2S} \int_{-S}^{S} v_s (v_{s_0} f - f)(x)\, ds \right) dt = 0.$$

Of course, if h is v_s-invariant, then the expression

$$\frac{1}{4TS} \int_{-T}^{T}\int_{-S}^{S} u_t v_s h(x)\, ds\, dt = \frac{1}{2T} \int_{-T}^{T} u_t h(x)\, dt$$

converges almost everywhere by Birkhoff's theorem. Thus pointwise convergence holds on a dense subspace, and using the maximal inequality, also for every $f \in L^p(X)$, $1 < p \leqslant \infty$.

(3) The identification of the limit is obtained as follows. For any $f \in L^p(X)$:

$$\left\| \frac{1}{4TS} \int_{-T}^{T} \int_{-S}^{S} u_t v_s f \, ds \, dt - \mathcal{U}\mathcal{V} f \right\|_p$$
$$\leqslant \left\| \frac{1}{2T} \int_{-T}^{T} u_t \left(\frac{1}{2S} \int_{-S}^{S} v_s f \, ds - \mathcal{V} f \right) dt \right\|_p + \left\| \frac{1}{2T} \int_{-T}^{T} u_t \mathcal{V} f \, dt - \mathcal{U}\mathcal{V} f \right\|_p .$$

Using the norm convergence to $\mathcal{V} f$ of the averages $\frac{1}{2S} \int_{-S}^{S} v_s f \, ds$, and the fact that each operator $\frac{1}{2T} \int_{-T}^{T} u_t \, dt$ is a contraction we can estimate the norm of the first summand. The norm of the second summand is estimated by the norm convergence of $\frac{1}{2T} \int_{-T}^{T} u_t h \, dt$ to $\mathcal{U} h$, $h = \mathcal{V} f$. We conclude that the limit of the foregoing expression as $\min(S, T) \to \infty$ is 0, and the proof of convergence in L^p is complete.

(4) For the proof of pointwise convergence for $f \in L(\log L)(X)$ we refer the reader to the original argument in [156]. □

REMARK 7.2 (*On the identification of the limit in Theorem* 7.1).

(1) It is obvious that Theorem 7.1 and its proof admit extensive generalizations. Thus, for example, we can replace U and V by any Abelian (not necessarily connected) Lie (semi)group, and replace the interval averages by any other Følner averages satisfying the pointwise and maximal ergodic theorem. Furthermore, similar conclusions holds for any finite sequence $U_1, \ldots, U_k$ of Abelian (semi)groups. We will comment on this fact and some further generalizations below.

(2) Anticipating some arguments that will occur later on, note that in the proof of part (2), the asymptotic invariance of the intervals $[-S, S] \subset V$ under translation plays an essential role in the proof. It provides an estimate in the pointwise convergence argument for the bounded functions in question which is *uniform* in T, and depends only on the size of S, allowing the argument to proceed.

(3) A key problem in utilizing Theorem 7.1 is to make the identification of the possible limits in Theorem 7.1 more precise. For example, if the intersection of the ranges of the two projections $\mathcal{U}$ and $\mathcal{V}$ reduces to the constant functions, when is it true that the limit in the theorem is the projection on the constants, namely $\int_X f \, dm$? Note that $\mathcal{U} f$ is not necessarily $\mathcal{V}$-invariant if f is.

REMARK 7.3 (*Unrestricted convergence and* $L(\log L)$*-results*).

(1) It was established already in [156] that the existence of the pointwise limit stated in part (2) of Theorem 7.1 holds for functions $f \in L(\log L)^{k-1}(X)$, in the case of k one-parameter flows. We refer to [50] for a more general result.

(2) On the other hand, we note that it is essential for the argument given in the proof of Theorem 7.1 that we consider the strong maximal inequality—given by a *norm* inequality in L^p. Weak-type inequalities cannot be treated by the same argument,

and indeed there is no weak-type $(1,1)$ maximal inequality for general two non-commuting flows. For general multi-parameter flows, and in fact even for commuting ones, pointwise convergence in L^1 does not always hold. For a counterexample (for an action of $\mathbb{Z}^d$) see [88, §6.1]. This phenomenon is due to the fact that in Theorem 7.1 we allow *unrestricted convergence*, meaning that in the rectangle averages the side lengths T and S are chosen independently. Such general rectangles behave differently than squares—for which pointwise convergence in L^1 does hold for commuting flows, by Wiener's theorem. The situation here is analogous to the classical discussion of maximal inequalities for unrestricted rectangle averages on $\mathbb{R}^d$, and the maximal inequality of Theorem 7.1 can be viewed as a non-commutative generalization of the Jessen–Marcinkiewicz–Zygmund maximal inequality—see, e.g., [50] or [141, Chapter X, §2.2] for a discussion.

Let us make the following simple observation, which will be found useful below.

REMARK 7.4. Under the condition of Theorem 7.1, if at least one of the groups U or V is ergodic on (X,m) then the following pointwise ergodic theorem holds, for $f \in L^p(X)$, $1 < p < \infty$, and almost every $x \in X$:

$$\lim_{\min(T,S)\to\infty} \frac{1}{4TS} \int_{-T}^{T}\int_{-S}^{S} u_t v_s f(x)\, dt\, ds = \int_X f\, dm.$$

Indeed, the assumption amounts to the fact that $\mathcal{V}$ or $\mathcal{U}$ coincide with the projection $f \mapsto \int_X f\, dm$. Since in any case $\int_X \mathcal{U} f\, dm = \int_X \mathcal{V} f\, dm = \int_X f\, dm$, the conclusion follows.

We now note the following natural generalization of Theorem 7.1, as well as [60, Theorem 7.1], which will be used below.

THEOREM 7.5 [112]. *Let U and V be lcsc group, and assume that V is amenable. Let $E_T \subset U$ and $F_S \subset V$ be compact sets of finite Haar measure. Assume that F_S is a Følner family, and that both families of normalized averages satisfy the pointwise ergodic theorem and strong maximal inequality in L^p, $1 < p < \infty$. Let U and V act by measure-preserving transformations on a probability space (X,m). Then*

(1) *The strong L^p-maximal inequality, $1 < p \leqslant \infty$, holds for the averages given by*

$$\left\| \sup_{T,S>0} \left| \frac{1}{m_U(E_T) m_V(F_S)} \int_{E_T}\int_{F_S} uvf\, dm_U(u)\, dm_V(v) \right| \right\|_p \leqslant C_p \|f\|_p.$$

(2) *Let $\mathcal{U}$ (resp. $\mathcal{V}$) be the conditional expectation w.r.t. the σ-algebra of sets invariant under every $u \in U$ (resp. every $v \in V$). Then for every $f \in L^p(X)$, $1 < p < \infty$, and for almost every $x \in X$*

$$\lim_{\min(T,S)\to\infty} \frac{1}{m_U(E_T) m_V(F_S)} \int_{E_T}\int_{F_S} uvf(x)\, dm_U(u)\, dm_V(v) = \mathcal{U}\mathcal{V} f(x)$$

and the convergence is also in the L^p-norm.

(3) *If either U or V act ergodically, then the limit satisfies $\mathcal{UV} f(x) = \int_X f\, dm$.*
(4) *If both E_T and F_S satisfy the weak-type maximal inequality in L^1, then pointwise convergence holds also for $f \in L(\log L)(X)$.*

REMARK 7.6. The proof of Theorem 7.5 proceeds using an argument similar to the one given in the proof of Theorem 7.1 (and Remark 7.4). We note however that the Følner property of V plays a crucial role in the pointwise result. If we are interested only in norm convergence, then the assumptions that V is amenable and F_S are Følner are not necessary, and F_S can be any family of sets on V for which the mean ergodic theorem holds. This fact follows easily using the argument in the proof of part (3) of Theorem 7.1.

We also note that if both families of averages E_T and F_S satisfy the weak-type maximal inequality in L^1, then the combined averages satisfy the maximal inequality in $L(\log L)(X)$. This is a general property of the composition of two maximal operators both satisfying the weak-type maximal inequality, which is due to [50]. We refer to [112] for the details.

7.2. *The ergodic theory of semidirect products*

Continuing now with our theme of establishing pointwise ergodic theorems for group actions, assume now that the two groups U and V that figure in Theorem 7.5 generate an lcsc group. Thus we let $G = UV$ be a semidirect product, where $V \lhd G$ is a closed normal subgroup, and U a closed subgroup of G, so that $G \cong U \ltimes V$. The multiplication in $U \ltimes V$ is given by $(u_1, v_1)(u, v) = (u_1 u, u^{-1} v_1 u \cdot v)$, and the explicit isomorphism between $U \ltimes V$ and G is given by $\psi : (u, v) \mapsto uv$. The left Haar measure on $U \ltimes V$ is given by the following well-known recipe.

LEMMA 7.7. *The map $\psi : U \ltimes V \to G$ is homeomorphism that maps the product of the two left Haar measures $m_U \times m_V$ on the product space $U \times V$ to a left Haar measure on $G \cong U \ltimes V$.*

PROOF. The claim amount to the fact that for $f \in C_c(G)$, writing $f(uv) = F(u, v)$, the integral given by

$$\int_{G=UV} f(uv)\, d\psi_*(m_U \times m_V)(uv) = \int_{U \ltimes V} F(u, v)\, dm_U(u)\, dm_V(v)$$

is invariant under left translations. But since V is normalized by U, and m_U, m_V are left-invariant, translating by $g = u_1 v_1$ we get:

$$\begin{aligned}&\int_{UV} f\big((u_1 v_1 uv)\big)\, d\psi_*(m_U \times m_V)\\ &\quad = \int_U \left(\int_V F\big(u_1 u, u^{-1} v_1 u \cdot v\big)\, dm_V(v) \right) dm_U(u)\end{aligned}$$

$$= \int_U \left(\int_V F(u_1 u, v)\, dm_V(v) \right) dm_U(u)$$
$$= \int_{U\times V} F(u,v)\, dm_U\, dm_V = \int_{UV} f(uv)\, d\psi_*(m_U \times m_V). \qquad \square$$

Lemma 7.7 allows us to state a pointwise ergodic theorem for averages defined by certain sets of product type on $G = UV$, taken with the normalized left Haar measure as follows.

THEOREM 7.8 (Pointwise ergodic theorem for semidirect products with normal amenable subgroup [112]). *Let $G = UV$ be a semidirect product as above, where $V \lhd G$ is an amenable subgroup. Let $E_T \subset U$ and $F_S \subset V$ be as in Theorem* 7.5. *Assume that G acts ergodically by measure-preserving transformations on a probability space (X, m). Then for every $f \in L^p(X)$, $1 < p < \infty$, and almost every $x \in X$ we have*

$$\lim_{\min(T,S)\to\infty} \frac{1}{m_U(E_T) m_V(F_S)} \int_{u\in E_T} \int_{v\in F_S} uvf(x)\, dm_U(u)\, dm_V(v)$$
$$= \int_X f\, dm = \lim_{\min(T,S)\to\infty} \frac{1}{m_G(Q_{T,S})} \int_{Q_{T,S}} gf(x)\, dm_G(g),$$

where the convergence is also in the L^p-norm. Here $Q_{T,S} \subset G$ is the image of $E_T \times F_S \subset U \times V$ under the multiplication map ψ, and the average is w.r.t. the restriction of left-invariant Haar measure on G to $Q_{T,S}$.

Furthermore, if the families E_T and F_S satisfy the weak-type maximal inequality in L^1, then the averages on $Q_{T,S}$ satisfy the maximal inequality in $L(\log L)$.

PROOF. The fact that the pointwise (and the norm) limit exist follows from Theorem 7.5, and the limit is given by $\mathcal{U}\mathcal{V}f$. The only remaining issue is to identify the limit. But since V is a normal subgroup of G, for every function h invariant under V, its translate $\pi(g)h$ is also invariant under V. Therefore the space $\mathcal{I}_V$ of V-invariant functions is G-invariant, and hence it is also invariant under the projection $\mathcal{U}$, which is the limit in the strong operator topology of $\frac{1}{m_U(E_T)} \int_{E_T} u\, dm_U(u)$. Hence $\mathcal{U}$ maps the space $\mathcal{I}_V$ into the space $\mathcal{I}_V \cap \mathcal{I}_U$ of functions invariant under both U and V. Since G is ergodic, the latter space consists of the constant functions only, and hence the function $\mathcal{U}\mathcal{V}f \in \mathcal{I}_U \cap \mathcal{I}_V$ is a constant. Its value must be $\int_X f\, dm$ since the operators $\pi(u)$ and $\pi(v)$ are measure-preserving, and $\mathcal{U}$ and $\mathcal{V}$ are conditional expectations. The assertion regarding $Q_{T,S}$ follows from Lemma 7.7. For the last assertion stated, see Remark 7.6. $\square$

We note that Theorem 7.8 generalizes [60, Theorem 7.1], which considers the case where G is amenable, and identifies the limit only when Q_{T_n,S_n} is assumed to be a Følner sequence. In the case of amenable semidirect products it is possible to use Følner families on the constituent groups to construct explicitly a Følner family on G, a fact due to F. Greenleaf [59, Theorem 5.3], whose formulation follows.

THEOREM 7.9 (Construction of Følner sequence in semidirect products [59, Theorem 5.3]). *Assume G is an amenable lcsc group, and $G = UV$ is a semidirect product,*

where U and V are closed subgroups with V normal. Let $E_T \subset U$ and $F_S \subset V$ be Følner families. Then

(1) *There exist subsequences T_n and S_n such that the sets $J_n = Q_{T_n,S_n} \subset G$ associated with E_{T_n} and F_{S_n} constitute a Følner sequence in G.*
(2) *J_n can be chosen to be an increasing sequence whose union covers G, provided E_T and F_S have the same properties.*

7.3. *Structure theorems and ergodic theorems for amenable groups*

Theorem 7.8 and Theorem 7.9 can be exploited together with some structure theorems for various classes of lcsc groups to produce a variety of pointwise and maximal ergodic theorems. The first such result was established by F. Greenleaf and W. Emerson for connected amenable lcsc groups in [60, Theorem 3.1], and was later extended to all connected lcsc groups by A. Tempelman [145, Chapter 6, Theorem 8.4]). We will give a streamlined account of the connections between the structure theory of some classes of lcsc groups and the ergodic theorems they satisfy, and this will allow us to derive extensions of these results in several directions. Succinctly put, we can establish pointwise ergodic theorems for groups of the form $G = KP$, where K is compact and S solvable and normal, and also more generally for groups $G = KP$ where K is compact and P amenable, but not necessarily normal.

To begin with, we have the following corollary of Theorem 7.8.

THEOREM 7.10 (Pointwise ergodic theorem for algebraically connected amenable algebraic groups and connected amenable Lie groups [112]). *Let G be an amenable lcsc group, and $H = UV$ be any decomposition into a semidirect product of two closed subgroup, where V is normal. Assume that* (*keeping the notation of Theorem* 7.8) *$E_T \subset U$ and $F_S \subset V$ satisfy the pointwise ergodic theorem in L^p, $1 < p < \infty$, and that F_S is Følner.*

(1) *The family $Q_{T,S} \subset H$ also satisfies the pointwise ergodic theorem in L^p, $1 < p < \infty$, as $\min\{T, S\} \to \infty$ independently, namely in every ergodic H-space*

$$\lim_{\min\{T,S\}\to\infty} \frac{1}{m_G(Q_{T,S})} \int_{Q_{T,S}} gf(x)\, dm = \int_X f\, dm.$$

(2) *Every algebraically connected amenable algebraic group G over a locally compact non-discrete field admits a closed normal co-compact subgroup H with a semidirect product structure $H = UV$, with $E_T \subset U$ and $F_S \subset V$ satisfying the conclusion in* (1). *For a certain compact subgroup $K \subset G$, the normalized averages on the compact sets* (*of positive Haar measure in G*) *$Q'_{T,S} = KE_TF_S$, satisfy the conclusion in* (1).
(3) *The same conclusion holds for a connected amenable Lie group G, provided its maximal compact normal subgroup C is trivial. In general, the conclusion holds for the averages $Q''_{T,S} \subset G$ which are the inverse images of $Q'_{T,S} \subset G/C$.*

PROOF. The convergence statement in part (1) is of course an immediate consequence of Theorem 7.8.

Parts (2) and (3) state, first, that the groups G in the relevant category satisfy a structure theorem. For algebraic groups over F, this fact follows without difficulty from the fact that G has homomorphic image with finite kernel, which has a finite index subgroup isomorphic to a subgroup L of a proper parabolic subgroup of $GL(n, F)$ (see, e.g., [64, Theorem IV.2]. For L a decomposition of the form $L = KAN$ is valid, where N is the unipotent radical (which is nilpotent), A is Abelian (and thus $S = AN$ is solvable), and K compact. The existence of E_T and F_S with the required properties on the component groups (or simply on G itself) is thus clear as soon as the averages F_S are shown to exist on unipotent groups. This can be done by induction on the dimension, for instance.

As to connected Lie groups, assume that the maximal compact normal subgroup is trivial. Then G is well-known to be a compact extension of a solvable Lie group, i.e. $G = KS$, K compact, S closed normal solvable and co-compact (see, e.g., [64, Theorem IV.3], or [60]). The existence of a semidirect product decomposition for S with E_T and F_S as required follows, e.g., as in [60].

To complete the proof of (2) and (3), assume that the averages constructed from E_T and F_S on a co-compact normal subgroup $H = UV$ converge pointwise to an H-invariant function. If $G = KH$, K compact, then we claim that the averages $Q'_{T,S}$ on G converge pointwise to a G-invariant function. This follows since H is normal, and hence its space of invariants is invariant under G, and hence under K. Thus if f is an H-invariant limit of the averages on H, the average of its translates by K is still an H-invariant function. Thus f is a G-invariant function, hence the constant $\int_G f\,dm$ when G acts ergodically.

Finally, to complete the case of connected Lie groups, note that clearly any limit of the averages $Q''_{T,S} = m_C * Q''_{T,S} * m_C$ is a C-invariant function. We therefore consider the space of C-invariant functions, which is a G-invariant subspace on which G acts via G/C. The preceding argument applies, and any limit of $Q''_{T,S}$ is indeed invariant under G. □

REMARK 7.11.

(1) We note that by Theorem 7.9 a subsequence of sets Q'_{T_n,S_n} can be chosen which is an increasing sequence of Følner sets J_n in G, whose union covers G, and which satisfies the pointwise ergodic theorem in L^p, $1 < p < \infty$. Thus Theorem 7.10 generalizes the pointwise ergodic theorem for connected amenable Lie groups due to F. Greenleaf and W. Emerson [60].

(2) In the proof given in [60, Theorem 3.1], the identification of the limit is achieved only by restricting the averages, and choosing sequences $T_n \to \infty$ and $S_n \to \infty$ where the averages Q_{T_n,S_n} can be guaranteed to form a Følner sequence J_n. Then the strong limit of this sequence of averages must be projection onto G-invariant function (by Proposition 6.7), hence the constant $\int_X f\,dm$ in the ergodic case. The pointwise ergodic theorem is asserted in [60, Theorem 3.1] only for averaging along such Følner sequences in connected amenable Lie groups.

(3) Nevertheless, in fact pointwise convergence to the ergodic mean holds more generally, for the unrestricted averages on $Q_{T,S}$, as Theorem 7.10 show. Furthermore the assumption of amenability of G is also superfluous, as Theorem 7.8 shows. Indeed, as we saw the identification of the limit as a product of conditional expectations that the Dunford–Zygmund method provides, can replace the assumption of the exis-

tence of global Følner sets in G. Together with the existence of a normal amenable subgroup, this allows the conclusion that the limit must be the ergodic mean.

(4) Note also that the *existence problem* for such a Følner sequence in an *arbitrary* lcsc amenable group (not necessarily connected Lie) has been solved by the pointwise ergodic theorem for tempered Følner sequences, which holds in fact in L^1 (and not only L^p, $1 < p < \infty$), as stated in Theorem 6.10.

The Dunford–Zygmund method provides a great deal of useful information even in the case where the group G in question does not have any semidirect product structure, for example if G is a simple group. We now turn to discuss this possibility.

7.4. *Structure theorems and ergodic theorems for non-amenable groups*

Let us first recall the following well-known result regarding Haar measure (see, e.g., [52, Chapter V, §3, Proposition 12]), which generalizes Lemma 7.7. We remark that for Lie groups (and by the same argument for algebraic groups over locally compact non-discrete fields) a very simple proof is given, e.g., in [54, Chapter 2, §2.4], see also [84, Chapter V, §6].

LEMMA 7.12 (Haar measure on general products). *Let G be an lcsc group, and let P and K be two closed subgroups, such that $P \cap K$ is compact. Assume that $G = PK$. Then for every $f \in C_c(G)$ we have*

(1) $$\int_G f(g)\,dm_G(g) = \int_{P\times K} f(pk)\frac{\Delta_G(k)}{\Delta_K(k)}\,dm_P(p)\,dm_K(k),$$

where m_G, m_P and m_K are left-invariant Haar measures, and Δ_K and Δ_G are the modular functions of K and G.

(2) *In particular, if G is unimodular, then*

$$\begin{aligned}\int_G f(g)\,dm_G(g) &= \int_{P\times K} f(pk)\,dm_P(p)\Delta_K\left(k^{-1}\right)dm_K(k)\\ &= \int_{P\times K} f(pk)\,dm_P(p)\,d\eta_K(k),\end{aligned}$$

where $d\eta_K = \Delta_K(k^{-1})\,dm_K$ is right-invariant Haar measure on K.

(3) *If in addition K is unimodular, then $dm_G = dm_P\,dm_K$.*

Perhaps the simplest family of groups which are not semidirect products and for which Lemma 7.12 applies is the family of groups G with an Iwasawa decomposition. Namely G contains two closed subgroup P and K, with P amenable and K compact, such that $G = PK$, and neither subgroups is normal. This family however is extremely important, since as is well known, it contains all connected lcsc groups and all algebraically connected algebraic groups over locally compact non-discrete fields.

Together with Theorem 7.8, Lemma 7.12 can be used to derive a variety of pointwise and maximal ergodic theorems for averaging on compact sets with respect to Haar-uniform measure on them. Let us start with the following result, which utilizes also Theorem 7.10. We note that the third and fourth part of the following theorem generalize results of A. Tempelman [145, Chapter 6, Theorems 8.6, 8.7].

THEOREM 7.13 (Pointwise ergodic theorem for groups with an Iwasawa decomposition [112]). *Let G be a non-amenable lcsc group with an Iwasawa decomposition, $G = PK$, with P amenable and closed, and K compact. Let $E_T \subset P$ be a family of sets, giving rise to averages (w.r.t. left Haar measure on P) satisfying the maximal and pointwise ergodic theorem in L^p, $1 < p < \infty$. Then*

(1) *The Haar uniform averages on the sets $R_T = E_T K \subset G$ (normalized by left Haar measure on G) satisfy the maximal inequality in L^p, $1 < p < \infty$.*
(2) *In every probability G-space in which P acts ergodically, the averages in* (1) *converge pointwise to the ergodic mean, for every $f \in L^p$, $1 < p < \infty$.*
(3) *Every algebraically connected algebraic group over a locally compact non-discrete field, and every connected lcsc group, admit an Iwasawa decomposition $G = PK$, and a family $E_T \subset P$ such that in every ergodic action of G and for every $f \in L^p$, $1 < p < \infty$,*

$$\lim_{T\to\infty} \frac{1}{m_G(R_T)} \int_{R_T} gf(x)\, dm_G(g) = \int_X f\, dm, \quad \text{for almost all } x \in X.$$

(4) *Under the same assumption as in* (3), *pointwise convergence and the weak-type $(1,1)$-maximal inequality hold in fact for $f \in L^1(X)$, provided that the averages on $E_T \subset P$ satisfy the same properties. Such averages E_T can be chosen in every algebraically connected semisimple algebraic group.*

SKETCH OF PROOF. Parts (1) and (2) are immediate consequences of Lemma 7.12 and Theorem 7.5. (In fact Theorem 7.5 is superfluous in the present simple case, which can be deduced directly.)

Part (3) involves three ingredients. First, every algebraically connected semisimple algebraic group has an Iwasawa decomposition, and in particular, so does every connected semisimple Lie group with finite center. A general algebraic group is a semidirect product of a semisimple component M without compact factors, and an amenable (and algebraic) radical. Taking the product of the amenable radical and the minimal parabolic subgroup of the semisimple component M we clearly obtain an Iwasawa decomposition. Similarly, every connected non-amenable lcsc group G has a compact normal subgroup K_0 such that G/K_0 is a connected semisimple Lie group without compact factors and with trivial center. Clearly the inverse image of a minimal parabolic subgroup of G/K_0 again gives rise to an Iwasawa decomposition in G.

The second ingredient is the fact that the subgroup P figuring in the Iwasawa decomposition described above does admit sets E_T which satisfy the pointwise and maximal ergodic theorem in L^p, $1 < p < \infty$. This of course follows in both cases considered here from Theorem 7.10.

The third ingredient is the fact that P in the Iwasawa decomposition above does indeed always act ergodically if G does. This fact follows from the Howe–Moore mixing theorem [74] (see also [75]), as follows. The vanishing at infinity of matrix coefficients of unitary representations without invariant vectors of a simple algebraic group H implies that every element whose powers are not confined to a compact subgroup of H acts ergodically in ergodic H-spaces. This can be used to show that P as defined above is always ergodic in ergodic G-spaces, in both cases. Indeed, every P-invariant function is invariant under the solvable radical S of P. The latter is a normal subgroup of G, and so G leaves invariant the space of S-invariant functions. On the latter space G acts vis its factor group G/S, which is a semisimple group. Factoring further by the maximal compact normal factor of G/S we are thus reduced to the case of a semisimple group without compact factors, where the Howe–Moore mixing theorem can be applied.

As to part (4), we are considering the composition of the averages associated with $E_T \subset P$ (taken with normalized left-invariant Haar measure on P), composed with the constant bounded operator given by averaging on the compact set K. It is clear by definition that if the maximal function associated with averaging on E_T satisfies the weak-type $(1,1)$ maximal inequality in its own right, then it will still have this property when composed with a fixed bounded operator. Now for algebraically connected algebraic groups, $G = PK = ANK$, where A is a split torus. It follows that A admits a family of Følner sets with a pointwise and weak-type maximal theorem in L^1, and thus we can compose the averages on A with a fixed average on a compact set in N and averaging on K. The integration functional $f \mapsto \int_X f\,dm$ is invariant under such operators, and so the limit is still the ergodic mean. We refer to [112] for more details. □

EXAMPLE 7.14 (*Groups of graph automorphisms with an Iwasawa decomposition*). We note that the class of Iwasawa groups is very extensive, and contains many lcsc groups which are not algebraic. For example, consider a closed non-compact boundary-transitive subgroup G of the group of automorphisms of a bi-regular tree, or more generally of the automorphism group of a locally finite graphs with infinitely many ends. The stability group P of a point in the boundary is a closed amenable subgroup, and it has a compact complement K with $G = KP$. It is not hard to show that P itself has the structure of a semidirect product $P = \mathbb{Z} \ltimes K_P$, where K_P is compact (see [104] for the details). Hence P fulfills the hypotheses of Theorem 7.10 and so has Følner sets E_T satisfying the pointwise and maximal theorems in L^p, $1 < p < \infty$. By Theorem 7.13, $R_T = E_T K$ satisfies the maximal and pointwise theorem in G-actions in which P acts ergodically. However, by [95], typically for such groups G the subgroup P is indeed ergodic, and even mixing in every ergodic action of G, so the conclusion of part (3) of Theorem 7.13 holds in the context as well. A similar conclusion applies in other contexts as well, e.g., for certain groups of automorphisms of a product of bi-regular trees, which act transitively on the product of the boundaries of the trees.

REMARK 7.15. We note that it is possible to use Theorem 7.8 to deduce a completely different pointwise and maximal ergodic theorems for the algebraic groups in question. Indeed, represent an algebraic group $G = ML$ as a semidirect product of a semisimple algebraic group without compact factors and an amenable radical L. Take the sets $F_S \subset$

L provided by Theorem 7.10. If we can find $E_T \subset M$ which satisfy the maximal and pointwise ergodic theorem in L^p, $1 < p < \infty$, then by Theorem 7.8, the same will hold for the sets $Q_{T,S} \subset G$. Now, as we will see below, in every semisimple algebraic group without compact factors, the natural ball averages on G, bi-invariant under a maximal compact subgroup, do indeed satisfy the pointwise and maximal theorems. Thus another pointwise ergodic theorem is obtained from Theorem 7.8, different than Theorem 7.13. We refer to [112] for more details.

REMARK 7.16. The result stated in Theorem 7.13 can be greatly improved in many cases. For example, consider simple algebraic groups over a locally compact non-discrete field. Then for many ergodic actions of such a group (in fact for all actions if its split rank is at least two), it is possible to establish the following property, which is in striking contrast to the ergodic theorems we have been considering thus far. The convergence of the horospherical averages (see Section 12.2) associated with an Iwasawa decomposition to the ergodic mean takes place at an exponentially fast rate for almost every point. Furthermore this rate can be described explicitly, and (for groups of split rank at least two) is independent of the action altogether, namely it depends only on the group and the averages chosen. We will discuss this phenomenon in greater detail below in Theorem 12.6, and we refer to [108] and [112] for full details.

7.5. *Groups of bounded generation*

As noted in Remark 7.2, the Dunford–Zygmund theorem applies to any finite sequence of (say) one-parameter flows. Thus the Dunford–Zygmund method produces pointwise and maximal theorems for averages on subsets of Lie groups which admit global coordinates of the second kind. A natural extension of this concept here is that G can be parametrized by a product of (say) Abelian closed subgroups (none of which need to be normal or connected). Thus the method applies to an extensive class of lcsc groups, namely all lcsc groups of bounded generation, i.e. allowing global coordinates with coordinate subgroups isomorphic to $\mathbb{R}$ or $\mathbb{Z}$. By the same token, we could allow any finite sequence of amenable subgroups to be taken here, provided we take Følner averages satisfying the maximal and pointwise ergodic theorem.

It should be noted however, that in the present generality, the averages that are constructed by the Dunford–Zygmund method are usually no longer Haar-uniform averages on compact subsets with an explicit geometric or algebraic description. Giving up both the assumption of a *semidirect* product decomposition, as well as unimodularity of the component groups, implies that the averages defined in terms of global coordinates will involve certain densities w.r.t. Haar measure, and will be supported on certain compact sets arising from the product. Both the sets and the densities may be difficult to compute, in general.

To illustrate the point, let us formulate the following consequence of the Dunford–Zygmund method for subsets of lcsc groups admitting global coordinates of the second kind (compare [145, Chapter 6, Theorem 8.4]).

THEOREM 7.17 (Pointwise ergodic theorem for boundedly generated lcsc groups). *Let G be an lcsc group, and suppose that $F \subset G$ contains n closed amenable subgroups U_i,*

$1 \leqslant i \leqslant n$, (for example, isomorphic to $\mathbb{Z}$, $\mathbb{R}$, $\mathbb{Q}_p$ or $\mathbb{Q}_p^*$) such that the map $\psi : U_1 \times \cdots \times U_n \to F$ given by $\psi(u_n, \ldots, u_1) = u_n \cdots u_1$ is surjective. Let $(X, \mathcal{B}, m)$ be a standard Borel probability G-space. Let $E^i_{R_i} \subset U_i$ be Følner sets satisfying the pointwise and maximal ergodic theorem in L^p, $1 < p < \infty$. Then as $\min\{R_i, 1 \leqslant i \leqslant n\} \to \infty$, *for every $f \in L^p$, $1 < p < \infty$,*

$$T(R_n, \ldots, R_1) f(x) = \frac{\int_{E^n_{R_n}} \cdots \int_{E^1_{R_1}} \pi(u_n \cdots u_1) f(x)\, dm_{U_n}(u_n) \cdots dm_{U_1}(u_1)}{m_{U_n}(E^n_{R_n}) \cdots m_{U_1}(E^1_{R_1})} \longrightarrow \mathcal{U}_n \cdots \mathcal{U}_1 f(x)$$

for m-almost all $x \in X$, and in the L^p-norm. The corresponding maximal function satisfies a strong maximal inequality in L^p, $1 < p < \infty$. Furthermore, pointwise convergence also holds for $f \in L(\log L)^{n-1}(X)$, provided the averages on $E^i_{R_i}$ satisfy the weak-type maximal inequality in L^1 for $1 \leqslant i \leqslant n$.

We note that discrete groups of bounded generation, where each component group is cyclic, have attracted quite a bit of attention. It was shown by [21] that for any ring of integers $\mathcal{O}$ of an algebraic number field, the group $SL_n(\mathcal{O})$ is of bounded generation, provided $n \geqslant 3$. In fact, one can take all the cyclic subgroups to be generated by elementary matrices, whose number is estimated by an explicit function of n (for an estimate in the case of $SL_n(\mathbb{Z})$ see [3]). The problem of determining which arithmetic groups are boundedly generated has been shown to be closely connected to the congruence subgroup problem (see [129] and [94]). It was established in [144] that it is typically the case that S-arithmetic lattices in connected simply-connected absolutely simple algebraic groups defined over a number field, which have split rank at least two have the property of bounded generation.

Thus Theorem 7.17 gives the corollary that on all of these lattices the averages described in Theorem 7.17 satisfy a maximal inequality and converge pointwise. If at least one of the cyclic subgroups act ergodically, then the limit in Theorem 7.17 is $\int_X f\, dm$. This is the case for example if the action of the lattice $\Gamma \subset G$ is a restriction to Γ of an ergodic action of G, as follows from the Howe–Moore mixing theorem.

REMARK 7.18. Let us note that the following question remains unresolved by the discussion of the present section. Let G be an lcsc group with the structure indicated in Theorem 7.17. Does the family of operators $T(R_1, \ldots, R_n)$ of Theorem 7.17 satisfy the unrestricted pointwise ergodic theorem in L^p, $1 < p < \infty$, namely

$$\lim_{\min(R_i) \to \infty} T(R_1, \ldots, R_n) f(x) \to \int_X f\, dm$$

pointwise almost everywhere and in L^p-norm, in every ergodic action of G? We remark that in the original Dunford–Zygmund formulation, the flows need not satisfy any relation, and thus no further information can be expected on the resulting limit operator. Here we assume that the component groups in question at least all lie in one and the same lcsc

group. This does indeed imply that the limit operator is invariant under the group in favorable cases, but the general case remains unresolved. This problem is unresolved even for the case of connected amenable Lie groups which admit an iterated semidirect product structure of the form $U_1 \ltimes (U_2 \ltimes (\cdots (\ltimes U_n)))$, see [60]. It constitutes an interesting challenge even for simply-connected nilpotent Lie groups, particularly since the Emerson–Greenleaf theorem can be used in this context as an important ingredient in the proof by M. Ratner [130] of strict measure rigidity for actions of unipotent subgroups of solvable groups.

Let us also note that for a connected Lie group G, the averages given by $T(R_n, \ldots, R_1)$ in Theorem 7.17 and in Theorem 7.5 are usually very different than balls w.r.t. an invariant Riemannian metric. Whether there exists some choices of the parameters R_i which will give a family of averages which are comparable to balls w.r.t. an invariant Riemannian metric does not seem to be known, in general.

7.6. *From amenable to non-amenable groups: some open problems*

We have focused in the present section on pointwise ergodic theorems for amenable and non-amenable lcsc groups which have a common origin, namely the Dunford–Zygmund method. In the succeeding chapters we will return to the theme of establishing pointwise and maximal theorems for radial (and other geometric) averages for group actions. In contrast to Section 5 which considered groups of polynomial volume growth, our emphasis below will be on non-amenable groups (and thus with exponential volume growth), and particularly on connected semisimple Lie group, and more generally semisimple algebraic groups. Before continuing with the pointwise theory of radial averages, however, let us note the following fundamental problems related to mean ergodic theory and equidistribution, as well as certain pointwise convergence problems, all of which are unresolved.

(1) *The Mean Ergodic Theorem*.

Given any strongly continuous isometric representation $\tau : G \to \mathrm{Iso}(V)$ to the isometry group of a Banach space V, we can consider the operator $\tau(\mu) = \int_G \tau(g)\, d\mu(g) \in \mathrm{End}(V)$, which constitutes a convex average of the isometric operators $\tau(g)$, $g \in \operatorname{supp}\mu$. If there exists a projection $\mathcal{E} : V \to V^I$ where V^I is the closed subspace of $\tau(G)$-invariant vectors, then a result establishing that $\|\tau(\mu_t) f - \mathcal{E} f\| \to 0$ as $t \to \infty$ is called a mean ergodic theorem for the family μ_t in the representation τ. The study of such results for averages on $G = \mathbb{R}$ or $G = \mathbb{Z}$ is an extensive field, which forms one of the classical themes of operator ergodic theory and fixed point theory. It was also to some extent pursued for Følner averages on amenable groups, see, e.g., [119]. We note however that this problem is largely unresolved for many of the natural averages μ_t on groups G with exponential volume growth, even in the case of unitary representation in Hilbert spaces. For example, even for connected amenable Lie groups with exponential volume growth and the ball averages w.r.t. an invariant Riemannian metric this problem is completely open.

(2) *The equidistribution problem.*

When X is a compact metric space and G acts continuously, it is of course natural to ask under what conditions can we conclude that equidistribution holds for *all* orbits w.r.t. the invariant measure. Namely, when does

$$\lim_{t\to\infty} \pi(\mu_t) f(x) = \int_X f \, dm$$

for every continuous function $f \in C(X)$ and *every* $x \in X$ hold ? When μ_t are Følner averages on an amenable group equidistribution holds if and only if m is the unique G-invariant probability measure on X. This fact, which can be proved in the same way as in the classical case of $\mathbb{Z}$, clearly accounts for at least some of the popularity that Følner averages enjoy as an averaging method along the orbits. Results of this type have not been established, or disproved, for any non-amenable group, for any family of averages μ_t. The main source of the difficulty lies of course in the fact that taking the weak* limit of a subsequence of the measures $\mu_t * \delta_x$, it is usually not possible to show that the limiting measure is invariant under the group. In the special case of homogeneous spaces of connected Lie groups, an extensive theory of equidistribution has been developed. For a recent comprehensive discussion of equidistribution of orbits of non-amenable groups acting on homogeneous algebraic varieties we refer to [57].

(3) *Amenable groups of exponential volume growth.*

The ball averaging problem is completely open when the group G is a connected amenable Lie group of exponential volume growth. Namely, it is even unknown whether β_t, defined w.r.t. an invariant Riemannian metric, converges pointwise in L^2 (or in any L^p). We note that since ball averages do not form a regular family, do not have the Følner property, and in general are not comparable to the product averages of Theorem 7.17, none of the methods discussed thus far applies. In fact as noted in (1) above even the mean ergodic theorem has not been established.

(4) *Pointwise convergence: the L^1-problem.*

It is unknown if for a left-invariant Riemannian metric on *any* connected Lie group of exponential volume growth. the normalized ball averages β_t satisfy the pointwise ergodic theorem in L^1. This has not been established even in one case, as far as we know. Furthermore, in Theorem 7.10 the averages to which the Dunford–Zygmund method applies are shown to converge only for $f \in L(\log L)$. Thus in the class of exponential solvable Lie groups for example, the only pointwise ergodic theorem in L^1 established so far is for a tempered sequence of Følner averages—see Section 6.3.

8. Spherical averages

In the following three sections, we will again concentrate on radial analysis, namely consider averages on balls associated with an invariant metric on G. However, we will now concentrate on the case when G has exponential volume growth, and our prime examples will be non-compact semisimple Lie and algebraic groups. These groups being non-amenable, they do not admit an asymptotically invariant sequence and no transfer principle

has been established for them. The volume of balls obviously does not satisfy the doubling property, and so the Wiener covering argument, even for convolutions, does not apply. Thus none of the arguments that were useful in the polynomial volume growth case is relevant here.

Nevertheless it is possible to develop a systematic theory of radial averages which elucidates the basic analytic facts about them (such as maximal inequalities for convolutions) and to establish ergodic theorems satisfied in general measure-preserving actions of the group.

Two key ideas that will be employed in our analysis are as follows. First, for simplicity of exposition, let us assume that the volume of the balls has exact exponential growth in terms of the radius (as is indeed the case for split rank one groups). Clearly the volume of a shell of unit width, namely $B_{t+1} \setminus B_t$, occupies a fixed proportion of the volume of the ball B_{t+1}. It then follows that the ball averages and the shell averages have equivalent maximal operators, so that we can restrict the discussion to the shell averages. On the other hand, the maximal function for the shell averages in Euclidean space is in fact equivalent to the maximal function for the *singular* spherical averages (see [111] for more details). As we shall see below, the first key idea is that it is the ideas and techniques of classical singular integral theory, and particularly Tauberian theory, that are most suitable for the analysis of shell (and hence ball) averages on semisimple groups with exponential volume growth. The second key idea is to apply spectral methods based on the unitary representation theory of the group in question to prove maximal inequalities and pointwise convergence theorems. This is indeed possible in the case of radial averages on semisimple Lie and algebraic groups, but also in many other cases, namely whenever the singular sphere averages all commute under convolution. Our methods will thus apply in principle to all lcsc groups that admit a radial commutative convolution structure, even to amenable ones, and in fact give rise to some interesting results regarding singular averages also on groups of polynomial volume growth.

As we shall see below, the geometric reduction from ball averages to shell averages together with the use of spectral methods from singular integral theory and the unitary representations theory of the groups restricted to the commutative convolution subalgebra will serve to replace the growth and Følner conditions on the group used in the polynomial volume growth case, and the existence of the algebraic semidirect product structure and the Dunford–Zygmund method used otherwise. Before turning to a discussion of the analytic tools involved, we present some basic examples which will serve to motivate our analysis and demonstrate its scope, and formulate some of the pertinent results which will be proved later on. We begin with the following fundamental result on singular averages on $\mathbb{R}^n$.

8.1. *Euclidean spherical averages*

Our first example of an ergodic measure preserving action of $\mathbb{R}$ was the flow given by an irrational line on the 2-torus (see Example 3.1). Let us now note that there is an equally basic problem related to geometric averaging on the plane, as follows.

EXAMPLE 8.1 (*Circles in* $\mathbb{R}^2$ *and spheres in* $\mathbb{R}^n$). Let us denote by S_t a circle of radius t and center 0 in $\mathbb{R}^2$. Let σ_t denote the normalized rotation-invariant measure on S_t, and

we consider the radial averages (= spherical means) that σ_t define on $\mathbb{R}^2$. If f a function on $\mathbb{R}^2$, define:

$$\pi(\sigma_t) f(v) = \int_{w \in S_t} f(v+w)\, d\sigma_t(w).$$

In analogy with Example 3.1 in Section 3.1, we focus on the action of the spherical means as operators on function spaces on $\mathbb{T}^2$, namely the action on $\mathbb{Z}^2$-periodic functions. We can then consider the natural problems of equidistribution, pointwise almost everywhere convergence to the space average, and (singular!) differentiation. These problems have all been resolved, and we recall the following results, which we formulate for the action of the spherical means on the n-torus, for use in later comparisons.

(1) *Equidistribution of spheres.*
$\forall v \in \mathbb{T}^n, \forall f \in C(\mathbb{T}^n)$

$$\lim_{t\to\infty} \pi(\sigma_t) f(v) = \int_{\mathbb{T}^2} f\, dm \quad \text{(Exercise!)}.$$

(2) *Maximal inequality for sphere averages.*
$\forall f \in L^p(\mathbb{T}^n)$, $p > \frac{n}{n-1}$,

$$\left\| \sup_{t>0} \left|\pi(\sigma_t) f\right| \right\|_{L^p(\mathbb{T}^n)} \leqslant C_p(n) \|f\|_{L^p(\mathbb{T}^n)}.$$

(3) *Singular spherical differentiation.*
$\forall f \in L^p(\mathbb{T}^n)$, $p > \frac{n}{n-1}$,

$$\lim_{t\to 0} \pi(\sigma_t) f(v) = f(v), \quad \text{for almost all } v \in \mathbb{T}^n.$$

(4) *Pointwise Ergodic Theorem for sphere averages.*
$\forall f \in L^p(\mathbb{T}^n)$, $p > \frac{n}{n-1}$,

$$\lim_{t\to\infty} \pi(\sigma_t) f(v) = \int_{\mathbb{T}^n} f dm, \quad \text{for almost all } v \in \mathbb{T}^n.$$

REMARK 8.2.

(1) The identification of the limit in (1) and (4) as the space average of the function (namely the mean ergodic theorem), is a simple exercise in spectral theory as will also be seen more generally below. The equidistribution theorem can be proved using a variant of the classical argument of Weyl for the action of a translation on $\mathbb{T}^n$—it requires only the vanishing at infinity of the characters of the convolution algebra of radial averages, which is a consequence of the Riemann–Lebesgue lemma.

(2) The fundamental result underlying the pointwise convergence theorems (3) and (4) is the maximal inequality (2), which is due to E. Stein (see [142]) for $n \geqslant 3$ and to J. Bourgain [14] for $n = 2$. The range of p stated is best possible.

(3) As usual, the pointwise convergence in the ergodic theorem (and the differentiation theorem) follows if it holds on a dense subspace. The existence of such a subspace was first established for $n \geqslant 3$ in [80], and for $n = 2$ in [90].
(4) The spherical averages σ_t being singular measures on $\mathbb{R}^n$ accounts for the non-trivial restriction on the range p for which the pointwise ergodic theorem holds. This phenomenon was not encountered in the discussion of the absolutely continuous averages that appeared thus far in the previous sections.
(5) The discrete analog of the spherical means are the averages over integer points lying on a discrete sphere $S_k \cap \mathbb{Z}^n$. Maximal inequalities for the convolution operators defined by such averages on $\mathbb{Z}^n$ were recently established for $n \geqslant 5$ in [96].

REMARK 8.3 (*Measurability of singular maximal functions*). Since the averages σ_t are singular, it is not clear why the maximal function f_σ^* is well-defined and measurable even for one function class in $L^p(\mathbb{T}^n)$, including the function class 0. The maximal inequality and pointwise theorem should be interpreted as asserting, in particular, that for any two representatives f and f' of a given function class, there exists a co-null set such that for v in this set, both $\pi(\sigma_t)f(v)$ and $\pi(\sigma_t)f'(v)$ exist *for all* $t > 0$ simultaneously and are equal. Then the supremum in the maximal inequality and the limit in the ergodic theorem are indeed well-defined.

This material problem is discussed in detail, e.g., in [141, Chapter XI, §3.5] or alternatively in [36, II.4], and we will content ourselves here with noting that $\sup_{t>0} |\sigma_t f(v)|$ is clearly defined for all v and constitutes a measurable function if f is continuous on $\mathbb{T}^n$. The strong L^p-maximal inequality for continuous functions (which is sometimes referred to as an a-priori inequality) can then be used to define and prove the measurability and the L^p boundedness of the maximal function for all L^p-functions. This remark applies to all other maximal inequalities that will appear in the sequel, since the spaces we consider can always be assume to be compact metric with the G-action continuous—see [106] for details.

Another basic set-up in which spherical averages appear naturally is furnished by the Heisenberg groups.

EXAMPLE 8.4 ($\mathbb{C}^n$*-spheres in the Heisenberg group*). Let $H_n = \mathbb{C}^n \times \mathbb{R}$ denote the Heisenberg group, and let σ_t denote the normalized rotation invariant measure on the sphere $S_t \subset \mathbb{C}^n$ with center 0 (which we call $\mathbb{C}^n$-spheres). Let $H_n(\mathbb{Z})$ be the discrete subgroup of integer points of the Heisenberg group. Consider the homogeneous space given by $U_n = H_n(\mathbb{Z}) \setminus H_n$, which is compact nilmanifold with a transitive (right) H_n-action. We then have the following results

(1) *Equidistribution of* $\mathbb{C}^n$*-spheres in* U_n.
$\forall v \in U_n, \forall f \in C(U_n)$

$$\lim_{t\to\infty} \pi(\sigma_t) f(v) = \int_{U_n} f \, dm.$$

(2) *Maximal inequality for* $\mathbb{C}^n$*-sphere averages.*

$\forall f \in L^p(U^n)$, $n > 1$, $p > \frac{2n}{2n-1}$,

$$\left\| \sup_{t>1} |\pi(\sigma_t) f| \right\|_{L^p(U^n)} \leqslant C_p(n) \| f \|_{L^p(U^n)}.$$

(3) *Singular $\mathbb{C}^n$-sphere differentiation on U^n.*
$\forall f \in L^p(U^n)$, $p > \frac{2n}{2n-1}$, $n > 1$

$$\lim_{t\to 0} \pi(\sigma_t) f(v) = f(v), \quad \text{for almost all } v \in U^n.$$

(4) *Pointwise Ergodic Theorem for $\mathbb{C}^n$-sphere averages.*
$\forall f \in L^p(U^n)$, $n > 1$, $p > \frac{2n}{2n-1}$,

$$\lim_{t\to \infty} \pi(\sigma_t) f(v) = \int_{U_n} f \, dm, \quad \text{for almost all } u \in U_n.$$

REMARK 8.5.

(1) The mean ergodic theorem and also the equidistribution theorem here can be proved spectrally, again in a manner analogous to Weyl's classical equidistribution theorem on $\mathbb{T}^n$. A necessary ingredient in this approach is thus the classification of all the characters of the commutative convolution algebra generated by the $\mathbb{C}^n$-sphere averages on the Heisenberg group.

(2) The p-range in (2) and (4) are the best possible for the action on U^n, and in fact for general probability-preserving actions of the reduced Heisenberg group. This result was established in [116].

(3) The maximal inequality for convolutions on the Heisenberg group itself, namely for the operator $\sup_{t>0} |f * \sigma_t|$, was recently established in [101] for $p > \frac{2n}{2n-1}$, excluding the case $n = 1$. This implies the differentiation theorem, and by the transfer principle for singular averages stated in Theorem 6.2 the same maximal inequality holds for any probability-preserving action of the Heisenberg group, and the pointwise theorem holds as well. Another proof of this result was given by [102]. The range $p > \frac{2n-1}{2n-2}$ for $n > 1$ was established earlier in [116].

(4) Note that in the result above σ_t are singular averages supported on *subvarieties of codimension* 2, rather than codimension one as in the case of ordinary spheres (see [36] for the corresponding result for the spheres of codimension one associated with the natural homogeneous norm). Results for even more singular averages on certain nilpotent Lie groups appear in [101].

Let us now pass from the considerations above regarding spherical averages in the familiar setting of nilpotent groups, and consider the ergodic theory of spherical averages for lcsc groups which are of exponential volume growth, and non-amenable. We will start with the simplest example, namely that of the isometry groups of hyperbolic space, acting on homogeneous spaces with finite volume.

8.2. *Non-Euclidean spherical averages*

EXAMPLE 8.6 (*Spheres in hyperbolic space*). Let M be a compact (or finite volume) Riemann surface, and $\Gamma = \pi_1(M)$ its fundamental group. Γ is naturally identified with a lattice subgroup of the isometry group $G = PSL_2(\mathbb{R}) = \mathrm{Iso}(\mathbb{H}^2)$ of the hyperbolic plane $\mathbb{H}^2$. G acts by translations on the homogeneous space $\Gamma \setminus G$, and if $K \subset G$ is the maximal compact subgroup of rotations fixing a point x, then M can be identified with the double coset space $\Gamma \setminus G/K$. Denote the unique probability measure supported on a sphere of radius t and center p in hyperbolic space $\mathbb{H}^2 = G/K$, which is invariant under the rotations fixing x, by $\tilde{\sigma}_t(x)$. Given any (continuous, say) function on M, we can lift it to a Γ-periodic function on $\mathbb{H}^2$ and average it w.r.t. $\tilde{\sigma}_t(x)$. We denote the result of this operation by $\pi(\sigma_t) f(x)$.

More generally, if $\mathbb{H}^n$ denotes hyperbolic n-space, and Γ is a lattice in $G = \mathrm{Iso}(\mathbb{H}^n)$, we can consider the homogeneous space $M = \Gamma \setminus G/K$, K a maximal compact subgroup fixing $x \in \mathbb{H}^n$. M is an n-dimensional Riemannian manifold of constant negative sectional curvature provided Γ is torsion free. We then define the averaging operators $\pi(\sigma_t) f(x)$ corresponding to the spherical means $\tilde{\sigma}_t(x)$, acting on Γ-periodic functions on $\mathbb{H}^n = G/K$. We denote the unique G-invariant probability measure on $\Gamma \setminus G$ by m, as well as its projection onto $M = \Gamma \setminus G/K$. We can now state the following results (for general dimension n).

(1) *Equidistribution of spheres in compact hyperbolic homogeneous spaces.*
When M is compact, $\forall x \in M$, $\forall f \in C(M)$

$$\lim_{t\to\infty} \pi(\sigma_t) f(x) = \int_M f\, dm.$$

(2) *Maximal inequality for sphere averages in hyperbolic finite-volume homogeneous spaces.*
For any M of finite volume, $\forall f \in L^p(M)$, $p > \frac{n}{n-1}$, $n \geqslant 2$, and for almost every $x \in M$

$$\left\| \sup_{t>0} \left| \pi(\sigma_t) f \right| \right\|_{L^p(M)} \leqslant C_p(n) \| f \|_{L^p(M)}.$$

(3) *Singular spherical differentiation in hyperbolic homogeneous spaces.*
$\forall f \in L^p(M)$, $p > \frac{n}{n-1}$, $n \geqslant 2$ and for almost every $x \in M$

$$\lim_{t\to 0} \pi(\sigma_t) f(x) = f(x).$$

(4) *Pointwise Ergodic Theorem for sphere averages in hyperbolic finite-volume homogeneous spaces.*
For any M of finite volume, $\forall f \in L^p(M)$, $p > \frac{n}{n-1}$, and almost every $x \in X$

$$\lim_{t\to\infty} \pi(\sigma_t) f(x) = \int_M f\, dm.$$

REMARK 8.7.

(1) The equidistribution of spheres was first proved by G.A. Margulis, who used the mixing property of the geodesic flow. Also the same result holds for any hyperbolic homogeneous space of finite volume, not only compact ones, provided we restrict the function f to be continuous of compact support. Further proofs of these results were also given by [82,136] and [85]. We note that mixing of the geodesic flow in the case of constant negative curvature is fairly straightforward, as it follows from the general fact that the matrix coefficients of irreducible non-trivial unitary representations of the isometry group of hyperbolic space vanish at infinity. This fact was noted already by Fomin and Gelfand [55], and was later generalized to the Howe–Moore mixing theorem [74].

(2) The L^2-maximal inequality in (2) and the pointwise ergodic theorem in (4) were proved in [106] for $\dim M > 2$. Interpolation arguments were used in [107] and [115] to establish the range as $p > \frac{n}{n-1}$, which is best possible. The case of $n = 2$ is treated in [110], and is based on the results of [76].

(3) The differentiation theorem and the maximal inequality for convolutions on $\mathbb{H}^n$, $n > 2$, is due to [45], and for $n = 2$ to [76]. Using the transfer principle of Theorem 6.2 (but for the local operator $\sup_{0<t\leqslant 1} |\pi(\sigma_t) f(x)|$ only!) this result holds also on M.

8.3. *Radial averages on free group*

EXAMPLE 8.8 (*Spheres in the free group*). Let $\mathbb{F}_2$ be the free group on two generators $\mathbb{F}_2 = \mathbb{F}_2(a, b)$, where a and b are two free generators. The symmetric generating set $S = \{a, b, a^{-1}, b^{-1}\}$ determines a word length on $\mathbb{F}_2$ given by (see the discussion in Section 4.1)

$$|w|_S = \min\{k;\ w = s_{i_1} \cdots s_{i_k},\ s_{i_j} \in S, 1 \leqslant j \leqslant k\}.$$

$d(u, v) = |u^{-1} \cdot v|_S$ is a metric on $\mathbb{F}_2$, which is invariant under the action of $\mathbb{F}_2$ on itself by left translations. The word metric determines the corresponding spheres and balls, and thus also the normalized averaging operators σ_n and β_n on them.

Now consider the sphere $\mathbb{S}^d \subset \mathbb{R}^{d+1}$, where $d \geqslant 2$, and let A and B be two orthogonal linear transformations on $\mathbb{S}^d$. Clearly, the assignment $a \mapsto A$ and $b \mapsto B$ extends uniquely to a homomorphism $\mathbb{F}_2 \mapsto O_d(\mathbb{R})$, which defines an action of $\mathbb{F}_2$ by orthogonal transformations on $\mathbb{S}^d$. This action preserves, in particular, the rotation-invariant probability measure m on the sphere. Let us assume for simplicity that the image of $\mathbb{F}_2$ is contained in the connected component of the identity in $O_d(\mathbb{R})$.

(1) *Equidistribution of spherical averages on free group orbits in the unit sphere.* $\forall x \in \mathbb{S}^d$, $\forall f \in C(\mathbb{S}^d)$

$$\lim_{t\to\infty} \pi(\sigma_t) f(x) = \int_{\mathbb{S}^d} f\, dm.$$

(2) *Mean ergodic theorem for spherical averages on free group orbits in the unit sphere.* $\forall f \in L^2(\mathbb{S}^d)$

$$\lim_{t\to\infty} \left\| \pi(\sigma_t) f - \int_{\mathbb{S}^d} f \, dm \right\|_{L^2(\mathbb{S}^d)} = 0.$$

(3) *Maximal inequality for spherical averages on free group orbits in the unit sphere.* $\forall f \in L^p(\mathbb{S}^d)$, $p > 1$, and for almost every $x \in \mathbb{S}^d$

$$\left\| \sup_{t>0} \left| \pi(\sigma_t) f \right| \right\|_{L^p(\mathbb{S}^d)} \leqslant C_p \| f \|_{L^p(\mathbb{S}^d)}.$$

(4) *Pointwise Ergodic Theorem for spherical averages on free group orbits in the unit sphere.*
$\forall f \in L^p(\mathbb{S}^d)$, $p > 1$, and almost every $x \in X$

$$\lim_{t\to\infty} \pi(\sigma_t) f(x) = \int_{\mathbb{S}^d} f \, dm.$$

REMARK 8.9.

(1) The equidistribution result was proved by V. Arnold and A. Krylov in [4], where in fact they consider an arbitrary connected homogeneous space of a compact Lie group, rather than just the sphere. Furthermore, the question of generalizing von Neumann mean ergodic theorem and Birkhoff's pointwise ergodic theorem for the sphere and ball averages on the free group is raised explicitly in [4]. The analogous problems for averages on the group of isometries of hyperbolic n-space in also raised there.

(2) The mean ergodic theorem for general probability-preserving action of the free group on r generators was proved by Y. Guivarc'h [66]. The formulation is slightly different, asserting the convergence (in the strong operator topology) of the operators $\sigma_n' = \frac{1}{2}(\sigma_n + \sigma_{n+1})$. This modification is necessary because in a general ergodic action a function f might satisfy $\pi(\sigma_n) f = (-1)^n f$, a situation that does not arise on $\mathbb{S}^d$ because of our density assumption. For more on this periodicity phenomenon, see Section 10.5 below.

(3) The maximal inequality and the pointwise ergodic theorem for σ_n' acting on L^2 functions were established in [105], for all measure-preserving ergodic actions of the free groups. The extension to L^p, $p > 1$, was established in [114].

All the foregoing examples fall under our general theme of study, which is the analysis of averaging operators arising from families of probability measures μ_t on an lcsc group G in a probability-measure-preserving action of G.

(1) In Example 3.1 the group of course is the real line $\ell \cong \mathbb{R}$, acting by translations on $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$, $\mu_t = \beta_t =$ ball averages on $\mathbb{R}$.

(2) In Example 8.1, the group is $\mathbb{R}^2$, again acting by translation on $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$, and $\mu_t = \sigma_t =$ the normalized circle averages on $\mathbb{R}^2$.

(3) In Example 8.3 the group is the Heisenberg group $H_n = \mathbb{C}^n \times \mathbb{R}$, acting by translations on the homogeneous space $U_n = H_n(\mathbb{Z}) \setminus H_n$ and $\mu_t = \sigma_t$ are the $\mathbb{C}^n$-sphere averages supported in $\mathbb{C}^n \subset H_n$.
(4) In Example 8.5 the group is $\mathrm{Iso}(\mathbb{H}^n)$, the measures are the unique bi-K-invariant measures σ_t on G projecting to the normalized rotation-invariant measure $\tilde{\sigma}_t$ on a sphere of radius t and center $[K]$ in $\mathbb{H}^n = G/K$, and the action is on the homogeneous space $\Gamma \setminus G$, with its unique invariant probability measure.
(5) In Example 8.8 the group is $\mathbb{F}_2$ and the space is $\mathbb{S}^d$ with its unique isometry-invariant probability measure m.

9. The spectral approach to maximal inequalities

We now turn to an exposition of a spectral approach to ergodic theorems for certain lcsc groups, including semisimple Lie and algebraic groups, and some of their lattices. We will start by demonstrating the method for the basic case of ball and sphere averages in the most accessible connected Lie groups, as follows.

9.1. *Isometry groups of hyperbolic spaces*

Our basic set-up and notation will be as follows:
(1) $\mathbb{H}^n$ = hyperbolic n-dimensional space, with connected isometry group $G = \mathrm{Iso}^0(\mathbb{H}^n)$.
(2) $S_t(x)$ = sphere of radius t with center $x \in \mathbb{H}^n$.
(3) $\tilde{\sigma}_t(x)$ = normalized measure on $S_t(x)$, invariant under the group of rotations fixing x.
(4) σ_t = the spherical averages on G. These are given by $\sigma_t = m_K * \delta_{a_t} * m_K$, where $\{a_t, t \in \mathbb{R}\}$ satisfies $d(a_t o, o) = |t|$, namely its orbit through $o = [K]$ in the symmetric space $\mathbb{H}^n = G/K$ is a geodesic. σ_t is the unique bi-K-invariant probability measure on G projecting onto $\tilde{\sigma}_t$.
(5) (X, m) a compact metric space with a continuous G-action, where m is a G-invariant probability measure.
(6) Radial averages (= spherical means) are defined by, for $f \in C(X)$,

$$\pi(\sigma_t) f(x) = \int_G f\left(g^{-1}x\right) d\sigma_t(g).$$

The basic example discussed in Section 8.2 for the set-up above arises when choosing Γ to be a discrete group of isometries with fundamental domain of finite volume. Then for $f \in C(\Gamma \setminus G)$, namely a Γ-periodic function on $\mathbb{H}^n$, the averages above give

$$\pi(\sigma_t) f(x) = \text{ average of } f \text{ on } S_t(x) \text{ w.r.t. } \tilde{\sigma}_t(x).$$

REMARK 9.1. We note that the assumption above that our basic Borel G-space with G-invariant probability measure is a compact metric G-space is without loss of generality,

as noted in [106]. In that case for each $F \in C_c^\infty(G)$, and every $f \in C(X)$, the function $h = \pi(F)f$ has the property that $g \mapsto h(g^{-1}x)$ is a C^∞-function on G, for every $x \in X$. Clearly the space of such functions is dense in $C(X)$ in the uniform norm, and also in every $L^p(X)$, in the L^p-norm, $1 \leqslant p < \infty$. We will thus consider below differentiation operators applied to functions in $C_c^\infty(G) * C(X)$ in a general action without further comment.

9.2. *Commutativity of spherical averages*

To analyze the spherical means, we use the following basic observation, originating with Gelfand and Selberg in the 1950's.

PROPOSITION 9.2. *The spherical averages $\sigma_t = \sigma_t(o)$ on* $\mathrm{Iso}(\mathbb{H}^n)$ *commute with one another under convolution.*

PROOF. Consider a one-parameter group of isometries $A = \{a_t, t \in \mathbb{R}\}$, whose orbit through a given point o forms a geodesic in $\mathbb{H}^n$, namely $d(a_t o, o) = |t|$. Now the probability measure $m_K * \delta_{a_t} * m_K$ on G projects under $G \to G/K$ to the unique rotation invariant probability measure on G/K, supported on a sphere of radius t and center $o = [K]$. Since $d(a_t o, o) = d(o, a_{-t} o)$, and K is transitive on each sphere with center o, we have $a_t = k a_{-t} k'$, and so $K a_t K = K a_{-t} K$. Hence the inversion map $g \mapsto g^{-1}$ restricts to the identity on bi-K-invariant sets, functions and measures, but also reverses the order of convolution. Now $\sigma_t * \sigma_s$ is bi-K-invariant since σ_t and σ_s are, and hence

$$(\sigma_t * \sigma_s)^\vee = \sigma_t * \sigma_s = \sigma_s^\vee * \sigma_t^\vee = \sigma_s * \sigma_t. \qquad \square$$

NOTATION. We denote the algebra of bounded complex bi-K-invariant Borel measures on G by $M(G, K)$.

Given any strongly continuous unitary representation π of G on a Hilbert space $\mathcal{H}$, each element μ of $M(G, K)$ is mapped to a bounded operator $\pi(\mu)$. The map $\mu \mapsto \pi(\mu)$ is a continuous algebra homomorphism, commuting with the involutions on $M(G, K)$ and on $\mathrm{End}\,\mathcal{H}$. When $M(G, K)$ is commutative, we denote by $\mathcal{A}$ the closure in the operator norm topology of $\pi(M(G, K))$, which is a commutative algebra closed under the adjoint operation, and so a commutative C^*-algebra. Thus we can appeal to the following fundamental result.

SPECTRAL THEOREM. *Let $\mathcal{A}$ be a commutative norm-closed algebra of bounded operators on a Hilbert space $\mathcal{H}$, closed under taking adjoints. Consider the $*$-spectrum $\Sigma^*(\mathcal{A})$ of $\mathcal{A}$, consisting of continuous complex $*$-characters of $\mathcal{A}$, with the w^*-topology inherited from the dual $\mathcal{A}^*$ of $\mathcal{A}$. Every $f \in \mathcal{H}$ determines a spectral measure ν_f on $\Sigma^*(\mathcal{A})$, and the action of an operator $\mu \in \mathcal{A}$ is given by the formula*:

$$\langle \pi(\mu) f, f \rangle = \int_{\varphi \in \Sigma^*(\mathcal{A})} \varphi(\mu)\, d\nu_f(\varphi).$$

Functional calculus. The spectral theorem for $\mathcal{A}$ implies that for any bounded measurable function F on $\Sigma^*(\mathcal{A})$, $\pi(F)$ can be interpreted as a bounded operator on $\mathcal{H}$, given by the formula

$$\langle \pi(F) f, f \rangle = \int_{\varphi \in \Sigma^*(\mathcal{A})} F(\varphi)\, d\nu_f(\varphi).$$

Furthermore, the functional calculus can be extended to more general distributions on the spectrum, including measures and also derivative operators. We will make extensive use of these facts below.

9.3. *Littlewood–Paley square functions*

We now turn to a proof of the following

THEOREM 9.3 (Pointwise ergodic theorem in L^2 for sphere and ball averages on $\mathrm{Iso}(\mathbb{H}^n)$, $n > 2$, [106]). *The sphere averages σ_t and the ball averages β_t on* $\mathrm{Iso}(\mathbb{H}^n)$ *satisfy the pointwise ergodic theorem and the strong maximal inequality in $L^2(X)$, if $n > 2$, for any probability measure-preserving action of G.*

For simplicity of exposition, we will consider the following

Model case: *Proof of the pointwise ergodic theorems for σ_t and β_t in $SL_2(\mathbb{C})$-actions on compact metric spaces.*

The proof proceeds along the following steps (we suppress the notation π for the representation for ease of notation):

(1) First, consider the *uniform average* μ_t of the spherical measures σ_s, $0 < s \leqslant t$, and use Proposition 9.2 to write:

$$\mu_t = \frac{1}{t}\int_0^t \sigma_s\, ds = m_K * \frac{1}{t}\int_0^t \delta_{a_s}\, ds * m_K.$$

Since $\frac{1}{t}\int_0^t \delta_{a_s}\, ds$ are the Birkhoff averages on $\mathbb{R} \cong A$, they satisfy a strong maximal inequality in every L^p, $1 < p < \infty$. It follows immediately that also $\|f_\mu^*\|_p \leqslant C_p \|f\|_p$.

(2) Next, compare σ_t to their uniform average μ_t. Using Remark 9.1, for every function $f \in C_c^\infty(G) * C(X)$, the function $g \mapsto f(g^{-1}x)$ is a C^∞ function on G (and thus $s \mapsto \sigma_s f(x)$ is C^∞ on $\mathbb{R}_+$), and we can write:

$$\sigma_t f(x) - \mu_t f(x) = \frac{1}{t}\int_0^t s \frac{d}{ds}\sigma_s f(x)\, ds.$$

By the Cauchy–Schwarz inequality:

$$\left|\sigma_t f(x) - \mu_t f(x)\right| \leqslant \frac{1}{t}\left(\int_0^t s\, ds \int_0^t s \left|\frac{d}{ds}\sigma_s f(x)\right|^2 ds\right)^{1/2}.$$

(3) Estimating, we have:

$$\sup_{t\geqslant 0}\left|\sigma_t f(x)\right| \leqslant \sup_{t\geqslant 0}\left|\mu_t f(x)\right| + R(f,x),$$

where we define:

$$R(f,x)^2 = \int_0^\infty s\left|\frac{d}{ds}\sigma_s f(x)\right|^2 ds.$$

$R(f,x)$ is called the Littlewood–Paley square function.

(4) We have by (3):

$$\left\|f_\sigma^*\right\|_2 \leqslant \left\|f_\mu^*\right\| + \left\|R(f,\cdot)\right\|_2.$$

We now compute the norm of the square function by the spectral theorem, namely by going over to the Fourier–Gelfand transform side. We obtain, recalling that Σ^* denotes the $*$-spectrum of $\mathcal{A} = M(G,K)$, and ν_f the spectral measure determined by f on Σ^*:

$$\begin{aligned}\left\|R(f,\cdot)\right\|_2^2 &= \int_X\int_0^\infty s\left|\frac{d}{ds}\sigma_s f(x)\right|^2 ds\,dm(x)\\ &= \int_0^\infty s\left\|\frac{d}{ds}\sigma_s f\right\|_{L^2(X)}^2 ds = \int_0^\infty s\int_{\Sigma^*}\left|\frac{d}{ds}\varphi_z(\sigma_s)\right|^2 d\nu_f(z).\end{aligned}$$

Here we have obtained a spectral expression for the distribution $\frac{d}{ds}\sigma_s$, using the functional calculus in the commutative algebra $\mathcal{A}$ of spherical averaging operators. We refer to [106, §6, Lemma 4] for more on this argument.

(5) We can conclude that if the expression:

$$\Phi(z) = \int_0^\infty s\left|\frac{d}{ds}\varphi_z(\sigma_s)\right|^2 ds$$

has a *uniform spectral estimate*, namely a bound independent of z, as φ_z varies over the spectrum Σ^*, then the strong L^2-maximal inequality is proved.

(6) We recall (see [70, Chapter IV, §5] for full details) that for $G = SL_2(\mathbb{C})$ the characters of $M(G,K)$ are given by

$$\varphi_z(\sigma_t) = \frac{\sinh(zt)}{z\sinh(t)}, \qquad \varphi_0(\sigma_t) = \frac{t}{\sinh t}.$$

The continuous (i.e. bounded) $*$-characters are parametrized by:

(i) $z = i\lambda$, λ real; *Principal series*, or

(ii) $z = a$ where a is real and $0 < a \leqslant 1$; *Complementary series*.

(7) For the principal series, part (6) implies immediately that the $*$-characters decay exponentially in the distance, *uniformly in* λ, with fixed rate. Furthermore for $SL_2(\mathbb{C})$ the explicit expression above yields

$$\left|\frac{d}{dt}\varphi_{i\lambda}(\sigma_t)\right| \leqslant C(1+t)\exp(-t).$$

Therefore $\Phi(i\lambda) \leqslant C\Phi(0) < \infty$, $\forall \lambda \in \mathbb{R}$.

Note however the foregoing estimate fails for the second derivative, namely the second derivative is *not* bounded uniformly in λ.

(8) For the complementary series, the $*$-characters decay arbitrarily slowly, and in fact, if $a = 1-\varepsilon$, then $\varphi_a(\sigma_t) \cong c_\varepsilon \exp(-\varepsilon t)$. Furthermore, it can easily be proved directly from the formula in (6) that here

$$\left|\frac{d}{dt}\varphi_a(\sigma_t)\right| \leqslant \varepsilon \exp(-\varepsilon t)$$

and therefore

$$\int_0^\infty s\left|\frac{d}{ds}\varphi_a(\sigma_t)\right|^2 ds \leqslant \varepsilon^2 \int_0^\infty s\exp(-2\varepsilon s)\,ds \leqslant C < \infty.$$

For future reference we also note that in fact for every derivative of the complementary series characters

$$\left|\frac{d^k}{dt^k}\varphi_{1-\varepsilon}(\sigma_t)\right| \leqslant C_k\varepsilon^k \exp(-\varepsilon t).$$

(9) Thus we have established that $\Phi(z) \leqslant C < \infty$ as φ_z ranges over the spectrum Σ^*, and this suffices to prove the (a-priori) strong L^2-maximal inequality for the sphere averages σ_t. Of course, it follows immediately that the ball averages β_t satisfy the same maximal inequality, being convex averages of the sphere averages.

(10) The mean ergodic theorem for $\mathrm{Iso}(\mathbb{H}^n)$ is of course a consequence of the Fomin–Gelfand result [55], or more generally the Howe–Moore mixing theorem [74], which in particular establishes decay of non-trivial continuous characters of the algebra $\mathcal{A}$. In other words, $\lim_{t\to\infty}\varphi_z(\sigma_t) = 0$, and by the spectral theorem it follows immediately that when G is ergodic

$$\langle \pi(\sigma_t)f, f\rangle = \int_{\varphi\in\Sigma^*} \varphi(\sigma_t)\,dv_f(\varphi) \longrightarrow \int_X f\,dm.$$

In the $SL_2(\mathbb{C})$-case, the latter conclusion also follows upon inspection of the explicit form of the characters. The mean ergodic theorem for the balls is an easy consequence.

(11) According to the recipe of Section 2.3, the last step left is to prove the pointwise convergence of $\pi(\sigma_t)f(x)$ for a dense set of functions f. Here, since the balls, and certainly the spheres, are not Følner sets, the variants of Riesz's argument used in the amenable case cannot be applied. A more technical argument has to employed, described briefly as follows. Decompose the spectrum as a union $\Sigma' = \bigcup_{\varepsilon>0} \Sigma_\varepsilon$, where $\Sigma_\varepsilon = \{\varphi_z;\ |z| \leqslant 1-\varepsilon\}$. We can then use the fact that each of the characters appearing in Σ_ε has (exponential) decay in t of a fixed positive rate $\delta(\varepsilon)$. Consider now functions f which are sufficiently L^2-smooth, and whose spectral measure ν_f is supported in Σ_ε. Such functions can be shown to satisfy the desired conclusion, namely $\pi(\sigma_t)f(x)$ converges almost everywhere, using Sobolev-space arguments. Finally, the set of all such functions as $\varepsilon \to 0$ is dense in the set of K-invariant functions, and this completes the proof in the case of sphere averages, and the case of ball averages easily follows. We refer to [106] for more details on these arguments.

9.4. *Exponential volume growth: ball versus shell averages*

Theorem 9.3 established the strong maximal inequality and pointwise ergodic theorem in L^2 for sphere and ball averages. Of course, for the absolutely continuous ball averages one would expect a similar result in L^1, or at least in L^p, $1 < p < \infty$. The L^1-problem is still open, and we now describe the proof of the following.

THEOREM 9.4 (Pointwise ergodic theorem in L^p, $1 < p < \infty$, for ball averages on $\mathrm{Iso}(\mathbb{H}^n)$ [107,115]). *The ball averages β_t satisfy the pointwise ergodic theorem and strong maximal inequality in $L^p(X)$, $p > 1$, for all dimensions $n \geqslant 2$.*

Model case: *Proof of the pointwise ergodic theorems for β_t on $SL_2(\mathbb{C})$.*

(1) The ball averages β_t satisfy, for $t \geqslant 1$:

$$\beta_t = \frac{\int_0^t (\sinh s)^2 \sigma_s \, ds}{\int_0^t (\sinh s)^2 \, ds} \leqslant C_1 e^{-2t} \int_0^t e^{2s} \sigma_s \, ds.$$

(2) Consider the shell average, for $t \geqslant 1$, and the corresponding maximal function, given by:

$$\gamma_t = \int_0^1 \sigma_{t-s} \, ds, \qquad f_\gamma^*(x) = \sup_{t \geqslant 1} \left| \pi(\gamma_t) f(x) \right|.$$

(3) We now use the exponential volume growth of balls in G, in order to bound f_β^* by f_γ^*. Since γ_t is the uniform average of spheres with radius in $[t-1, t]$ (here $t \geqslant 1$), this amounts to comparing the average on a ball of radius r to the maximum of the averages on annuli of width one and radii bounded by r. Thus, using the foregoing

estimate of the densities, for $f \geqslant 0$ we have

$$\begin{aligned}\beta_t f(x) &\leqslant C_1 e^{-2t} \int_0^t e^{2s} \sigma_s f(x)\, ds \\ &\leqslant C_1 e^{-2t} \left(\sum_{k=0}^{[t]-1} \int_{t-k-1}^{t-k} e^{2s} \sigma_s f(x)\, ds + \gamma_1 f(x) \right) \\ &\leqslant C_1 \sum_{k=0}^{[t]-1} e^{-2k} \gamma_{t-k} f(x) + C_1 \gamma_1 f(x) \\ &\leqslant C_1 \left(\sum_{k=0}^{\infty} e^{-2k} \right) \sup_{1 \leqslant s \leqslant t} \gamma_s f(x) + C_1 \gamma_1 f(x) \leqslant 4 C_1 f_\gamma^*(x).\end{aligned}$$

Hence it suffices to prove the maximal inequality for the shell averages. One advantage that γ_t offers is that the exponential density that weighs the sphere averages in β_t no longer appears, and we can use classical methods of Fourier analysis to estimate the Gelfand transform without difficulty. In the next section we will estimate the rate of decay of the transform, and then apply some analytic interpolation techniques to estimate certain maximal functions associated with the (regularized) shell averages γ_t. These techniques will allow us to convert L^2-boundedness results for square functions associated with the derivative of (regularized) shell averages, to L^p-boundedness results for the maximal function associated with the shells themselves.

This completes the proof of Theorem 9.4 for ball averages, provided we prove the corresponding result for the shell averages. □

9.5. *Square functions and analytic interpolation*

We will presently show how to pass from the norm boundedness of the square functions in L^2 to best-possible L^p results, via analytic interpolation. This technique is most easily implemented for the shell averages, a fact that motivates their introduction to our discussion. However, by the discussion of Section 9.4, to complete the proof of Theorem 9.4, it suffices indeed to consider the shell averages and prove the following.

THEOREM 9.5 (Pointwise ergodic theorem for shell averages γ_t on $\mathrm{Iso}(\mathbb{H}^n)$, $n \geqslant 2$, [107, 115]). *The shell averages γ_t satisfy the pointwise ergodic theorem and the strong maximal inequality in $L^p(X)$, $p > 1$, for all dimensions $n \geqslant 2$.*

Model case: *Proof of the pointwise ergodic theorems for γ_t on $SL_2(\mathbb{C})$.*

(1) First, let us smooth the shell averages by the usual procedure, and define, for $t \geqslant 1$:

$$\tilde{\gamma}_t = \int_{\mathbb{R}} \psi(t-s) \sigma_s \, ds,$$

where ψ is a positive smooth function identically one on $[0,1]$, vanishing outside $[-1,2]$.

(2) We would like now to consider all the derivative operators $\frac{d^k}{ds^k}\tilde{\gamma}_s$, and bound the expressions

$$\Phi_k(z) = \int_1^\infty s^{2k-1} \left| \frac{d^k}{ds^k} \varphi_z(\tilde{\gamma}_s) \right|^2 ds$$

independently of z, namely uniformly on the spectrum $\Sigma^*(\mathcal{A})$, as in Section 9.2.

(3) For shell averages (unlike the case of the singular sphere averages σ_s) the smoothing allows uniform control (as $\lambda \to \infty$) of any given derivative of *principal series* characters. Indeed, recall that for $SL_2(\mathbb{C})$ these characters are given by $\frac{\sin \lambda t}{\lambda \sinh t}$. Thus the desired boundedness is an easy consequence of classical 1-dimensional Fourier theory, since here the problem reduces to estimating the decay in λ of the Fourier transform of a smooth compactly-supported function on the line. As is well-known, the decay in this case is faster than any polynomial in λ, and the boundedness of the integral above follows.

(4) Arbitrarily high derivatives of the *complementary series* characters can also be controlled, in fact even when evaluated on σ_t, and therefore also for $\tilde{\gamma}_t$. This was already noted in Section 9.3, part (8), and can be proved by differentiating the explicit form of the characters.

(5) Therefore, using the functional calculus for the distributions corresponding to higher derivatives, the kth-order square functions

$$R_k(f,x)^2 = \int_1^\infty s^{2k-1} \left| \frac{d^k}{ds^k} \tilde{\gamma}_s f(x) \right|^2 ds$$

have an L^2-norm bound.

(6) Now use the Riemann–Liouville fractional integral family of operators (see, e.g., [140]), and embed $\tilde{\gamma}_t$ in an analytic family of operators T_z, $z \in \mathbb{C}$. By the analytic interpolation theorem, an L^2-norm bound for the derivative of $\tilde{\gamma}_t$ appearing in a square function can be converted to an L^p-norm bound for the operator $\sup_{t \geqslant 1} \tilde{\gamma}_t$. This is done by interpolating against the maximal inequality that the uniform averages $\frac{1}{t}\int_0^t \tilde{\gamma}_s \, ds$ satisfy in every L^p, $p > 1$. The latter result was noted for the averages $\mu_t = \int_0^t \frac{1}{t}\sigma_s \, ds$ in Section 9.3, part (1), and of course it follows in exactly the same way in the present case. We refer to [107] and [115] for the details.

9.6. *The L^p-theorem for sphere averages on* $\mathrm{Iso}(\mathbb{H}^n)$

It is of course natural to complete also the discussion of the maximal inequality for the sphere averages, whose boundedness was established in L^2, and prove the best possible results in L^p. Indeed, a variant of the method of analytic interpolation via the Riemann–Liouville fractional integrals can be applied to σ_t also, in order to embed it in an analytic family of operators. This method yields the following:

THEOREM 9.6 (Pointwise ergodic theorem for the sphere averages σ_t on $\mathrm{Iso}(\mathbb{H}^n)$, $n > 2$, [107,115]). *The sphere averages σ_t on $\mathrm{Iso}(\mathbb{H}^n)$, $n > 2$, satisfies the pointwise ergodic theorem and strong maximal inequality in L^p, $p > \frac{n}{n-1}$, which is the best possible range.*

PROBLEM 9.7. The pointwise ergodic theorem and L^p, $p > 2$ maximal inequality for sphere averages in general measure-preserving actions of $\mathrm{Iso}(\mathbb{H}^2)$ is an open problem.

REMARK 9.8.

(1) The constraint on the L^p-range of the maximal inequality for spheres in hyperbolic space is the same as the Euclidean constraint. The constraint is determined by the rate of decay (in the spectral variable) of the Fourier–Gelfand transform of the spherical measure on a sphere of radius one. We note that the counter-example in the Euclidean case for boundedness of the maximal operator in $L^{\frac{n}{n-1}}$ can be taken to be a local one [142]. Namely, it is given by a function of compact support and a singularity at the origin (say), and thus the L^p-constraint arises already from the local operators $\sup_{0<t\leqslant 1}\sigma_t$. Now since hyperbolic spheres (in the ball model) are just off-center Euclidean spheres, the Euclidean local counter-example is also a hyperbolic counter-example [107, §5.4]. The point is thus to show that there are no further constraints in the hyperbolic case.

(2) The complementary series poses a serious challenge in the analysis employed to prove the maximal inequality $\sup_{1\leqslant t<\infty}\sigma_t$ for the sphere averages on the groups $\mathrm{Iso}(\mathbb{H}^n)$. This is the result of the arbitrarily slow rate of decay of the complementary series characters. To control the norm of square functions, it is necessary to prove derivative estimates for the spherical functions of the complementary series which are significantly better than the standard Harish Chandra estimates. This problem takes a sizable part of the effort in [107] and [115].

(3) Let us emphasize that the behavior of the averages γ_t is in marked contrast to the Euclidean case. In the hyperbolic set-up, the maximal operators associated with γ_t and β_t are in fact equivalent. But the Euclidean shell averages satisfy the same maximal inequalities as the sphere averages, and not the same maximal inequalities as the ball averages. This fact is a reflection of the difference between polynomial and exponential volume growth, and for more on this matter we refer to [113] and [111].

(4) Recalling our comments in the introduction to Section 8, we note that the only available proof of the maximal inequality for the ball averages β_t in L^p, namely the proof described in Sections 9.3–9.5 above, makes use of differentiation theory of the singular averages σ_t. This is also in marked contrast to the Euclidean (or polynomial volume growth) case, where the singular sphere averages did not play any role.

(5) A more geometric approach to obtain the L^p, $p > 1$, maximal inequality for balls in every dimension greater than two, is to start with $SL_2(\mathbb{C})$ and use an analog of the "method of rotations". Namely, embed $\mathrm{Iso}(\mathbb{H}^3) \subset \mathrm{Iso}(\mathbb{H}^n)$ as the stability group of a totally geodesic subspace. Then the Cartan polar coordinates decompositions $G = KAK$ in the two groups can be aligned. Since A is one-dimensional, the maximal inequalities for the ball averages in $\mathrm{Iso}(\mathbb{H}^n)$ follows from those of σ_t in $\mathrm{Iso}(\mathbb{H}^3)$. However, this leaves out $\mathrm{Iso}(\mathbb{H}^2)$, where the spherical functions estimate are the

hardest case. For spheres, the range of p where the strong maximal inequality holds improves with dimension, so this method does not give optimal results.

10. Groups with commutative radial convolution structure

The methods outlined in the previous section have a wide scope of applications, and have been developed into a systematic spectral approach to the proof of pointwise ergodic theorems for radial averages on lcsc groups admitting a commutative radial convolution structure. In the present section we will indicate some of the results obtained in this direction, and comment on some of the open problems.

10.1. *Gelfand pairs*

Let us recall the following well known definition

DEFINITION 10.1 (*Gelfand pairs*). A Gelfand pair (G, K) consists of an lcsc group G, and a compact subgroup $K \subset G$, such that the algebra $M(G, K)$ of bounded Borel measures on G which are bi-K-invariant is commutative, or equivalently, the convolution algebra $L^1(G, K)$ of bi-K-invariant L^1-functions is commutative. We remark that G is then necessarily unimodular.

EXAMPLE 10.2 (*Some examples of Gelfand pairs*).

(1) G a connected semisimple Lie group, K a maximal compact subgroup. This extensive family includes:
   (i) $G = SO(n, 1) = \mathrm{Iso}(\mathbb{H}^n)$ the isometry group of the simply connected Riemannian n-manifold of constant negative curvature, $K = SO(n)$ the group of rotations fixing a point. Thus here $M(G, K)$ is the usual algebra of radial averages on hyperbolic space.
   Similarly, $G = SU(n, 1)$, the isometry group of complex hyperbolic space, K the maximal compact subgroup fixing a point.
   (ii) $G = O(p, q)$, the isometry group of simply-connected pseudo-Riemannian manifold of signature (p, q) and constant curvature, $K = O_p(\mathbb{R}) \times O_q(\mathbb{R})$.
   (iii) $G = SL_n(\mathbb{C})$, the general Linear group, which is the isometry group of the space of positive definite matrices, $K = SU_n(\mathbb{C})$ the unitary group.
   (iv) $G = Sp_n(\mathbb{R})$, the symplectic group, $K = Sp(n)$.

(2) $S = K \ltimes \mathfrak{p}$, a Cartan motion group. Here K is a maximal compact subgroup of a semisimple Lie group G, and $\mathfrak{p}$ the fixed-point-subspace of a Cartan involution on the semisimple Lie algebra $\mathfrak{g}$. Examples include:
   (i) The Euclidean motion group $S = O_n(\mathbb{R}) \ltimes \mathbb{R}^n$, $K = O_n(\mathbb{R})$. Here $M(G, K)$ is the usual algebra of radial measures on $\mathbb{R}^n$.
   (ii) The Heisenberg motion group, $S = U_n(\mathbb{C}) \ltimes H_n$, $K = U_n(\mathbb{C})$. Here $M(G, K)$ is the algebra of radial measures on H_n generated by the $\mathbb{C}^n$-spheres.

(iii) $S = O_n(\mathbb{R}) \ltimes Sym_n(\mathbb{R})$, $K = O_n(\mathbb{R})$, where $Sym_n(\mathbb{R})$ is the space of $n \times n$ symmetric matrices, and the action of $O_n(\mathbb{R})$ is by conjugation.

(3) G a connected semisimple algebraic Chevalley group over a locally compact non-discrete field, $K =$ a good compact open subgroup. Examples include:
   (i) $G = PGL_2(\mathbb{Q}_p)$, $K = PGL_2(\widehat{\mathbb{Z}}_p)$. Here G/K is a $(p+1)$-regular tree, $M(G, K)$ the algebra of radial averages on the regular tree.
   (ii) $G = SL_n(\mathbb{Q}_p)$, $K = SL_n(\widehat{\mathbb{Z}}_p)$. Here G acts by isometries of an affine building, and $M(G, K)$ is a commutative subalgebra of the convolution Hecke algebra of double cosets of an Iwahori subgroup.

(4) $G = Aut(T_{r_1,r_2})$ the automorphism group of the semihomogeneous bi-partite tree of valencies r_1 and r_2, K_i the maximal compact subgroup fixing a vertex of valency r_i. Further examples of Gelfand pairs (G, K) arise from certain closed non-compact boundary-transitive subgroups of the group of automorphisms of a finite product of such trees.

The foregoing list is a very partial one (for a discussion of some more examples of amenable Gelfand pairs see, e.g., [7,8]). However already the groups mentioned (together with other groups possessing a radial convolution structure, e.g., the free groups) give rise to a large collection of interesting measure-preserving actions. We mention briefly only the following.

EXAMPLE 10.3 (*Some examples of measure-preserving actions*).

(1) $SL_2(\mathbb{R})$ acts (transitively) on the unit tangent bundle of a compact Riemann surface $M = \pi_1(M) \setminus SL_2(\mathbb{R})$, preserving a volume form of finite total mass.

(2) $SL_n(\mathbb{R})$ acts (transitively) on the space of unimodular lattices $\mathcal{L}_n$ in $\mathbb{R}^n$, namely $\mathcal{L}_n = SL_n(\mathbb{Z}) \setminus SL_n(\mathbb{R})$, preserving a volume form of finite total mass. Of course, any subgroup of $SL_n(\mathbb{R})$ also acts on $\mathcal{L}_n$.

(3) More generally, any subgroup H of a semisimple algebraic group G acts on the probability space $\Gamma \setminus G$, where Γ is a lattice subgroup, e.g.,
   (i) $G = SL_2(\mathbb{R}) \times SL_2(\mathbb{R})$ and $\Gamma = SL_2(\mathbb{Z}[\sqrt{d}])$ under the skew diagonal embedding $\gamma \mapsto (\gamma, \tau(\gamma))$, τ the Galois automorphism of $\mathbb{Z}[\sqrt{d}]$, d a square free positive integer.
   (ii) $G = SL_2(\mathbb{R}) \times SL_2(\mathbb{Q}_p)$, $\Gamma = SL_2(Z[\frac{1}{p}])$ an irreducible lattice in G.

(4) $\Gamma = \mathbb{F}_k$ the free group, for example embedded as a lattice in $G = PGL_2(\mathbb{Q}_p)$, as $\mathbb{Z}[\frac{1}{p}]$-points in an appropriate quaternion algebra, and X any action of G, e.g., on a compact locally symmetric space.

(5) $\Gamma = SL_2(\mathbb{Z})$, and the action by group automorphisms of $\mathbb{T}^2$, or on $SL_2(\mathbb{R})/\Gamma$, Γ a lattice.

Again, this list is very partial, but it already clearly demonstrates that establishing a pointwise (or mean) ergodic theorem (preferably with error term), a strong maximal inequality or a differentiation theorem gives rise to diverse applications, depending on the family μ_t to which it applies, and the group and action involved. We will mention here some applications only very briefly, just to motivate our discussion, and without attempt-

ing to explain them further. Rather, we will concentrate below on the proof of the ergodic theorems themselves.

EXAMPLE 10.4 (*Some applications of ergodic theorems and maximal inequalities*).

(1) Boundedness properties of natural singular integrals on homogeneous spaces.
(2) Integral geometry on locally symmetric spaces, e.g., singular spherical differentiation, and pointwise equidistribution of spheres and other singular subvarieties in homogeneous spaces.
(3) Evaluation of the main term in counting lattice points on homogeneous algebraic varieties.
(4) Estimating error terms in problems of Diophantine approximation on homogeneous algebraic varieties and homogeneous spaces.

10.2. *Pointwise theorems for commuting averages: general method*

We now assume (G, K) is a Gelfand pair, and as usual (X, m) denotes an ergodic probability measure preserving action of G. The spectral approach outlined in Section 9 to pointwise convergence is based on applying certain geometric properties of the convolution structure of $M(G, K)$, together with the tools of harmonic analysis on Abelian Banach algebras, in order to prove pointwise convergence when the elements of the algebra are represented as averaging operators on $L^p(X)$. Since we consider representations arising from a measure-preserving and thus unitary action of G, the homomorphism $\pi : M(G, K) \to \operatorname{End} L^2(X)$ is a $*$-homomorphism. Furthermore, an algebra representation arising from a unitary representation of the group, gives rise to $*$-characters (also called (G, K)-spherical functions), which can all be identified with *positive-definite* continuous bounded functions on the group—see [54, Chapter I] and [70, Chapter IV]. The basic measures in $M(G, K)$ are $\sigma_g = m_K * \delta_g * m_K$, $g \in G$. Every other bi-K-invariant probability measure on G is a convex combination of these. Note that in the discussion of Section 9, the one-parameter group $A = \{a_t,\ t \in \mathbb{R}\}$ was a fixed group of hyperbolic isometries, and there it was enough to consider $\sigma_t = m_K * \delta_{a_t} * m_K$. Many other possibilities for the choice of the averages arise in practice, and it is not necessary to restrict to averages associated with one-parameter groups, or even to one parameter family of averages—we refer to [113] for some examples and more details. Nevertheless, for simplicity of exposition, let us choose a family of bi-K-invariant probability measures ν_t, $t \in \mathbb{R}$ (not necessarily of the form $m_K * \delta_{a_t} * m_K$), and explain briefly the ingredients sufficient for a proof of the ergodic theorems and maximal inequalities in this case. We will also comment briefly in the next section on the various problem that arise along the way. The recipe for the proof proceeds along the following steps.

(1) Identify the positive-definite $*$-spectrum $\Sigma^*(\mathcal{A})$. Prove that for every non-trivial positive-definite $*$-character $\varphi_z \in \Sigma^*(\mathcal{A})$ (i.e. every non-constant positive-definite spherical function)

$$\lim_{t\to\infty} \varphi_z(\nu_t) = 0.$$

Conclude that ν_t satisfy the *mean ergodic theorem*:

$$\lim_{t\to\infty} \left\| \nu_t f - \int_X f\, dm \right\|_{L^2(X)} = 0.$$

(2) Analyze the behavior in z of $\varphi_z(\nu_t)$ and $\frac{d^k}{dt^k}\varphi_z(\nu_t)$ as $t \to \infty$ (for a pointwise ergodic theorem) and as $t \to 0$ (for a differentiation theorem).

(3) Establish the existence of a dense set of functions in L^2 where $\nu_t f(x)$ converges pointwise almost everywhere, using the spectral decomposition of $\mathcal{A}$ in $L^2(X)$, estimates of the positive-definite $*$-characters, and Sobolev space arguments (see [106]).

(4) Establish a maximal inequality in L^p, $p > 1$, for an averaged version μ_t of ν_t, for example the uniform averages $\mu_t = \frac{1}{t}\int_0^t \nu_s\, ds$. This may be achieved using a number of methods, depending on the case at hand, as follows.

(I) First possibility [106]: Use a maximal inequality for the radial components of ν_t, namely for the averages ν_t determine on the Abelian group A, when a Cartan polar decomposition $G = KAK$ is available. Namely use the representation of ν_t as a convex combination of the basic measures $\sigma_g = m_K * \delta_a * m_K$, a the Cartan component of g.

(II) Second possibility [114]: Use a central limit theorem for the transient random walk associated with ν_1 and conclude that (for fixed positive constants c and C)

$$\mu_n \leqslant \frac{C}{cn+1} \sum_{k=0}^{cn} \nu_1^{*k}.$$

Then use the Hopf–Dunford–Schwarz maximal inequality for uniform averages of ν_1^{*k}. Finally, argue that for some constant B, $\mu_t \leqslant B\mu_{[t]+1}$, $[t]$ the integer part of t [99].

(III) Third possibility [105]: Establish the following subadditive convolution inequality given by (for fixed positive constants c and C)

$$\mu_t * \mu_s \leqslant C(\mu_{ct} + \mu_{cs}),$$

using convolution estimates on the group. The subadditive convolution inequality in the group algebra is sufficient to deduce strong maximal inequality in L^2, in every action of the group (see also Section 10.5 below). A strong maximal inequality in L^p, $p > 1$, can be deduced if an iterated form of the subadditive inequality is established, for the products $\mu_{t_1} * \mu_{t_2} * \cdots * \mu_{t_n}$.

(5) Estimate the difference $|\sigma_t - \mu_t|$ using an appropriate Littlewood–Paley square function. For the particular case of uniform averages μ_t, as noted in Section 9.3:

$$\nu_t - \mu_t = \frac{1}{t}\int_0^t s\frac{d}{ds}\nu_s\, ds$$

and

$$\left|\nu_t f(x)\right| \leqslant \left|\mu_t f(x)\right| + \left(\int_0^\infty s \left|\frac{d}{ds}\nu_s f(x)\right|^2 ds\right)^{1/2}.$$

So $f_\nu^*(x) \leqslant f_\mu^*(x) + R(f,x)$.

(6) Transfer the estimate of the L^2-norm of the square function $R(f,x)$ to the Fourier–Gelfand transform side, using the functional calculus in the commutative algebra $\mathcal{A}$. Namely, show that

$$\left\|R(f,\cdot)\right\|_2^2 \leqslant \|f\|_2^2 \sup_{z\in\Sigma^*(\mathcal{A})} \Phi(z)^2,$$

where:

$$\Phi(z)^2 = \int_0^\infty s \left|\frac{d}{ds}\varphi_z(\nu_s)\right|^2 ds.$$

(7) Use the estimates of the characters and their derivative in (2) to show that $\Phi(z)$ has a bound *independent of* z. Do the same for the square functions $R_k(f,x)$ corresponding to higher derivatives.

(8) To convert L^2-norm bounds for the square functions associated with the derivative operators $\frac{d^k}{ds^k}\nu_s$, to an L^p-norm bound for the maximal function $f_\nu^*(x) = \sup_{t>0}|\nu_t f(x)|$, embed ν_t and their derivatives (and integrals) in an analytic family of operators. For example, we have utilized in Section 9.4(6) the Riemann–Liouville fractional integral operators in the case of μ_t (for the use of other families see [115]). This allows the use of the analytic interpolation theorem to interpolate between the maximal inequality for the derivative operators in L^2, and the L^p-maximal inequality for the uniform averages μ_t, for $p > 1$. The latter maximal inequalities are a consequence of the methods indicated in part (4) of the present recipe.

10.3. *Pointwise theorems and the spectral method: some open problems*

Each ingredient in the recipe outlined in Section 10.2 above poses certain difficulties, depending on the Gelfand pair and the family of averages under consideration.

We indicate briefly some of the open problems that arise, taking the example of semisimple Lie and algebraic groups.

(1) The classification of positive-definite spherical functions (namely the $*$-characters of $L^1(G,K)$ that arise in the representations under consideration) on semisimple Lie (and algebraic) groups is far from complete, for many infinite families of groups.

(2) Even when the classification is complete, the best estimates for the derivatives of the spherical functions are far from sufficient to prove the ergodic theorems for (say) the sphere averages. It is thus necessary to establish decay estimates uniformly over the positive-definite $*$-spectrum, which considerably improve Harish Chandra's classical estimates, for example.

(3) In semisimple Lie (or algebraic) groups of real (or split) rank greater than two radial averages more singular that Riemannian spheres occur very naturally. For very singular averages of high codimension, optimal results require very precise estimates of the spherical functions, including along the different singular directions in the radial variable $H \in \mathfrak{a} = Lie(A)$, where $G = KAK$ is the Cartan decomposition. Such estimates are usually not available for higher rank groups.

(4) The latter phenomenon manifests itself even in the case of product groups, where spheres and balls are a convex average of spheres and balls on the component groups, taken with exponential weights. The complexity of the convolution structure makes it difficult to estimate the spherical functions and their derivatives, when evaluated on these averages. In particular, the best possible range of L^p maximal inequalities for sphere averages is not known.

(5) For sphere averages on $\mathrm{Iso}(\mathbb{H}^2)$, the maximal inequality in not valid in L^2 (as in the case of the Euclidean plane). A similar problems arises for Cartan motion groups, for example, the Heisenberg group H_1. This makes the spectral methods much less effective, and indeed in these cases, the pointwise ergodic theorem for sphere averages in general actions of the group has not been established. Note that the convolution case for $\mathrm{Iso}(\mathbb{H}^2)$ has been settled in [76], but in the absence of a transfer principle for non-amenable groups, this result has no bearing on the case of general ergodic actions.

(6) Spectral methods and analytic interpolation theory do not give maximal inequalities and ergodic theorems in L^1. This is an open problem even for the ball averages, on all semisimple groups, and in particular, for $G = SL_2(\mathbb{C})$. Note that a general weak-type $(1,1)$-maximal inequality for ball averages on semisimple groups, acting by convolutions on the symmetric space, has been established in [143]. Again, however, the absence of a transfer principle renders this result irrelevant for the case of general ergodic actions.

While the list of problems above certainly poses some formidable challenges, in many interesting situations these challenges can be surmounted, and the general spectral method outlined in Section 10.2 can be implemented (for sphere averages, say). This is true particularly in the case of the action by convolution, which is much more explicit than a general action. This fact is demonstrated by the examples discussed in Section 8.1, namely the maximal inequalities for convolution with spheres in Euclidean spaces [142], Heisenberg groups [36,116,102], and hyperbolic spaces of dimension $n \geqslant 3$ [45], or dimension $n = 2$ [76]. By the discussion in Section 9, maximal inequalities for sphere averages can also be established for general ergodic actions of $G = \mathrm{Iso}(\mathbb{H}^n)$, $n > 2$. More generally, the spectral method gives the best possible range of p where the maximal inequality and pointwise ergodic theorem for sphere averages hold, for any simple Lie group of real rank one [115] (except $SL_2(\mathbb{R})$).

We will present in the following two sections further examples where the spectral method can be fully or partially implemented for sphere averages, and in the following section, further examples where it can be implemented for ball averages. We first turn to the case of singular averages on higher real-rank semisimple groups, which is far from completely solved, and exhibits many of the problems referred to in Section 10.3. The only available results on singular averages on higher-real rank groups were obtained in the

case of complex group, which we discuss in Section 10.4. In Section 10.5 we will consider some totally disconnected Gelfand pairs, as well as some of their discrete lattice subgroups, to which the spectral method applies. In particular we will prove the ergodic theorems of spheres in the free groups $\mathbb{F}_k$, noted in Section 8.3.

10.4. *Sphere averages on complex groups*

We begin by a brief reminder of the basic relevant set-up and notation. Let G denote a connected semisimple Lie group with finite center and without non-trivial compact factors, $\mathfrak{g}$ its Lie algebra. Let θ denote a Cartan involution on G and $\mathfrak{g}$, and let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ be the corresponding Cartan decomposition, so that $\mathfrak{k}$ is the Lie subalgebra corresponding to a connected maximal compact subgroup K. Let $\mathfrak{a} \subset \mathfrak{p}$ denote a maximal Abelian subalgebra, and let $\Phi(\mathfrak{a}, \mathfrak{g}) = \Phi \subset \mathfrak{a}^*$ denote the (real) root system of $\mathfrak{a}$ in $\mathfrak{g}$. Let $\mathfrak{g}_\alpha$ denote the root space corresponding to $\alpha \in \Phi$, and $\mathfrak{g} = \mathfrak{m} \oplus \mathfrak{a} \oplus \sum_{\alpha \in \Phi} \mathfrak{g}_\alpha$ the root space decomposition. Fix a system of simple roots $\Delta \subset \Phi$, the corresponding ordering of $\mathfrak{a}^* = \hom(\mathfrak{a}, \mathbb{R})$ and the system of positive roots Φ_+, and let ρ denote half the sum of the positive roots. Let $W = W(\mathfrak{a}, \mathfrak{g})$ denote the Weyl group of the root system, $\mathfrak{a}_+$ the positive Weyl chamber, and $\overline{\mathfrak{a}_+}$ its closure. Let $A = \exp \mathfrak{a}$ denote the Lie subgroup corresponding to $\mathfrak{a}$, $A_+ = \exp \mathfrak{a}_+$ and $\overline{A_+}$ its closure. The Cartan (or polar coordinates) decomposition in G is given by $G = K\overline{A_+}K$ and $g = k_1 e^{H(g)} k_2$, where $H(g)$ is the $\overline{\mathfrak{a}_+}$ component of g. Let $\langle , \rangle$ denote the Killing form on $\mathfrak{g}$, and let d denote the induced Riemannian metric on the symmetric space G/K. Then the restriction of $\langle , \rangle$ to $\mathfrak{a}$ is an inner product, and we have $d(\exp(H)o, o) = \sqrt{\langle H, H \rangle}$ for all $H \in \mathfrak{a}$, where $o = [K]$ denotes our choice of origin in G/K. We recall that the Cartan polar coordinates decomposition yield the following integration formula for Haar measure on G (see [70, p. 186] or [54]):

$$\int_G f(g)\, dm_G(g) = \int_K \int_{\overline{\mathfrak{a}_+}} \int_K f(ke^H k') \xi(H)\, dm_K(k)\, dH\, dm_K(k').$$

Here $\xi(H) = \prod_{\alpha \in \Phi_+} (\sinh \alpha(H))^{m_\alpha}$, $H \in \overline{\mathfrak{a}_+}$, $m_\alpha = \dim_{\mathbb{R}} \mathfrak{g}_\alpha$, and m_G, m_K denotes Haar measures on G and K, dH denotes Lebesgue measure on $\mathfrak{a}$.

We can now formulate the following result for complex semisimple Lie groups.

THEOREM 10.5 (Pointwise ergodic theorem for sphere averages on complex groups [40]). *Let G be a connected complex semisimple Lie group with finite center. Fix a regular direction $H \in \mathfrak{a}$, and let $\sigma_t^H = m_K * \delta_{\exp tH} * m_K$. Let σ_t denote the Riemannian sphere averages. In every measure-preserving action of G, we have*

(1) *The averages σ_t^H satisfy the pointwise ergodic theorem, strong maximal inequality and the singular differentiation theorem in L^p, $p > p_G$. Here p_G is an explicit computable constant, and, e.g., for $G = SL_n(\mathbb{C})$, $p_G = \frac{2n-1}{2n-2}$.*
(2) *The Riemannian sphere averages σ_t satisfy the pointwise ergodic theorem, maximal inequality and singular differentiation theorem in L^2.*

We remark that it is not known whether the range of p stated in Theorem 10.5(1) is optimal.

Theorem 10.5 is proved using the general method described in Section 10.2. A key ingredient is thus to establish the right spectral estimates that allow control of the Littlewood–Paley square function. The relevant spectral estimate are given as follows.

THEOREM 10.6 (Uniform spectral estimates for positive-definite spherical functions on complex groups [40]). *For G as in Theorem* 10.5, *the following holds for non-trivial positive-definite spherical functions φ_λ, and for $H \in \overline{\mathfrak{a}_+}$*

$$\left|\frac{d^k}{dt^k}\varphi_\lambda\big(\exp(tH)\big)\right| \leqslant C_k(G,H)\big(1+\|\lambda\|\big)^{k-\gamma}\exp\big(-\kappa t\rho(H)\big).$$

Here $\gamma = \gamma_G$ depend only on G. Furthermore κ depends only on G and is strictly positive, provided G has no $SL_2(\mathbb{C})$-factors (*equivalently, G is complex and satisfies property T*). *Otherwise κ is still strictly positive but depends also on λ.*

Theorem 10.6 implies that the kth Littlewood–Paley square function associated with the operators σ_t^H satisfies $\|R_k(f,\cdot)\|_2 \leqslant C_k\|f\|_2$, as long as $k \leqslant \gamma$. Thus the spectral method of Section 10.2 applies, and this proves Theorem 10.5.

The proof of Theorem 10.6 uses the explicit form of the spherical functions given by Harish Chandra's formula in the complex case. It is based on a method of descent, which allows writing a spherical function on G as a sum of multiples of spherical functions on lower-dimensional complex subgroups. Such a decomposition is defined for every $\lambda \in \mathfrak{a}^*$, and the multiples that occur depend on λ. Choosing the optimal decomposition for each λ, a rate of decay in λ is obtained, which depends only on the root system. We remark that the classical Harish Chandra estimate only provides an estimate which amounts to *polynomial growth* in the spectral parameter, rather than *polynomial decay* as in Theorem 10.6. Thus the Harish Chandra estimate cannot be used to establish the uniform spectral estimates necessary in order to bound the Littlewood–Paley square functions, that the spectral method presented above calls for.

The exponential decay in t established in Theorem 10.6 holds uniformly for all positive-definite spherical functions, and is a consequence of the results of M. Cowling and R. Howe on matrix coefficients of unitary representations with a spectral gap of semisimple groups, which we will describe further in the next section.

Let us now turn to discuss maximal inequalities and ergodic theorems in the completely different set-up of totally disconnected lcsc groups, and some of their lattice subgroups. In the next section we consider groups of tree automorphisms, and in 11.3 we will discuss higher-rank groups and lattices.

10.5. *Radial structure on lattice subgroups: a generalization of Birkhoff's theorem*

The general spectral method presented in Section 10.2 has some remarkable applications to certain countable groups which are not in themselves Gelfand pairs. In particular these application leads to a very natural generalization of Birkhoff's pointwise ergodic theorem as well of Hopf's maximal inequality. To explain it, consider the regular tree T_k of constant valency $k \geqslant 3$, and its group of graph automorphisms $G = Aut(T_k)$, which acts transitively

on the tree. The group K of automorphisms stabilizing a given vertex o is compact, and T_k can be identified with G/K. The orbits of K in $T_k = G/K$ are precisely the spheres $S_n(o)$ with o as a center. A simple direct computation shows that if $\mathcal{S}_n$ denotes the operator of averaging a function on the tree T_k on a sphere of radius n, then this sequence of operators satisfy the recurrence relation $\mathcal{S}_1\mathcal{S}_n = \frac{1}{k}\mathcal{S}_{n-1} + \frac{k-1}{k}\mathcal{S}_{n+1}$. It follows immediately that each $\mathcal{S}_n$ is a polynomial in $\mathcal{S}_1$, and thus the algebra generated by these operators is cyclic and commutative. It is easy to see that each operator $\mathcal{S}_n$ can be identified with a (right) convolution operator on G, namely with convolution by the double coset KgK corresponding the sphere $S_n(o)$ (a K-orbit in G/K). Thus it follows that (G, K) is a Gelfand pair, and since $L^1(G, K) = M(G, K)$ is cyclic, and satisfies the second-order recurrence relation with constant coefficients given by the identity above, it is not difficult to give its characters and $*$-characters in explicit form, as we shall see below.

Now note that the when $k = 2r$ is even, the Cayley graph $X(\mathbb{F}_r, S)$ of the free group on r generators determined by a set S of r free generators and their inverses, has the structure of a $2r$-regular tree. The free group in question acts as a group Γ of automorphisms of the Cayley graph, via its action by left translations and thus embeds in G. Furthermore the Γ-action is of course simply transitive, and Γ intersects trivially with the stability group of every vertex. It is easily seen that the operator $\mathcal{S}_n$ of averaging on a sphere with radius n in the Cayley graph can be given as an operator of (right) convolution on the group Γ, namely convolution by the measure σ_n, the uniform probability measure on a sphere of radius n in the free group, w.r.t. the word metric defined by S. Indeed:

$$\begin{aligned}\rho_\Gamma(\sigma_n) f(w) &= f * \sigma_n(w) = \frac{1}{|S_n|} \sum_{|x|_S = n} f\left(wx^{-1}\right) = \frac{1}{|S_n|} \sum_{d(w,u)=n} f(u) \\ &= \mathcal{S}_n f(w).\end{aligned}$$

Thus we conclude that the algebra generated by the averaging operators on spheres in the tree T_{2r} embeds as a *commutative subalgebra* $\mathcal{A}$ of the convolution algebra $\ell^1(\mathbb{F}_r)$ (which in itself is of course, as non-commutative as a group algebra can be). It follows that every unitary representation of $\mathbb{F}_r$ gives rise to a $*$-representation of the commutative algebra $\mathcal{A}$, and thus the general spectral method explained in Section 10.2 applies.

Note that in fact the identification explained above of the algebra generated by the operators of averaging on spheres in the Cayley graph, with the algebra generated in $\ell^1(\Gamma)$ by the operators of right convolution by σ_n holds more generally for every discrete group Γ with generating set S. Thus the method of Section 10.2 applies whenever the automorphism group G of the Cayley graph, together with the stability group K of a vertex form a Gelfand pair.

Continuing with the case of the free groups, we recall that the spectrum of $\mathcal{A}(\mathbb{F}_r)$ is given as follows (where we define $q = 2r - 1$). The solutions to the second-order recurrence relation satisfied in the algebra are

$$\varphi_z(\sigma_n) = \mathbf{c}(z) q^{-nz} + \mathbf{c}(1-z) q^{-n(1-z)}, \quad z \neq \frac{1}{2} + \frac{ij\pi}{\log q},$$

$$\mathbf{c}(z) = \frac{q^{1-z} - q^{z-1}}{(q+1)(q^{-z} - q^{z-1})}$$

and:

$$\varphi_z(\sigma_n) = \left(1 + n\frac{q-1}{q+1}\right)(-1)^{jn} q^{-n/2}, \quad z = \frac{1}{2} + \frac{ij\pi}{\log q}.$$

A necessary and sufficient condition for φ_z to be a continuous character (on the closed subalgebra of $\ell^1(\mathbb{F}_r)$ generated by the spheres) is that it be bounded, and this condition is equivalent to $0 \leqslant \operatorname{Re} z \leqslant 1$. The unitary representation of $\mathbb{F}_r$ in $L^2(X)$, extended to $\ell^1(\mathbb{F}_r)$, assigns to σ_1 a self-adjoint operator. Consequently, the values $\varphi_z(\sigma_1) = \gamma(z)$ are real, for those φ_z that occur in the spectrum of σ_1 in $L^2(X)$. It is easily verified that $\gamma(z)$ is real iff $\operatorname{Re} z = \frac{1}{2}$, or $\operatorname{Im} z = \frac{ij\pi}{\log q}$. The image of this set under γ is the $*$-spectrum of $A(\mathbb{F}_r)$. Note that for z and $1-z$ the same character obtains, so we can assume that $0 \leqslant \operatorname{Re} z \leqslant \frac{1}{2}$. Note also that the characters corresponding to $z = s$ and to $z = s + \frac{ij\pi}{\log q}$ differ by sign only: $\varphi_{s+\frac{ij\pi}{\log q}}(\sigma_n) = (-1)^{jn}\varphi_s(\sigma_n)$. In particular, the sign character ε, given by $\varepsilon(\sigma_n) = (-1)^n$, is obtained at the points $z = \frac{i(2j+1)\pi}{\log q}$.

The $*$-spectrum is naturally divided to the principal series characters φ_z where $\operatorname{Re} z = \frac{1}{2}$, and the complementary, where $z = s + \frac{ij\pi}{\log q}$, where s is real and $0 \leqslant s \leqslant \frac{1}{2}$.

The comparison with the case of the spherical functions on the group $SL_2(\mathbb{C})$ is evident (see Section 9.3(6)), save of course for the fact that $\mathcal{A}(\mathbb{F}_r)$ has a unit, and thus its spectrum is compact.

In order to demonstrate some of the phenomena that arise here, let us consider in addition to the free group also the groups $\Gamma(r,h) = G_1 * G_2 \cdots * G_r$, the free product of r finite groups each of order h, where $r \geqslant 2$, $h \geqslant 2$, $r + h > 4$, with generating set $S = \bigcup_{i=1}^{r} G_i \setminus \{e\}$. Here we define $q(\Gamma(r,h)) = (r-1)(h-1)$.

The sphere averages on $\Gamma(r,h)$ also commute, and also satisfy a second-order recurrence relation with constant coefficients. Indeed, again this algebra is closely related to a Gelfand pair, this time associated with the group of automorphisms of a semihomogeneous tree (see [104] and [105] for more details).

The spectrum of $\Gamma(r,h)$ can thus be analyzed similarly, with one significant difference compared to that of $\mathbb{F}_r$. At the point $i\pi/\log q = i\zeta$ the special character that obtains is of the form (see [105, §2.3])

$$\varphi_{i\zeta}(\sigma_n) = c_{i\zeta}(-1)^n + c_{1-i\zeta}(-1)^n q^{-n}.$$

Here $c_{i\zeta} = 1$ and $c_{1-i\zeta} = 0$ if and only if $h = 2$. Indeed $c(i\zeta) = \frac{q-(r-1)^{-1}-h+2}{r(h-1)(1-q^{-1})}$, and so $c(i\zeta) = 1$ iff $\Gamma = \Gamma(r,2)$ or $\Gamma = F_k$. We define $c_{i\zeta} = c(\Gamma)$.

Given a probability-preserving action of one of the groups $\Gamma(r,h)$ or $\mathbb{F}_r$, let us denote by $\mathcal{E}'$ the orthogonal projection on the subspace of $L^2(X)$ given by $\ker(\pi(\sigma_1) - \varphi_{i\zeta}(\sigma_1)I)$. Note that the latter subspace consists of functions satisfying: $\pi(\sigma_n)f = (-1)^n f$ in the case of the free groups and $\Gamma(r,2)$, but not otherwise.

As usual let $\mathcal{E}$ be the projection on the space of Γ-invariant functions ($\mathcal{E}f = \int_X f\,dm$ in the ergodic case). Recall that we defined $\sigma_n' = \frac{1}{2}(\sigma_n + \sigma_{n+1})$, and let (X,m) be a Γ-space where $\mathcal{E}' \neq 0$.

We can nor formulate the following result.

THEOREM 10.7 (Ergodic theorems for radial averages on free products [105]). *For $\Gamma = \Gamma(r,h)$ or $\Gamma = \mathbb{F}_k$ and (X,m) (as above), the following holds*:

(1) *σ_n and β_n satisfy the maximal inequality in $L^2(X)$, but are not mean* (*and hence not pointwise*) *ergodic sequences in $L^2(X)$.*
(2) *The sequences σ_n' is a pointwise ergodic sequence in $L^2(X)$.*
(3) *σ_{2n} converges to $\mathcal{E} + c(\Gamma)\mathcal{E}'$, which is a conditional expectation operator w.r.t. a Γ-invariant sub-σ-algebra iff $\Gamma = \Gamma(r,2)$ or $\Gamma = \mathbb{F}_k$.*
(4) *β_{2n} converges to $\mathcal{E} + c(\Gamma)\frac{q(\Gamma)-1}{q(\Gamma)+1}\mathcal{E}'$, which is not a conditional expectation operator on a Γ-invariant sub-σ-algebra.*

The convergence is for each function $f \in L^2(X)$, pointwise almost everywhere and in the norm of $L^2(X)$.

REMARK 10.8. We have chosen to focus on L^2 for simplicity of exposition. The strong maximal inequality for σ_n (and thus σ_n' and β_n) holds in fact in every L^p, $1 < p < \infty$, as can be verified using the argument in [114].

REMARK 10.9 (*The ball averaging problem: some counterexamples*). Recalling the ball averaging problem in ergodic theory stated in Section 4.3, we see that Theorem 10.7 paints a rather complicated picture of the possibilities. First, note that the ball averages β_n do *not* form a mean (and of course, pointwise) ergodic sequences, and neither does the sequence of spheres σ_n. This fact was originally observed in [10], by constructing an ergodic action of $\mathbb{F}_2$ where the special character $\sigma_n \mapsto (-1)^n$ of $\mathcal{A}$ is realized by an joint eigenfunction of the algebra $\mathcal{A}$ acting in L^2. Thus the ball average problem does not have a positive solution for general groups. However, Theorem 10.7 shows that this periodicity phenomenon is the only obstruction, and in fact σ_{2n} and β_{2n} do converge pointwise. Furthermore, note that the limit of σ_{2n} is the conditional expectation w.r.t. the Γ-invariant σ-algebra of sets invariant under a subgroup of index at most two in the case where Γ is a free group, but this is not the case for the groups $\Gamma(r,h)$, $h > 2$. For the ball averages β_{2n} the limit is not a projection operator (and thus not a conditional expectation) at all. Thus it appears that the identification of the exact possible limits of subsequences of σ_n and β_n is a rather delicate problem, which seems inaccessible at this time, for general word-hyperbolic groups, say. The only exception thus far are those groups for which the spectral considerations above (or some variants, see [105]) apply.

The situation just described is of course a reflection of non-amenability, since the balls do not have the Følner property of asymptotic invariance, and so a limit of a subsequence of β_n need not posses any invariance properties w.r.t. the Γ-action.

We remark that it was conjectured in [18, §9] that for every word-hyperbolic group Γ, and every symmetric set of generators the averages $\sigma_{2n} f$ converge pointwise to a function f' invariant under the group Γ_2 generated by all words of even length. However consider the groups $\Gamma = \Gamma(r,h)$, $h > 2$, and a function $f \in L^2(X)$ which realizes the character $\varphi_{i\zeta}$ of the algebra $\mathcal{A}$. Then according to the formula above for $\varphi_{i\zeta}$

$$\lim_{n\to\infty} \pi(\sigma_{2n}) f(x) = \lim_{n\to\infty} \varphi_{i\zeta}(\sigma_{2n}) f(x) = c_{i\zeta} f(x).$$

Thus the limit does exist, but f is not a function invariant under Γ_2, being an eigenfunction of each $\pi(\sigma_{2n})$ with eigenvalue different than 1.

REMARK 10.10 (*Generalization of Birkhoff's theorem*). Given an arbitrary invertible measure preserving transformation T on a probability space X, Birkhoff's pointwise ergodic theorem asserts that for any $f \in L^1(X)$, the averages of f along an orbit of T, namely the expressions $\frac{f(T^{-n}x)+\cdots+f(T^n x)}{2n+1}$ converge, for almost all $x \in X$, to the limit $\tilde{f}(x)$, where $\tilde{f}$ is the conditional expectation of f w.r.t. the σ-algebra of T-invariant sets. Part of our quest to establish ergodic theorems for group actions can thus be motivated by the following obvious and natural question which presents itself. Given *two* arbitrary invertible measure preserving transformations T and S, find a geometrically natural way to average a function f along the orbits of the group generated by T and S, so as to obtain the same conclusion.

Of course if T and S happen to commute, then, according to the discussion in Section 5, the expressions $\frac{1}{(2n+1)^2}\sum_{-n\leqslant n_1,n_2\leqslant n} f(T^{n_1}S^{n_2}x)$ converge for almost all $x \in X$, for any $f \in L^1(X)$, and again the limit is the conditional expectation of f w.r.t. the σ-algebra of sets invariant under T and S. In other words, the pointwise ergodic theorem holds for finite-measure-preserving actions of the free Abelian group on two generators, namely $\mathbb{Z}^2$. However, it is clear that when choosing generically two volume-preserving diffeomorphisms of a compact manifold, or two orthogonal transformations of the Euclidean unit sphere, or in general two measure preserving maps of a given measure space, the group generated by them is not Abelian, and in fact, it is generically free.

The answer to the problem above is then to find an averaging sequence satisfying a pointwise ergodic theorem for finite-measure-preserving actions of the free non-Abelian group on two generators. The first choice that one would consider by direct analogy with Birkhoff's and Wiener's theorems (for $\mathbb{Z}$ and $\mathbb{Z}^d$), would be the normalized ball averages w.r.t. a set of free generators. For the free group this problem has been settled, using the spectral methods described above, by the following result.

THEOREM 10.11 (Generalization of Birkhoff's theorem [114]). *Consider the free group $\mathbb{F}_r$, $r \geqslant 2$, with symmetric free generating set S. Let (X, m) be a probability-preserving ergodic action. Then*

(1) *The sequence $\sigma'_n = \frac{1}{2}(\sigma_n + \sigma_{n+1})$ satisfies the strong maximal inequality and is a pointwise ergodic sequence in L^p, for $1 < p < \infty$.*
(2) *The sequence β_n satisfies the pointwise ergodic in L^p, $1 < p < \infty$, if and only if $L^2(X)$ does not contains a non-zero function f_0 satisfying $\pi(w) f_0 = (-1)^{|w|} f_0$ for every $w \in \mathbb{F}_r$.*
(3) *If such an eigenfunction f_0 is present then it is unique, has constant absolute value, and $\beta_n f_0$ does not converge. For any $f \in L^p(X)$, $1 < p < \infty$,*

$$\lim_{n\to\infty} \beta_{2n} f(x) = \int_X f\, dm + \frac{r-1}{r}\int_X f\overline{f_0}\, dm \cdot f_0(x)$$

pointwise and in the L^p-norm.

PROBLEM 10.12. We note that the weak-type maximal inequality in L^1 for the sphere averages (or equivalently, the ball averages) on the free group is an open problem.

REMARK 10.13. Note that the pointwise ergodic theorem for the free finitely generated group $\mathbb{F}_k$ implies a corresponding one for any factor group of $\mathbb{F}_k$, namely for any finitely generated group, just as Wiener's theorem for $\mathbb{Z}^d$ implies the corresponding result for any finitely generated Abelian group. However, the weights that must be taken on the factor group are those induced by the canonical factor map, and these usually bear little resemblance to the intrinsic ball and sphere averages on the factor group.

REMARK 10.14 (*Sphere averages on free algebras*). We note that the same spectral methods that were employed for sphere averages in the group algebra of the free group can be used more generally for other free algebras in various varieties. For example, consider the free associative algebra on r non-commutative elements. This algebra has of course a natural length function, and clearly the algebra of radial elements is commutative (under convolution), and satisfy a first-order recurrence relation. It is thus a simple exercise to develop the spectral theory of the $*$-representations of the subalgebra of radial elements. These include representation where each generator is mapped to bounded self-adjoint contraction on a Hilbert space, and we thus obtain a pointwise ergodic theorem for the powers of the self-adjoint operator which is the uniform average of the r contractions. Similarly, we can consider the free algebra on r non-commuting idempotents, where again we have a commutative subalgebra of radial elements. The $*$-spectrum can be determined here too, again by solving a second-order recurrence relation with constant coefficients. The $*$-representations include those where each generator is mapped to a self-adjoint projection. We thus obtain in particular a pointwise ergodic theorem for the radial averages of (non-commuting) conditional expectations on a probability space.

In Theorem 10.7 we already considered the convolution algebras of the free products $\mathbb{Z}_p * \cdots * \mathbb{Z}_p$ which are the free groups generated by r elements of order p. The $*$-representations here are given by the unitary representations of the groups, and the spherical functions can again be explicitly determined from a second-order recurrence relation (see, e.g., [105] for their description).

There are further examples of free algebras in other varieties, where the radial elements form a commutative subalgebra whose $*$-spectrum can be determined using a recurrence relation. In all of these cases a mean and pointwise ergodic theorem for sphere averages in $*$-representations is obtained, with the sole obstruction given by periodicity phenomena, when they occur.

REMARK 10.15 (*Boundary transitive subgroups of tree automorphisms*). Let us note that the group algebra of the simple algebraic group $PGL_2(\mathbb{Q}_p)$ also contains an isomorphic copy of the algebra $\mathcal{A}$ of even radial averages on the tree T_{p+1}. This follows from the fact that the group has a faithful representation as a group of automorphisms of the regular tree, which is transitive on the boundary. Again the averaging operators on the tree can be represented as convolution operators on $PGL_2(\mathbb{Q}_p)$. However, here the Howe–Moore mixing theorem [74] applies, namely the matrix coefficients of unitary representations without invariant unit vectors vanish at infinity on an algebraically connected simple group.

This implies that the sign character of the algebra $\mathcal{A}$ cannot appear in ergodic actions of $PSL_2(\mathbb{Q}_p)$, and both β_{2n} and σ_{2n} converge pointwise and in norm to the ergodic mean, in every L^p, $1 < p < \infty$, as follows from [105] and [114]. A similar analysis holds for many closed non-compact boundary-transitive subgroups of the group of automorphisms of a semihomogeneous tree [105]. The analog of the Howe–Moore theorem here was proved in [95].

REMARK 10.16 (*Non-commutative Hecke algebras*). We note that further natural algebra structures appear in the setting of groups of automorphisms of semihomogeneous tree, and more generally groups of automorphisms of the Bruhat–Tits buildings of semisimple algebraic groups over locally compact totally disconnected non-discrete fields. These are the Hecke algebras $L^1(Q \setminus G/Q)$ of double cosets of a compact open subgroup Q under convolution, which are non-commutative in general, but have the property that all of their irreducible $*$-representations have a uniformly bounded degree [9]. Particularly significant among these algebras is the Iwahori algebra, consisting of double cosets of an Iwahori subgroup. In the case of the group $Aut(T_{r_1,r_2})$, $(r_1 \neq r_2)$ for example, the Iwahory subgroup can be identified with the stability group of an edge, and thus it is contained in the two non-conjugate maximal compact open subgroup stabilizing one of the vertices of the edge. Thus the commutative algebra associated with a Gelfand pair structure on the group is contained as a subalgebra of the Iwahori algebra in this case (and others). The spectrum of the Iwahori algebra on a semihomogeneous tree can be easily determined (see, e.g., [104]), and it appears naturally when analyzing the spectrum of some natural convolution algebras in certain lattices of $Aut(T_{r_1,r_2})$ $(r_1 \neq r_2)$. These include for example the groups $\mathbb{Z}_p * \mathbb{Z}_q$, and so also the group $PSL_2(\mathbb{Z}) = \mathbb{Z}_2 * \mathbb{Z}_3$. Thus it is possible also to use non-commutative harmonic analysis on Hecke algebras to derive ergodic theorems for discrete groups. This possibility was explored in the case of semihomogeneous trees in [105], but it is natural to expect that it can be developed much further.

REMARK 10.17. We note that an alternative proof of the ergodic theorems for spheres on the free group was developed by A. Bufetov [18]. The method is based on the theory of Markov processes rather than on spectral theory, and will be discussed further in Section 12.4, together with some other ergodic theorems on free groups and other Markov groups.

11. Actions with a spectral gap

We now turn to a discussion of a fundamental phenomenon that appears in the study of non-amenable algebraic groups, and which has no analog in the theory of amenable groups. Utilizing it, we will be able to greatly expand the scope of the radial ergodic theorems on semisimple algebraic groups, obtain quantitative estimates in the pointwise ergodic theorems, and also obtain a host of results on a diverse array of non-radial averages.

The phenomenon in question is the existence of properly ergodic (i.e. non-transitive) actions with a spectral gap. We define the latter property in the form most convenient for our purposes here, as follows.

DEFINITION 11.1 (*Spectral gaps*).

(1) A strongly continuous unitary representation π of an lcsc group G is said to have a spectral gap if $\|\pi(\mu)\| < 1$, for some (or equivalently, all) absolutely continuous symmetric probability measure μ whose support generate G as a group.
(2) Equivalently, π has a spectral gap if the Hilbert space does not admit an asymptotically-G-invariant sequence of unit vectors, namely a sequence satisfying $\lim_{n\to\infty} \|\pi(g)v_n - v_n\| = 0$ uniformly on compact sets in G.
(3) A measure preserving action of G on a σ-finite measure space (X, m) is said to have a spectral gap if the unitary representation of G in the space orthogonal to the space of G-invariant functions has a spectral gap. Thus in the case of an ergodic probability-preserving action, the representation in question is on the space $L^2_0(X)$ of function of zero integral.
(4) An lcsc group G is said to have *Kazhdan's property T* [83] provided every strongly continuous unitary representation which does not have G-invariant unit vectors has a spectral gap.

REMARK 11.2.

(1) The equivalence between (1) and (2) is a standard argument in spectral theory and can be found, e.g., in [99].
(2) The phenomenon of spectral gaps does not occur for properly ergodic probability-preserving actions of amenable groups. Indeed, in any such action X, there exists a non-trivial sequence of sets $A_n \subset X$ whose measures satisfy $0 < c < m(A_n) < C < 1$ for all n, which is asymptotically invariant, namely $\lim_{n\to\infty} m(gA_n \Delta A_n) = 0$ uniformly on compact sets in G [132]. It follows immediately that $\|\pi(\mu)\|_{L^2_0(X)} = 1$ for *every* probability measure μ on G.
(3) We recall the well-known fact that amenability of an lcsc group (namely the existence of a Følner sequence) can be characterized by the condition that the left regular representation λ_G satisfies $\lambda_G(\mu) = 1$, for at least one (or equivalently, all) absolutely continuous symmetric probability measure μ whose support generate G as a group.

In the realm of non-amenable algebraic groups, actions preserving a σ-finite measure which have a spectral gap are quite ubiquitous, and we briefly indicate some examples.

EXAMPLE 11.3 (*Some examples of actions with a spectral gap*).

(1) G a non-amenable group, $X = G$, and the action is by left translations, w.r.t. Haar measure.
(2) G a connected semisimple Lie group, and the action is by isometries of the Riemannian symmetric space G/K. More generally, the action on a reductive symmetric space of the form G/H, where H is the fixed-point-group of an involutive automorphism of G.
(3) G a simple algebraic group, and the action by translation on the homogeneous space G/L, where L is a proper algebraic unimodular subgroup [65], e.g., the action of $SL_n(\mathbb{R})$ on the space of symmetric matrices, or the action of $Sp(n, \mathbb{R})$ on $\mathbb{R}^{2n}$.

(4) Similarly, for G a simple algebraic group the action by translations on the homogeneous space G/Λ where $\Lambda \subset G$ is a discrete subgroup which is not Zariski dense (see, e.g., the discussion in [77]).
(5) $\tau : G \to H$ a representation of a semisimple algebraic group G in a simple algebraic group H, $\Delta \subset H$ a lattice subgroup, and G acts on $X = \Delta \setminus H$, a locally symmetric space of finite volume via τ.
(6) If G is a simple algebraic group of split-rank at least two then G has property T (see, e.g., [97]), and then of course *any* unitary representation without invariant unit vectors has a spectral gap.

11.1. *Pointwise theorems with exponentially fast rate of convergence*

The utility of a spectral gap in a given representation π is in the fact that typically, given a natural family μ_t of probability measures on G, not only do we have $\|\pi(\mu_t)\| < 1$ for each $t > 0$, but in fact (as we shall see presently) the far stronger conclusion that the norms decay exponentially in t holds, namely:

$$\left\|\pi(\mu_t)\right\| \leqslant C_\mu \exp(-\delta_\mu t),$$

where $\delta_\mu > 0$ depends on the family μ_t.

When $G = SL_2(\mathbb{C})$, for example, the spectral gap condition is equivalent with the condition that the $*$-spectrum that arise in the representation π of $M(G, K)$ contains only complementary series characters φ_a with parameter satisfying $a \leqslant 1 - \theta$, $\theta = \theta(X) > 0$ (except for the trivial character). The exponential decay of the operator norms of σ_t, for example, follows immediately upon evaluating the characters φ_z on the sphere S_t. It also follows easily for γ_t and β_t upon integrating φ_z against these measures, as we will see below.

The exponentially decaying norm estimate above is a most useful fact, which as we shall see gives rise to an interesting new phenomenon in ergodic theory, namely pointwise ergodic theorems with an explicit exponentially fast rate of convergence to the ergodic mean, for properly ergodic actions. The validity of the norm estimate follows from spectral estimates that we will consider in more detail in Section 11.2. But to illustrate the point, we now turn to our first use of such estimates, namely the following pointwise ergodic theorem with an error term for the bi-K-invariant averages β_t on a simple Lie group, and to its proof. We recall that the averages β_t we shall consider are defined as the K-invariant lifts to G of the K-invariant probability measures on balls B_t w.r.t. the Killing form on the symmetric space G/K, with center $[K]$ and radius t. This family generalizes the case of hyperbolic space considered in Section 9.

THEOREM 11.4 (Pointwise ergodic theorem with exponentially fast rate of convergence for ball averages on semisimple Lie groups in actions with a spectral gap [99]). *Let G be connected non-compact semisimple Lie group with finite center, and let the G-action on X*

have a spectral gap. Then the ball averages β_t converge pointwise exponentially fast to the ergodic mean. More precisely $\forall f \in L^p(X)$, $p > 1$, and for almost every $x \in X$:

$$\left|\beta_t f(x) - \int_X f\, dm\right| \leqslant B_p(f,x)\exp(-\theta_p t), \quad \theta_p > 0,$$

where θ_p depends on p, G and X.

Furthermore, the integer-radius sphere averages σ_n also converge pointwise exponentially fast to the ergodic mean.

Model case: *Proof for ball averages β_t in ergodic actions of $SL_2(\mathbb{C})$ with a spectral gap.*

Recall that we are assuming here that the $*$-spectrum determined by the representation π_0 of $G = SL_2(\mathbb{C})$ and $M(G,K)$ on $L^2_0(X)$ (the space of functions with zero integral), contains only complementary series characters φ_a with parameter satisfying $a \leqslant 1-\theta$, $\theta = \theta(X) > 0$. We denote the spectrum by $\Sigma^*(\pi_0)$. The proof proceeds along the following steps.

(1) Since $a \leqslant 1-\theta < 1$, we have, using the explicit form of the characters (see Section 9.3):

$$\left|\varphi_a(\sigma_t)\right| \leqslant B\exp(-\theta t)$$

and hence we have the following exponential decay estimate on the norm of the sphere averages

$$\|\sigma_t\|_{L^2_0 \to L^2_0} = \sup_{z \in \Sigma^*(\pi_0)} \left|\varphi_z(\sigma_t)\right| \leqslant B\exp(-\theta t).$$

(2) Similarly, for the ball averages, using the estimate of their density:

$$\begin{aligned}\|\beta_t\|_{L^2_0 \to L^2_0} &\leqslant C\exp(-2t)\int_0^t \exp(2s)\|\sigma_s\|_{L^2_0 \to L^2_0}\, ds \\ &\leqslant C\,B\exp(-2t)\int_0^t \exp\big((2-\theta)s\big)\, ds \leqslant C'\exp(-\theta t).\end{aligned}$$

(3) For integer radius balls (and similarly, spheres) acting in L^2_0, we have:

$$\sum_{n=0}^{\infty}\left\|\exp\left(\frac{\theta}{2}n\right)\beta_n f\right\|_2^2 \leqslant \sum_{n-0}^{\infty}\exp(-\theta n)\|f\|_2^2 < \infty,$$

equivalently:

$$\sum_{n=0}^{\infty}\int_X \left|\exp\left(\frac{\theta}{2}n\right)\beta_n f(x)\right|^2 dm < \infty$$

and so

$$\sup_{n\geqslant 0}\left|\exp\left(\frac{\theta}{2}n\right)\beta_n f(x)\right|^2 \leqslant \sum_{n=0}^{\infty}\left|\exp\left(\frac{\theta}{2}n\right)\beta_n f(x)\right|^2 = C(f,x)^2 < \infty$$

for almost all $x \in X$.

(4) We can therefore conclude:
   (A) β_n (and σ_n) satisfies the *exponential-maximal inequality* in $L_0^2(X)$ stated in (3), and
   (B) β_n (and σ_n) *converges pointwise exponentially fast* to the ergodic mean for $f \in L^2$, i.e. for almost every $x \in X$:

$$\left|\beta_n f(x) - \int_X f\,dm\right| \leqslant C\left(f - \int_X f\,dm, x\right)\exp\left(-\frac{\theta n}{2}\right).$$

(5) The natural generalization of statements (A) and (B) in (4) above are of course true in every L_0^p, $p > 1$, by the Riesz–Thorin interpolation theorem.

(6) Now note that the foregoing arguments show in fact a more general result, namely that the same conclusion holds for every sequence β_{t_k} with $\sum_{k\in\mathbb{N}}\exp(-\frac{1}{2}\theta t_k) < \infty$. Fix such a sequence t_k, and given a point t, choose the closest point to it (which we assume is at a distance at most one), and denote it by t_n. Now write *for f bounded*:

$$\left|\beta_t f(x) - \int_X f\,dm\right| \leqslant \left|\beta_t f(x) - \beta_{t_n} f(x)\right| + \left|\beta_{t_n} f(x) - \int_X f\,dm\right|.$$

The second term is bounded (using part (3) above, and replacing the integers by the sequence $t_k \in \mathbb{R}$) by

$$C'\left(f - \int_X f\,dm, x\right)\exp\left(-\frac{1}{2}\theta t_n\right),$$

where

$$\left\|C'\left(f - \int_X f\,dm, \cdot\right)\right\|_2 \leqslant B\|f\|_2 \leqslant B\|f\|_\infty.$$

As to the first term, note that the family β_t is uniformly (locally) Lipschitz continuous (w.r.t. the $L^1(G)$-norm), and the function f is bounded. It follows that the first term is bounded by $|t - t_n|\|f\|_\infty$.

(7) Let us choose the sequence t_k fine enough, so that it satisfies $|t - t_n| \leqslant \exp(-\frac{1}{4}\theta n)$. This can be achieved by dividing the interval $[n, n+1]$ to $2 + [\exp(\theta n/4)]$ equally spaced points, and the resulting sequence still satisfies the condition $\sum_{k\in\mathbb{N}}\exp(-\frac{1}{2}\theta t_k) < \infty$ stated in (6). We can then use the argument of (6) for the sequence t_k to estimate both the first and the second term.

The estimate in (6) immediately implies the following, which we call an (L^∞, L^2)-exponential maximal inequality:

$$\left\| \sup_{t>0} \exp\left(\frac{1}{4}\theta t\right) \left| \beta_t f(x) - \int_X f\, dm \right| \right\|_{L^2} \leqslant C_2 \|f\|_\infty.$$

(8) We can now conclude that for every bounded function f the expression

$$\sup_{t>0} \exp\left(\frac{1}{4}\theta t\right) \left| \beta_t f(x) - \int_X f\, dm \right| = B(f,x)$$

is finite almost everywhere, and the conclusion of Theorem 11.4 is thus established for f in L^∞.

(9) We now note that the strong maximal inequality for ball averages of Theorem 9.4 namely

$$\left\| \sup_{t>0} \beta_t f \right\|_p \leqslant C_p \|f\|_p, \quad p > 1,$$

can be established very easily for balls with exact exponential growth in actions with a spectral gap. Indeed, clearly exact exponential volume growth implies that for a fixed constant B we have $\beta_t \leqslant B\beta_{[t]+1}$, $t \geqslant 1$, as measures on G. Thus the maximal inequality for the family of all balls follows from its validity for the sequence of balls with integer radii. The boundedness of the latter maximal function is an elementary conclusion of exponential decay of the operator norm, as follows from an obvious variation on the arguments presented in parts (3) and (4).

(10) Finally, we can use the analytic interpolation theorem again. This time we interpolate between the exponential maximal inequality stated in (7) (from L^∞ to L^2) and the strong maximal inequality (from L^p to L^p, $p > 1$) for the ball averages, stated in (9). We then obtain an (L^p, L^r)-exponential maximal inequality (in the obvious notation) and hence exponential pointwise convergence to the ergodic mean, for every $f \in L^p$, $1 < p < \infty$. We refer to [99] for the details. This concludes the outline of the proof of Theorem 11.4.

The proof above is complete for actions of $SL_2(\mathbb{C})$ with a spectral gap, and similar arguments also yields the general case of ball averages on semisimple groups in actions with a spectral gap. One uses spectral estimates of spherical functions, resulting in the exponential decay of the operator norm (see the discussion in the following section). Further, the monotonicity property $\beta_t \leqslant C\beta_{[t]+1}$ of the ball averages is valid here and follows from strict $t^q \exp ct$-volume growth of the balls, which is a relatively straightforward consequence of the structure theory of semisimple Lie groups. The monotonicity is utilized to deduce the strong maximal inequality in L^p, $1 < p < \infty$, of the operator $\sup_{t>0} \beta_t$ from the discrete version $\sup_{n\in\mathbb{N}} \beta_n$ (see more on this argument in Section 12). It is also necessary to establish that the ball averages are uniformly locally Lipschitz continuous in the $L^1(G)$-norm. We refer to [99,108] for the details.

11.2. *The spectral transfer principle*

Spectral estimates for spherical functions are crucial to the successful implementation of the spectral approach to maximal inequalities and pointwise ergodic theorems that was outlined in the preceding section. The uniform derivative estimates necessary for the proof of the maximal inequalities for spheres (and other singular averages) are usually very difficult and have not been established in general. However the basic exponential decay estimate for the operator norm of radial averages such as spheres and balls (acting in an irreducible unitary representation, say) hold in great generality and depend only on basic structural features that hold for all simple algebraic groups over locally compact non-discrete fields.

These estimates were developed in various forms by M. Cowling [35] and R. Howe [73], as well as C.C. Moore [74], U. Haagerup [39] and Borel and Wallach [13]. An elegant exposition to the case of $SL_n(\mathbb{R})$ appears in [75], and the strategy outlined there was used by H. Oh [117,118] to obtain definitive quantitative results for semisimple algebraic groups. We summarize some of these results as follows.

THEOREM 11.5 (Decay estimate and L^p-integrability of matrix coefficients).

(1) [35] *Let G be a simple non-compact connected Lie group with finite center. For every irreducible non-trivial unitary representation $(\pi, \mathcal{H})$ of G, and every two K-finite vectors $u, v \in \mathcal{H}$, the associated matrix coefficient $\psi_{u,v}(g) = \langle \pi(g)u, v\rangle$ has the following two properties.*
   (a) *The matrix coefficient satisfies an exponential decay estimate along the group G:*

$$\big|\psi_{u,v}(g)\big| = \big|\langle \pi(g)u, v\rangle\big| \leqslant C_{u,v} \exp\big(-\delta_\pi d(K, gK)\big),$$

   where d denotes the distance function associated with the Riemannian metric given by the Killing form on the symmetric space G/K.
   (b) *The matrix coefficient $\psi_{u,v}(g) = \langle \pi(g)u, v\rangle$ belongs to $L^p(G)$, for some $p = p(\pi) < \infty$.*

(2) [117, Theorem 1.1] *The same estimate holds for infinite-dimensional irreducible unitary representations of any simple algebraic group over a locally compact non-discrete field F with Char $F \neq 2$, where K is a good maximal compact subgroup of G (see* [22, §3.5]*) and $d(gK, K)$ the metric induced on G/K by its inclusion in the Bruhat–Tits building of G.*

(3) [35], [118, §5.7] *When G has property T, the same estimates hold in both cases for K-finite vectors in any unitary representation of G, which does not have G^+-invariant unit vectors.*

REMARK 11.6.

(1) We note that in this set-up, G can be defined to have property T if and only if $p(\pi) \leqslant p(G) < \infty$ for all irreducible non-trivial unitary representations, i.e. the spectral estimate depends only on G and holds uniformly for all the representations. Equivalently $\inf_\pi \delta_\pi > 0$ for all representations π with a spectral gap.

(2) Property T holds for all simple algebraic groups of split-rank at least two, but also for some simple real Lie groups of real-rank one—see [97] for a discussion.
(3) We remark that even more precise quantitative estimate for the decay of positive definite spherical functions have been developed and refer the reader to the reference cited above.
(4) Finally, G^+ is the group generated by the split unipotent subgroups of G (see [97, Chapter I, §§1.5, 2.3] for a discussion). It is normal and co-compact in G, and is of finite index in G whenever the characteristic of the field is zero. In particular, it coincides with G when G is a connected semisimple Lie group without compact factors.

Note that for any $n \geqslant p/2$, Theorem 11.5 implies that $\psi_{u,v}(g)^n \in L^2(G)$. This fact implies, much as in the Peter–Weyl theorem for compact groups, the following result due to M. Cowling in the real semisimple case, and R. Howe and C.C. Moore in general.

THEOREM 11.7 (Spectral Transfer Principle [35,74]). *Let G be as in Theorem* 11.5. *If the representation π of G has a spectral gap, there exists $n = n(\pi)$ such that*

$$\pi^{\otimes n} \subset \infty \cdot \lambda_G,$$

where $\infty \cdot \lambda_G$ denotes the direct sum of countably many copies of the regular representation of G, $\pi^{\otimes n}$ the n-fold tensor power of the representation π, and $\subset$ denote a unitary isomorphism onto a subrepresentation. If G has property T then there exists a uniform bound $n(\pi) \leqslant n(G) < \infty$ for all representations π with spectral gap (*and conversely*).

This result has the following explicit spectral estimate as a corollary:

THEOREM 11.8 (Uniform norm estimate of arbitrary measures on G in arbitrary representations with a spectral gap [108, Theorem 1.1]). *Let G be a group as in Theorem* 11.5. *Let π be any unitary representation of G with a spectral gap. Let μ be any probability measure on G. Then*

$$\left\|\pi(\mu)\right\| \leqslant \left\|\lambda_G(\mu)\right\|^{1/n(\pi)}.$$

In particular, if (X, m) is a probability-preserving ergodic action of G with a spectral gap, then

$$\left\|\pi(\mu)\right\|_{L^2_0(X)} \leqslant \left\|\lambda_G(\mu)\right\|^{1/n(\pi_0)}.$$

Note that it follows easily from the spectral estimate of Theorem 11.5(1) that typically, given a family μ_t of radial probability measures on G satisfying mild natural growth conditions, we have an *exponential decay estimate on the convolution norm*:

$$\left\|\lambda_G(\mu_t)\right\| \leqslant \exp(-\delta t), \quad \delta = \delta(\mu) > 0.$$

In particular this visibly holds for sphere averages, balls averages, a host of variations of shell averages [113], and many other radial averages μ_t on G.

As we shall see in Section 12, it is possible to use Theorem 11.8, in conjunction with the arguments of Section 11.1 to establish exponential-maximal inequalities for a wide class of non-radial families μ_t, as well as exponential pointwise convergence to the ergodic mean. But before turning to non-radial averages let us mention an application of the spectral transfer principle in the radial case, namely establish the pointwise ergodic theorems in the case of totally disconnected simple algebraic groups.

11.3. *Higher-rank groups and lattices*

We have encountered in Section 10.5 totally disconnected Gelfand pairs which appeared as groups acting on semihomogeneous trees. A more general natural set of geometries to consider is that of affine Bruhat–Tits buildings, in particular those associated with semi-simple algebraic groups over locally compact totally disconnected non-discrete fields. The case of totally disconnected simple algebraic groups of (split) rank one reduces to that of closed boundary-transitive group of automorphisms of semihomogeneous trees. Thus the pointwise ergodic theorems for sphere and ball averages on them are completely resolved by Theorem 10.7 together with Remark 10.8 and Remark 10.15.

Consider now the case of a simple algebraic group over locally compact totally disconnected non-discrete field F of split rank at least two. Such a group G has a good compact open subgroup K giving rise to a Gelfand pair structure (see, e.g., [22, §3.5 and Theorem 4.1] and the references there). Furthermore, as follows immediately from Theorem 11.5, the spherical functions associated with (G, K) decay exponentially with a uniform bound, when the group is simple of split rank greater than one. More precisely, positive-definite spherical functions associated with irreducible non-trivial representations are bounded by a function which decays exponentially fast to zero as a function of $d(gK, K)$, where d is the G-invariant distance on the building, as is the case when G/K is a Riemannian symmetric space. The same holds true for matrix coefficients of unitary representations without invariant unit vectors. It follows that the spectral norm of the convolution operators associated with the bi-K-invariant ball averages β_n on G decays exponentially in n. Together with the spectral transfer principle for these groups, which follows from the estimate just described, Theorem 11.8 implies that the norm of $\pi_0(\beta_n)$ in $L^2_0(X)$ for any ergodic action X decays exponentially also, with a fixed bound, independent of X. Using the arguments brought in the proof of Theorem 11.4 (for integer radius only, this time!) the exponential norm decay estimate gives a proof of the following result.

THEOREM 11.9 (Pointwise ergodic theorem with exponentially fast rate of convergence for simple algebraic groups). *Let G and K be as in Theorem* 11.5. *Let β_n be the bi-K-invariant ball averages and σ_n the sphere averages. Assume G has split rank at least two, or more generally, property T. Then for any action of G on a probability space which is ergodic under G^+, for all $f \in L^p(X)$, $1 < p < \infty$, and almost every $x \in X$*

$$\left| \pi(\beta_n) f(x) - \int_X f \, dm \right| \leqslant C_p(f, x) \exp(-n\delta_p),$$

where $\delta_p > 0$ depends only on G and p. The same result holds also for the sphere averages σ_n. Furthermore $\|C_p(f,\cdot)\|_{L^p(X)} \leqslant B_p\|f\|_{L^p(X)}$, and so the maximal functions associated with the spheres (and balls) satisfy an exponential-maximal inequality in $L_0^p(X)$.

Finally, the same holds true for any action of a group as in Theorem 11.5*, provided that it has a spectral gap, but with δ_p here depending also on the action.*

Thus we see that the ball and sphere averages converge exponentially fast to the ergodic mean from almost every starting point, with a fixed rate independent of the starting point as well as the action.

Recall that our discussion in Section 10.5 of totally disconnected Gelfand pairs extended also to some of their lattices. Note that the free group appeared in our discussion there as a group of automorphisms of the tree which acts simply transitively on the vertices. The fact that the group algebra of the free group contained an isomorphic copy of the algebra of radial averaging operators on the tree followed without difficulty.

The discussion is not limited of course to groups acting on semihomogeneous trees, and this set-up can be expanded considerably, as follows. In [23] the authors have constructed groups acting simply transitively on the vertices of certain affine buildings of rank 2, namely the $\tilde{A}_2$-buildings. Furthermore, such groups have been constructed for $\tilde{A}_n$-buildings (which are of rank n) for any n in [25]. As noted already, a simple algebraic groups G always contains a good compact open subgroup such that (G, K) is a Gelfand pair (see, e.g., [22]), and thus for appropriately chosen lattices one can expect to find an isomorphic copy of the Gelfand pair algebra $L^1(G, K)$ in the group algebra $\ell^1(\Gamma)$ of the lattice. In [24] the authors have succeeded in showing that such a commutative algebra does occur in the group algebras of some of the discrete groups constructed in [23]. Furthermore they analyzed its structure and $*$-representations, even in the case where the discrete groups do not arise as lattices in simple algebraic groups. An interesting new feature that arise here is the fact that the discrete groups in question satisfy property T, a fact that is proved in [24] directly from the representation theory of the commutative algebra in question. These results are very interesting in the context of ergodic theory, as they demonstrate the following rather remarkable phenomenon (based on the results of [24] and [128]).

THEOREM 11.10 (Uniform pointwise ergodic theorem for some groups with property T). *There exists a discrete group Γ with a finite generating set S, with the following property. For some fixed positive $\delta > 0$ depending only on (Γ, S), and every ergodic probability preserving action of Γ on (X, m), for all $f \in L^2$, and almost every $x \in X$*

$$\left|\pi(\sigma_n) f(x) - \int_X f\, dm\right| \leqslant C_2(f,x)\exp(-n\delta).$$

In fact, there are infinitely many $\tilde{A}_2$ groups satisfying the property above. The same result holds of course in L^p, $1 < p < \infty$, with $\delta_p > 0$ (but this is an open problem in L^1!).

This phenomenon is of course, in striking contrast to the behavior of averages on discrete amenable groups in classical ergodic theory. It raises the following intriguing problem, which however seems quite inaccessible at this time.

PROBLEM 11.11. Does the phenomenon described in Theorem 11.10 occur for *every* discrete group with property T, and for every set of generators on it?

12. Beyond radial averages

The spectral methods introduced in Section 11, together with some further arguments, can be used to prove a diverse variety of (not necessarily radial) pointwise ergodic theorems on semisimple Lie and algebraic groups. When the action has a spectral gap, it is possible to establish exponentially fast rate of convergence for general families of probability measures μ_t, which are decidedly non-radial. Such families often have a great deal of intrinsic geometric interest and occur naturally in applications. We now turn to an exposition of some of these results, referring for more details to [108] and [113].

12.1. *Recipe for pointwise theorems with rate of convergence*

12.1.1. *Estimating convolution norms.* When discussing the ergodic theory of a general family μ_t of probability measures on a semisimple Lie group G, the first step is to establish exponential decay estimates on the norms of the convolution operators $\lambda_G(\mu_t)$. This will be then converted by the spectral transfer principle (Theorem 11.7 and Theorem 11.8) to norm decay estimates in an arbitrary measure-preserving action with a spectral gap.

There exists a fundamental estimate of the convolution norms when the averages μ_t are absolutely continuous, and it is given in terms of the $L^p(G)$-norms of the densities of μ_t. The validity of such an estimate is called the Kunze–Stein phenomenon, established in [89] for $G = SL_2(\mathbb{R})$ and in [34] in general. The precise formulation is as follows.

THEOREM 12.1 (Kunze–Stein phenomenon [89,34]). *Given a connected semisimple Lie group G with finite center, for every $1 \leqslant p < 2$ there exists a constant K_p satisfying: $\|F * f\|_2 \leqslant K_p \|F\|_p \|f\|_2$, for every $F \in L^p(G)$ and $f \in L^2(G)$.*

We will refer to any lcsc group satisfying the estimate of Theorem 12.1 a Kunze–Stein group.

12.1.2. *Monotonicity, Hölder continuity, and norm decay.* Suppose then that indeed for some $1 < r < 2$ we have $\|\mu_t\|_{L^r(G)} \leqslant C \exp(-\theta t)$, $\theta > 0$, for a family μ_t of absolutely continuous measures on a semisimple Lie group G. It follows from Theorem 12.1 and Theorem 11.8 that in any probability-preserving action of G with a spectral gap, the sum $\sum_{n=0}^{\infty} \|\mu_n\|_{L^2_0}^2$ is finite. It follows easily that the operator $\sup_{n\in\mathbb{N}} |\mu_n f(x)|$ satisfies the strong maximal inequality in L^p, $1 < p < \infty$. (In fact, it satisfies an exponential maximal inequality in $L^2_0(X)$, see Section 11.1.) We would like to use the boundedness of the maximal function for the sequence μ_n in order to prove a strong L^p-maximal inequality for μ_t, $t \in \mathbb{R}$. To that end, recall the use we made in Section 11.1 of the estimates for the volume growth of the balls, namely the fact that we could dominate the measure β_t by

$C\beta_{[t]+1}$, C fixed. Thus it is natural to introduce the following conditions on general probability measures μ_t [99,108], generalizing the conditions used already for the ball averages in Section 11.

DEFINITION 12.2. A family μ_t of probability measures on an lcsc group G is

(1) Monotone, if $\mu_t \leqslant C\mu_{[t]+1}$, as measures on G (where C is fixed, independent of $t > 0$).
(2) Uniformly locally Hölder continuous, if for $F \in L^\infty(G)$ and $t > 0$

$$\left|\mu_{t+s}(F) - \mu_t(F)\right| \leqslant Cs^a \|F\|_\infty, \quad 0 < s \leqslant 1.$$

We can generalize the arguments used in Section 11 and formulate a recipe for proving pointwise convergence with exponentially fast rate, using the monotonicity, local Hölder continuity and exponential decay of norms, as follows. (We refer to Section 11.1 for an example of the method, and [99,108] and [113] for more details on its use and applications.)

Recipe for pointwise ergodic theorems with exponentially fast rate of convergence. To prove the strong maximal inequality for μ_t, we follow the following steps.

(1) Let $f \in L^2(X)$ be a non-negative function on X. Then, since $\pi(\mu_t)f(x) \leqslant C\pi(\mu_{[t]+1})f(x)$, we have

$$\left\| \sup_{t>0} \pi(\mu_t)f(x) \right\|^2_{L^2(X)} \leqslant C^2 \left\| \sup_{n\in\mathbb{N}} \pi(\mu_n)f(x) \right\|^2_{L^2(X)} \leqslant C^2 \|f\|^2_{L^2(X)}.$$

(2) The previous argument clearly extends to every L^p, $1 < p < \infty$, using the Riesz–Thorin interpolation theorem.
(3) Using the estimate $\|\pi(\mu_t)\|_{L^2_0(X)} \leqslant C\exp(-\theta t)$, the argument in (1) can be used to prove an exponential maximal inequality for the operators $\exp(\frac{1}{2}\theta t_k)\mu_{t_k}$ in L^2_0, where t_k is a sequence such that the sum of the norms converges. Repeat the argument in $L^p_0(X)$.
(4) Now distribute $\exp(\frac{1}{4}\theta n)$ equally spaced points in the interval $[n, n+1]$. Then approximate $\pi(\mu_t)f$ by $\pi(\mu_{t_n})f$ using the closest point t_n to t in the sequence t_k. Estimate the difference using the exponential maximal inequality for the entire sequence μ_{t_k}, and the local Hölder regularity of the family μ_t, applied when f is a bounded function.
(5) The previous argument gives an (L^∞, L^2)-exponential maximal inequality, which says that the exponential-maximal function for f bounded has an L^2 norm bound in term of the L^∞-norm of f. Now interpolate against the usual strong maximal inequality in L^p proved in step (1), using the analytic interpolation theorem.

Thus the recipe above establishes the following result.

THEOREM 12.3 (Pointwise ergodic theorem with exponentially fast rate of convergence for general averages on semisimple Lie groups [108]). *Let G be a connected semisimple Lie group with finite center (or any Kunze–Stein group). Let $f_t \in L^1(G)$ satisfy $f_t \geqslant 0$, and $\int_G f_t(g)\,dg = 1$. Assume that the family of probability measures μ_t with density f_t form a*

monotone and uniformly locally Hölder continuous family. Assume that $\|f_t\|_{L^r(G)} \leqslant Ce^{-\theta t}$ *for some* $1 < r < 2$ *and some* $\theta > 0$. *Let* (X, m) *be a probability-preserving action of* G, *and assume that the unitary representation* π_0 *of* G *on* $L^2_0(X)$ *satisfies* $\pi_0^{\otimes n} \subset \infty \cdot \lambda_G$. *Then* $\pi(\mu_t)$ *satisfies a pointwise ergodic theorem with exponentially fast rate of convergence to the ergodic mean in* L^p, $1 < p < \infty$. *In particular, for every* $f \in L^p(X)$, *and for almost every* $x \in X$,

$$\left|\pi(\mu_t) f(x) - \int_X f\, dm\right| \leqslant B_p(x, f) \exp\left(-\frac{\theta_p}{2n} t\right), \quad \theta_p > 0.$$

Furthermore an exponential (L^p, L^r)*-maximal inequality holds in every* L^p, $1 < p < \infty$.

As we shall now demonstrate, Theorem 12.3 applies to some interesting geometric averages, as follows.

12.2. *Horospherical averages*

Theorem 12.3 requires in its assumptions an estimate on the convolution norm $\|\lambda_G(\mu_t)\|$. Let us therefore note that the majorization principle due to C. Herz ([71], see [37] for a discussion) has a corollary which is very useful in this regard, due to M. Cowling, U. Haagerup and R. Howe [39]. The corollary in question allows us to estimate the norm of the convolution operator $\lambda_G(f)$ on $L^2(G)$, by radialization, as follows.

THEOREM 12.4 (Estimating convolution norms by radialization [39]). *Let* G *be a* (*non-compact*) *semisimple algebraic group over a locally compact non-discrete field. Then the following holds for every measurable function* F

$$\left\|\lambda_G(F)\right\| \leqslant \int_G \left(\int_K \int_K \left|F(kgk')\right|^2 dk\, dk'\right)^{1/2} \Xi(g)\, dg,$$

where $\Xi(g)$ *is the Harish Chandra* Ξ *function, namely the fundamental positive-definite positive spherical function on* G.

REMARK 12.5.

(1) For $SL_2(\mathbb{C})$, $\Xi(a_t) = \varphi_0(a_t) = \frac{t}{\sinh t}$, and in particular, it decays exponentially in the distance t in hyperbolic space.

(2) In general, the Harish Chandra Ξ-function on connected semisimple Lie groups has the same behavior: it decays exponentially in the distance on the symmetric space G/K. More precisely $\Xi(g) \leqslant C \exp(-c|g|)$ $(c > 0)$, where $|g| = d(gK, K)$, d the invariant distance on G/K derived from the Riemannian structure given by the Killing form. The same holds true for general semisimple algebraic groups, where instead of the Riemannian distance on the symmetric space we consider the natural distance on the Bruhat–Tits building.

The estimate of Theorem 12.4 yields an estimate of the convolution norm $\lambda_G(f)$ of a non-radial function in terms of an associated radial function, which is much easier to control. In particular, together with Theorem 12.3 it gives a simple and explicit integral criterion for a family $f_t \in L^1(G)$, $t \in \mathbb{R}_+$, to satisfy a strong exponential-maximal inequality in every $L^p(X)$, $1 < p \leqslant \infty$.

A particularly interesting example to consider is that of horospherical averages, defined as follows.

Let $G = KAN$ be an Iwasawa decomposition of a connected semisimple Lie group with finite center. Recall from Section 10.4 that Haar measure m_G can be normalized so that in horospherical coordinates it is given by (see [54] or [70, Chapter I, Proposition 5.1])

$$\int_G f(g)\,dg = \int_{K\times A\times N} f(ke^H n)e^{2\rho(H)}\,dk\,dH\,dn,$$

where ρ is half the sum of the positive roots. Now let h_t denote the absolutely continuous probability measure on G whose density is given by $\chi_{U_t}/m_G(U_t)$, where $U_t = \{ke^H n \mid k \in K,\ \|H\| \leqslant t,\ n \in N_0\}$. Here N_0 is a fixed compact neighborhood of the identity in N. Applying Theorem 12.4 and Theorem 12.3, we obtain the following sample result.

THEOREM 12.6 (Pointwise ergodic theorem with exponentially fast rate of convergence for horospherical averages in action with a spectral gap [108]). *Let G be a connected semisimple Lie group with finite center, h_t the horospherical averages. Let (X, m) be a probability-preserving action whose unitary representation π_0 in $L^2_0(X)$ has a spectral gap. Then*

(1) $\|\lambda_G(h_t)\| \leqslant C\exp(-\theta t)$, $\theta > 0$.

(2) *If $\pi_0^{\otimes n} \subset \infty \cdot \lambda_G$, then for every $f \in L^p(X)$, $1 < p < \infty$, and almost every $x \in X$*

$$\left|\pi(h_t)f(x) - \int_X f\,dm\right| \leqslant B_p(x, f)\exp\left(-\frac{\theta_p}{2n}t\right),$$

where $\theta_p > 0$.

REMARK 12.7.

(1) Again h_t actually satisfies a more precise estimate, namely an exponential (L^p, L^r)-maximal inequality (see [108] for details).

(2) Theorem 12.6 can be established also for other averages defined by horospherical coordinates. As an example, consider the case of a real-rank-one group. Then the group $A = \{a_t;\ t \in \mathbb{R}\}$ figuring in the Iwasawa decomposition is one-dimensional, and let I_1 denote the unit interval in $\mathbb{R}$. Then $t + I_1$ is a unit interval with center t, and let $J_t = \{ka_s n \mid k \in K,\ s \in t + I_1,\ n \in N_0\}$. Then the Haar-uniform averages j_t on J_t satisfy the conclusion Theorem 12.6.

(3) We note that it is interesting to compare the horospherical averages h_t and j_t to the averages constructed in the proof of Theorem 7.13. Note that here the pointwise ergodic theorem is strengthened to yield an explicit exponentially fast rate of convergence to the ergodic mean.

(4) Another interesting comparison is between j_t and the averages $m_K * \delta_{a_t}$ appearing in the mean ergodic theorem of W. Veech [150].
(5) Similar results can of course be established for semisimple algebraic groups.

12.3. *Averages on discrete subgroups*

We have commented throughout our discussion on the problem of comparing between the ball averages on an lcsc group, and the discrete ball averages on its lattice subgroups. In the case homogeneous of nilpotent groups such as of $\mathbb{Z}^d \subset \mathbb{R}^d$ and $H_n(\mathbb{Z}) \subset H_n$, the comparison was completely straightforward: proofs of maximal inequalities on the Lie group can be easily adapted to prove maximal inequalities on the discrete lattice, and then the transfer principle implies that they hold in *every* measure-preserving action (see Section 5 for the details).

In Sections 10.5 and 11.3 we have encountered a select group of extra symmetric lattices in some simple algebraic groups over local fields, and other totally disconnected Gelfand pairs (G, K). In such a lattice Γ, the group algebra $\ell^1(\Gamma)$ contains an isomorphic copy of the commutative convolution algebra $M(G, K)$, and most of the basic problems in spectral theory and ergodic theory pertaining to these averages on Γ are reducible to the corresponding problems for $M(G, K)$. This gives a satisfactory solution to the basic questions in spectral and ergodic theory for the corresponding averages on Γ, provided that the representation theory of the commutative algebra $M(G, K)$ is sufficiently well understood.

However, in the set-up of general lattices Γ in non-amenable Gelfand pairs, and even semisimple algebraic groups, there is usually no direct connection between $M(G, K)$ and $\ell^1(\Gamma)$, and the basic problems in ergodic theory cannot be resolved using this method. Note that the absence of a transfer principle implies that a result on a maximal inequality for convolutions on the discrete lattice has no bearing on the case of a general action. Furthermore, the natural discrete ball averages (w.r.t. a word metric) on the lattice do not commute in general, and so there is no natural commutative algebra whose spectral theory can be used to establish even a mean ergodic theorem for the averages. Thus establishing a pointwise, or even mean, ergodic theorem for the discrete uniform ball averages in *arbitrary* measure preserving action of Γ is a challenging goal, outside the short (but interesting) list of examples in Sections 10.5 and 11.3. Nevertheless, it is possible to make considerable progress on this problem, at least for certain natural averages on Γ, although these are usually not comparable to balls w.r.t. a word metric. These very recent results are based on three principle, namely induction of actions from the lattice Γ to the group G, a duality principle which controls discrete averages on the lattice Γ by certain (non-radial) absolutely continuous averages on the group G, and ergodic theorems for sufficiently general non-radial averages on the group G. These results will be reported in [56].

For a general lattice of a semisimple Lie group, we will content ourselves here with the following partial result, which applies the spectral transfer principle, and establishes a pointwise ergodic theorem for the lattice in actions of the group. It exemplifies the natural procedure of estimating discrete convolution operators in terms of absolutely continuous ones, by reducing the problem to geometric comparison for the translation action on G, and thus obtaining an estimate of convolution norm.

More precisely, let $\beta_k = \frac{1}{vol\, B_k}\chi_{B_k}(g)$ (where B_k is the bi-K-invariant lift of a ball of radius k in G/K). Let Γ be a lattice subgroup of G, let $B_k(\Gamma) = B_k \cap \Gamma$, and let $b_k = \frac{1}{|B_k(\Gamma)|}\sum_{\gamma \in B_k(\Gamma)} \gamma$. We have:

THEOREM 12.8 (Pointwise ergodic theorems for lattice averages in G-actions with a spectral gap [108]). *Let $\Gamma \subset G$ be a lattice in a connected semisimple Lie group G with finite center. Then the sequence $b_k \in \ell^1(\Gamma)$ satisfies*:

(1) $\|\lambda_\Gamma(b_k)\| \leqslant C\exp(-\theta k)$. *Here*

$$0 < \theta < \theta_\beta(G) = \lim_{t\to\infty} -\frac{1}{t}\log\|\lambda_G(\beta_t)\|.$$

(2) *In any Γ-action satisfying $\pi_0^{\otimes n} \subset \infty \cdot \lambda_\Gamma$, and in particular in any action of G with a spectral gap, for any $f \in L^p(X)$, $1 < p < \infty$, and for almost every $x \in X$*

$$\left| b_k f(x) - \int_X f\, dm \right| \leqslant C(x, f)\exp\left(-\frac{\theta}{2n}k\right).$$

Clearly, the spectral estimate for convolutions stated in (1), together with norm estimate provided by the spectral transfer principle (see Theorem 11.8) implies the exponentially fast pointwise convergence in (2). Such spectral estimates can be established for many other discrete groups, and not only for lattices. In fact, the unitary representation of Γ in $L^2(G)$ is equivalent with countably many copies of the unitary representation of Γ in $\ell^2(\Gamma)$. Thus any sequence of measures $\nu_n \in \ell^1(\Gamma)$ with $\|\lambda_\Gamma(\nu_n)\|_{\ell^2(\Gamma)} \leqslant \exp(-\theta n)$ will satisfy the conclusion of Theorem 12.8.

The challenge of establishing spectral norm estimates for convolution operators on discrete groups has attracted considerable attention, and one particularly interesting approach was to establish the even stronger property of rapid decay [67,79], one of whose formulations is as follows. Assume d is a word metric, and let $L(\gamma) = d(e, \gamma)$. Then for some $s = s(\Gamma, d) \geqslant 0$, some $C = C(\Gamma, d) > 0$, and every finitely supported function $f \in \ell^1(\Gamma)$,

$$\|\lambda_\Gamma(f)\| \leqslant C\|f \cdot (1 + L)^s\|_{\ell^2(\Gamma)}.$$

From this estimate it follows that if in addition the spheres (or balls) have strict $t^q \exp ct$-exponential growth (with $c > 0$) then the spectral norm decays exponentially fast. We note that strict $t^q \exp ct$-volume growth has been established in the case of word metrics on word-hyperbolic groups in [31], so that every discrete hyperbolic subgroup of a connected semisimple Lie group satisfies the conclusions on Theorem 12.8, w.r.t. balls defined by a word metric on it (see [108, Corollary 3.3]).

The rapid decay property has been established in a number of interesting cases, including all word-hyperbolic groups [79], uniform lattices in $SL_3(\mathbb{R})$ and $SL_3(\mathbb{C})$ [91], certain lattices acting on rank-two Bruhat–Tits building [128], and groups acting properly on cube complexes [27]. However strict $t^q \exp ct$-volume growth has not been established in general for uniform lattices or cube complex groups, and constitutes a completely open problem for general discrete (sub)groups.

13. Weighted averages on discrete groups and Markov operators

In this section we will consider briefly some applications of the general theory of Markov operators to ergodic theorems for group actions. We divide the section into three parts, dealing with applications of ergodic and maximal theorems for the following sequences:

(1) uniform averages of powers of a single Markov operator,
(2) subadditive sequences of self-adjoint Markov operators,
(3) powers of a single self-adjoint Markov operator.

13.1. *Uniform averages of powers of a Markov operator*

In our discussion so far we have put a great deal of emphasis on proving ergodic theorems for Haar-uniform averages on geometrically significant sets. A different approach, which has long roots in ergodic theory and the theory of Markov processes is to relax the requirement of uniform averages and allow weighted ergodic theorems. Thus we can consider for a discrete group a sequences of averages of the form $\sum_{\gamma\in\Gamma}\nu(\gamma)\delta_\gamma$, where ν is a general probability measure on Γ. One ergodic theorem that can be obtained here is for the sequence of averages $\nu_n=\sum_{k=0}^n \nu^{*n}$, namely the uniform average of convolution powers of a single measure. It follows immediately from the general theory of non-negative contractions developed by Hopf and Dunford and Schwartz (see, e.g., [44] for an extensive exposition) that the sequence $\pi(\nu_n)$ satisfies the weak-type $(1,1)$ maximal inequality in L^1, and converges pointwise to a $\pi(\nu)$-invariant function, in every probability-preserving action of Γ. In particular, if the support of ν generates γ as a group, then the limit is a Γ-invariant function, namely $\int_X f\,dm$ when Γ acts ergodically.

The ergodic theorem for the uniform averages of powers of a Markov operator can be used to obtain weighted ergodic theorems for actions of several transformations also in another manner, namely using the construction of skew product actions. This useful idea has been introduced already by S. Kakutani in his proof of the random ergodic theorem [81] for almost all sequences of transformations chosen independently (or according to a Markov measure) from a finite set (say). Thus for a probability-preserving action of FS_r, the free semigroup on r non-commuting elements, one can form the skew product $\Omega_r\times X$, where Ω_r is the topological Markov chain consisting of infinite words in the free generators of the free semigroup. Consider the transformation $T(\omega,x)=(S\omega,\omega_1 x)$ where $S:\Omega_r\to\Omega_r$ is the forward shift, and ω_1 the first letter of ω. One takes the natural probability measure p on Ω_r associated with uniform weights on the generators S (or any other Markov measure on Ω_r), and the measure $p\times m$ on $\Omega_r\times X$. Ergodicity of T and the FS_r-action on X are equivalent [81], and thus the uniform averages of powers of the operator T satisfy the pointwise ergodic theorem, by the Hopf–Dunford–Schwartz theorem, so that

$$\lim_{n\to\infty}\frac{1}{n+1}\sum_{k=0}^{n}F\bigl(S^k\omega,\omega_k\omega_{k-1}\cdots\omega_1 x\bigr)=\int_{\Omega_r\times X}F(\omega,x)\,dp\,dm.$$

Now apply the foregoing result to the functions $F(\omega,x)=f(x)$ on the skew product $\Omega_r\times X$ which are lifted from X, and take expectations w.r.t. the probability p

on Ω_r. Then one obtains a pointwise ergodic theorem in $L^1(X)$ for the uniform averages $\mu_n = \frac{1}{n}\sum_{k=0}^{n}\sigma_k$ of the sphere averages on the free semigroup (or more general weights determined by an arbitrary Markov measure on Ω_r).

Note however that the passage to expected values implies that we can not conclude that the weak-type $(1,1)$-maximal inequality in valid for f_μ^*.

This application of Kakutani's random ergodic theorem in the context of ergodic theorems for free groups is due to R. Grigorchuk [61,62]. The idea of using skew products and the theory of Markov operators is developed further in [62] and [16,17]. Indeed, the method is suitable for operator averages for Banach space representations which do not arise from measure-preserving action [61,16], requires the measure m on X to be merely stationary, and not necessarily invariant [62], and applies also when the weights taken along the group orbit depend on the starting point [62]. However, in general these results establish weighted ergodic theorems, which bear no discernible relation to the uniform averages of spheres w.r.t. a word metric. The only exception is in the case where the Markov measure is associated with all the generators in S having equal probability. We remark that the same analysis applies also to the free group, and not only the free semigroup.

We note that the maximal and pointwise ergodic theorem in L^2 for the uniform averages of the spheres averages was proved in [105], and for $f \in L^1$ in [114]. The proof in [114] is also probabilistic and quite elementary, and does not require spectral theory. It uses the standard estimates of the central limit theorem for convolution powers of a binomial distribution on $\mathbb{N}$, and Hopf's maximal inequality (see a particularly simple proof by Garcia of the latter in [53]). Indeed, it is shown that the sum of convolution powers of σ_1 dominates the uniform average of spheres, or more precisely:

$$\mu_n = \frac{1}{n}\sum_{k=0}^{n}\sigma_k \leqslant \frac{C}{3n+1}\sum_{k=0}^{3n}\sigma_1^{*k}.$$

Thus it follows immediately that the maximal function f_μ^* satisfies the weak-type $(1,1)$-maximal inequality, by the maximal inequality for the average of powers. Furthermore, it is clear that on functions of the form $h = f - \pi(\sigma_1)f$ the sequence $\pi(\mu_n)h$ converges pointwise, and this implies the pointwise ergodic theorem as usual by the recipe of Section 2.4.

We record the facts described above as follows:

THEOREM 13.1. *The sequence $\mu_n = \frac{1}{n+1}\sum_{k=0}^{n}\sigma_k$ of uniform averages of spheres on the free group satisfied the weak-type $(1,1)$-maximal inequality in L^1 and is a pointwise ergodic sequence in L^p, for all $1 \leqslant p < \infty$.*

Thus it appears that the sequence of uniform averages of the sphere averages $\mu_n = \frac{1}{n}\sum_{k=0}^{n}\sigma_k$ is a natural sequence of weighted averages to consider, and we may inquire for which discrete groups Theorem 13.1 is satisfied for a word metric. We next turn to consider a method which establishes at least the maximal inequality in L^2 for μ_n, for a certain class of groups. This method was first employed in [105] to prove the maximal and pointwise ergodic theorem in L^2 for the averages μ_n on $\mathbb{F}_k$. It also does not use spectral considerations, and is based on a general subadditive maximal inequality, which we consider below.

13.2. *Subadditive sequences of Markov operators, and maximal inequalities on hyperbolic groups*

Let us introduce the following definition.

DEFINITION 13.2 (*Subadditive sequences*, see [105]). A sequence T_n of operators on $L^2(X)$ will be called a *subadditive sequence of self-adjoint Markov operators* if it satisfies the following:

(1) $T_n = T_n^*$, $\|T_n\| \leqslant 1$.
(2) $T_n f \geqslant 0$ if $f \geqslant 0$, $T_n 1 = 1$.
(3) There exist a constant $C_0 > 0$, a positive integer k, and a fixed non-negative bounded operator B on $L^2(X)$ such that:

$$T_n T_m f(x) \leqslant C_0\big(T_{kn} f(x) + T_{km} f(x)\big) + Bf(x)$$

for all bounded and nonnegative $f \in L^2$.

We can now state the following subadditive maximal inequality, proved independently in [5] and [105].

THEOREM 13.3 (Subadditive maximal inequality [5,105]). *Let T_n be a subadditive sequence of self adjoint Markov operators. Define $f^*(x) = \sup_{n\geqslant 0} |T_n f(x)|$. Then $\|f^*\|_2 \leqslant C\|f\|_2$ for all $f \in L^2$. We can take $C = 2C_0 + \|B\|$.*

We note that the subadditive maximal inequality of Theorem 1 generalizes similar results due to E.M. Stein [139,140] and B. Weiss [153]. In particular, it was applied in [139,140] to prove a pointwise ergodic theorem for the even powers of positivity-preserving self-adjoint contractions on L^2. Also, it is noted in these references that it implies the pointwise convergence of martingales in L^2, as well as Birkhoff's pointwise ergodic theorem in L^2. The origin of this maximal inequality is attributed in [140,153] to A. Kolmogoroff and G. Seliverstoff [87], and to R.E.A.C. Paley [122].

It is reasonable to expect that for a large class of discrete groups Γ, the sequence of averages $\mu_n \in \ell^1(\Gamma)$ will satisfy the subadditive inequality given by $\mu_n * \mu_m \leqslant C(\mu_{kn} + \mu_{km}) + b$. When this inequality holds in $\ell^1(\Gamma)$, then $T_n = \pi(\mu_n)$ is a subadditive sequence of self-adjoint Markov operators on L^2 in any finite-measure-preserving action, and therefore will satisfy the strong maximal inequality in L^2.

This is indeed the case at least in the following context.

THEOREM 13.4 (Subadditive inequality for word-hyperbolic groups [51]). *Let Γ be a non-elementary word-hyperbolic group, S a finite symmetric generating set. Then there exist constants $1 < q < \infty$ and $0 < C < \infty$, depending only on (Γ, S), such that the following inequalities hold*:

(1) $\sigma_t * \sigma_s \leqslant C \sum_{j=0}^{2s} q^{-(s-\frac{1}{2}j)} \sigma_{t-s+j}$ *if* $t \geqslant s$.
(2) $\mu_n * \mu_m \leqslant C(\mu_{2n} + \mu_{2m})$.

Combining the foregoing results, we have

THEOREM 13.5 (Maximal inequality for word-hyperbolic groups [51]). *Let Γ be a word-hyperbolic group, S a finite symmetric generating set. Then the sequence μ_n satisfies the strong maximal inequality in $L^2(X)$, i.e. $\|f_\mu^*\|_2 \leqslant C(\Gamma, S)\|f\|_2$ for every $f \in L^2(X)$.*

We remark that an elementary word-hyperbolic group is a finite extension of $\mathbb{Z}$, and the subadditivity of μ_n in this case (for free generators) is of course easily verified (see [140] for the case of $\mathbb{N}$).

As usual, given the maximal inequality for μ_n in L^2, to complete the proof of the pointwise ergodic theorem it suffices (see Section 2.4) to find a dense set of functions $f \in L^2$ where $\pi(\mu_n)f(x) \to \int_X f\,dm$, almost everywhere and in the L^2-norm. For hyperbolic groups, a sufficient condition is the existence of a dense set of functions $f \in L^2$ satisfying the following exponential mixing condition: $|\langle \pi(\gamma)f, f\rangle| \leqslant C_f \exp(-c_f|\gamma|)$, for some $c_f > 0$ (see [51]). However, in general these issues are far from being resolved, and we thus formulate the following.

PROBLEM 13.6 (*Analogs of von Neumann and Birkhoff theorems for word-hyperbolic groups*).

(1) Is the mean ergodic theorem valid for the averages μ_n on every word-hyperbolic group? Is the pointwise ergodic theorem valid for μ_n in L^2, or even L^1?
(2) Is there any finitely generated group for which the subadditive convolution inequality fails for the averages μ_n?

13.3. *The powers of a self-adjoint Markov operators*

For a *self-adjoint* Markov operator a result much sharper than the Hopf–Dunford–Schwartz ergodic theorem was subsequently proved independently by E. Stein [104] and J.C. Rota [133], using two entirely different methods. In both cases, the results proved imply as a special case that when the measure ν is symmetric, then not only does the sequence of uniform averages $\pi(\mu_n) = 1/(n+1)\sum_{k=0}^n \pi(\nu_k)$ converge, but already the sequence of powers $\pi(\nu^{*2n})$ converge, pointwise almost everywhere, and in the L^p-norm, at least for $1 < p < \infty$. The limit is a $\pi(\nu^{*2})$-invariant function, and thus invariant under the group generated by the support of ν^{*2}. We note that the passage to even powers reflect the same periodicity phenomenon that was encountered in Section 10.5 with regard to the free groups. Namely it reflects the fact that $\pi(\nu)$ may have an eigenfunction f_0 with eigenvalue (-1), or equivalently that the commutative convolution algebra generated by the powers of ν may have the an eigenfunction $f_0 \in L^2(X)$ realizing the sign character of the algebra $\nu^{*n} \mapsto (-1)^n$.

Both results mentioned above apply to general self-adjoint Markov operators P and not only to convolution powers, but in this generality, it was established by D. Ornstein [120] that pointwise convergence usually fails to hold in L^1 for the powers of P^2.

Stein's method is spectral, and can be viewed as a special case of the general method described in Section 10.2, where the commutative algebra is taken to be simply the con-

volution algebra $\ell^1(\mathbb{N})$, corresponding to the algebra generated by powers of a single operator. The continuous $*$-representation are given here simply by self-adjoint contraction operators, and the $*$-characters are given by $\varphi_\lambda(k) = \lambda^k$, where $\lambda \in [-1, 1]$. This point of view was developed in [105], where Stein's method was generalized to apply, e.g., to the radial algebra on free groups.

Rota's method depends on probabilistic considerations, and reduces the pointwise convergence theorem for the even powers of a self-adjoint Markov operator to the pointwise convergence theorem for martingales, on an auxiliary probability space constructed from the Markov operator in question. In was shown in [133] that in fact pointwise convergence holds for $f \in L(\log L)(X)$.

In [18] A. Bufetov has constructed, for a given action of the free group $\mathbb{F}_r$ on a space (X, m), and a given free generating set S, a Markov operator P on the set $X \times S$, which is not self-adjoint, but such that the operators $(P^*)^n P^n$ are comparable to the action of the sphere averages on $\mathbb{F}_r$ on the space X. Rota's theorem applies to such sequence and the pointwise convergence of $\sigma_{2n} f$ to a function invariant under words of even length follows, in $L(\log L)(X)$.

14. Further developments

14.1. *Some non-Euclidean phenomena in higher-rank groups*

In the present section we return to the set-up of connected semisimple Lie groups, and we would like to demonstrate the fact that exponential volume growth on semisimple Lie groups (and more general lcsc groups) can be used to derive a number of non-standard maximal inequalities and pointwise convergence results which have no Euclidean analog. This phenomenon occurs for actions of semisimple algebraic groups of split rank at least two, or for direct products of lcsc groups. To illustrate some of these results, let us consider the simplest case where $G = L_1 \times L_2$ is a product of two real-rank one groups, say $L_1 = L_2 \cong PSL_2(\mathbb{R})$ for definiteness. We attempt to prove maximal inequalities for the product group G based on maximal inequalities of its factors L_i. The ball averages on the factor groups L_i are given in this case by (choosing the curvature on $\mathbb{H}^2$ appropriately)

$$\beta_t = \frac{\int_0^t \sinh s \sigma_s \, ds}{\int_0^t \sinh s \, ds}.$$

Let us define $D_t = \{(u, v);\ u^2 + v^2 \leqslant t^2\}$, and let $\gamma_t^{(i)}$ (resp. $\sigma_t^{(i)}$) be the unit shell (resp. sphere) averages on the real-rank one group L_i. Then the ball averages β_t^G on the product group $G = L_1 \cdot L_2$ satisfy the following estimate:

$$\beta_t^G = \frac{\int_{D_t} |\sinh u \sinh v| \sigma_{|u|}^{(1)} \sigma_{|v|}^{(2)} \, du \, dv}{\int_{D_t} |\sinh u \sinh v| \, du \, dv} \leqslant B \frac{\sum_{D_t \cap \mathbb{N}^2} \exp n \exp m \gamma_n^{(1)} \gamma_m^{(2)}}{\sum_{D_t \cap \mathbb{N}^2} \exp n \exp m}.$$

On the left-hand side we have the integral w.r.t. the G-invariant Riemannian volume, which in geodesic polar coordinates is given by the integral w.r.t. the density $|\sinh u \sinh v|$

on a disc of radius t in $\mathbb{R}^2$. Using the Weyl group symmetry, this integral is equal to four times its value on $D_t \cap \mathbb{R}^2_+$. On the right-hand side the integral was then further replaced by the weighted sum of integrals (w.r.t. the uniform Euclidean measure) on unit squares whose lower left-hand corner is a lattice point of norm at most t. The weight attached to each unit square is the obvious one: we estimate the function $\sinh u \sinh v$ on the square whose lower left-hand corner is (n, m) by $C \exp n \exp m$. Since the shell averages on L_i satisfy the strong maximal inequality in L^p, $1 < p \leqslant \infty$, by Theorem 9.6, it follows that the same holds for the operators $\beta_t^G = \beta^{L_1 \times L_2}$.

The support of $\gamma_t^{(1)} \gamma_s^{(2)}$, $(t, s) \in \mathbb{R}^2_+$, has radial coordinates in $\mathfrak{a}_+ \cong \mathbb{R}^2_+$ which constitute (the W-orbit of) a square of unit side length, whose lower left hand corner is the element (tH_1, sH_2). But now note that already the family of unit square averages $\gamma_t^{(1)} \gamma_s^{(2)}$ (for arbitrary non-negative s and t) satisfy a strong maximal inequality in L^p, $p > 1$. Indeed, each square average as above is the product of two one-dimensional shell (or "interval", in this context) averages on real rank one groups, and Theorem 9.5 applies to the shell averages on L_1 and L_2. As a result, the strong maximal inequalities which hold for the unit square averages in an arbitrary measure-preserving action of G, also hold for an *arbitrary family of sets which admit a reasonable covering by unit squares*, not only for the discs described above (which correspond to the radial coordinates of ball averages w.r.t. the Riemannian Killing metric).

It is possible to greatly expand the scope and generality of results of this kind, and apply them to all higher-rank semisimple groups. In particular they yield a pointwise ergodic theorem for uniform averages supported on an arbitrary sequence of bi-K-invariant sets which are reasonably covered by "cube averages" of a fixed size, provided they leave eventually any given compact set. In particular these arguments apply to a wide array of bi-K-invariant averages which occupy an exponentially decaying fraction of the volume of a ball. This improves the results we previously described for the shells, which occupy a *fixed* proportion of the volume of the ball. To illustrate these points, let us define the following simple geometric property of sets in the vector space $\mathfrak{a}$.

DEFINITION 14.1 (*Condition* $A_{(c,C)}$). A measurable set $E \subset \mathfrak{a}$ (of finite measure) satisfies condition $A_{(c,C)}$ if for every $H \in E$, there exists $H' \in E$, satisfying the conditions $\|H - H'\| \leqslant C$, and $b_c(H') \subset E$. Here $b_c(H')$ is a ball of radius c and center H' in $\mathfrak{a}$.

Clearly, any union of sets satisfying condition $A_{(c,C)}$ also satisfies it. Equally clearly, all balls of radius at least a fixed constant satisfy condition $A_{(c,C)}$ for some (c, C). The same holds for Euclidean cells (= cubes) of fixed side length. Thus any union of such sets (with fixed (c, C)) also satisfies the condition.

Now view $\mathfrak{a}$ as the Lie algebra of a split Cartan subgroup of a connected semisimple Lie group with finite center. For a Weyl group invariant set $E \subset \mathfrak{a}$ let $K \exp(E) K = R(E)$ be the radialization of E, and let ν_E be the normalized average on $R(E)$, namely the normalized restriction of Haar measure to this set. We can now formulate the following

THEOREM 14.2 (Pointwise ergodic theorem for general sequences of bi-K-invariant averages [113]). *Let G be a connected semisimple Lie group with finite center. Then*

(1) *The maximal function*

$$\mathcal{A}^* f(x) = \sup\{|\pi(v_E) f(x)|;\ E \subset \mathfrak{a},\ \text{and } E \text{ satisfies condition } A_{(c,C)}\}$$

satisfies the strong maximal inequality in every L^p, $1 < p < \infty$.

(2) *Let* E_n *be any sequence of measurable sets satisfying condition* $A_{(c,C)}$. *Assume that* $\lim_{n\to\infty} vol_G(E_n \cap b_t)/vol_G(E_n) = 0$, *for every* $t > 0$, *where* b_t *is a ball of radius* t *in* $\mathfrak{a}$ *with center* 0. *Assuming* G *simple, the averages* v_{E_n} *satisfy the pointwise ergodic theorem in* L^p, $1 < p < \infty$. *For* G *semisimple, the same conclusion holds, provided we assume the previous condition also for the projection of* E_n *to any factor group*.

The phenomenon described in Theorem 14.2 allows for a choice of very general sequences, and does not seem to have any Euclidean analog. We refer to [113] for further details.

14.2. *Best possible rate of convergence in the pointwise theorem*

The pointwise ergodic theorems with exponentially fast speed of convergence for ball averages can be considerably sharpened, and the best possible rate of convergence can be determined. Furthermore, exponentially fast pointwise convergence can be proved even for the sphere averages on real-rank one groups (in the range of p where pointwise convergence actually holds for L^p functions!). This requires spectral arguments which are considerably more elaborate than those we have presented here, and the details can be found in [110].

14.3. *Added in proof*

We note the following very recent developments that have taken place since the final version of the present survey was submitted.

14.3.1. *Exact polynomial volume growth.* E. Breuillard has completed the proof of the following remarkable result.

THEOREM 14.3. *On every lcsc group of polynomial volume growth, the balls w.r.t. any word metric have exact polynomial volume growth. In fact, the same holds true for every invariant asymptotically geodesic pseudo-metric on the group*.

This result provides an alternative proof of the general case of the localization conjecture, namely the fact that the balls are asymptotically invariant under translations, from which the pointwise ergodic theorem in L^1 follows—see the discussion in Sections 4 and 5. We note that in the case of connected Lie groups, precise results are obtained by Breuillard regarding the actual geometric shape of the balls, and not just their volume asymptotics. The passage from the Lie case to the general case uses the results of Guivarc'h [64] and

Jenkins [79] on growth, and of Gromov [63] and Losert [93] on the structure of groups of polynomial volume growth. Thus the theorem considerably sharpens Proposition 5.11 of Section 5.5, where we have used these results to deduce strict polynomial growth. Details can be found in the preprint "The asymptotic shape of metric balls in groups of polynomial growth" by E. Breuillard.

14.3.2. *Exact exponential volume growth.* Given a linear representation $\tau : G \to GL_n(\mathbb{R})$ of a semisimple group G, and a vector space norm on $M_n(\mathbb{R})$, one can consider (say) the distance function $\log \|g\|$. It was noted in Theorem 4.7 of Section 4.4 that in [57] it was shown that the balls of radius t w.r.t. such a distance function have exact $t^q \exp(ct)$ growth. F. Maucourant has developed an alternative approach to this result which yields that in fact after scaling by the volume growth one obtains w^* convergence to a limiting Radon measure on $M_n(\mathbb{R})$. Details can be found in the preprint "Homogeneous asymptotic limits of Haar measure of semisimple linear groups and their lattices", by F. Maucourant.

14.3.3. *The ergodic theory of lattice subgroups.* It is possible to develop general pointwise ergodic theorems for *arbitrary actions* of a lattice Γ in a semisimple group G. Such theorems apply to the discrete averages on lattice points in balls w.r.t. a distance functions on the group G. We note here a sample result formulated in the simplest case, and refer to [56] for a full discussion.

Let $\beta_k = \frac{1}{\mathrm{vol}\, B_k} \chi_{B_k}(g)$, where B_k is the bi-K-invariant lift of a ball of radius k in G/K. Let Γ be a lattice subgroup of G, let $B_k(\Gamma) = B_k \cap \Gamma$, and let $b_k = \frac{1}{|B_k(\Gamma)|} \sum_{\gamma \in B_k(\Gamma)} \gamma$. We have:

THEOREM 14.4 (Pointwise ergodic theorem for general lattice actions [56]). *Let $\Gamma \subset G$ be a lattice in a connected semisimple Lie group G with finite center. Then the sequence $b_k \in \ell^1(\Gamma)$ satisfies, in every ergodic probability preserving action of Γ, for any $f \in L^p(X)$, $1 < p < \infty$, and for almost every $x \in X$:*

(1) $\lim_{k\to\infty} b_k f(x) = \int_X f\, dm$.

(2) *In any ergodic action of Γ with a spectral gap (and thus in every ergodic action if Γ has property T), there exists $\delta_p = \delta_p(X) > 0$, such that for any $f \in L^p(X)$, $1 < p < \infty$, and for almost every $x \in X$*

$$\left| b_k f(x) - \int_X f\, dm \right| \leqslant C_p(x, f) \exp(-\delta_p k).$$

If Γ has property T, then δ_p depends only on Γ, and not on X.

In reference to Theorem 14.4, note that in Theorem 12.8 a similar conclusion is asserted, but only for those actions of Γ arising from actions of G. However the rate of exponential convergence obtained in Theorem 12.8 is in general faster. Also note that in Theorem 11.10 the best possible rate of exponential convergence is obtained for balls w.r.t. a *word metric* on the lattice, whereas b_k are not associated with a word metric. However in Theorem 11.10 the lattices (and the metrics) are severely restricted.

We note that the same result holds for balls (and many of their subsets) w.r.t. a large class of distance functions on the group, and allows the calculation of the main term as well as an estimate of an error term in many lattice point counting problems. We refer to [56] for details.

14.3.4. *The ball averaging problem for word metrics on semisimple groups.* As noted in Theorem 4.11, word metrics on semisimple groups are coarsely isometric to norm-like metrics. It is possible to utilize this fact and obtain a solution of the ball averaging problem in this context. We formulate for simplicity the following basic special case.

THEOREM 14.5 (Pointwise ergodic theorem for word metric balls on simple Lie groups). *The balls defined by any word metric on a connected simple Lie group with finite center satisfy the pointwise ergodic theorem in* L^p, $1 < p < \infty$.

In fact, a similar result holds for all algebraically connected semisimple algebraic groups, at least in actions where G^+ acts ergodically. The convergence is exponentially fast almost surely, if the action has a spectral gap. Details will appear in the paper "On the ball averaging problem in ergodic theory", currently under preparation.

14.3.5. *Further reading.* The present survey aimed for the most part to indicate just the bare outlines of the relevant arguments appearing in the proofs of the ergodic theorems cited, and most details are of course left out. The reader wishing to learn more about some of these arguments is referred to the detailed comprehensive exposition in the forthcoming book "Théorèmes Ergodiques pour les Actions de Groupes". This book is the result of a collaborative effort initiated by the late Martine Babillot, with the participation of C. Anantharaman, J.-Ph. Anker, A. Batakis, A. Bonami, B. Demange, F. Havard, S. Grellier, Ph. Jaming, E. Lesigne, P. Maheux, J.-P. Otal, B. Schapira and J.-P. Schreiber. The book contains a wealth of information on ergodic theorems for group actions, including an elaboration of a number of the topics mentioned in the survey.

References

Surveys in volume 1A

[1] B. Hasselblatt and A. Katok (eds), *Handbook of Dynamical Systems*, Vol. 1A, Elsevier, Amsterdam (2002).

Other sources

[2] H. Abels and G.A. Margulis, *Coarsely geodesic metrics on reductive groups*, Modern Dynamical Systems and Applications, Cambridge University Press, Cambridge (2004), 163–183.

[3] S.I. Adian and J. Mennicke, *On bounded generation of* $SL_n(\mathbb{Z})$. Internat. J. Algebra Comput. **2** (1992), 357–365.

[4] V.I. Arnold and A.L. Krylov, *Uniform distribution of points on a sphere and some ergodic properties of solutions of linear ordinary differential equations in the complex plane*, Soviet Math. Dokl. **4** (1962), 1–5.

[5] J. Barrionuevo, *Estimates for some Kakeya-type maximal operators*, Trans. Amer. Math. Soc. **335** (1993), 667–682.

[6] H. Bass, *The degree of polynomial growth of finitely generated groups*, Proc. London Math. Soc. **25** (1972), 603–614.
[7] C. Benson, J. Jenkins and G. Ratcliff, *On Gelfand pairs associated with solvable Lie groups*, Trans. Amer. Math. Soc. **321** (1990), 85–116.
[8] C. Benson, J. Jenkins and G. Ratcliff, *The orbit method and Gelfand pairs associated with nilpotent groups*, J. Geom. Anal. **9** (1999), 569–582.
[9] I.N. Berenstein, *All reductive p-adic groups are tame*, Funct. Anal. Appl. **8** (1974), 91–93.
[10] T. Bewley, *Sur l'application des théorèmes ergodiques aux groupes libres de tranformations: un contre exemple*, C. R. Acad. Sci. Paris, Sér. A **270** (1970), 1533–1534.
[11] T. Bewley, *Extension of the Birkhoff and von Neumann ergodic theorems to semigroup actions*. Ann. Inst. H. Poincaré **VII** (1971), 283–291.
[12] G.D. Birkhoff, *Proof of the ergodic theorem*, Proc. Nat. Acad. Sci. U.S.A. **17** (1931), 656–660.
[13] A. Borel and N. Wallach, *Continuous Cohomology, Discrete Subgroups and Representations of Reductive Groups*, Ann. of Math. Stud., Vol. 94, Princeton University Press (1980).
[14] J. Bourgain, *Averages in the plane over convex curves and maximal operators*, J. Anal. Math. **47** (1983), 69–85.
[15] E. Breuillard, *The asymptotic shape of metric balls in Lie groups of polynomial growth*, Preprint (September 2004).
[16] A.I. Bufetov, *Operator ergodic theorems for actions of free semigroups and groups*, Funct. Anal. Appl. **34** (2000), 239–251.
[17] A.I. Bufetov, *Markov Averaging and Ergodic Theorems for Several Operators*, Amer. Math. Soc. Transl., Vol. 202, Amer. Math. Soc. Publications (2001).
[18] A.I. Bufetov, *Convergence of spherical averages for actions of free groups*, Ann. of Math. **155** (2002), 929–944.
[19] A. Calderon, *A general ergodic theorem*, Ann. of Math. **57** (1953), 182–191.
[20] A. Calderon, *Ergodic theory and translation invariant operators*, Proc. Nat. Acad. Sci. U.S.A. **59** (1968), 349–353.
[21] D. Carter and G. Keller, *Bounded elementary generation of $SL_n(\mathcal{O})$*, Amer. J. Math. **105** (1983), 673–687.
[22] P. Cartier, *Representations of p-adic groups: A survey*, Proc. Sympos. Pure Math. **33** (1979), 111–155.
[23] D.I. Cartwright, A.M. Mantero, T. Steger and A. Zappa, *Groups acting simply transitively on the vertices of an $\tilde{A}_2$-building*, Geom. Dedicata **47** (1993), 143–166.
[24] D.I. Cartwright, W. Mlotkowski and T. Steger, *Property T and $\tilde{A}_2$-groups*, Ann. Inst. Fourier (Grenoble) **44** (1994), 213–248.
[25] D.I. Cartwright and T. Steger, *A family of $\tilde{A}_n$-groups*. Israel J. Math. **103** (1998), 125–140.
[26] J. Chatard, *Applications des propriétés de moyenne d'un groupe localement compact à la théorie ergodique*, Ann. Inst. H. Poincaré **VI** (1970), 307–326.
[27] I. Chatterji and K. Ruane, *Some geometric groups with rapid decay*, Geom. Funct. Anal. **15** (2005), 311–339.
[28] M. Christ and D. Müller, *On L^p-spectral multipliers for a solvable Lie group*, Geom. Funct. Anal. **6** (1996), 860–876.
[29] J.-L. Clerc and E.M. Stein, *L^p-multipliers for noncompact symmetric spaces*, Proc. Nat. Acad. Sci. U.S.A. **71** (1974), 3911–3912.
[30] R. Coifman and G. Weiss, *Transference Methods in Analysis*, CBMS Regional Conferences in Mathematics, Vol. 31, Amer. Math. Soc., Providence, RI (1976).
[31] M. Coornaert, *Mesures de Patterson–Sullivan sur le bord d'un espace hyperbolique au sens de Gromov*, Pacific J. Math., **159** (1993), 241–270.
[32] L. Corwin and F.P. Greenleaf, *Representation of Nilpotent Lie Groups and Their Applications, Part 1*, Cambridge Studies in Advanced Mathematics, Vol. 18, Cambridge University Press (1990).
[33] M. Cotlar, *A unified theory of Hilbert transform and ergodic theorems*, Rev. Mat. Cuyana **1** (1955), 105–167.
[34] M. Cowling, *The Kunze–Stein phenomenon*, Ann. Math. **107** (1978), 209–234.
[35] M. Cowling, *Sur les coefficients des représentations unitaires des groupes de Lie simples*, Analyse harmonique sur les groupes de Lie. Séminaire Nancy–Strasbourg 1975–77, Lecture Notes in Math., Vol. 739, Springer (1979), 132–178.

[36] M. Cowling, *On Littlewood–Paley–Stein theory*, Suppl. Rend. Circ. Mat. Palermo **1** (1981), 21–55.
[37] M. Cowling, *Herz's "principe de majoration" and the Kunze–Stein phenomenon*, CMS Conf. Proc., Vol. 21, Amer. Math. Soc. (1997), 73–88.
[38] M. Cowling, A. Dooley, A. Koranyi and F. Ricci, *An approach to symmetric spaces of rank one via groups of Heisenberg type*, J. Geom. Anal. **8** (1998), 199–237.
[39] M. Cowling, U. Haagerup and R. Howe, *Almost L^2 matrix coefficients*, J. Reine Angew. Math. **387** (1988), 97–110.
[40] M. Cowling and A. Nevo, *Uniform estimates for spherical functions on complex semisimple Lie groups*, Geom. Funct. Anal. **11** (2001), 900–932.
[41] S.G. Dani, *Flows on homogeneous spaces: A review*, London Math. Soc. Lecture Notes Series, Vol. 228, Cambridge University Press (1996), 63–112.
[42] P. de la Harpe, *Topics in Geometric Group Theory*, Chicago Lectures in Mathematics Series, University of Chicago Press (2000).
[43] N. Dunford, *An individual ergodic theorem for non-commutative transformations*, Acta Sci. Math. Szeged **14** (1951), 1–4.
[44] N. Dunford and J.T. Schwartz, *Linear Operators, Part I: General Theory*, 4th edn, Interscience Publishers (1967).
[45] A. El-Kohen, *Maximal operators on hyperboloids*, J. Operator Theory **3** (1980), 41–56.
[46] W.R. Emerson, *The pointwise ergodic theorem for amenable groups*, Amer. J. Math. **96** (1974), 472–487.
[47] W.R. Emerson, *Large symmetric sets in amenable groups and the individual ergodic theorem*, Amer. J. Math. **96** (1974), 242–247.
[48] A. Eskin and H. Masur, *Asymptotic formulas on flat surfaces*, Ergodic Theory Dynamical Systems **21** (2001), 443–478.
[49] A. Eskin and C. McMullen, *Mixing, counting and equidistribution in Lie groups*, Duke Math. J. **71** (1993), 143–180.
[50] N.A. Fava, *Weak-type inequalities for product operators*, Studia Math. **XLII** (1972), 271–288.
[51] K. Fujiwara and A. Nevo, *Maximal and pointwise ergodic theorems for word-hyperbolic groups*, Ergodic Theory Dynamical Systems **18** (1998), 1–16.
[52] S.A. Gaal, *Linear Analysis and Representation Theory*, Springer (1973).
[53] A. Garcia, *Topics in Almost Everywhere Convergence*, Lectures in Advanced Mathematics, Vol. 4, Markham Publishing Co., Chicago (1970).
[54] R. Gangolli and V.S. Varadarajan, *Harmonic Analysis of Spherical Functions on Real Reductive Groups*, A Series of Modern Surveys in Mathematics, Vol. 101, Springer (1988).
[55] S.V. Fomin and I.M. Gelfand, *Geodesic Flows on Manifolds of Constant Negative Curvature*, Amer. Math. Soc. Transl., Vol. 1 (1955), 49–65.
[56] A. Gorodnik and A. Nevo, *The ergodic theory of lattice subgroups*, in preparation.
[57] A. Gorodnik and B. Weiss, *Distribution of lattice orbits on homogeneous varieties*, Preprint (June 2004).
[58] F.P. Greenleaf, *Invariant Means on Topological Groups and Their Applications*, Van Nostrand Math. Studies, Vol. 16 (1969).
[59] F.P. Greenleaf, *Ergodic theorems and the construction of summing sequences in amenable locally compact groups*, Comm. Pure. Appl. Math. **26** (1973), 29–46.
[60] F.P. Greenleaf and W.R. Emerson, *Group structure and the pointwise ergodic theorem for connected amenable groups*, Adv. Math. **14** (1974), 153–172.
[61] R.I. Grigorchuk, *Ergodic theorems for the actions of a free group and a free semigroup*, Math. Notes **65** (1999), 654–657.
[62] R.I. Grigorchuk, *An ergodic theorem for action of a free semigroup*, Proc. Steklov Inst. Math. **231** (2000), 113–127.
[63] M. Gromov, *Groups of polynomial growth and expanding maps*, Pub. Math. I.H.E.S. **53** (1981), 53–78.
[64] Y. Guivarc'h, *Croissance polynomiale et periodes des fonctions harmoniques*, Bull. Soc. Math. France **101** (1973), 353–379.
[65] Y. Guivarc'h, *Sur la loi des grands nombres et la rayon spectral d'une marche alèatoire*, Astérisque **74** (1980), 47–98.
[66] Y. Guivarc'h, *Gèneralisation d'un thèoréme de von-Neumann*, C. R. Acad. Sci. Paris **268** (1969), 1020–1023.

[67] U. Haagerup, *An example of a non-nuclear C^*-algebra which has the metric approximation property*, Invent. Math. **50** (1979), 279–293.
[68] G.H. Hardy and J.E. Littlewood, *A maximal theorem with function theoretic applications*, Acta Math. **54** (1930), 81–116.
[69] W. Hebisch and A. Sikora, *A smooth subadditive homogeneous norm on a homogeneous group*, Studia Math. **96** (1990), 231–236.
[70] S. Helgason, *Groups and Geometric Analysis*, Pure and Applied Math., Vol. 113, Academic Press (1984).
[71] C. Herz, *Sur le phénomène de Kunze–Stein*, C. R. Acad. Sci. Paris **271** (1970), 491–493.
[72] C. Herz, *The theory of p-spaces with an application to convolution operators*, Trans. Amer. Math. Soc. **154** (1971), 69–82.
[73] R.E. Howe, *On a notion of rank for unitary representations of the classical groups*, Harmonic Analysis and Group Representations, C.I.M.E., 2° ciclo, Ed. A Figa Talamanca, Liguori, Naples, Italy (1980), 223–232.
[74] R.E. Howe and C.C. Moore, *Asymptotic properties of unitary representations*, J. Funct. Anal. **32** (1979), 72–96.
[75] R.E. Howe and E.C. Tan, *Non-Abelian Harmonic Analysis*, Springer (1992).
[76] A. Ionescu, *Fourier integral operators on non-compact symmetric spaces of real rank one*, J. Funct. Anal. **174** (2001), 274–300.
[77] A. Iozzi and A. Nevo, *Algebraic hulls and the Følner property*, Geom. Funct. Anal. **6** (1996), 666–688.
[78] J.W. Jenkins, *Growth of connected locally compact groups*, J. Funct. Anal. **12** (1973), 113–127.
[79] P. Jolissaint, *Rapidly decreasing functions in reduced C^*-algebras of groups*, Trans. Amer. Math. Soc. **317** (1990), 167–196.
[80] R. Jones, *Ergodic averages on spheres*, J. Anal. Math. **61** (1993), 29–45.
[81] S. Kakutani, *Random ergodic theorems and Markov processes with stable distributions*, Proc. 2nd Berkeley Sympos. Math. Stat. and Probab. (1951), 247–261.
[82] A. Katok, *Four applications of conformal equivalence to geometry and dynamics*, Ergodic Theory Dynamical Systems **8** (1998), 139–152.
[83] D.A. Kazhdan, *On a connection between the dual space of a group and the structure of its closed subgroups*, Funct. Anal. Appl. **1** (1967), 63–65.
[84] A.W. Knapp, *Representation Theory of Semisimple Groups*, Princeton Math. Series, Vol. 36, Princeton University Press (1986).
[85] G. Knieper, *Spherical means on compact Riemannian manifolds of negative curvature*, Diff. Geom. Appl. **4** (1994), 361–390.
[86] G. Knieper, *On the asymptotic geometry of non-positively curved manifolds*, Geom. Funct. Anal. **7** (1997), 755–782.
[87] A. Kolmogoroff and G. Seliverstoff, *Sur la convergence des sèries de Fourier*, C. R. Acad. Sci. Paris **178** (1924), 303–306.
[88] U. Krengel, *Ergodic Theorems*, de Gruyter Studies in Math., Vol. 6, Walter de Gruyter, Berlin (1985).
[89] R. Kunze and E. Stein, *Uniformly bounded representations and harmonic analysis of the 2×2 unimodular group*, Amer. J. Math. **82** (1960), 1–62.
[90] M. Lacey, *Ergodic averages on circles*, J. Anal. Math. **67** (1995), 199–206.
[91] V. Lafforgue, *A proof of property (RD) for cocompact lattices of $SL_3(\mathbb{R})$ and $SL_3(\mathbb{C})$*, J. Lie Theory **10** (2000), 255–267.
[92] E. Lindenstrauss, *Pointwise ergodic theorems for amenable groups*, Invent. Math. **146** (2001), 256–295.
[93] V. Losert, *On the structure of groups with polynomial growth*, Math. Z. **195** (1987), 109–117.
[94] A. Lubotzky, *Subgroup growth and congruence subgroups*, Invent. Math. **119** (1995), 267–295.
[95] A. Lubotzky and S. Mozes, *Asymptotic properties of unitary representations of tree automorphisms*, Harmonic Analysis and Discrete Potential Theory (Frascati, 1991), Plenum, New York (1992), 289–298.
[96] A. Magyar, E. Stein and S. Wainger, *Discrete analogs in harmonic analysis: Spherical means*, Ann. Math. **155** (2002), 189–208.
[97] G.A. Margulis, *Discrete Subgroups of Semisimple Lie Groups*, A Series of Modern Surveys in Mathematics, Vol. 17, Springer (1991).
[98] G.A. Margulis, *On Some Aspects of the Theory of Anosov Systems*, Springer Monographs in Mathematics, Springer (2004).

[99] G.A. Margulis, A. Nevo and E.M. Stein, *Analogs of Wiener's ergodic theorems for semi-simple Lie groups II*, Duke Math. J. **103** (2000), 233–259.
[100] P. Milne, *Counterexample to a conjecture of Greenleaf*, Canad. Math. Bull. **13** (1970), 497–499.
[101] D. Müller and A. Seeger, *Singular spherical maximal operators on a class of step two nilpotent Lie groups*, Israel J. Math. **141** (2004), 315–340.
[102] E.K. Narayanan and S. Thangavelu, *An optimal theorem for the spherical maximal operator on the Heisenberg group*, Israel J. Math. **144** (2004), 211–220.
[103] C. Nebbia, *Groups of isometries of a tree and the Kunze–Stein phenomenon*, Pacific J. Math. **133** (1988), 141–149.
[104] A. Nevo, *Boundary theory and harmonic analysis on boundary-transitive graphs*, Amer. J. Math. **116** (1994), 243–282.
[105] A. Nevo, *Harmonic analysis and pointwise ergodic theorems for non-commuting transformations*, J. Amer. Math. Soc. **7** (1994), 875–902.
[106] A. Nevo, *Pointwise ergodic theorems for radial averages on simple Lie groups I*, Duke Math. J. **76** (1994), 113–140.
[107] A. Nevo, *Pointwise ergodic theorems for radial averages on simple Lie groups II*, Duke Math. J. **86** (1997), 239–259.
[108] A. Nevo, *Spectral transfer and pointwise ergodic theorems for semi-simple Kazhdan Lie groups*, Math. Res. Lett. **5** (1998), 305–325.
[109] A. Nevo, *On discrete groups and pointwise ergodic theory*, Random Walks and Discrete Potential Theory, Symposia Mathematica, Vol. XXXIX, INDAM, Cambridge University Press (1999), 279–305.
[110] A. Nevo, *Best possible speed of convergence in pointwise ergodic theorems on semisimple groups*, Preprint (2003).
[111] A. Nevo, *Radial geometric analysis on groups*, Contemp. Math. **347** (2004), 221–244.
[112] A. Nevo, *Group structure theorems and ergodic theorems: The Dunford–Zygmund method revisited*, Preprint.
[113] A. Nevo, *Exponential volume growth, maximal functions on symmetric spaces, and ergodic theorems for semisimple Lie groups*, Ergodic Theory Dynamical Systems, to appear.
[114] A. Nevo and E.M. Stein, *A generalization of Birkhoff's pointwise ergodic theorem*, Acta Math. **173** (1994), 135–154.
[115] A. Nevo and E.M. Stein, *Analogs of Wiener's ergodic theorems for semi-simple Lie groups I*, Ann. of Math. **145** (1997), 565–595.
[116] A. Nevo and S. Thangavelu, *Pointwise ergodic theorems for radial averages on the Heisenberg group*, Adv. Math. **127** (1997), 307–334.
[117] H. Oh, *Tempered subgroups and representations with minimal decay of matrix coefficients*, Bull. Soc. Math. France **126** (1998), 355–380.
[118] H. Oh, *Uniform pointwise bounds for matrix coefficients of unitary representations and applications to Kazhdan constants*, Duke Math. J. **113** (2002), 133–192.
[119] J.M. Ollagnier, *Ergodic Theory and Statistical Mechanics*, Lecture Notes, Vol. 1115, Springer (1985).
[120] D. Ornstein, *On the pointwise behavior of iterates of a self-adjoint operator*, J. Math. Mech. **18** (1968), 473–477.
[121] D. Ornstein and B. Weiss, *Entropy and isomorphism theorems for actions of amenable groups*, J. Anal. Math. **48** (1987), 1–142.
[122] R.E.A.C. Paley, *A proof of a theorem of averages*, Proc. London Math. Soc. **31** (1930), 289–300.
[123] P. Pansu, *Croissance des boules et des géodésiques fermeés dans les nilvariétés*, Ergodic Theory Dynamical Systems **3** (1983), 415–445.
[124] A.L.T. Paterson, *Amenability*, Mathematical Surveys and Monographs, Vol. 29, Amer. Math. Soc. (1988).
[125] J.-P. Pier, *Amenable Locally Compact Groups*, Pure and Applied Math., Wiley-Interscience, New York (1984).
[126] H.R. Pitt, *Some generalizations of the ergodic theorem*, Proc. Cambridge Philos. Soc. **38** (1942), 325–343.
[127] Ch. Pittet, *The isoperimetric constant of homogeneous Riemannian manifolds*, J. Differential Geom. **54** (2000), 255–302.
[128] J. Ramagge, G. Robertson and T. Steger, *A Haagerup inequality for $\tilde{A}_1 \times \tilde{A}_1$ and $\tilde{A}_2$ buildings*, Geom. Funct. Anal. **8** (1998), 702–731.

[129] A.S. Rapinchuk, *Congruence subgroup problem for algebraic groups: Old and new*, Astérisque **209** (1992), 73–84.
[130] M. Ratner, *Strict measure rigidity for unipotent subgroups of solvable groups*, Invent. Math. **101** (1990), 449–482.
[131] F. Riesz, *Some mean ergodic theorems*, J. London Math. Soc. **13** (1938), 274–278.
[132] J. Rosenblatt, *Uniqueness of invariant means for measure-preserving transformations*, Trans. Amer. Math. Soc. **265** (1981), 623–636.
[133] J.C. Rota, *An "Alternierende Verfahren" for general positive operators*, Bull. Amer. Math. Soc. **68** (1962), 95–102.
[134] D. Rudolph and B. Weiss, *Entropy and mixing for amenable group actions*, Ann. Math. **151** (2001), 1119–1150.
[135] K. Schmidt, *Dynamical Systems of Algebraic Origin*, Progress in Mathematics, Vol. 128, Birkhäuser, Boston (1995).
[136] Y. Sinai, *Geodesic flows on manifolds of negative curvature*, Algorithms, Fractals and Dynamics, Y. Takahashi, ed., Plenum, New York (1995).
[137] M. Stoll, *On the asymptotics of the growth of* 2*-step nilpotent groups*, J. London Math. Soc. **58** (1998), 38–48.
[138] A. Starkov, *Dynamical Systems on Homogeneous Spaces*, Transl. Math. Monographs, Vol. 190, Amer. Math. Soc., Providence, RI (2001).
[139] E.M. Stein, *On the maximal ergodic theorem*, Proc. Nat. Acad. Sci. U.S.A. **47** (1961), 1894–1897.
[140] E.M. Stein, *Topics in Harmonic Analysis Related to the Littlewood–Paley Theory*, Ann. of Math. Stud., Vol. 63, Princeton University Press (1970).
[141] E.M. Stein, *Harmonic Analysis*, Princeton Mathematical Series, Vol. 43, Princeton University Press (1993).
[142] E.M. Stein and S. Wainger, *Problems in harmonic analysis related to curvature*, Bull. Amer. Math. Soc. **84** (1978), 1239–1295.
[143] J.-O. Stromberg, *Weak type* L^1 *estimates for maximal functions on noncompact symmetric spaces*, Ann. of Math. **114** (1981), 115–126.
[144] O.I. Tavgen, *Bounded generation of Chevalley groups over ring of algebraic S-integers*, Math. USSR-Izv. **36** (1991), 101–128.
[145] A. Tempelman, *Ergodic Theorems for Group Actions*, Mathematics and Its Applications, Vol. 78, Kluwer Academic Publishers (1992).
[146] A. Tempelman, *Ergodic theorems for general dynamical systems*, Soviet Math. Dokl. **8** (1967), 1213–1216.
[147] A. Tempelman, *Ergodic theorems for general dynamical systems*, Trudy Moskov. Mat. Obshch. **26** (1972), 95–132.
[148] R. Tessera, *Asymptotic volume of spheres in metric measured spaces and in polynomial groups*, Preprint (December 2004).
[149] V.S. Varadarajan, *Lie Groups, Lie Algebras, and Their Representations*, Graduate Texts in Math., Vol. 102, Springer (1984).
[150] W.A. Veech, *Siegel measures*, Ann. Math. **148** (1998), 895–944.
[151] S. Wainger, *Averages and Singular Integrals over Lower Dimensional Sets*, Beijing Lectures in Harmonic Analysis, Ann. of Math. Studies, Vol. 112, Princeton University Press (1986).
[152] T.E. Walker, *Ergodic theorems for free group actions on von-Neumann algebras*, J. Funct. Anal. **150** (1997), 26–47.
[153] B. Weiss, *Positive cones in Hilbert space and a maximal inequality*, Inequalities III, O. Shisha, ed., Academic Press (1972).
[154] B. Weiss, *Actions of amenable groups*, London Math. Soc. Lecture Notes Series, Vol. 310, Cambridge University Press (2003), 226–262.
[155] N. Wiener, *The ergodic theorem*, Duke Math. J. **5** (1939), 1–18.
[156] A. Zygmund, *An individual ergodic theorem for non-commutative transformations*, Acta Sci. Math. Szeged **14** (1951), 103–110.

CHAPTER 14

Global Attractors in PDE

A.V. Babin

Department of Mathematics, University of California, Irvine, CA 92697-3875, USA
E-mail: *ababine@math.uci.edu*

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

0. Introduction

Partial differential equations and systems describe evolution of time-dependent functions and vector fields $u(x,t)$ where x is a spatial variable and t is time. We consider $u(x,t)$ with a fixed t as an element of a function space E and obtain a vector-function $u(t)$. Therefore, a partial differential equation or system can be written in the form

$$\partial_t u = \mathcal{F}\big(u(t)\big), \tag{1}$$

where the operator $\mathcal{F}(u)$ includes partial derivatives of u with respect to spatial variables $x = (x_1, \dots, x_n)$. This equation looks like an ordinary differential equation and one may try to use methods from the theory of finite-dimensional dynamical systems to study the dynamics generated by (1). Dynamics can be studied locally and globally. The local theory of equilibria, periodic solutions and their perturbations is very rich and includes their stability, bifurcations, theory of local invariant manifolds through them (see [99,203,230, 291,344,360]). Here we mostly consider global aspects of dynamics.

The dynamics generated by (1) with initial data in a function space E can be described by the solution semigroup

$$S_t : u(0) \mapsto u(t)$$

that acts in the space E. When $\mathcal{F}(u(\cdot))$ does not depend on t explicitly, the solution operators S_t satisfy the semigroup identity

$$S_{t+\tau} = S_t S_\tau, \quad t \geqslant 0,\ \tau \geqslant 0, \qquad S_0 = 1. \tag{2}$$

The long-time behavior of solutions of such equations can be adequately described in terms of global attractors of the equations. In many problems the influence of initial data has vanished after a long time has elapsed, therefore permanent regimes are of importance. The simplest permanent regimes are described by time-independent functions that are solutions of the equation $\mathcal{F}(u) = 0$. Such regimes are important but very special and it is widely believed that time-dependent permanent regimes are of importance, in particular they describe turbulence in hydrodynamics (see [102,341]). Time-dependent regimes may include time-periodic, time quasiperiodic and chaotic regimes; their common feature is that they are defined for all times, positive and negative. A mathematically rigorous description of such regimes and related questions of asymptotic behavior and stability is given by the theory of attractors. The theory of global attractors of PDE is developed in works of many mathematicians, see the list of references and in particular the books [55, 98,209,270,338,353,363,371] and references therein. Here we give a brief sketch of basic ideas, approaches and directions of research in this field. We also try to complement recent reviews [331] and [336] on related subjects from this series.

The central concept of the theory we discuss here is a global attractor. Since the terminology used in the theory of global attractors of PDE was changing with time we give a brief review of the history of related concepts. A discussion of the concept of an attractor in the theory of finite-dimensional dynamical systems is given by Milnor [308]. Usually an

attractor of a semigroup (a semiflow in different terminology) is understood as an invariant set that attracts its neighborhood, it equals the omega-limit set of a neighborhood of the attractor (see [308] for different variants of this definition). Here we call such an attractor a local attractor. A dynamical system may have several local attractors, for example several stable equilibria or stable periodic solutions with different domains of attraction. In the dynamical systems generated by PDE local attractors are often considered, see [102, 230,291,341,378]. Sometimes a smaller attractor is considered, namely a set which attracts most of the points of the neighborhood, such an object is called a minimal attractor by Milnor [308], where the reader can find exact definitions. Before 1982 in the research on global dynamics of PDE, in particular in the works of Ladyzhenskaya [261,262], Foias and Temam [181], Henry [230], the attracting sets were presented as omega-limit sets of a large ball and characterized as maximal invariant bounded sets. An absorbing ball in connection with a description of the long-time dynamics of the two-dimensional Navier–Stokes system was found by Foias and Prodi [175]. The invariant set that is the omega-limit set of an absorbing ball was constructed by Ladyzhenskaya [261,262] for the two-dimensional Navier–Stokes system. One of results of [37,40] is that the invariant set constructed by Ladyzhenskaya is the global attractor in the modern terminology, namely it attracts all bounded sets in the norm-induced topology of the energy space. The seminal work of Ladyzhenskaya [262] is the first work where a global attractor of a PDE was constructed and its important properties described; in particular, the invertibility of dynamics on the attractor was proven. Ladyzhenskaya [262] also proved that a trajectory on the attractor is uniquely determined by its finite-dimensional projection, this theorem is the first in the important direction of research of finite-dimensionality of attractors of PDE; the research was continued by Mallet-Paret [286], Foias and Temam [181], Mañe [289], Foias, Temam, Manley and Treve [172], Babin and Vishik [39,42] and in many subsequent papers; for more details and references see Section 2.1.

Dynamical systems generated by PDE have their specifics. The description of dynamics usually is given in terms of inequalities that are formulated in terms of function norms, this makes them uniform in corresponding normed spaces; the inequalities describe uniformly behavior of solutions with initial data from a bounded set in such a space. A natural description of dynamics should take into account these features. The following definition of a maximal attractor in terms of attraction of all bounded sets was given and was used as a basis for a systematic approach to the study of global dynamics of parabolic, damped hyperbolic equations and the Navier–Stokes system in a series of papers of Babin and Vishik published in 1982–1983 [37,38,40,39,42] and in many subsequent papers. In these works the existence of maximal attractors was proven for general multidimensional parabolic systems, two-dimensional Navier–Stokes system and damped wave equations; the basic properties of the attractors were described; in particular, upper and lower estimates of the Hausdorff dimension of attractors were obtained and a regular structure of attractors for parabolic and hyperbolic equations with a global Lyapunov function was described. We quote in the introduction the definition from [39,42], the earlier definitions in [37,38,40] did not include the closedness (or compactness) as a requirement.

DEFINITION. A *maximal attractor* of a semigroup $\{S_t\}$ in a Banach space E is a bounded closed set $\mathcal{A}$ with the following two properties:

(i) $\mathcal{A}$ is invariant, that is $S_t\mathcal{A} = \mathcal{A}$ for all $t \geqslant 0$;
(ii) $\mathcal{A}$ attracts all bounded sets in E, that is $\delta_E(S_t B, \mathcal{A}) \to 0$ as $t \to \infty$ for every bounded set B.

This definition explicitly describes the domain of attraction, that is the whole Banach space E and, more important, explicitly specifies the attraction of $S_t u_0$ to $\mathcal{A}$. Namely, the attraction is assumed to be uniform with respect to a bounded $u_0 \in B$. Compared with the concept of a maximal invariant set that was used before in the dynamical theory of PDE this definition explicitly includes the topology of the attraction. This distinction is important in the infinite-dimensional case when the same space may be endowed with two non-equivalent topologies, for example the norm-induced and the weak topology of a Hilbert space. The maximal invariant set can be the same, but the attraction is understood in different ways and this difference is a major point of research, especially when the dynamics generated by equations in unbounded domains and damped hyperbolic problems is considered; very often the same set with the attraction in the weak topology is called a weak attractor. Before 1982–1983 in the literature on dynamical properties of PDE the attractors were considered (as omega-limit sets) but the attraction as such was not discussed.

In addition to properties of dynamics in PDE mentioned above there is the following motivation for this definition. Firstly, the maximal attractor is determined uniquely by the semigroup $\{S_t\}$, that is by the operator $\mathcal{F}$ in (1) and by the space E. Secondly, the definition does not include a specific construction of the attractor.

After 1983 the above definition of a maximal attractor or its minor variations became a standard definition in the theory of global attractors of PDE (see [55,363,209,101,270, 98,353,336] and references therein) but the name *global attractor* is now used more often. Sometimes this object is called a *universal attractor* (see [363]). We originally used the term maximal attractor to point out that the domain of attraction is maximal (namely the whole space) and that it is a maximal invariant set. Note that under natural assumptions the maximal attractor is a maximal invariant bounded set and a minimal closed set that attracts all bounded sets; the latter property is not in a perfect match with the name maximal attractor, but wise people say that nothing is perfect. The term *minimal closed B-global attractor* used by Ladyzhenskaya [270] for the same object is very precise but seems to be too long. One has to take into account that originally in the theory of infinite-dimensional dynamical systems the definition of a global attractor given in [212] did not include the attraction of bounded sets, and the global attractor was defined as a set that attracted $S_t u_0$ for all $u_0 \in E$; this terminology was used until 1984, see [214, p. 46]. Note that a set which was called a global attractor in the old terminology is usually smaller than the maximal attractor (or the global attractor in the modern terminology). By 1981 the general theory of maximal invariant sets of infinite-dimensional semigroups was developed by Billotti and La Salle [66], Hale, La Salle and Slemrod [212], Massatt [292,293]. Important concepts of asymptotically smooth semigroups were introduced by Hale, La Salle and Slemrod [212] and existence of maximal invariant sets of asymptotically smooth semigroups was proved; non-trivial sufficient conditions for the asymptotic smoothness were found; relations between different concepts of attraction were studied; see [206,208,209,336] for details and references. This theory in particular includes theorems on existence of maximal bounded invariant sets that attract all bounded sets, see [214]. One has to note though that before

1984 the attraction of bounded sets in the literature on abstract semigroups in infinite-dimensional spaces was considered among other properties such as attraction of points, attraction of compact sets and their neighborhoods and was not a subject of special interest (see, for example, [207, Chapter 4], [214]). The main application of the general theory was the dynamics of retarded functional differential equations, we could not find in the literature on Partial Differential Equations published before 1982 a paper where a theorem on attraction of every bounded set to an attractor of an equation with partial derivatives was formulated or proved.

In this review we try to pay attention to the aspects of the dynamics of PDE which distinguish this subject from the theory of finite-dimensional dynamical systems and from the abstract theory of infinite-dimensional dynamical systems.

The theory of infinite-dimensional systems generated by PDE includes technical complications that are absent in the finite-dimensional theory:

- Semigroup operators S_t often are defined only for $t \geqslant 0$ and cannot be extended for $-\infty < t < \infty$.
- Infinite-dimensional function spaces are not locally compact.
- Dynamics in infinite-dimensional spaces for given initial data as a rule does not allow an explicit description, therefore only a collective description is available, usually in terms of inequalities.
- Solutions with bounded energy can blow-up in a finite time.
- Uniqueness of solutions may be difficult to establish (3D Navier–Stokes system).
- The dependence on initial data may be non-smooth even when non-linear operators are polynomial thanks to infinite-dimensional effects (strongly non-linear monotonic parabolic equations).

More importantly, the dynamics generated by PDE has completely new features:

- Dimension of the global attractor can be considered as a large parameter, this allows to study the asymptotic behavior of the dimension.
- The spatial variables allow one to classify functions from invariant sets according to their geometric properties:
  - (i) number of zeros;
  - (ii) homotopy type;
  - (iii) symmetry properties.
- Interaction of spatial and temporal behavior (dependence of the dimension of the attractor and the fragmentation complexity of the attractor on the volume of the spatial domain).

Therefore, the central problems studied in the theory of global attractors of PDE include:

- Reduction in some sense of infinite-dimensional systems to finite-dimensional.
- Characterization of the attraction in different topologies, exponential attraction, tracking property.
- Interconnection of spatial properties of solutions and their dynamical properties.
- Expression of characteristics of attractors in terms of physical parameters of the problems.
- Relation of the properties of dynamics (for example, the existence of a global attractor) with the number of spatial variables and the growth of non-linearities.

One has to take into account that there are obvious similarities between the infinite-dimensional and finite-dimensional cases. For example, the construction of a global attractor as an omega-limit set works in both cases. The theory of local invariant manifolds and foliations is similar to the finite-dimensional theory. Though the semigroups generated by parabolic operators are not invertible (cannot be extended to negative times) the technical difficulties that arise in many cases can be solved and do not lead to significant differences.

We pay here more attention to the aspects of the theory which are specific to the infinite-dimensional case. There are completely new phenomena, for example, the dimension of the global attractors tends to infinity when the viscosity tends to zero; such behavior and its asymptotics makes sense only in an infinite-dimensional situation. Another phenomenon that has no simple analogues in the finite-dimensional case is the presence of a spatial variable in addition to the time variable. Relations between spatial and time variables manifest themselves most clearly in the case of an unbounded or a very large domain, for example the growth of the dimension of attractor and its fragmentation complexity when the domain increases, or the trivialization of dynamics on the attractor of the Navier–Stokes system in unbounded channels near spatial infinity. Many aspects of the theory of attractors are important for applications, in particular to geophysics and meteorology (see [279,280, 278]). In particular, the dimension of attractor estimates the number of degrees of freedom of the dynamical system which describes long time behavior of a physical system. A global attractor also contains all the information on the instability of the dynamical system (see [55]).

The purpose of this chapter is to give a sketch of the core of the classical theory of attractors with a minimum of technicalities and to point to major directions in which the theory develops. We do not intend to give the most general results, but rather we want to show the ideas in the simplest possible way. We prefer to present results with simple formulations rather than the most general results and give references to the literature for possible generalizations. We do not give here detailed proofs; if the formulations of results are very technical, we refer to original papers for details. Since this review reflects scientific interests of the author, inevitably not all directions in the theory of global attractors of PDE are represented with the same degree of detail. The author apologizes that many interesting papers are not discussed in this review.

1. Global attractors of semigroups

Here we discuss basic concepts related to dynamics in infinite-dimensional spaces.

1.1. *Basic definitions and existence of attractors*

Absorption and attraction. Let E be a complete metric space with distance $\rho(x_1, x_2)$ and a semigroup of (non-linear) operators $\{S_t, t \geqslant 0\}$ act in E:

$$S_t : E \to E, \quad t \geqslant 0.$$

Everywhere we assume that operators $S_t u$ are continuous with respect to u in the metric of E and are bounded. We introduce a non-symmetric distance (sometimes it is called a semidistance) $\delta_E(B_1, B_2)$ from a set B_1 to a set B_2,

$$\delta_E(B_1, B_2) = \sup_{x_1 \in B_1} \inf_{x_2 \in B_2} \rho(x_1, x_2). \tag{3}$$

A set B is *invariant* if $S_t B \subset B$, for all $t \geqslant 0$; often a set with this property is called *positive invariant*. A set B is strictly invariant if $S_t B = B$ for all $t \geqslant 0$; if $B \subset S_t B$ the set B is called *negative invariant*.

DEFINITION 1.1.1. A bounded set B_0 is called an *absorbing set* of $\{S_t\}$ if for every bounded set B there exists T such that $S_t B \subset B_0$ for all $t \geqslant T$.

Semigroups that possess a bounded absorbing set are often called *bounded dissipative*; when the statement holds only for one-point (or compact) sets B a semigroup is called *point* (*compact*) *dissipative*. Often a semigroup that has a bounded absorbing set is called *dissipative*. For a discussion of the terminology and general concepts see [336].

When $\{S_t\}$ has a bounded absorbing set B_0 and the operators S_t are continuous and bounded, the set

$$B_{0T} = \operatorname{closure}_E \left(\bigcup_{t \geqslant T} S_t B_0 \right) \tag{4}$$

is also absorbing and is invariant. Therefore existence of a bounded absorbing set is equivalent to existence of a closed bounded invariant absorbing set.

A set B_0 is called an *attracting set* (in a space E) if for every bounded set B

$$\lim_{t \to \infty} \delta_E\big(S_t(B), B_0\big) = 0. \tag{5}$$

DEFINITION 1.1.2. A set $\mathcal{A}$ is called the global attractor of $\{S_t\}$ in E if it has the following three properties:

(i) $\mathcal{A}$ is compact;

(ii) $\mathcal{A}$ is strictly invariant: $S_t \mathcal{A} = \mathcal{A}$ for all t;

(iii) $\mathcal{A}$ is an attracting set for $\{S_t\}$ in E, that is

$$\delta_E\big(S_t(B), \mathcal{A}\big) \to 0 \quad \text{as } t \to \infty$$

for every bounded set B.

The following properties characterize global attractors. A global attractor is the minimal set among all compact sets which attract all bounded sets. A global attractor is the maximal set among all bounded strictly invariant sets (see [55,209,270,353,363]). Sometimes global attractors are called maximal attractors or minimal attractors. Note that if a global attractor exists it is unique.

REMARK. We most often use the above definition and the theorems which we formulate below in the situation when E is a Banach space with the distance $\rho(x_1, x_2) = \|x_1 - x_2\|_E$. The general definition and theorems are useful when we use the weak topology of a ball in a separable Hilbert space or when functions are defined on unbounded domains and we use the topology generated by convergence on bounded subdomains of the unbounded domain (for example, such topology is widely used in the theory of trajectory attractors). Another important example is the case when a semigroup is defined only on an invariant subset of a Banach space which does not coincide with the entire space. See [55,57,209,228,363,353, 270,336] for details and examples. In particular, one may find a very detailed discussion of different aspects of general theorems on existence of attractors and the history of the question in [353]. A similar definition of attraction can be given in terms of topology of the function space rather than its metric, see [44,55].

When E is a separable reflexive Banach space and we use the topology generated by the weak convergence (or a metric ρ on an absorbing ball which induces the weak convergence) we call the global attractor (E, E_w)-attractor or a weak attractor to distinguish it from the attractor in the norm-induced topology.

A curve $u(t)$, $-\infty < t < \infty$, is called a trajectory of $\{S_t\}$ if $S_{t_1} u(t_2) = u(t_1 + t_2)$ for all $-\infty < t_2 < \infty$, $0 < t_1 < \infty$.

The following important property of a global attractor is equivalent to its strict invariance.

For every point $a \in \mathcal{A}$ there exists a bounded trajectory $u(t)$ of $\{S_t\}$ defined for all $-\infty < t < +\infty$ such that $u(0) = a$.

Existence theorems. We formulate basic existence theorems from the theory of global attractors. More details are given in the books and reviews [55,363,353,336,270,228]. Detailed treatments of general aspects of the theory of existence of attractors of operator semigroups is given in [209,353,336].

We consider an operator semigroup $\{S_t\}$ in a complete metric space E. The operators S_t are everywhere assumed to be continuous bounded (non-linear) operators in E. Most often in the applications presented in this chapter the complete metric space E is a Banach space.

THEOREM 1.1.3. *Let a semigroup $\{S_t\}$ of continuous operators have a bounded absorbing set B_0 such that the set $\bigcup_{t \geqslant 0} S_t(B_0)$ is bounded and let $\{S_t\}$ have a compact attracting set. Then $\{S_t\}$ has a global attractor $\mathcal{A}$. The attractor $\mathcal{A}$ is defined as an omega-limit set of B_0 by the following formula*

$$\mathcal{A} = \bigcap_T \widetilde{B}(T), \quad \text{where } \widetilde{B}(T) = \text{closure}_E \left(\bigcup_{t \geqslant T} S_t B_0 \right). \tag{6}$$

(For proofs and variants see [55,209,270,336,353,228,59].) Theorem 1.1.3 is sufficient in many applications, but its conditions can be relaxed, below we give a more general theorem from which it follows.

Following Ladyzhenskaya [265] we call a semigroup $\{S_t\}$ *asymptotically compact* if for any bounded set B such that the set $\{S_t B, t \geqslant \tau\}$ is bounded for some τ every sequence $S_{t_n} z_n$ with $z_n \in B$ and $t_n \to +\infty$ is relatively compact.

A semigroup is called *asymptotically smooth* if for any bounded invariant set B there exists a compact set $B_1 = B_1(B)$ such that $\lim_{t\to\infty} \delta_E(S_t(B), B_1) = 0$ (see [212]). The condition of being asymptotically compact is equivalent to being asymptotically smooth, see [336]; see [336,353] for a discussion of general properties of such semigroups. We prefer the term asymptotically compact by two reasons: first, this property is not related to the differentiability of operators S_t; second, in unbounded domains higher smoothness of functions does not imply compactness.

THEOREM 1.1.4. *Let a semigroup $\{S_t\}$ of continuous operators satisfy the following conditions*:

(i) *$\{S_t\}$ is asymptotically compact*;

(ii) *$\{S_t\}$ is point dissipative* (*that is there exists a bounded set B_0 such that for every point z $S_t z \in B_0$ for all $t \geqslant t_0(z)$*);

(iii) *for any bounded set B there exists $\tau = \tau(B)$ such that the set $\{S_t B, t \geqslant \tau\}$ is bounded*.

Then $\{S_t\}$ has a global attractor $\mathcal{A}$.

As a simple corollary of Theorem 1.1.3 we obtain the following statement.

COROLLARY 1.1.5. *Let $\{S_t\}$ have a bounded absorbing ball. Let operators S_t for $t \geqslant 0$ be continuous and uniformly bounded on bounded sets. Let S_t be compact for every $t > 0$. Then $\{S_t\}$ has a global attractor.*

REMARK. For parabolic equations in bounded domains it is sufficient to use Corollary 1.1.5. Applications to damped hyperbolic equations, degenerate parabolic equations, mixed parabolic–hyperbolic systems, equations in unbounded domains when the semigroup operators are not compact require more general Theorem 1.1.3.

REMARK. When E is a Banach space the attractor defined by (6) is a connected set and contains an equilibrium, see [209,270,353]. Trivial examples (a set E formed by two points and a set E that coincides with a circle) show that there exist global attractors in metric spaces which are not connected sets and there exists a global attractor in a metric space which does not contain an equilibrium.

More on the attraction. A semigroup $\{S_t\}$ that acts in a separable Hilbert space H may be studied from different points of view. One may consider the topology in H generated by the Hilbert norm. Another choice is to use the weak topology in H, this topology restricted to a bounded ball is metrizable; the space H endowed with the weak topology we denote by H_w. The same set $\mathcal{A}$ can be a global attractor in the norm-induced topology or in a weak topology (a weak attractor) in the same space; it can also be an attractor in a norm that is stronger than the norm of H (see below the definition of (E, E_1)-attractors). Since the norm-induced topology is stronger, the attraction in the norm-induced topology

$\delta_H(S_t B, \mathcal{A}) \to 0$ implies that $\delta_{H_w}(S_t B, \mathcal{A}) \to 0$ as $t \to \infty$, therefore the attraction in the norm-induced topology contains more information. One has to note that bounded sets are precompact in the weak topology of H_w and this makes construction of a compact attracting set more straightforward. There are examples of semigroups (for example, semigroups generated by non-linear monotonic operators with a monotonic principal part, see [55], or by general parabolic systems in unbounded domains in spaces of functions that do not decay at infinity, see [57]) with non-compact operators for which the existence of attractors is proven only in the weak topology. One has to take into account that the proof of continuity of operators S_t in the weak topology H_w in many cases is not much harder than the proof of continuity of operators S_t in the norm-induced topology H.

In many cases the attraction property of semigroups generated by parabolic equations and systems is proven in a stronger norm than the norm of the space H where the semigroup acts. For example, the semigroup defined in a Sobolev space $H_0(\Omega)$ can attract bounded in $H_0(\Omega)$ sets in the stronger norm of $H_2(\Omega)$. To describe such a situation Babin and Vishik [44,55] introduced the following definition.

DEFINITION 1.1.6. Let E_1, E be two metric spaces and $E_1 \subset E$ or $E \subset E_1$. A strictly invariant set $\mathcal{A}$ is called (E, E_1)-attractor if $\delta_{E_1}(S_t B, \mathcal{A}) \to 0$ as $t \to \infty$ for any bounded in E set B.

See [44,55] for examples of (E, E_1)-attractors. In particular, when $E = H$, $E_1 = H_w$ we have a weak attractor. Note that the space E defines bounded sets and E_1 defines the topology, therefore in this definition E_1 can be a topological space (not necessarily a Banach space or a metric space, see [44,55]).

It is worth noticing that when the domain Ω is bounded, the norm-induced topology in $H_{2-\epsilon}(\Omega)$ with arbitrary small $\epsilon > 0$ is weaker than the weak topology of $H_2(\Omega)_w$. For example, the existence of the global attractor $\mathcal{A}$ in $H_1(\Omega)$ in the norm-induced topology implies much weaker attraction to $\mathcal{A}$ than the existence of $(H_1(\Omega), H_2(\Omega)_w)$-attractor, therefore the weakness of a weak attractor is relative.

REMARK. Using a pair of spaces as in the definition of (E, E_1)-attractor often is instrumental in dealing with equations in unbounded domains, see Mielke and Schneider [305], Mielke [303].

REMARK. Pointwise attraction in the norm different from the norm of the Banach space E was considered already in [212], a systematic study of the attraction to global attractors of PDE in stronger (or weaker) norms or in the weak topology in terms of (E, E_1)-attractors was done in [44].

REMARK. Usually the smoothness of functions on the attractor is determined by the smoothness of the forcing term and by the boundary. We note that the attraction can be in a stronger norm than boundedness of solutions (see [55]) since the difference of two functions from $\mathbf{H}_1$ may belong to $\mathbf{H}_2$. Regularity of functions on attractors and attraction in stronger norms are studied in [44,55,200]. In particular, when f and $\partial\Omega$ are infinitely smooth the attractor $\mathcal{A}$ consists of infinitely smooth functions (see [363]).

Unbounded attractors. In the definition of a global attractor we have imposed the compactness condition. When a semigroup acts in a Banach space one may consider unbounded, locally compact attractors. Non-trivial examples of unbounded, locally compact attractors of PDE are given by Chepyzhov and Goritskii [92].

1.2. *Equilibria and local invariant manifolds*

A global attractor always contains all equilibria and unstable manifolds through them. This fact was used in [39,202] (see also [55]) to obtain lower estimates of dimension of global attractors.

DEFINITION 1.2.1. A point z is an *equilibrium point* of $\{S_t\}$ if $S_t z = z$ for all t.

Since z does not depend on t, for semigroups defined by (1), z satisfies the equation $\mathcal{F}(z) = 0$.

DEFINITION 1.2.2. An *unstable manifold* $M^{\mathrm{un}}(z)$ through an equilibrium point z of S_t is the set of all points $v \in E$ such that $S_t v$ is defined for all $t \leqslant 0$ and $S_t v \to z$ in E as $t \to -\infty$.

If $\{S_t\}$ has a global attractor $\mathcal{A}$ and z is an equilibrium point of $\{S_t\}$, then $M^{\mathrm{un}}(z) \subset \mathcal{A}$.

DEFINITION 1.2.3. A *stable manifold* $M_-(z)$ through an equilibrium point z of S_t is the set of all points $v \in E$ such that $S_t v$ is defined for all $t \geqslant 0$ and $S_t v \to z$ in E as $t \to +\infty$.

The behavior of a dynamical system near an equilibrium is described by the theorem on stable and unstable manifolds of semigroups in Banach spaces; this theorem is fairly similar to the finite-dimensional theorem. We formulate the theorem skipping technical details, in particular the differentiability conditions (see [55,230] for details; see also [61, 62,89] for more details, generalizations and more references).

Let S_t be a non-linear differentiable (of class $\mathcal{C}^\alpha$, $\alpha \geqslant 1$) semigroup in a Banach space E. Let a point z be an equilibrium of S_t, that is $S_t z = z$ for all $t \geqslant 0$. The differentials $S_t'(z)$ form a semigroup of linear bounded operators in E. The properties of this semigroup play important role; the behavior of S_t near z is in many respects similar to that of $S_t'(z)$. The most important assumption is the existence of a circular gap in the spectrum of $S_t'(z)$. Namely, we assume that the spectrum of $S_t'(z)$ does not contain a circle $|\zeta| = \rho^t$ in the complex plane. We conclude that the spectrum is divided by the circle into two parts: external σ_+ and internal σ_-. Therefore, the Banach space E splits into two complementary invariant subspaces $E_+(\rho)$ and $E_-(\rho)$, $S_t'(z)E_-(\rho) \subset E_-(\rho)$, $S_t'(z)E_+(\rho) = E_+(\rho)$ for all $t \geqslant 0$. We assume that $E_+(\rho)$ is finite-dimensional.

Under these conditions, the non-linear semigroups have *local invariant manifolds* $M_+(z, \rho)$ and $M_-(z, \rho)$ through a point z in a neighborhood of z which are tangent respectively to $E_+(\rho)$ and $E_-(\rho)$ (the local manifolds may be non-unique). A set M is

called local invariant (in a neighborhood of z) if the assumptions $u \in M$ and $S_\tau u$ stays in the neighborhood of z for $0 \leqslant \tau \leqslant t$ imply that $S_t u \in M$.

When $\rho = 1$ $M_+(z, \rho)$ is called a local unstable manifold of S_t, $M_-(z, \rho)$ is called a local stable manifold of S_t. When $|\zeta| = 1$ is in the spectrum, and $\rho < 1$ $M_+(z, \rho)$ is called a *center-unstable manifold* of S_t.

THEOREM 1.2.4. *There exists a local unstable manifold $M_+(z, \rho)$ which in a neighborhood of z is a graph of a function of class C^α from $E_+(\rho)$ to $E_-(\rho)$, $M_+(z, \rho)$ is tangent to $E_+(\rho)$ at z. In the neighborhood of z, $M_+(z, \rho)$ is locally invariant with respect to S_t. The following attraction estimate*

$$\delta_E\big(S_t u, M_+(z, \rho)\big) \leqslant c'(\rho - \epsilon)^t, \quad 0 \leqslant \tau \leqslant t, \tag{7}$$

holds when $S_\tau u$ is in the neighborhood of z for $0 \leqslant \tau \leqslant t$. When $\rho \geqslant 1$, S_t is extended inside $M_+(z, \rho)$ to negative t and

$$\delta_E(S_t u, z) \leqslant c''(\rho + \epsilon)^t, \quad t \leqslant 0. \tag{8}$$

If there are many circular gaps $|\zeta| = \rho_i^t$ with points of the spectrum between them, picking different ρ_i one can find many different local invariant manifolds of S_t near z. Intersections of these manifolds are also smooth local invariant manifolds. Therefore, local non-linear dynamics near z is in many respects similar to the linear dynamics of $S_t'(z)$. Note that the spectrum of $S_t'(z)$ in applications equals the exponent of the spectrum of $\mathcal{F}'(z)$ where $\mathcal{F}'(z)$ is the differential of $\mathcal{F}$ in (1) and in (18). A circular gap in the spectrum of $S_t'(z)$ corresponds to a gap $\{\ln \rho - \epsilon_1 \leqslant \operatorname{Re} \lambda \leqslant \ln \rho + \epsilon_1\}$ in the spectrum of $\mathcal{F}'(z)$.

A common way to study the local behavior near an equilibrium z is by using cut-off functions. Namely, (1) is replaced by the following equation

$$\partial_t v = \mathcal{F}'(z)v + \varphi\big(L\|v\|_E\big)\big(\mathcal{F}(z + v) - \mathcal{F}'(z)v\big), \tag{9}$$

where a smooth cut-off function $\varphi(s) = 1$ when $|s| \leqslant 1$, $\varphi(s) = 0$ when $|s| \geqslant 2$. Equation (9) generates a semigroup $\{S_t^0\}$ that coincides with the original semigroup $\{S_t\}$ near z. The semigroup $\{S_t^0\}$ is close to the linear semigroup $\{S_t'(z)\}$ when $M \gg 1$, since $\varphi(L\|v\|_E) = 0$ for $\|v\|_E \geqslant 2/L$ and, roughly speaking, for large L one has to consider small v and observe that $\varphi(L\|v\|_E)(\mathcal{F}(z + v) - \mathcal{F}'(z)v) = \mathrm{O}(L^{-2})$, its Lipschitz constant is $\mathrm{O}(vL^{-1})$ and one may expect the Lipschitz constant to be small. These statements can be made precise for typical parabolic and hyperbolic equations if the space E is chosen in a proper way, see, for example, [55,230].

DEFINITION 1.2.5. If the unit circle $\rho = 1$ is not in the spectrum of $S_t'(z)$ the equilibrium point z is called hyperbolic.

When z is hyperbolic the local unstable manifold $M_+(1)$ coincides with the intersection of the unstable manifold $M_+(z)$ with a neighborhood of z.

REMARK. To prove the existence of an unstable manifold of a hyperbolic equilibrium point one can treat directly S_t not using (9). This is important when non-smooth norms are used, like the norm in $\mathcal{C}^\alpha(\Omega)$. See [42] and Section 5.3 in [55] for details.

Tracking property. If a trajectory $u(t) = S_t u_0$ spends long time near an equilibrium z, then there exists a trajectory $\tilde{u}(t)$ that lies on a finite-dimensional invariant manifold $M_+(z, \rho)$, this trajectory approximates $u(t)$ very well (with an exponential error estimate). The approximation $\tilde{u}(t)$ is called a tracking trajectory (or a shadowing trajectory).

THEOREM 1.2.6. *Let S_t be close to a linear semigroup near an equilibrium z. Let $|\zeta| = \rho$, $\rho < 1$ be not in the spectrum of $S_1'(z)$. Then there exists a small neighborhood O of the point z and a constant C that have the following property. If $u(t) = S_t u_0 \in O$ for $0 \leqslant t \leqslant T$ then there exists $\tilde{u}_0 \in M_+(z, \rho)$ and $\tilde{u}(t) = S_t \tilde{u}_0 \in M_+(z, \rho)$ that satisfies the inequality*

$$\left\|\tilde{u}(t) - u(t)\right\|_E \leqslant C\rho^t, \quad \text{for } 0 \leqslant t \leqslant T. \tag{10}$$

In particular, if $u(t) \to z$ as $t \to \infty$, there exists $\tilde{u}(t) = S_t \tilde{u}_0 \in M_+(z, \rho)$ such that

$$\left\|\tilde{u}(t) - u(t)\right\|_E \leqslant C\rho^t, \quad \text{for } 0 \leqslant t < \infty. \tag{11}$$

Note that if the spectrum of $S_t'(z)$ has many circular gaps, one may find different $\tilde{u}(t)$ for different ρ. In particular, for parabolic problems in bounded domains one may take ρ arbitrary small, drastically increasing the accuracy of approximation at the expense of increasing the dimension of $M_+(z, \rho)$.

1.3. *Inertial manifolds*

The fact that the global attractor $\mathcal{A}$ is finite-dimensional is quite remarkable. One though has to take into account that this set may have a non-regular structure. To describe the dynamics on the attractor one may try to include it into a larger but still finite-dimensional exponentially attracting manifold M which is invariant with respect to S_t and contains the attractor. If this is done, the dynamics on M is given by a system of ODE and such a system could be used for numerical simulations. The best possible solution is to describe M as a graph of a smooth (or, at least, Lipschitz) function defined on a finite-dimensional subspace of the function space. Such an object is called an inertial manifold. An inertial manifold is a global variant of the local invariant manifold $M_+(z, \rho)$ in Theorem 1.2.4. Inertial manifolds exist when there is a wide enough gap in the spectrum of the linear part of the equation. Existence of inertial manifolds is proven for a number of systems, in particular for low-dimensional parabolic equations and systems, for the Kuramoto–Sivashinsky equation (see for details Foias, Nicolaenko, Sell and Temam [173], Foias, Sell and Titi [180], Foias, Sell and Temam [178,179], Temam [363], Sell and You [353]). The theory of inertial manifolds is well developed, see [134,353] and references therein. Here we give only the definition and we try to illustrate this object comparing it with local invariant manifolds discussed above. We give the definition of the inertial manifold following [363,119,353].

DEFINITION 1.3.1. An *inertial manifold* $\mathcal{M}$ is a graph of a Lipschitz function which domain is a finite-dimensional linear subspace of E. It is positive invariant, $S_t\mathcal{M} \subset \mathcal{M}$. The manifold $\mathcal{M}$ attracts all trajectories of the semigroup exponentially.

Since the definition is global, to apply this definition to semigroups generated by PDE usually one has to modify the non-linearity outside an absorbing ball as in (9) to make it globally Lipschitz. To prove the existence of an inertial manifold instead of using smallness of the non-linearity that is used in local theorems one may use other parameters. A natural large parameter is the dimension of the manifold. Analysis of the construction of local invariant manifolds shows that when ρ lies in a wide enough gap in the spectrum of $S_1'(z)$ (equivalently, $\log \rho$ lies in a wide enough gap in the spectrum of $\mathcal{F}'(z)$) one can take a small L in (9) and still prove existence of an invariant manifold. In fact, one does not need z to be an equilibrium, one can take, for example, $z = 0$. The gap condition on consecutive eigenvalues λ_N, λ_{N-1} of $\mathcal{F}'(0)$ is of the form

$$\lambda_N - \lambda_{N-1} \geqslant c(\lambda_N + \lambda_{N-1})^p, \tag{12}$$

where c is a constant and p, $0 \leqslant p < 1$, depends on the properties of the non-linearity, see for details [363,353]. Typically $\lambda_N \sim N^q$, and one has to assume $q - 1 > qp$ and to take large N to satisfy (12).

The invariant inertial manifold enjoys all the properties of $M_+(z, \rho)$, $\rho < 1$. In particular the tracking property holds, that is for any trajectory $u(t) = S_t u_0$ there exists a tracking trajectory $\tilde{u}(t)$ on the inertial manifold $\mathcal{M}$ such that (11) holds (see [180,353] and references therein).

The gap should be large relative to the Lipschitz constant of the non-linearity. Clearly, in the local theory, since the Lipschitz constant is small for large L any spectral gap will do. One though has to take into account difficulties which arise when the non-linearity includes unbounded operators.

When L is small, the semigroup S_t^0 generated by (9) coincides with the original semigroup S_t in a large ball $\|z - u\| \leqslant L^{-1}$. For small L this ball contains an absorbing ball and the attractor of S_t, therefore the long-time dynamics of S_t is described by the modified equation. The large spectral gap condition is restrictive, but there are important equations that satisfy this condition (see [287,353,363]). Eigenvalues λ_N, $N = 1, 2, \ldots$, of a second-order elliptic differential operator A (for example, $A = -\Delta$ where Δ is the Laplacian with Dirichlet boundary condition in a bounded d-dimensional domain Ω) have asymptotical behavior $\lambda_N \sim CN^{2/d}$ as $N \to \infty$. Therefore in a generic case one may expect gaps to behave like $C[(N+1)^{2/d} - N^{2/d}]$ and, generally speaking, large gaps are absent when $d \geqslant 2$. Nevertheless, existence of inertial manifolds is proven in some high-dimensional cases by Mallet-Paret and Sell [287].

An approach to approximate long-time dynamics constructively is based on the construction of approximate inertial manifolds, see [179,185,186,195,363] and references therein. For the approach based on construction of approximating algebraic and analytic sets see [185].

Inertial manifolds and approximate inertial manifolds are constructed for many equations in mathematical physics, see in particular [67,110,322,353,367,384].

1.4. *Exponential attractors*

Another important notion of the theory of infinite-dimensional dynamical systems is an *exponential attractor*, also called an *inertial set* (see [134,133,131,34]).

DEFINITION 1.4.1. A set $\mathcal{E} \subset E$ is called an exponential attractor of the semigroup $\{S_t\}$ in the Banach space E if the following three conditions hold:

(i) $\mathcal{E}$ is compact and has a finite fractal dimension;
(ii) $\mathcal{E}$ is invariant (not strictly) $S_t\mathcal{E} \subset \mathcal{E}$ for all $t \geqslant 0$;
(iii) there exist positive constants c and c' such that for all $t \geqslant 0$ and for every bounded set $B \subset E$

$$\delta_E(S_t B, \mathcal{E}) \leqslant c' \exp[-ct].$$

Note that points (i) and (iii) of this definition are more restrictive than the corresponding points of Definition 1.1.2 and point (ii) is less restrictive. For the definition of a fractal dimension see Section 3. If an exponential attractor exists, it always contains the global attractor, $\mathcal{A} \subset \mathcal{E}$. An exponential attractor is non-unique. The simplest examples of exponential attractors are given by regular attractors of semigroups with a Lyapunov function, since the tracking property includes exponential attraction, see [42,54,55]. It is remarkable that the existence of an exponential attractor can be proven for very general systems.

Existence of an exponential attractor for 2D Navier–Stokes system, reaction–diffusion systems and damped wave equations is proven by Eden, Foias and Nicolaenko [132]; see also [133,134,131,125]. The theory of exponential attractors in Hilbert spaces is expanded to Banach spaces by Dung and Nicolaenko [131]. Exponential attractors of reaction–diffusion systems in unbounded domains are constructed by Babin and Nicolaenko [34], see also Efendiev, Miranville and Zelik [137]. Fabrie, Galusinski and Miranville [142] study behavior of exponential attractors when damped wave equation degenerates into a parabolic equation. Exponential attractors of generalized Cahn–Hilliard equation are constructed by Miranville [310]. Exponential attractors for non-autonomous evolution equations are constructed by Miranville [309,311] based on the theory of trajectory attractors of Chepyzhov and Vishik [97].

1.5. *Hausdorff and fractal dimension*

A fundamental characteristic of an attractor of a dynamical system is its dimension. The physical meaning of the dimension of an attractor is, roughly speaking, the number of degrees of freedom required to describe the large-time dynamics of a dynamical system. The attractor may be a very complicated set, so a definition of dimension has to be applicable to general sets.

First, we give the definition of the *Hausdorff dimension* of a set in a Banach space.

If K is a compact set, we consider finite coverings C_K of K by balls $B_{r_i}(x_i)$ of radius r_i centered at x_i, $B_{r_i}(x_i) = \{u\colon \|u - x_i\|_E < r_i\}$. We denote by $|C_K|$ the maximum of r for the covering C_K. Let

$$\mu_{\alpha,\epsilon}(K) = \inf_{|C_K| \leqslant \epsilon} \sum_i r_i^\alpha. \tag{13}$$

We define the Hausdorff measure by the formula

$$\mu_\alpha(K) = \lim_{\epsilon \to 0} \mu_{\alpha,\epsilon}(K), \tag{14}$$

the measure $\mu_\alpha(K)$ equals ∞ for small α and equals 0 for large α. The Hausdorff dimension is defined as

$$\dim_{\mathrm{H}}(K) = \inf\bigl\{\alpha\colon\ \mu_\alpha(K) = 0\bigr\}. \tag{15}$$

Note that since E is infinite-dimensional, $\dim_{\mathrm{H}}(K)$ may be infinity.

Now, we define *the fractal dimension* of a compact set K.

We consider all finite coverings of K by balls $B_\varepsilon(x_i)$ of radius ε centered at x_i with a fixed radius ε. We denote by $n(\varepsilon, K)$ the minimum number of balls in such a covering. The number

$$\mathbb{H}_\varepsilon(K) = \log_2\bigl(n(\varepsilon, K)\bigr) \tag{16}$$

is called Kolmogorov ε-*entropy* of the set K. The box-counting dimension (fractal dimension) of K is

$$\dim_{\mathrm{F}}(K) = \limsup_{\varepsilon \to 0} \frac{\mathbb{H}_\varepsilon(K)}{\log_2(1/\varepsilon)}. \tag{17}$$

We always have

$$\dim_{\mathrm{H}}(K) \leqslant \dim_{\mathrm{F}}(K).$$

Note that if K is a smooth compact d-dimensional manifold (or a piecewise smooth manifold, or a manifold with a boundary) that lies in E, then $d = \dim_{\mathrm{H}}(K) = \dim_{\mathrm{F}}(K)$. An important property of the fractal dimension is the existence of Mañé's projection (see Mañé [289]). Namely, if a set K has dimension $\dim_{\mathrm{F}}(K)$, there exists a projection onto a linear subspace with dimension $d < 2\dim_{\mathrm{F}}(K) + 1$ which is one-to-one on K. We give the following refinement of the Mañé's theorem (see [134,174]):

THEOREM 1.5.1. *Let H be a Banach space and X be a compact subset of E with fractal dimension $\dim_{\mathrm{F}} X < d/2$ with an integer d. Let $E_{(d)}$ be a d-dimensional linear subspace of E. Then the set of linear projections P from E onto $E_{(d)}$ which are one-to-one from X onto $PE \subset E_{(d)}$ is a residual set in the operator topology* (*a set is residual if it is the complement of a countable union of nowhere dense sets*).

REMARK. Conditions for the inverse Mañé projection to be Hölder continuous from PX to X are given by Eden, Foias, Nicolaenko and Temam [134], Foias and Olson [174], Hunt and Kaloshin [236].

A fundamental problem is to estimate the dimension of the attractor or, more generally, of an invariant set of a semigroup $\{S_t\}$. Methods for estimating its dimension essentially use the properties of the linear operators S'_t obtained by the differentiation of $S_t u_0$ with respect to u_0. When the operators $S_t u_0$ are Fréchet differentiable with respect to u_0, the differential $S'_t(u_0)$ is defined by the formula $v(t) = S'_t(u_0)v_0$ where $v(t)$ is the solution of the variation equation

$$\partial_t v(t) = \mathcal{F}'\big(u(t)\big)v(t), \qquad v(0) = v_0, \tag{18}$$

where $u(t) = S'_t u_0$; (18) is obtained by formal differentiation of (1) (see for examples and details [55]).

DEFINITION 1.5.2. Let H be a Banach space. An operator S is uniformly quasidifferentiable on a set $X \subset H$ and a linear operator $S'(u)$ is a quasidifferential of S on X at a point $u \in X$ if

$$\sup_{\substack{u,v\in X\\ 0<|u-v|\leqslant\epsilon}} \frac{\|Su - Sv - S'(u)(v-u)\|_H}{\|v-u\|_H} \to 0 \quad \text{as } \epsilon \to 0. \tag{19}$$

When X is an open set $S'(u)$ is the differential and S is called differentiable. In applications, $S'(u)$ usually coincides with the operator S' obtained by the formal differentiation of S. We call $S'(u)$ a quasidifferential because $v \in X$ may be not an arbitrary element of H. Note that in some cases estimate (19) holds for $u, v \in X \subset H$ (and not for arbitrary $u, v \in H$) since X in applications consists of more regular functions than a general function from H (for example, we may apply it when $H = H_0(\Omega)$ and X is bounded in $H_1(\Omega)$). We say that quasidifferentials $S'(u)$ are uniformly bounded on X when

$$\sup_{u\in X} \big\|S'(u)\big\|_{\mathcal{L}(H,H)} < \infty,$$

where $\|S'(u)\|_{\mathcal{L}(H,H)}$ is the operator norm.

The following theorem of Mañé [289] shows that under natural conditions a compact invariant set has a finite fractal dimension. Let $\mathcal{L}_1(H, H)$ be the set of bounded linear operators S' from H to H which can be split in the following way: $S' = S'_1 + S'_2$ where S'_1 is compact and $\|S'_2\| < 1$.

THEOREM 1.5.3. *Let $X \subset H$ be a compact, negatively invariant set of an operator S, $X \subset SX$, let S be differentiable in a neighborhood of X and the differential $S'(u)$ continuously depend on u. Let $S'(u) \in \mathcal{L}_1(H, H)$ for any $u \in X$. Then* $\dim_{\mathrm{F}}(X) < \infty$.

One of the central problems of the theory of global attractors is to estimate the dimension of the attractor in terms of parameters of the problem. The most precise estimations of Hausdorff and fractal dimension of invariant sets of semigroups which do not possess a global Lyapunov function are based on the study of the evolution of volume elements under the dynamics.

Let H be a Hilbert space. If $E_d \subset H$ is a d-dimensional linear subspace, then a linear operator L maps a d-dimensional ellipsoid $B_d \subset E_d$ into a d-dimensional ellipsoid $L(B_d) \subset L(E_d)$. In a Hilbert space, a volume $\mathrm{vol}_d(B_d)$ of a d-dimensional ellipsoid is well-defined. For a bounded operator L in a Hilbert space H, the quantity $\omega_d(L)$, which measures the changes of d-dimensional volumes under action of L, is defined as

$$\omega_d(L) = \sup_{B_d} \frac{\mathrm{vol}_d(L(B_d))}{\mathrm{vol}_d(B_d)},$$

the supremum being over all d-dimensional ellipsoids. If B_d is a ball, $\omega_d(L) = \alpha_1 \dots \alpha_d$, where α_j is the length of jth axis of the ellipsoid $L(B_d)$. Clearly, α_j^2 coincides with jth eigenvalue of the operator L^*L when the spectrum of L^*L is discrete. For details, see [55, 363]. We only note here that when L is compact, $\omega_d(L) \to 0$ as $d \to \infty$. A global version of $\omega_d(S'(u))$ is

$$\bar{\omega}_d = \sup_{u \in X} \omega_d\big(S'(u)\big). \tag{20}$$

We will give a sketch of the theory of estimates of fractal and Hausdorff dimension of global attractors of semigroups of PDE in Section 2.1.

Instability dimension of a semigroup. When a dynamical system is globally stable (that is it has a one-point global attractor) we have

$$\sup_{t \geqslant 0} \big\| S_t(u) - S_t(v) \big\| \to 0 \quad \text{if } \|u - v\| \to 0.$$

We say that the stabilization dimension of $\{S_t\}$ is not greater than d if there exists a d-dimensional subspace $E_{(d)}$ and a linear projection P_d in E onto $E_{(d)}$ such that

$$\sup_{t \geqslant 0} \big\| S_t(u) - S_t(v) \big\|_E \to 0 \quad \text{if } \|u - v\|_E + \sup_{t \geqslant 0} \big\| P_d\big(S_t(u) - S_t(v)\big) \big\|_E \to 0. \tag{21}$$

We call the minimum of such d that (21) holds *instability dimension* of the semigroup $\{S_t\}$ and denote it $\dim_S(\{S_t\})$.

The following theorem shows a relation between the stability properties of the semigroup and properties of its global attractor. It shows that the fractal dimension of the global attractor not only estimates the number of the degrees of freedom of the permanent regimes but also the number of unstable directions of the semigroup.

THEOREM 1.5.4. *Let H be a Banach space. Let $\{S_t\}$ be a semigroup in H with operators $S_t u$, $t \in [0, T]$, uniformly continuous in u on every bounded set for every T and $\{S_t\}$ possesses a global attractor $\mathcal{A}$ and $2\dim_F(\mathcal{A}) < d$ where d is an integer. Then* $\dim_S(\{S_t\}) \leqslant d$.

The proof follows from observations made by Babin and Vishik [49] (see also Theorem 8.1.2 of [55]) and from Mañé's Theorem 1.5.1. Namely, the attraction property of $\mathcal{A}$ implies that $\|S_t v - w_1(t)\|_E + \|S_t u - w_2(t)\|_E \leqslant \epsilon$ for $t \geqslant T(\epsilon)$, where $w_i(t) \in \mathcal{A}$. From continuity we see that $\|S_t v - S_t u\| \leqslant \epsilon$ when $t \leqslant T(\epsilon)$, $\|v - u\| \leqslant \delta(\epsilon)$. If $\sup_{t \geqslant 0} \|P_d(S_t(u) - S_t(v))\|_E \leqslant \epsilon$ then

$$\begin{aligned}
&\big\|P_d\big(w_1(t) - w_2(t)\big)\big\|_E \\
&\quad\leqslant \big\|P_d\big(S_t(u) - S_t(v)\big)\big\|_E + \|P_d\|\big(\big\|S_t v - w_1(t)\big\|_E + \big\|S_t u - w_2(t)\big\|_E\big) \\
&\quad\leqslant \big(2\|P_d\| + 1\big)\epsilon.
\end{aligned}$$

Since any one-to-one mapping P_d from a compact set $\mathcal{A}$ into $E_{(d)}$ has a continuous inverse we conclude that $\|w_1(t) - w_2(t)\| \leqslant \epsilon_2(\epsilon)$, $\epsilon_2(\epsilon) \to 0$ as $\epsilon \to 0$. Therefore for $t \geqslant T(\epsilon)$

$$\big\|S_t(u) - S_t(v)\big\|_E \leqslant \big\|w_1(t) - w_2(t)\big\|_E + \epsilon \leqslant \epsilon_2(\epsilon) + \epsilon$$

and for all $t \geqslant 0$

$$\begin{aligned}
&\big\|S_t(u) - S_t(v)\big\|_E \leqslant \epsilon_2(\epsilon) + \epsilon \\
&\quad\text{if } \|v - u\| \leqslant \delta(\epsilon),\ \sup_{t \geqslant 0}\big\|P_d\big(S_t(u) - S_t(v)\big)\big\|_E \leqslant \epsilon
\end{aligned}$$

which implies (21).

Entropy of infinite-dimensional sets. Sometimes attractors have infinite dimension. In particular, global attractors of equations in unbounded domains and attractors of nonautonomous equations may have infinite dimension. When the attractor has infinite dimension it still can be much thinner than a ball in a Banach space. One still can use Kolmogorov ε-entropy $\mathbb{H}_\varepsilon(K)$ to characterize a compact K even when it has infinite dimension. Estimates of Kolmogorov entropy of attractors of parabolic equations in unbounded domains are discussed in Section 5. Note that according to (17) the fractal dimension of K is finite when $\mathbb{H}_\varepsilon(K)$ tends to infinity not faster than $C\log_2(1/\varepsilon)$ with a finite C. When $\mathbb{H}_\varepsilon(K)$ tends to infinity faster one may introduce different characteristics of the growth or to write explicit estimates. In particular, following [257] and [98] one may introduce functional dimension $\mathrm{adf}(K)$ and metric order $\mathbf{q}(K)$

$$\mathrm{adf}(K) = \limsup_{\varepsilon \to 0+} \frac{\log_2 \mathbb{H}_\varepsilon(K)}{\log_2 \log_2(1/\varepsilon)}, \qquad \mathbf{q}(K) = \limsup_{\varepsilon \to 0+} \frac{\log_2 \mathbb{H}_\varepsilon(K)}{\log_2(1/\varepsilon)}. \tag{22}$$

1.6. *Fragmentation complexity of attractors*

Though a global attractor is a connected set, dynamics on it can be fragmented. We call a subset $X \subset \mathcal{A}$ a subattractor of $\mathcal{A}$ if X is a compact, stable, strictly invariant subset of $\mathcal{A}$ and there exists an open in topology of $\mathcal{A}$ invariant neighborhood $O(X) \subset \mathcal{A}$ of X which is invariant, $S_t O(X) \subset O(X)$, $t \geqslant 0$, and such that X attracts $O(X)$, namely $\delta_E(S_t O(X), X) \to 0$ as $t \to \infty$. For example, if $z \in \mathcal{A}$ is a stable equilibrium point, $\{z\}$ is a subattractor. Another examples include a stable cycle (a periodic trajectory), or a stable invariant torus. Now we define an intrinsic characteristic of the dynamics on $\mathcal{A}$, namely the *fragmentation number* $\mathsf{Fr}(\mathcal{A})$.

DEFINITION 1.6.1. Let $\vec{X} = \{X_j,\, j = 1, \ldots, N\}$ be a collection of subattractors of a global attractor $\mathcal{A}$, $X_j \subset \mathcal{A}$, $X_j \cap X_i = \emptyset$ when $j \neq i$, such a collection is called a subfragmentation of $\mathcal{A}$ of rank N. The fragmentation number $\mathsf{Fr}(\mathcal{A}) \leqslant \infty$ is a supremum of ranks N of all possible subfragmentations of $\mathcal{A}$. We call $\log_2 \mathsf{Fr}(\mathcal{A})$ *fragmentation complexity* of the attractor $\mathcal{A}$.

Obviously, one can take $X_1 = \mathcal{A}$, therefore $\mathsf{Fr}(\mathcal{A}) \geqslant 1$, $\log_2 \mathsf{Fr}(\mathcal{A}) \geqslant 0$.

REMARK. Note that a global attractor $\mathcal{A}$ of a semigroup in a metric space may have a high dimension and have zero fragmentation complexity when there exists a trajectory which is everywhere dense on the attractor. Though the temporal behavior of a solution may be complex, the attractor can be recovered with a high precision using only one trajectory. When the fragmentation complexity is large, one has to use many trajectories.

The following theorem follows from the invariance principle of La Salle which holds for semigroups with a Lyapunov functions (see Section 3.2 for the definitions).

THEOREM 1.6.2. *If $\{S_t\}$ has a global attractor $\mathcal{A}$ and a global Lyapunov function $\mathcal{L}$ which is continuous on the attractor and the set of equilibria is finite, then the fragmentation number* $\mathsf{Fr}(\mathcal{A})$ *coincides with the number of stable equilibria* (*local minima of* $\mathcal{L}$).

We give an elementary example when $\mathsf{Fr}(\mathcal{A}) = \infty$. An ordinary differential equation $\partial_t u = f(u)$ with

$$f(u) = -u^4 \sin\left(\frac{1}{u}\right) \quad \text{when } u \neq 0, \qquad f(0) = 0 \tag{23}$$

generates a semigroup in $\mathbb{R}$. The semigroup has the global attractor $\mathcal{A} = [-\frac{1}{\pi}, \frac{1}{\pi}]$. There are infinitely many stable points $z \in \mathcal{A}$ and $N(\mathcal{A}) = \infty$. Note though that a generic, arbitrary small $\mathcal{C}^1$ perturbation of $f(u)$ in (23) makes $\mathsf{Fr}(\mathcal{A})$ finite.

The above example shows that a robust characteristic of the complexity should include families of equations.

For a family of attractors $\mathcal{A}(\theta)$ where $\theta \in \Theta$ is a parameter (for example, $\theta = f$ where f is the function which determines the non-linearity) from a parameter space Θ we introduce the fragmentation number and complexity

$$\mathsf{Fr}\big(\mathcal{A}(\Theta)\big) = \inf_{\theta\in\Theta} \mathsf{Fr}\big(\mathcal{A}(\theta)\big), \qquad \mathrm{L}\big(\mathcal{A}(\Theta)\big) = \log_2 \mathsf{Fr}\big(\mathcal{A}(\Theta)\big). \tag{24}$$

Estimates from below of the fragmentation number for gradient systems of PDE are given in [17,18], lower bounds for the fragmentation complexity of the global attractor of a reaction–diffusion equation in a large domain $\Omega \subset \mathbb{R}^d$ for a given non-linearity are given in [24,3] and in Section 4 of this chapter, namely $\log_2 \mathsf{Fr}(\mathcal{A}) \geqslant c|\Omega|$, where $|\Omega|$ is d-dimensional volume of Ω.

If Ω_N, $N = 1, 2, \ldots$, is a one-parameter family of domains, $\Omega_N \subset \Omega_{N'}$ when $N < N'$ and $\bigcup_N \Omega_N = \mathbb{R}^d$, and the attractors $\mathcal{A}(\theta) = \mathcal{A}(\theta, \Omega_N)$ lie in a space of functions over Ω_N, we introduce the *average spatial complexity* of the attractor

$$\mathsf{cmp}(\theta) = \lim_{N\to\infty} \inf \frac{\log_2 \mathsf{Fr}(\mathcal{A}(\theta, \Omega_N))}{|\Omega_N|}, \qquad \mathsf{cmp}(\Theta) = \lim_{N\to\infty} \inf \frac{\log_2 \mathsf{Fr}(\mathcal{A}(\Theta, \Omega_N))}{|\Omega_N|}, \tag{25}$$

where $|\Omega_N|$ is the volume of the domain Ω_N.

Below in Section 4 we give estimates from below $\mathsf{cmp}(\Theta) \geqslant c > 0$ and estimates from above $\mathsf{cmp}(\Theta) \leqslant C$ for the average spatial complexity of attractors of parabolic and hyperbolic equations.

1.7. *Dependence on parameters*

Under very mild conditions a global attractor *upper semicontinuously* depends on parameters that are involved in the equations. First results on upper semicontinuous dependence on a parameter of global attractors for semigroups with a global Lyapunov function were obtained by Hale [208]. For general semigroups upper semicontinuity of global attractors was proven first by Babin and Vishik [46], similar results were obtained by Hale [209], Hale and Raugel [215]. For details and further results see [55,209,336]. Here we follow [46] (in a less general setting for simplicity), since the approach of [46] is geometrically transparent. Namely, the upper semicontinuity can be derived from the following principle: if a compact set lies in the product $E \times \Theta_0$, then its section $E \times \{\theta\}$, $\theta \in \Theta_0$, depends on θ upper semicontinuously. This observation is used to prove that $\mathcal{A}(\theta)$ upper semicontinuously depends on a parameter θ in the following way. If Equation (1) depends on a parameter θ,

$$\partial_t u = \mathcal{F}(u, \theta) \tag{26}$$

we have a semigroup $\{S_t(\theta)\}$ that depends on the parameter θ. We assume that parameter θ takes values in a compact subset Θ_0 of a metric space Θ and consider the action of S_t in

the product space $\widetilde{E} = E \times \Theta$. One may consider (u, θ) as a new variable and introduce a new semigroup $\{\widetilde{S}_t\}$ in the product space:

$$\widetilde{S}_t : (u, \theta) \mapsto \big(S_t(\theta)u, \theta\big).$$

Abusing notation we say that $S_t(\theta_0)$ is defined on $E \times \theta_0$ as well as on E.

The following general result of [46] shows that under natural assumptions global attractors $S_t(\theta)$ upper semicontinuously depend on θ.

THEOREM 1.7.1. *Let $\widetilde{S}_t$ have a global attractor $\widetilde{\mathcal{A}} \subset E \times \Theta_0$. Then the sets $\mathcal{A}(\theta_0) = \widetilde{\mathcal{A}} \cap \{\theta = \theta_0\}$ are global attractors of semigroups $S_t(\theta_0)$, and the sets $\mathcal{A}(\theta_0)$ upper semicontinuously in E depend on $\theta_0 \in \Theta_0$, that is*

$$\delta_E\big(\mathcal{A}(\theta), \mathcal{A}(\theta_0)\big) \to 0 \quad \text{as } \theta \to \theta_0. \tag{27}$$

Existence of the global attractor $\widetilde{\mathcal{A}}$ can be proven based on Theorems 1.1.3 and 1.1.4. The main condition of these theorems is the existence of a compact attracting set. We call a set B_0 uniformly (with respect to θ_0) attracting if for any bounded set $B \subset E$ and any open neighborhood $O(B_0)$ of B_0 there exists $T = T(B)$ such that for all $t \geqslant T(B)$, $\theta_0 \in \Theta_0$

$$S_t(\theta_0)\big(X(\theta_0) \cap B\big) \subset O(B_0).$$

Since Θ_0 is compact, if B_0 is uniformly attracting compact set for $S_t(\theta_0)$ then $B_0 \times \Theta_0$ is a compact attracting set for $\widetilde{S}_t$. From Theorem 1.1.3 we obtain the following result of [46].

THEOREM 1.7.2. *Assume that there exists a uniformly attracting compact set B_0. Assume also that $\widetilde{S}_t$ are continuous on $E \times \Theta_0$ and uniformly bounded on a neighborhood of B_0. Then there exists the global attractor $\widetilde{\mathcal{A}}$ of $\widetilde{S}_t$. For every $\theta \in \Theta_0$ there exists a global attractor $\mathcal{A}(\theta)$ of the semigroup $S_t(\theta)$ and $\mathcal{A}(\theta)$ upper semicontinuously depends on θ.*

The case of a sequence of equations with non-linearities $\mathcal{F}_n(u)$, $n = 1, \ldots$, which converge to a limit non-linearity $\mathcal{F}_n(u) \to \mathcal{F}_\infty(u)$ when $n \to \infty$ can be put in the above framework by setting $\theta_{0n} = \frac{1}{n}$, $\theta_\infty = 0$ $\mathcal{F}(u, \frac{1}{n}) = \mathcal{F}_n(u)$; in this case the compact set of parameters $\Theta_0 = \bigcup_{n=1}^{\infty} \{\frac{1}{n}\} \cup 0$.

Theorem 1.7.1 is sufficient if one needs to prove upper semicontinuous dependence on parameters for applications to strongly non-linear and semilinear parabolic equations and systems, damped hyperbolic problems and other problems, see [46,55] for details. These problems include, in particular, the Galerkin approximations of order N to the 2D Navier–Stokes system. It is proven by Babin and Vishik in [46] that the attractor $\mathcal{A}(N)$ of the Galerkin system upper semicontinuously depends on N when $N \to \infty$,

$$\delta_{H_1}\big(\mathcal{A}(N), \mathcal{A}(\infty)\big) \to 0 \quad \text{as } N \to \infty. \tag{28}$$

Here $\mathcal{A}(\infty)$ is the attractor of the 2D Navier–Stokes system; see [46,55] for details. Similar results for approximations of parabolic and damped hyperbolic equations were obtained by Hale, Lin and Raugel [213].

The case when the dependence on a parameter is singular is more difficult. As an example we consider the damped wave equation

$$\epsilon\partial_t^2 u + \gamma\partial_t u = \Delta u - f(u) - g(x), \qquad u|_{\partial\Omega} = 0 \tag{29}$$

with $\epsilon < 0$, when $\epsilon \to 0$ the hyperbolic equation (29) turns into a parabolic equation. The semigroup generated by (29) written in the form of a system similarly to (101) acts on two-component vectors (u, p). Its attractor $\mathcal{A}_\epsilon$ lies in the space E_1 defined in (102). When $\epsilon = 0$ the parabolic equation describes dynamics of u-component only, its attractor $\mathcal{A}$ lies in the space $u \in W_2^2(\Omega) \cap \{u|_{\partial\Omega} = 0\}$. Nevertheless, one can express $\partial_t u$ from the parabolic equation and set for $u \in \mathcal{A}$ $p = \frac{1}{\gamma}[\Delta u - f(u) - g]$, the pairs (u, p) lie in E_1. If we denote the set of pairs by $\mathcal{A}_0$ the following upper semicontinuity holds:

$$\delta_{E_{1-\epsilon}}(\mathcal{A}_\epsilon, \mathcal{A}_0) \to 0 \quad \text{as } \epsilon \to +0.$$

For more details see [50,215,55,336].

Non-trivial examples of upper semicontinuity in the case of a singular dependence on a parameter include results on dependence on the shape of domains when the domain is thin and changes its dimension in the limit. Upper semicontinuity for such problems is proven by Hale and Raugel [216,217,219,221]. More examples of singularly perturbed systems that couple hyperbolic and parabolic PDE or PDE and ODE are studied by Vishik and Skvortsov [372,373,357], Fitzgibbon, Parrott and You [166].

Lower semicontinuous dependence on parameters. Since the distance δ_E from one set to another is not symmetric, (27) does not imply that $\delta_E(\mathcal{A}(\theta_0), \mathcal{A}(\theta)) \to 0$ as $\theta \to \theta_0$, that is the dependence may be not lower semicontinuous. Lower semicontinuous dependence of attractors on a parameter is proven only for semigroups that possess a global Lyapunov function and have only hyperbolic equilibria, see [46,54,55,209,336]. Easily verifiable conditions that guarantee lower semicontinuity for general semigroups are not yet known. It is proven though that the lower semicontinuity is a generic property, see Babin and Pilyugin [35].

As Theorem 1.7.1 shows, upper semicontinuity requires very mild assumptions on the semigroup, so one usually may take for granted that $\mathcal{A}(\theta)$ depends on θ upper semicontinuously. When lower semicontinuity also holds, one has a continuous dependence on θ, that is

$$\delta_E\big(\mathcal{A}(\theta), \mathcal{A}(\theta_0)\big) + \delta_E\big(\mathcal{A}(\theta_0), \mathcal{A}(\theta)\big) \to 0 \tag{30}$$

as $\theta \to \theta_0$. Global attractors of semigroups that possess a global Lyapunov function and have only hyperbolic equilibria Holder continuously depend on θ, namely

$$\delta_E\big(\mathcal{A}(\theta), \mathcal{A}(\theta_0)\big) + \delta_E\big(\mathcal{A}(\theta_0), \mathcal{A}(\theta)\big) \leqslant C|\theta_0 - \theta|^q, \quad \eta > 0.$$

This fact is proven in [46], it directly follows from the non-trivial exponential attraction estimate (56) of [46] (see Theorem 2.3.9) to the attractor, and from the estimate

$$\big\| S_t(\theta)u - S_t(\theta_0)u_0 \big\|_E \leqslant C\big(\|u\|_E + \|u_0\|_E\big)\big(\|u - u_0\| + |\theta_0 - \theta|\big)e^{\alpha t}.$$

An elementary computation shows that $q = \eta/(\eta+\alpha)$ with η from (56), see for details [55].

1.8. *Toy models*

In this subsection we discuss several toy models which are much simpler than the equations we study later but still can be used to illustrate properties of global attractors.

Toy model 1.8.1. We consider a very simple ODE that generates a semigroup in $\mathbb{R}^1$

$$\partial_t y = y\big(1 - y^2\big). \tag{31}$$

Obviously,

$$\frac{1}{2}\partial_t |y|^2 = |y|^2 - |y|^4.$$

This equation implies that $\partial_t |y|^2 \leqslant 0$ when $|y|^2 \geqslant 1$. Therefore every segment $B_R = \{y\colon -R \leqslant y \leqslant R\}$ with $R > 1$ is an absorbing set for $\{S_t\}$ generated by (31) in $\mathbb{R}^1$. The semigroup has three equilibria $z_1 = -1$, $z_2 = -0$, $z_3 = 1$. The global attractor $\mathcal{A}$ of the semigroup coincides with the segment $B_1 = \{y\colon -1 \leqslant y \leqslant 1\}$. Note that every trajectory $y(t)$ of the semigroup tends to one of the three points z_i as $t \to \infty$. The attractor is a larger set since $S_t(B_R)$ with $R > 1$ must include points ± 1 and since $S_t(B_R)$ is a connected set it must include the whole segment B_1. The dimension of $\mathcal{A}$ is 1, the fragmentation number $\mathsf{Fr}(\mathcal{A}) = 2$. A similar equation

$$\partial_t y = -y\big[\big(1 - y^2\big)^2 + \theta\big], \quad \theta \geqslant 0, \tag{32}$$

demonstrates the general property of global attractors: the upper semicontinuous dependence of the global attractor $\mathcal{A}(\theta)$ on the parameter θ. When $\theta > 0$ the global attractor is one point $y = 0$. When $\theta = 0$ the global attractor is the segment $\{y\colon -1 \leqslant y \leqslant 1\}$. Obviously $\delta_{\mathbb{R}}(\mathcal{A}(\theta), \mathcal{A}(0)) \to 0$ when $\theta \to 0$ (in fact, $\delta_{\mathbb{R}}(\mathcal{A}(\theta), \mathcal{A}(0)) = 0$) but $\delta_{\mathbb{R}}(\mathcal{A}(0), \mathcal{A}(\theta)) = 1$ for $\theta > 0$.

Toy model 1.8.2. We consider in the plane $\mathbb{R}^2$ the Van der Pol system

$$\partial_t u = u + \beta u^{\perp} - |u|^2 u, \tag{33}$$

where $u = (x, y) \in \mathbb{R}^2$, $|u|^2 = (x^2 + y^2)$, $u^{\perp} = (-y, x)$, $\beta \geqslant 0$.

Multiplication of (33) by u yields the equation for the norm $|u|$

$$\frac{1}{2}\partial_t |u|^2 = |u|^2 - |u|^4. \tag{34}$$

This equation implies that $\partial_t |u|^2 \leqslant 0$ when $|u|^2 \geqslant 1$. Therefore every ball $B_R = \{u\colon |u| \leqslant R\}$ with $R > 1$ is an absorbing set for $\{S_t\}$ generated by (33) in $\mathbb{R}^2$. Note that (34)

implies that the circle $|u|^2 = 1$ is a strictly invariant set, it is a local attractor. The global attractor $\mathcal{A}$ of the semigroup coincides with the disc B_1

$$\mathcal{A} = \{u\colon\ |u| \leqslant 1\}.$$

When $\beta \neq 0$ there is one unstable equilibrium $u = 0$. When $\beta = 0$ the set of equilibria consists of $u = 0$ and of the circle $|u| = 1$. When $\beta = 0$ the semigroup $\{S_t\}$ has a global Lyapunov function

$$F(u) = -\frac{1}{2}|u|^2 + \frac{1}{4}|u|^4.$$

This function satisfies the inequality

$$\partial_t F(u) = -\big(|u|^2 - |u|^4\big)^2 \tag{35}$$

and $F(u)$ is strictly decreasing when $(|u|^2 - |u|^4)^2 \neq 0$. This is true for $\beta \neq 0$ too, but in the latter case the stationary set of $F(u)$, that is set $\{u\colon\ |u|^2 - |u|^4 = 0\}$ does not consist of equilibria and in this case we do not call such a Lyapunov function a global Lyapunov function. The dimension of the attractor is 2, the fragmentation number $\mathsf{Fr}(\mathcal{A}) = 1$, the fragmentation complexity $\log \mathsf{Fr}(\mathcal{A}) = 0$.

Toy model 1.8.3. To illustrate the introduced concepts in an infinite-dimensional situation we consider a model equation

$$\partial_t u = -A_0 u - f_0(u), \tag{36}$$

where A_0 is a self-adjoint operator in a Hilbert space H with the orthonormal basis e_j

$$A_0 e_j = \lambda_j e_j, \quad j = 1, \dots,$$

and for every $u \in H$ we have the eigenvector expansion

$$u = \sum_{j=1}^{\infty} u_j e_j.$$

One may take $A_0 = -\Delta + c$ where Δ is the Laplace operator with appropriate boundary conditions, $c > 0$ is a constant. The numeration of eigenvalues is in increasing order, $\lambda_{j+1} \geqslant \lambda_j$, $\lambda_j \to \infty$ as $j \to \infty$. Let N_0 be the number of non-positive eigenvalues

$$\lambda_{N_0} \leqslant 0, \qquad \lambda_{N_0+1} > 0.$$

The action of the non-linearity f_0 in the eigenbasis is written as follows

$$f_0(u) = \sum_{j=1}^{\infty} u_j^3 e_j.$$

Equation (36) can be written in the basis e_j in the form

$$\partial_t u_j = -\lambda_j u_j - u_j^3, \quad j = 1, \ldots,$$

which is similar to (31). A solution of (36) is given by

$$u(t) = \sum_{j=1}^{\infty} u_j(t) e_j.$$

It can be split into two parts

$$u(t) = u_+(t) + u_-(t),$$
$$u_+(t) = \sum_{j=1}^{N_0} u_j(t) e_j, \qquad u_-(t) = \sum_{j=N_0+1}^{\infty} u_j(t) e_j.$$

We have the inequality

$$\|u_-(t)\|_H \leqslant \|u(0)\|_H \exp(-\lambda_{N_0+1} t).$$

This estimate is uniform when initial data $u(0)$ belong to a bounded set B in H. In the invariant linear subspace E_{N_0} with the basis $e_1, \ldots, e_{N_0}$ the global attractor $\mathcal{A}_+ \subset E_{N_0}$ is determined by the inequalities

$$|u_j| \leqslant \sqrt{-\lambda_j}, \quad j = 1, \ldots, N_0.$$

One can easily estimate the attraction rate:

$$\delta\big(u_+(t), \mathcal{A}_+\big) \leqslant C\big(\|u_+(0)\|_H\big) e^{-2\lambda_{N_0} t} \quad \text{when } \lambda_{N_0} > 0, \tag{37}$$

$$\delta\big(u_+(t), \mathcal{A}_+\big) \leqslant C\big(\|u_+(0)\|_H\big) \frac{1}{1+\sqrt{t}} \quad \text{when } \lambda_{N_0} = 0. \tag{38}$$

The attractor $\mathcal{A} \subset H$ of (36) is given by

$$\mathcal{A} = \{u \in H\colon\ u_- = 0, u_+ \in \mathcal{A}_+\}.$$

The dimension of the attractor $\dim_H \mathcal{A} = \dim_F \mathcal{A} = N_0$, the fragmentation number $\mathsf{Fr}(\mathcal{A}) = 2^{N_0}$. When $\lambda_{N_0} > 0$ the following exponential attraction property holds:

$$\delta_H\big(S_t(B), \mathcal{A}\big) \leqslant C(B) \exp(-\eta t), \quad \text{where } \eta = \min(\lambda_{N_0+1}, 2\lambda_{N_0}), \tag{39}$$

for every bounded set B; when $\lambda_{N_0} = 0$ the rate of attraction is algebraic like in (38). Therefore, when $\lambda_{N_0} > 0$ $\mathcal{A}$ is an *exponential attractor*.

One may drastically increase the rate of convergence to an attractor slightly increasing it but losing the strict invariance. We take a larger subspace with the basis $e_1, \dots, e_{N_1}$, $N_1 > N_0$ and take a neighborhood $O_r(\mathcal{A}_+)$ in E_{N_0}. We set

$$\mathcal{E} = O_r(\mathcal{A}_+) + \sum_{j=N_0+1}^{N_1} u_j e_j, \qquad \sum_{j=N_0+1}^{N_1} |u_j|^2 \leqslant r^2, \quad \mathcal{E} \subset H,$$

with an arbitrary small r. Now we have

$$\delta_H\big(S_t(B), \mathcal{E}\big) \leqslant C(B)\exp(-\eta t), \quad \text{where } \eta = \lambda_{N_1+1}.$$

The set $\mathcal{E}$ is invariant, $S_t\mathcal{E} \subset \mathcal{E}$ but it is not strictly invariant, $S_t\mathcal{E} \neq \mathcal{E}$. This toy example shows that by expanding the global attractor a little and increasing its dimension one may drastically increase the rate of convergence to it and obtain an exponential attractor. A non-trivial generalization of this idea leads to the proof of existence of an exponential attractor of a dynamical system, see [134].

We use the set $\mathcal{A}$ constructed in this example as an illustration to one more important concept in the theory of dynamical systems, namely unstable and stable manifolds through an equilibrium point of a semigroup. Clearly, if $\lambda_{N_0} < 0$ the subspace $E_{N_0} = E_{0+}$ is the unstable manifold passing through zero of the linear semigroup S_t generated by (36). When $\lambda_{N_0} = 0$ the subspace E_{N_0} is called the center-unstable manifold of S_t through zero.

The center-unstable manifold of this semigroup is non-zero if $\lambda_1 \leqslant 0$, it has the basis $e_1, \dots, e_{N_0}$ with $\lambda_1, \dots, \lambda_{N_0} \leqslant 0$, $\lambda_{N_0+1} > 0$. The stable manifold of the linear semigroup through zero is a linear subspace with the basis $e_{N_0+1}, \dots$.

This toy model illustrates the simplest properties of global attractors: the set $\mathcal{A}$ is an attracting set, it is compact; moreover, it is finite-dimensional. It is invariant, that is on $\mathcal{A}$ every trajectory $u(t)$ can be extended to $-\infty < t < +\infty$. When $A_0 = -\Delta + cI$ where Δ is a negative operator, for $c = 0$ the attractor consists of only one point zero. The non-trivial attractor $\mathcal{A}$ is a result of the perturbation cI, with sufficiently large $c > 0$ that creates instability. Of course non-linear PDE are much more complicated than this toy model. Still for many semilinear cases one may consider their global attractor as a result of perturbation of the linear equation by the non-linear term and an external forcing. Finite-dimensional non-linear models (for example, the Lorenz system) show that the structure of such a set may be very complicated and this set does not look like a smooth manifold neither locally nor globally. Nevertheless, many observations are still true. The global attractor is a compact, finite-dimensional (in the sense which will be discussed below) set. This set is invariant, that is dynamics on the attractor is invertible. And this set uniformly attracts all solutions of the dynamical problem, with bounded initial data.

Toy model 1.8.4. Consider a system of ODE in the plane similar to (33) (we use the same notation)

$$\partial_t u = -\big(|u|^2 - 1\big)^2 u + \big(|u|^2 - 1\big)u^{\perp}. \tag{40}$$

The global attractor $\mathcal{A}$ of this system is a disc $|u| \leqslant 1$. Since

$$\partial_t |u|^2 = -2\big(|u|^2 - 1\big)^2 |u|^2,$$

every solution $u(t)$ with $|u(t)| > 1$ tends to $\mathcal{A}$, $|u(t)| \to 1$ as $t \to +\infty$. But the attraction is slow, $|u|^2 - 1 \geqslant |u(0)|^2 (2t + C)^{-1}$ when $t \to \infty$. Note that every point on the circle $\{|u| = 1\}$ is an equilibrium. At the same time, the polar angle θ of every solution satisfies $\partial_t \theta = |u|^2 - 1$. A simple analysis shows that the variation of the angle θ of a solution with $|u(t)| > 1$ is unbounded when $t \to \infty$ since

$$\int_0^\infty \big(|u|^2 - 1\big)\, dt \geqslant \big|u(0)\big|^2 \int_1^\infty (2t + C)^{-1}\, dt = \infty.$$

Therefore the omega-limit set of every non-equilibrium trajectory $u(t)$ coincides with the whole circle $\{|u| = 1\}$. Therefore $\mathsf{Fr}(\mathcal{A}) = 1$. Note that (40) has a Lyapunov function $\mathcal{L}(u) = |u|^2$.

2. Properties of attractors

2.1. *Upper and lower estimates of Hausdorff and fractal dimension*

Here we present theorems on the Hausdorff and fractal dimensions of strictly invariant sets. According to the definition of the dimension, to estimate it one has to use coverings of $\mathcal{A}$ by small balls. The possibility to prove the finite dimensionality of the global attractor for a semigroup in an infinite-dimensional space, in particular, for a semigroup corresponding to the 2D Navier–Stokes system, is based on three fundamental facts. The first fact is the invariance property $S_t \mathcal{A} = \mathcal{A}$ (or negative invariance $\mathcal{A} \subset S_t \mathcal{A}$) which implies that if one has a covering of $\mathcal{A}$ by a family of balls $B_r + u_j$ then $S_t(B_r + u_j)$ is also a covering of $\mathcal{A}$. The second fact is the differentiability of the operators $S_t u$ with respect to u. Thanks to the differentiability, one can locally approximate the action of S_t on a ball $B_r + u_j$ by $S_t'(u_j) B_r$. To describe the third fact we consider here the case of a semilinear operator $\mathcal{F}(u) = \nu A u + \mathcal{F}_0(u) - g$ in Equation (1) which takes the form

$$\partial_t u + \nu A u + \mathcal{F}_0\big(u(t)\big) = g, \tag{41}$$

where A is a linear operator, g is a given forcing term. The non-linear part $\mathcal{F}_0$ is assumed to be a differentiable in some sense operator, $\mathcal{F}_0'$ being its differential and

$$\mathcal{F}' = \nu A + \mathcal{F}_0'. \tag{42}$$

The differential $S_t'(u_0) v_0 = v(t)$ satisfies the variation equation (18), which for the semilinear equation takes the form

$$\partial_t v + \nu A v + \mathcal{F}_0'\big(u(t)\big) v = 0, \qquad v(0) = v_0, \tag{43}$$

with $u(t) = S_t u_0$. The third fact is that in many cases the linear operators $S'_t(u_0)$ have the smoothing property and therefore are compact (though it is not necessary for dimension estimates). The eigenvalues $\beta_1^{2t}, \beta_2^{2t}, \dots$ of the operator $S_t'^* S_t$ are positive and, since S'_t is compact, $\beta_n \to 0$ as $n \to \infty$. Therefore for a d-dimensional measure $\mu_d(S'_t(u_0)B_r) \leqslant C\beta_1^t\beta_2^t \cdots \beta_d^t \mu_d(B_r)$ and, if d is large, $\mu_d(S'_t(u_0)B_r) \leqslant C\beta^t \mu_d(B_r)$ with $\beta < 1$. Therefore, for large t, $\mu_d(S'_t(u_0)B_r) \leqslant \mu_d(B_r)/4$ and, if the linear approximation works for small r, we get $\mu_d(\mathcal{A}) = \mu_d(S_t\mathcal{A}) \leqslant \mu_d(\mathcal{A})/2$ for large t. This is possible only for $\mu_d(\mathcal{A}) = 0$. This argument shows that d cannot be very large, therefore we have an estimate of the dimension of $\mathcal{A}$, $d \leqslant d^*$. The ideas sketched above can be made absolutely rigorous; this was proven by Mallet-Paret [286].

The following theorem that allows to get quantitative estimates is proven by Douady and Oesterle [130].

THEOREM 2.1.1. *Let $SX = X$, and let S be uniformly quasi-differentiable on X, with uniformly bounded on X quasidifferentials $S'(u)$. Let d be such that for some $k < 1$ the quantity $\bar{\omega}_d$ defined by* (20) *satisfies the inequality*

$$\bar{\omega}_d \leqslant k < 1. \tag{44}$$

Then the Hausdorff dimension of X is finite and is not greater than d.

Theorems on the fractal dimension were proven by Constantin, Foias and Temam [120], see also Constantin and Foias [117]. We give here the result of Chepyzhov and Ilyin [95].

THEOREM 2.1.2. *Under the hypotheses of Theorem* 2.1.1 *suppose that the quasidifferentials $S'(u)$ continuously depend on $u \in X$ in the operator norm. Then the fractal dimension of X is not greater than d.*

The statements of Theorems 2.1.1 and 2.1.2 can be illustrated as follows. Inequality (44) means that the mapping S strictly decreases the d-dimensional Hausdorff measure with a coefficient $k < 1$. Since $SX = X$, this is possible only when the Hausdorff measure of X is zero. This sketch can be made rigorous (see [55,363]).

If $S'_t(u_0)$ is the solution operator for the variation equation (18), then using a Liouville type formula for solutions of linear equations of the form (18), we get the estimate

$$\omega_d\big(S'_t(u_0)\big) \leqslant \sup_{E_d} \exp\left[\int_0^t \operatorname{tr}\big(\mathcal{F}'\big(u(\tau)\big)\Pi_{E_d(\tau)}\big)\,d\tau\right], \tag{45}$$

where $\Pi_{E_d(\tau)}$ is the orthoprojection in H onto $E_d(\tau) = S'_\tau(u_0)E_d$, the supremum being taken over all d-dimensional subspaces and $\operatorname{tr}(F'(u(\tau))\Pi_{E_d(\tau)})$ being the trace of the finite-dimensional operator $\mathcal{F}'_0(u(\tau))\Pi_{E_d(\tau)}$.

From (44) and (45), we obtain that $\dim_{\mathrm{H}} X \leqslant d$ if, for some $t > 0$ and $k < 1$, for all trajectories $u(t)$ on X

$$\sup_{E_d}\left[\int_0^t \operatorname{tr}\big(\mathcal{F}'\big(u(\tau)\big)\Pi_{E_d(\tau)}\big)\,d\tau\right] \leqslant \ln k < 0. \tag{46}$$

Note that by (42)

$$\operatorname{tr}\bigl(\mathcal{F}'\bigl(u(\tau)\bigr)\Pi_{E_d(\tau)}\bigr) = -\nu \operatorname{tr}(A\Pi_{E_d(\tau)}) + \operatorname{tr}\bigl(\mathcal{F}_0' u(\tau)\Pi_{E_d(\tau)}\bigr), \tag{47}$$

where A is the linear operator. For a second-order elliptic operator A in a bounded domain $\Omega \subset \mathbb{R}^N$ the typical estimate of eigenvalues and the trace takes the form

$$\lambda_j \geqslant C_A \lambda_1 j^{2/N}, \quad \operatorname{tr}(A\Pi_{E_d(\tau)}) \geqslant C_A' \lambda_1 d^{1+2/N}, \quad C_A' > 0. \tag{48}$$

A typical estimate (see [25,55,363]) of the trace of the operator $S_t'(u_0)(t)$ generated by (43), where $u(t) = S_t u_0$ takes values in the attractor $\mathcal{A} = X$, has the form

$$\limsup_{t\to\infty} \frac{1}{t} \int_0^t \operatorname{tr}\bigl(\mathcal{F}'\bigl(u(\tau)\bigr)\Pi_{E_d(\tau)}\bigr)\, d\tau \leqslant b_{\mathcal{A}} - \nu C_A' \lambda_1 d^{1+2/N}. \tag{49}$$

Here, $b_{\mathcal{A}}$ is a constant that depends on $\mathcal{F}_0'$, on the attractor and the parameters of the problem but not on a particular solution $u(t)$ on the attractor. Therefore, from (45) we get the estimate

$$\limsup_{t\to\infty} \sup_{u_0\subset X} \ln \omega_d\bigl(S_t'(u_0)\bigr) \leqslant -\nu C_A' \lambda_1 d^{1+2/N} + b_{\mathcal{A}}.$$

When d is large enough, namely when

$$\nu C_A' \lambda_1 d^{1+2/N} > b_{\mathcal{A}} \tag{50}$$

(we can take the minimum integer d that satisfy this condition) we have (46),

$$\limsup_{t\to\infty} \sup_{u_0\subset X} \ln \omega_d\bigl(S_t'(u_0)(t)\bigr) < 0$$

and condition (44) holds.

One can apply Theorems 2.1.1 and 2.1.2 to $S(t)$, $t \to \infty$. In addition, the right-hand side of (49) is convex with respect to d. This allows to apply the observation of Chepyzhov and Ilyin [94] that (50) is sufficient for both Hausdorff and fractal dimension to be not greater than d. (See Chepyzhov and Vishik [98] for optimization of estimates of fractal dimension.) Therefore, we obtain the following theorem.

THEOREM 2.1.3. *Let $S_t X = X$, S_t be quasidifferentiable on X and, for every t, the quasidifferentials $S_t'(u_0)$ be uniformly bounded on X. Let an integer d satisfy* (50). *Then the Hausdorff dimension of X is not greater than d. The fractal dimension of X is not greater than d.*

REMARK. The above results can be extended to non-integer values of d, see [55,94,363].

REMARK. Estimates of dimension of attractors can be made in terms of the Lyapunov exponents of the operators $S'_t(u_0)(t)$. We define, for $j \geqslant 2$, a global jth Lyapunov exponent by

$$\mu_j = \ln(\Lambda_j), \quad \Lambda_j = \lim_{t\to\infty}\left[\frac{\bar{\omega}_j(t)}{\bar{\omega}_{j-1}(t)}\right]^{1/t}$$

(see [363] for more detail). Condition (44) then is replaced by the following condition:

$$\mu_1 + \cdots + \mu_{n+1} < 0. \tag{51}$$

Lower bounds of dimension of the global attractors are based on the following observation of Babin and Vishik [39].

THEOREM 2.1.4. *Let $\mathcal{A}$ be the global attractor of a semigroup $\{S_t\}$, let $z \in \mathcal{A}$ be an equilibrium. Assume that for some $\rho \geqslant 1$ there exists a local invariant manifold $M_+(z, \rho)$ (see Subsection* 1.2) *which is tangent to the invariant subspace $E_+(\rho)$ of the linear semigroup $S'_t(z)$. Then Hausdorff and fractal dimension of $\mathcal{A}$ are not less than* $\dim(E_+(\rho))$.

2.2. *More aspects of finite dimensionality*

Parametrization of attractors. The first statements related to finite-dimensional parametrization of attractors of PDE were given by Foias and Prodi [175] and Ladyzhenskaya [262,261]. Foias and Prodi [175] proved that the asymptotic behavior as $t \to \infty$ of the solutions of the two-dimensional Navier–Stokes system in many cases is determined by the asymptotic behavior of their finite-dimensional projections. Ladyzhenskaya [262,261] constructed the attractor of the two-dimensional Navier–Stokes system and proved that a trajectory on the attractor is completely determined by its orthogonal projection onto the space spanned by the first n eigenfunctions of the Stokes operator if n is large enough.

There are several ways to parametrize attractors. The first one is to project the attractor onto a finite-dimensional linear subspace, and if the projection is injective, the subspace gives a parametrization of the attractor. By Mañé's theorem such a projection always exists. A related question, which arises in connection with spectral numerical methods is the following: do lower Fourier modes give the parametrization and how many modes does one have to take? A similar question, which arises in connection with finite-difference numerical methods, is the following: can one parametrize functions on the attractor by their values at given points (determining nodes) and how many nodes does one need? More general, when do values of a finite number of functionals determine the long-time dynamics? Such problems, important for computational applications, are addressed in papers by Chueshov [104,105,107], Chueshov and Kalantarov [109], Cockburn, Jones and Titi [111], Constantin, Foias, Manley and Temam [118], Foias, Manley, Temam and Trève [172], Foias and Temam [183], Jones and Titi [245,246], Ladyzhenskaya [266], Shao and Titi [354]. Recently Friz and Robinson [188] and Kukavica and Robinson [259] proved that, under

analyticity-type conditions, functions on a global attractor $\mathcal{A}$ with a finite fractal dimension $\dim_F(\mathcal{A})$ can be parametrized by their values on almost every finite set which includes at least $16\dim_F(\mathcal{A})+1$ nodes.

Approximation of attractors and dynamics. A global attractor is a very complicated object lying in an infinite-dimensional functional space. Exact analytic description of attractors is very difficult, even for low-dimensional systems of ordinary differential equations. Therefore, it is important to develop methods of finding approximate attractors and approximate dynamics. The simplest approach is to approximate infinite-dimensional system by a finite-dimensional one and study the attractor of the finite-dimensional system. It follows from (28) that the attractor of Nth-order Galerkin approximations for the two-dimensional Navier–Stokes system lies in a small neighborhood of the attractor of the original system when N is large. Therefore, the attractor of approximations lies near the exact attractor. But this property does not exclude that the exact attractor can be much larger than the attractors of approximations. Therefore, one has to consider the original equation and find ways to approximate its attractor and dynamics on it and near it. The approaches based on the concepts of inertial manifold and approximate inertial manifolds are briefly discussed in Subsection 1.3.

2.3. *Structure of attractors with a global Lyapunov function*

2.3.1. *Basic properties.* Semigroups that possess a global Lyapunov function are very special. Their dynamics admits in many cases a very detailed description; sometimes it is called gradient-like dynamics and their global attractors have very good properties. At the same time wide classes of PDE have global Lyapunov functions, for example any 1D scalar second-order parabolic semilinear equation and any multidimensional scalar parabolic second-order semilinear equation that does not include first-order derivatives.

We call $\mathcal{L}(u)$ a *global Lyapunov function* if $\mathcal{L}(S_t u)$ is a decreasing function of t for all $t \geqslant 0$ and is strictly decreasing if u is not an equilibrium, that is $\mathcal{L}(S_t u) = \mathcal{L}(u)$ for some $t > 0$ implies $S_t u = u$ for all $t \geqslant 0$.

Sometimes semigroups that possess a global Lyapunov function are called *gradient systems* since the simplest example is given by the equation $\partial_t u = \nabla\mathcal{L}(u)$.

Let $\mathcal{N}$ be the set of all equilibria of S_t. If $\{S_t\}$ has a global attractor, then $\mathcal{N} \subset \mathcal{A}$. For a positive half-trajectory $u(t)$, $t \geqslant 0$, and a negative half-trajectory $u(t)$, $t \leqslant 0$ (when it exists) their omega-limit set $\omega(u)$ and alpha-limit set $\alpha(u)$ are defined respectively by

$$\omega(u) = \bigcap_{\tau\geqslant 0} \operatorname{closure}_E \bigcup_{t\geqslant\tau} u(t), \qquad \alpha(u) = \bigcap_{\tau\leqslant 0} \operatorname{closure}_E \bigcup_{t\leqslant\tau} u(t).$$

THEOREM 2.3.1 (Invariance principle of La Salle). *Let X be an invariant set of $\{S_t\}$. Let $\{S_t\}$ have a global Lyapunov function that is continuous on X and operators S_t be continuous on X for every t. Then for every pre-compact positive or negative half-trajectory $u(t) = S_t u_0$ its omega-limit set or alpha-limit set lies in the set of equilibria,*

$$\omega(u) \subset \mathcal{N}, \qquad \alpha(u) \subset \mathcal{N}.$$

PROOF. One can easily prove that $\omega(u)$ and $\alpha(u)$ are invariant sets. By the monotonicity of $\mathcal{L}(u(t))$ it has a limit $\mathcal{L}(u(t)) \to \mathcal{L}_0$ as $t \to +\infty$. $\mathcal{L}$ equals the constant $\mathcal{L}_0$ on $\omega(u)$. Therefore every trajectory on $\omega(u)$ is an equilibrium. □

Toy model 1.8.4 shows that $\omega(u)$ may contain more than one equilibrium.

DEFINITION 2.3.2. If $\mathcal{N}$ is the set of equilibria, the unstable set $M^{\mathrm{un}}(\mathcal{N})$ is the set that includes all points $u_0 \in E$ such that there is a trajectory $u(t)$ defined for negative t such that

$$\delta_E\big(u(t), \mathcal{N}\big) \to 0 \quad \text{as } t \to -\infty.$$

COROLLARY 2.3.3. *If $\{S_t\}$ has a global Lyapunov function then the global attractor $\mathcal{A}$ coincides with the unstable set $M^{\mathrm{un}}(\mathcal{N})$.*

Indeed, for any $u_0 \in \mathcal{A}$ there exists a trajectory $u(t)$ that lies in the attractor such that $u(0) = u_0$. Since attractor is compact the trajectory is pre-compact. Therefore $\alpha(u) \subset \mathcal{N}$ and $\delta_E(u(t), \mathcal{N}) \to 0$ as $t \to -\infty$.

COROLLARY 2.3.4. *If the set $\mathcal{N}$ is finite, $\mathcal{N} = \{z_1, \dots, z_N\}$ and $u(t) = S_t u_0$ continuously depend on $t \geqslant 0$ for every $u_0 \in E$ then*

$$\mathcal{A} = \bigcup_{j=1}^{N} M^{\mathrm{un}}(z_j). \tag{52}$$

Moreover, every trajectory on the global attractor $\mathcal{A}$ is a connecting orbit of some two equilibria z_i, z_j:

$$\lim_{t\to\infty} u(t) = z_i, \qquad \lim_{t\to\infty} u(t) = z_j.$$

More properties of $\mathcal{A}$ are described in the following subsection.

REMARK. There are situations when a semigroup possesses a global Lyapunov function but the invariance principle of La Salle is not applicable, for example the Lyapunov function is not continuous on the space where the semigroup is defined or the half-trajectories are not compact. Nevertheless in many cases results on the structure of the attractors similar to the above can be obtained for such semigroups, see, for example, Section 2.2, see also [55].

REMARK. In many problems that admit a global Lyapunov function the set $\mathcal{N}$ is generically finite and consists of hyperbolic points (see [42]).

2.3.2. *Regular attractors of gradient-like systems.* A global attractor has a special structure (we call it regular) when $\{S_t\}$ has a global Lyapunov function, the equilibrium set $\mathcal{N}$ is finite and every point $z \in \mathcal{N}$ is hyperbolic. Then, in addition to (52), global attractors have the properties we describe below. Let us order the equilibria so that

$$\mathcal{L}(z_1) \leqslant \mathcal{L}(z_2) \leqslant \cdots \leqslant \mathcal{L}(z_N).$$

Let

$$\mathcal{M}_k = \bigcup_{j=1}^{k} M^{\mathrm{un}}(z_j), \qquad \mathcal{M}_0 = \emptyset. \tag{53}$$

DEFINITION 2.3.5. Following [42,55] we call $\mathcal{A}$ a *regular attractor* if

$$\mathcal{A} = \mathcal{M}_N, \tag{54}$$

and for $k = 1, \ldots, N$ the following statements hold

(1) $\mathcal{M}_k$ is closed and compact in E.
(2) For all $t \geqslant 0$ $\mathcal{M}_k$ is strictly invariant, $S_t \mathcal{M}_k = \mathcal{M}_k$.
(3) $\mathcal{M}_k$ is a stable set.
(4) The boundary $\partial M^{\mathrm{un}}(z_k) = \mathrm{closure}(M^{\mathrm{un}}(z_k)) \setminus M^{\mathrm{un}}(z_k)$ is invariant and $\partial M^{\mathrm{un}}(z_k) \subset \mathcal{M}_{k-1}$ (here $\partial M = \mathrm{closure}(M) \setminus M$).
(5) $S_t \partial(M^{\mathrm{un}}(z_k)) = M^{\mathrm{un}}(z_k)$ for all $t \geqslant 0$.
(6) For every compact set $K \subset \mathcal{M}_k \setminus z_k$, $\lim_{t \to +\infty} \delta_E(S_t K, \mathcal{M}_{k-1}) = 0$.
(7) $M^{\mathrm{un}}(z_j) \cap M^{\mathrm{un}}(z_i) = \emptyset$ when $i \neq j$.
(8) Every set $M^{\mathrm{un}}(z_k)$ is a $\mathcal{C}^1$ manifold of a finite dimension n_j; this manifold is diffeomorphic to $\mathcal{R}^{n_j}$ and the embedding $M^{\mathrm{un}}(z_k) \subset E$ is of class $\mathcal{C}^1$ in a neighborhood of any point $v \in M^{\mathrm{un}}(z_k)$.

A discussion of some of the above properties for the global attractor of the one-dimensional Chaffee–Infante parabolic equation

$$\partial_t u = \partial_x^2 u + a\big(u - u^3\big), \qquad u(0,t) = u(\pi,t) = 0$$

was given by Henry [230]. He used methods of the bifurcation theory varying a in the region $0 < a < 16$ when the dimension of $\mathcal{A}$ varies from 0 to 3 to give a detailed description of the attractor in terms of connecting orbits between equilibria; that was the first description of the structure of a global attractor of a PDE with a global Lyapunov function. Existence of regular attractors for general semilinear and strongly non-linear parabolic and semilinear damped hyperbolic problems was proven by Babin and Vishik in [42] (see also [55]).

REMARK. Hausdorff dimension of a regular attractor is given by the explicit formula $\dim_{\mathrm{H}}(\mathcal{A}) = \max_j \dim M^{\mathrm{un}}(z_j)$.

Global smoothness of a regular attractor. Regular attractors are represented as a union of locally smooth finite-dimensional manifolds. In a number of cases one can prove that the attractor is a subset of a smooth finite-dimensional manifold, for example of an inertial manifold. This is the case for 1D equations of the form (90) when the non-linearity does not include dependence on the $\partial_x u$. In this case the gap condition (12) holds with $p = 0$ for large enough N since $\lambda_N \sim N^2$. When the non-linearity includes the first derivative, there exists a finite-dimensional manifold that is a graph of a continuous function, see [296]. In general, for systems one cannot expect existence of a smooth manifold that contains the attractor. Mora and Sola-Morales [321] proved that the global attractor of a damped wave equation (101) with a small γ cannot be included into a $\mathcal{C}^1$ manifold.

2.3.3. *Global tracking property.* A global attractor is an approximation for a solution $u(t) = S_t u_0$ of the original equation at every fixed time, and the approximation becomes better as time tends to infinity. A natural question is: do solutions on the attractor approximate $u(t)$ as a function of t? Generally, the answer is negative. As an elementary example, consider a system (40) of ODE. Note that (40) has a global Lyapunov function $\mathcal{L}(u) = |u|^2$.

Such a behavior is impossible in the case when all equilibria are hyperbolic. Still we cannot assert that for every $u(t)$ there exists a solution $\tilde{u}(t)$ on the global attractor $\mathcal{A}$ such that (11) holds. The situation is more complicated than in the case of a local invariant manifold or an inertial manifold. Nevertheless, we can construct $\tilde{u}(t)$ such that (11) holds, but we must allow $\tilde{u}(t)$ to have a few jumps when it switches from the finite-dimensional unstable manifold $M^{\mathrm{un}}(z_i)$ to $M^{\mathrm{un}}(z_j)$. We give corresponding results of [51], for details see [53–55].

DEFINITION 2.3.6. We call $\tilde{u}(t)$ a *finite-dimensional composed trajectory* (f.d.c.t.) on the attractor $\mathcal{A}$ if we take a partition of the semiaxis $t \geqslant 0$ into $m+1$ ($m \leqslant N$) non-intersecting intervals $[t_j^0, t_{j+1}^0)$, $j = 0, \ldots, m$, such that $t_0^0 = 0$, $t_{m+1}^0 = +\infty$, $t_0^0 < t_1^0 < \cdots < t_m^0 < t_{m+1}^0$. For every t_j^0 we choose an equilibrium point z_i, $i = i(j)$, $j = 0, \ldots, m$. We take $\tilde{u}(t_j^0)$ in a small neighborhood $O(z_i)$ of one of $z_i, \tilde{u}(t_j^0) \in M_+(z_i, 1)$, and we set $\tilde{u}(t) = S_{t-t_j^0}\tilde{u}(t_j^0)$, $\tilde{u}(t) \in M^{\mathrm{un}}(z_i)$ for $t \in [t_j^0, t_{j+1}^0)$. The resulting finite-dimensional composed trajectory (f.d.c.t.) $\tilde{u}(t)$ lies on the attractor $\mathcal{A}$.

We formulate the results of [51,53,54] skipping technical conditions of the theorem (for details see [51,53–55]).

We say that the time of arrival of a bounded set B to the equilibrium set $\mathcal{N}$ is finite if for any $\delta > 0$ there exists T such that for any $v \in B$ there exists $t \in [0, T]$ such that $S_t v$ lies in the δ-neighborhood of $\mathcal{N}$. When the set $\mathcal{N}$ is finite and all equilibria are hyperbolic the time of arrival of any bounded set B to the equilibrium set $\mathcal{N}$ is finite for parabolic and damped hyperbolic equations and systems with a gradient non-linearity which satisfy usual conditions on the existence of a global attractor. It follows from Lemma 7.1′ of [55] for parabolic equations and from the proof of Lemma 3.4.1 of [55] for the hyperbolic equations.

THEOREM 2.3.7. *Let $\{S_t\}$ satisfy smoothness conditions and have a global Lyapunov function. Let the time of arrival of any bounded set B to the equilibrium set N be finite. Let the set $\mathcal{N}$ of equilibria be finite, $\mathcal{N} = \{z_1, \dots, z_N\}$ and all equilibria be hyperbolic. Then there exists $\eta > 0$ such that for any bounded set $B \subset E$ for any solution $u(t) = S_t u_0$ the following assertion holds. It is possible to choose t_j^0 and to define $\tilde{u}(t_j^0)$ so that the corresponding finite-dimensional composed trajectory satisfies the inequality*

$$\left\| u(t) - \tilde{u}(t) \right\|_E \leqslant C(B) e^{-\eta t} \quad \text{for all } t \geqslant 0. \tag{55}$$

Note that the estimate (55) is uniform, it does not depend on a particular solution and depends only on the norm of initial data in the space E. The right limit points $\tilde{u}(t_{j+1}^0 - 0)$ lie in $O(z_{i'}), i' = i(j+1)$. The constant η depends on spectral properties of $S_1'(z_i)$.

REMARK. Inequality (55) implies that under conditions of Theorem 2.3.7 the global attractor is an exponential attractor, namely the following estimate of [46] holds:

$$\delta_E(S_t B, \mathcal{A}) \leqslant C(B) e^{-\eta t}. \tag{56}$$

Examples of equations for which the statements of Theorem 2.3.7 hold are given in [53–55]; they include parabolic equations and systems and damped hyperbolic equations that have a global Lyapunov function.

Now we give a generalization of this theorem, see [54,55]. This generalization allows to increase the rate of decay in (55) at the expense of increasing the dimension of invariant sets $M_+(z_i, 1)$.

According to Theorem 1.2.4 when $S_1'(z)$ has a circular spectral gap of radius ρ there exists a local invariant manifold $M_+(z, \rho)_{\text{loc}}$. This manifold may be extended to a global invariant (not strictly when $\rho \leqslant 1$) finite-dimensional manifold

$$M_+(z, \rho) = \bigcup_{t \geqslant 0} S_t M_+(z, \rho)_{\text{loc}}.$$

Note that $M^{\text{un}}(z) \subset M_+(z, \rho)$ when $\rho \leqslant 1$. The dimension of $M_+(z, \rho)_{\text{loc}}$ equals the dimension of the invariant subspace of $S_1'(z)$ corresponding to the part of the spectrum outside the circle $|\varsigma| = \rho$.

When the set of equilibria is finite, the attractor $\mathcal{A}$ lies in the larger set

$$\mathcal{A} = \bigcup_j M^{\text{un}}(z_j) \subset \widetilde{M} = \bigcup_j M_+(z_j, \rho_j),$$

where all $\rho_j \leqslant 1$ (when all equilibria are hyperbolic and all $\rho_j = 1$ or are close to 1, $\mathcal{A} = \widetilde{M}$).

DEFINITION 2.3.8. Let ρ_i be in circular gaps of the spectrum of $S_1'(z_i)$ and $M_+(z_i, \rho_i)_{\text{loc}}$ be corresponding local invariant manifolds. We take a partition of the semiaxis $t \geqslant 0$ into $m+1$ ($m \leqslant N$) intervals $[t_j^0, t_{j+1}^0)$, $j = 0, \dots, m$, such that $t_0^0 = 0$, $t_{m+1}^0 = +\infty$, $t_0^0 < t_1^0 <$

$\cdots < t_m^0 < t_{m+1}^0$. For every t_j^0 we choose an equilibrium point z_i, $i = i(j)$, $j = 0, \dots, m$. We take $\tilde{u}(t_j^0)$ in a fixed small neighborhood $O(z_i)$ of one of z_i, $\tilde{u}(t_j^0)$ from the local invariant manifold $M_+(z_i, \rho_i)_{\mathrm{loc}}$, and on every interval $[t_j^0, t_{j+1}^0)$ we set $\tilde{u}(t) = S_{t-t_j^0}\tilde{u}(t_j^0)$, $\tilde{u}(t) \in M_+(z_i, \rho_i)$ for $t \in [t_j^0, t_{j+1}^0)$. We call the resulting piecewise continuous trajectory $\tilde{u}(t)$ a generalized finite-dimensional composed trajectory (g.f.d.c.t.), $\tilde{u}(t)$ lies on $\widetilde{M}$.

THEOREM 2.3.9 [54]. *Let conditions of Theorem* 2.3.7 *hold. For any bounded set $B \subset E$ for any solution $u(t) = S_t u_0$ the following assertion holds. It is possible to choose t_j^0 and to define $\tilde{u}(t_j^0)$ so that for the corresponding g.f.d.c.t. $\tilde{u}(t)$ estimate* (55) *holds. The exponent η in* (55) *is a decreasing function of ρ_i, $\eta \to \infty$ if* $\max_i \rho_i \to 0$.

An explicit expression for η is given in [46,55].

3. Dynamical systems in function spaces

In this section we use the concepts introduced above to describe dynamical properties of Partial Differential Equations.

3.1. *Function spaces and regularity of solutions*

3.1.1. *Function spaces.* Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. The space $L_p(\Omega)$ of Lebesgue integrable with pth power functions has the norm

$$\|u\|_{L_p(\Omega)} = \left(\int_\Omega |u|^p \, dx\right)^{1/p}, \quad p \geqslant 1 \tag{57}$$

(when u is a vector $|u|$ is its magnitude). The Sobolev space $W_p^s(\Omega)$ has the norm

$$\|u\|_{W_p^s(\Omega)} = \left(\sum_{|\alpha| \leqslant s} \int_\Omega |\partial^\alpha u|^p \, dx\right)^{1/p},$$

where $\partial^\alpha u = \partial_1^{\alpha_1} \cdots \partial_d^{\alpha_d} u$, $|\alpha| = \alpha_1 + \cdots + \alpha_d$. When $p = 2$ $W_p^s(\Omega) = H^s$ is a Hilbert space, this case is the most widely used in applications.

The space $\mathcal{C}^0(\Omega)$ of continuous functions has the norm

$$\|u\|_{\mathcal{C}^0(\Omega)} = \sup_{x \in \Omega} \big|u(x)\big|. \tag{58}$$

The space $\mathcal{C}^\alpha(\Omega)$, $0 < \alpha < 1$, of Hölder continuous functions has the norm

$$\|u\|_{\mathcal{C}^\alpha(\Omega)} = \sup_{x,y \in \Omega} \big|u(x) - u(y)\big| / |x - y|^\alpha + \|u\|_{\mathcal{C}^0(\Omega)}. \tag{59}$$

The space $\mathcal{C}^{k+\alpha}(\Omega)$, $0 \leqslant \alpha < 1$, with integer k has the norm

$$\|u\|_{\mathcal{C}^{k+\alpha}(\Omega)} = \sum_{|\beta| \leqslant s}^{s} \|\partial^\beta u\|_{\mathcal{C}^\alpha(\Omega)}. \tag{60}$$

By Sobolev embedding theorem for bounded domains with Lipschitz boundary

$$W_p^s(\Omega) \subset \mathcal{C}^l(\Omega) \quad \text{when } l \leqslant s - d/p$$

and the embedding is compact when the inequality is strict, see [359].

The space $\mathcal{C}^{2+\alpha,1+\alpha/2}(\Omega \times [0,T])$ consists of functions such that $\partial_t u \in \mathcal{C}^{\alpha,\alpha/2}(\Omega \times [0,T])$, $\partial_i \partial_j u \in \mathcal{C}^{\alpha,\alpha/2}(\Omega \times [0,T])$ where $\mathcal{C}^{\alpha,\alpha/2}(\Omega \times [0,T])$ is the space of Hölder continuous functions. Namely, a function from $\mathcal{C}^{\alpha,\alpha/2}(\Omega \times [0,T])$ satisfies Hölder condition

$$\bigl|u(x,t) - u(x',t')\bigr| \leqslant C\bigl(|x-x'|^\alpha + |t-t'|^{\alpha/2}\bigr).$$

The norm in the space $L_p([0,T];Y)$ of functions defined on $[0,T]$ with the target space Y is given by the formula

$$\|u\|_{L_p([0,T];Y)} = \left(\int_0^T \|u\|_Y^p \, dt\right)^{1/p}, \qquad \|u\|_{L_\infty([0,T],Y)} = \operatorname{vrai} \sup_{0 \leqslant t \leqslant T} \|u\|_Y.$$

In particular $L_p([0,T], W_p^1)$ consists of functions that have

$$\int_0^T \|u\|_{L_p(\Omega)}^p \, dt + \int_0^T \|\nabla u\|_{L_p(\Omega)}^p \, dt < \infty.$$

The norm in the space $\mathcal{C}([0,T];Y)$ of continuous functions defined on $[0,T]$ with the target space Y is given by the formula

$$\|u\|_{\mathcal{C}([0,T];Y)} = \sup_{0 \leqslant t \leqslant T} \bigl\|u(t)\bigr\|_Y.$$

3.1.2. *General framework.* Here we describe a general framework of infinite-dimensional dynamical systems which can be applied to wide classes of partial differential equations. We consider equations that can be written in the form (1)

$$\partial_t u = \mathcal{F}(u), \tag{61}$$

where $\mathcal{F}$ is a (non-linear) differential operator. The solution is assumed to satisfy the initial condition

$$u|_{t=0} = u_0, \tag{62}$$

where u_0 belongs to a function space E. It is assumed that the equation has a unique solution $u(t) = u(x,t)$, $t \geqslant 0$, in an appropriate class $\widetilde{E}$ of functions of the spatial variable x and time t. This class should be defined in such a way that for every fixed t_0 the restriction at $t = t_0$ of $u(x,t)$ produces a function of x which belongs to the functional space E. So we obtain a vector $u(t) \in E$ which depends on time t and initial data u_0. In concrete situations the operator $\mathcal{F}(u)$ has to be specified as well as classes of solutions u and initial data u_0. The definition of a solution should include appropriate boundary conditions and smoothness conditions.

Generally speaking, we consider here classes of non-linear PDE for which the initial value problem (61), (62) is locally well posed. Namely, for every $u_0 \in E$ there exists a time interval $(0,T)$, $T = T(u_0)$ and a unique solution $u(x,t)$ of (61) from $\widetilde{E}(0,T)$ such that $u(\cdot,t) \in E$ for every t, $0 \leqslant t < T$. Note that for finite-dimensional systems of ODE with a locally Lipschitz $F(u)$ initial value problem is always well-posed. For equations we consider in this chapter this question is not completely trivial. But for most problems arising in applications one usually can find a class $\widetilde{E}(0,T)$ and the space E such that the problem is locally well-posed; as a rule, the choice of the class is non-unique; sometimes properties of dynamics in different classes are different. Usually one has to take for $\widetilde{E}(0,T)$ and E spaces of regular enough functions of the spatial variables. One of reasons why regularity of solutions helps to get local solvability is the following. When E includes non-regular functions definition of $\mathcal{F}(u)$ may become non-trivial. For example, when $\mathcal{F}(u)$ includes $(\partial_x u)^3$ one cannot take functions $u(x)$ that have a jump discontinuity since multiplication of delta functions is not well-defined. Another reason why regular classes are more convenient is that very wide classes of solutions may lead to non-uniqueness. On the other side, the existence of an absorbing ball usually is easily available in a weaker norm, for example in the space $L_2(\Omega)$ of square integrable functions. Very regular solutions also may lead to non-linear compatibility conditions, see Subsection 2.1. We do not consider in the main part of this paper equations that do not allow locally well-posed initial value problems. Problems without uniqueness generate multi-valued semigroups, we consider such semigroups in Section 4.

Solution semigroup. Since the solution is unique, and the class $\widetilde{E}(0,T)$ is such that the restriction of $u(x,t)$ for a fixed t belongs to E, the following solution mapping is well-defined:

$$u_0 \mapsto u(t), \quad t \geqslant 0. \tag{63}$$

We denote this solution mapping by S_t, $u(t) = S_t u_0$, the mapping is defined on the set of initial data. When the equation does not include explicit dependence on time t, the mappings S_t satisfy the semigroup identity (2). We call the family of the operators $\{S_t\} = \{S_t, t \geqslant 0\}$ a (global) semigroup in E which corresponds to the evolution equation (61).

All partial differential equations we consider in this article possess some kind of dissipativity. That means that solutions after a long time elapses "forget" about their initial data. In particular, they forget about the size of their data. The latter property can be formulated rigorously as the existence of a bounded absorbing set.

Weak solutions. When the class $\widetilde{E}(0,T)$ includes functions which do not have all the derivatives which are included in the equation a solution is understood in the weak sense (or in the sense of distributions). For example, consider the parabolic equation

$$\partial_t u - \partial_x^2 u + f(u) = 0, \quad t > 0,\ x \in \mathbb{R}, \tag{64}$$

with 2π-periodic boundary conditions and initial condition $u(0) = u_0$. One can look for the solution the class $\widetilde{E}(0,T)$ which includes solutions $u(x,t)$ of (64) from $L_\infty([0,T]; L_2([0,T])) \cap L_2([0,T]; H^1([0,T]))$. The derivatives $\partial_t u$ and $\partial_x^2 u$ of such functions do not exist in the classical sense. To determine the concept of weak solution the following observation is used. Let $\varphi(x,t)$ be infinitely smooth 2π-periodic in x test function which equals zero at $t = T$. If $u \in \mathcal{C}^{2+\alpha,1+\alpha/2}([0,L] \times [0,T])$ we can multiply (64) by $\varphi(x,t)$ and integrate over $[0,L] \times [0,T]$. After integration by parts we obtain

$$\begin{aligned} &\int_0^T \int_0^L \big[\partial_t u - \partial_x^2 u + f(u)\big]\varphi \, dx \, dt \\ &\quad = -\int_0^L u_0(x)\varphi(x,0)\, dx + \int_0^T \int_0^L \big[-\partial_t \varphi - \partial_x^2 \varphi\big] u \, dx \, dt \\ &\qquad + \int_0^T \int_0^L \varphi f(u)\, dx \, dt. \end{aligned} \tag{65}$$

The right-hand side of this equality does not include derivatives of $u(x,t)$ and is defined for $u \in \widetilde{E}(0,T)$ as long as the non-linearity $f(u)$ satisfies certain *growth conditions*. A function $u(x,t)$ is called a solution of (64) in the sense of distributions (or a weak solution) if the right-hand side of (65) equals zero for every test function φ. Note that restriction of an arbitrary function $u(x,t)$ from the space $L_\infty([0,T]; L_2([0,T])) \cap L_2([0,T]; H^1([0,T]))$ to a fixed value $t = t_0$ is not well-defined; nevertheless, since Equation (64) gives an expression for $\partial_t u$ the restriction is well-defined for solutions of (64) from this space. For more details and applications of the concept of a solution in the sense of distributions to non-linear problems see [277,361,55].

Below we give typical examples of equations that generate global semigroups of operators. We do not intend to present the most general cases and give the most general conditions for the existence of semigroups. Our purpose is to introduce ideas and methods avoiding technicalities when possible.

3.1.3. *Regularity properties of semigroup operators of PDE.* Usually the dynamical approach is applied to PDE which have solutions which are defined for all non-negative times and bounded in some sense uniformly in time.

Uniform boundedness. Semigroups generated by equations and systems from mathematical physics and their generalizations usually have nice boundedness properties. When the system admits global solvability, namely for any initial data u_0 from a Banach space E the system has a global solution $u(t)$, $t \geqslant 0$, with $u(t) \in E$ for every t, usually estimates that are required to prove the existence of the solution are uniform with respect to bounded

u_0 and the operators S_t (may be multivalued if uniqueness is not proven) are bounded for every $t \leqslant T$ for every fixed T. When an appropriate dissipation condition is imposed the operators are uniformly bounded for all $0 \leqslant t < \infty$. Such a dissipation condition usually takes the form of a *sign condition* on the non-linearity $f(u)$. Examples are given in this chapter, for more examples see [55,363,98]. One of examples is the 3-dimensional Navier–Stokes system in a bounded domain Ω; its weak solutions which satisfy energy estimate are uniformly bounded in $L_2(\Omega)$ though their uniqueness is an open problem.

Continuity. Usually when the uniqueness of a solution is proven the proof gives some kind of continuity of $S_t u_0$ with respect to u_0. See [55,363,98] for numerous examples.

Smoothing property. Operators S_t which correspond to parabolic equations and systems have a smoothing property, namely when $t > 0$ $(S_t u)(x)$, $x \in \Omega$, has better smoothness properties with respect to x than $u(x)$. When the domain Ω is bounded this implies compactness of S_t.

Differentiability. Usually differentiability of operators corresponding to semilinear problems with a subcritical growth of the non-linearity can be proved under natural growth conditions on derivatives of the same type as required for the existence and uniqueness. But the verification is sometimes tedious. Detailed proofs of differentiability of semigroups corresponding to different types of PDE are given in Chapter 3 of [55].

There are examples when the differentiability of operators $S_t u$ with respect to u is not proven even when the operators $S_t u$ are continuous and the non-linearities are analytic. For example, semigroup operators that correspond to monotone parabolic systems from Subsubsection 3.2.2 are Lipschitz continuous but their differentiability is not proven.

Regularity of functions on the attractor. The attractors of parabolic and damped hyperbolic equations usually consist of functions that are more regular than the general functions from the space where the semigroups act. As a rule, they are as regular as the equilibria (time-independent solutions). Such results on regularity are proven in [44], see also [55,363]. Babin and Vishik [44] proved higher regularity of functions on the attractors of parabolic systems, scalar parabolic equations and hyperbolic equations and systems. For results on infinite smoothness see Temam [363]. Hale and Scheurle [224] and Hale [208] study time regularity of solutions on the attractor, see also [336].

3.2. *Non-linear equations with a strong non-linearity*

When the linear part of the differential operator is not dominant, one cannot reduce the differential equation to an integral equations using the variation of constants formula and the theory of sectorial operators, which is commonly used to study subcritical semilinear problems (see [209,336,353]). Nevertheless, a rather complete theory can be developed in many interesting cases, we discuss below some of them.

3.2.1. *Non-linear scalar parabolic equations.* We consider a scalar parabolic equation in a bounded domain Ω with a smooth boundary in the d-dimensional space $\mathbf{R}^d$. A general non-linear second-order scalar parabolic equation has the form

$$\partial_t u = G\big(\nabla^2 u, \nabla u, u, x\big) \tag{66}$$

with appropriate boundary conditions and with the ellipticity conditions assumed. Here we impose Dirichlet boundary condition

$$u|_{\partial\Omega} = 0. \tag{67}$$

Such equations and their attractors are considered in [36,55,268]. Note that when solutions of (66) are in $\mathcal{C}^{2+\alpha,1+\alpha/2}(\Omega \times [0,T])$ their restriction to the boundary together with (67) imply the compatibility condition

$$G\big(\nabla^2 u, \nabla u, 0, x\big)\big|_{x\in\partial\Omega} = 0. \tag{68}$$

This condition is non-linear; therefore one has to consider dynamics in a space which is not linear. If one wants to work in linear spaces, to make this condition linear one has to impose certain restrictions on the non-linearity, see [36,55] for details; existence of global attractors in $\mathcal{C}^{2+\alpha,1+\alpha/2}(\Omega)$ is proven there too. If such restrictions on the non-linearity are not imposed, one still can consider semigroups in linear spaces. To this end one considers wider classes of solutions such that $\nabla^2 u(x,t)$, $\nabla u(x,t)$ and $\partial_t u(x,t)$ are not continuous with respect to x up to the boundary $\partial\Omega$ and (68) does not have to hold. Such solutions in the semilinear case

$$\partial_t u = \nu\Delta u + b(x, u, \nabla u) \tag{69}$$

can be found in Sobolev classes $\widetilde{E}(0,T) = W_p^{2,1}(\Omega \times [0,T])$ with $p > d$. Note that functions from this space have first time derivatives and second-order spatial derivatives in $L_p(\Omega \times [0,T])$. Derivatives of functions from the space $W_p^1(\Omega)$ are understood in the sense of distributions. This may create difficulties when we have to consider non-linear functions of derivatives. These difficulties can be overcome when $p > d$. The existence of the solution operators S_t, the differentiability of the operators, the existence of global attractors for the equations is proven in [42,36,55,268].

3.2.2. *Monotone parabolic equations and systems.* The equation we consider here is interesting since its principal part is non-linear and it cannot be studied from the point of view of semilinear systems. For simplicity we consider here a specific example of a scalar equation, all results can be extended to more general equations and systems (see [55]). We consider in $\Omega \subset R^d$, $d \geqslant 3$, the equation with monotone non-linearity

$$\partial_t u = \sum_{i=1}^{d} \partial_i (\partial_i u)^p - |u|u^{p_0-2} + \lambda u + g(x) \tag{70}$$

with $p \geqslant 1$, $\lambda \geqslant 0$, $\frac{2d}{d-2} > p_0 > 2$. The boundary condition is (67). This equation is strongly non-linear and in contrast to semilinear equations cannot be considered as a perturbation of a linear equation. Note that Equation (70) is not strictly parabolic and in general case theorems on smoothness of solutions for such equations and systems are not known. We also cannot prove that operators S_t are compact in E. Nevertheless, as we will see, a rather complete theory can be built.

We assume that $g \in L_q(\Omega)$ where

$$\text{either} \quad 1/q + 1/p_0 \leqslant 1 \quad \text{or} \quad 1/q + 1/p_1 - 1/d \leqslant 1.$$

The space of solutions $\widetilde{E}(0,T) = L_p([0,T], W^1_p(\Omega)) \cap L_\infty([0,T], L_p(\Omega))$ consists of functions that satisfy the boundary condition (67) and the inequality

$$\|u\|_{L_p([0,T],W^1_p(\Omega))} + \|u\|_{L_\infty([0,T],L_2(\Omega))} < \infty.$$

THEOREM 3.2.1. *Given $u(0) \in H$ there exists a unique solution of* (70) *from $\widetilde{E}(0,T)$. Operators of corresponding semigroup $\{S_t\}$ are continuous bounded operators in the Hilbert space $H = L_2(\Omega)$ both in strong and in the weak topology. The space $E = L_{p_0}(\Omega) \cap W^1_p(\Omega)$ is a reflexive Banach space, it is an invariant set for S_t and the restriction of S_t to this space is continuous in the weak topology of E.*

For details and generalizations see [55, Section 2.3]. Note that operators $S_t u$ are Lipschitzian in H but it is not known if they are differentiable with respect to u.

We denote by $\mathcal{N}$ the set of time-independent solutions of (70) that satisfy (67), $\mathcal{N} \subset E$.

THEOREM 3.2.2. *The semigroup $\{S_t\}$ in H possesses a global attractor $\mathcal{A} \subset H$ in the weak topology, the attractor is bounded in E and compact in the weak topology E_w of E.*

The functional

$$\mathcal{F}(u) = \int_\Omega \left[\frac{1}{p+1} \sum_{i=1}^{d} |\partial_i u|^{p+1} + \frac{1}{p_0+1} |u|^{p_0+1} - \frac{\lambda}{2}|u|^2 + gu \right] dx$$

is a global Lyapunov function for S_t. The functional $\mathcal{F}(u)$ is not continuous neither in H nor in the weak topology of E, and this causes technical difficulties; in particular invariance principle of La Salle is not applicable. Nevertheless, one can prove the following theorem on the structure of the attractor $\mathcal{A}$.

THEOREM 3.2.3. *The attractor $\mathcal{A}$ consists of values of bounded in E trajectories $u(t)$, $-\infty < t < +\infty$, such that $\delta_{E_w}(u(t), \mathcal{N}) \to 0$ as $t \to -\infty$ or $t \to \infty$.*

The proof is given in [55], here we notice only that the proof uses the inequality

$$\mathcal{F}\big(u(T)\big) + \frac{1}{2} \int_0^T \int_D |\partial_t u|^2 \, dx \, dt \leqslant \mathcal{F}\big(u(0)\big).$$

This inequality implies that $\|\partial_t u(t)\|$ is small for large t, therefore $u(t)$ is close to a solution of (70) with $\partial_t u = 0$, that is to an equilibrium. For details and generalizations see [55, Sections 1.3, 1.4, 2.3 and 3.6]. Note that the theorem on the description of unstable manifolds of equilibria $z \in \mathcal{N}$ given in Subsection 1.2 is not applicable because of lack of differentiability and an interesting open problem is to give a description of unstable manifolds of equilibria z of equations of the type (70). For results on dynamics of monotone equations see also [101].

3.2.3. *Semilinear equations with a supercritical non-linearity.* Note that the methods typical for the theory of monotone operators are useful not only in the case (70) when the system includes non-linearity in the principal part but also when the system is semilinear of the form (72) but with a high growth rate of the vector-function $f(u)$, for example systems of the form

$$\partial_t u = \Delta u - |u|u^{p_0-2} + \lambda u + g(x) \tag{71}$$

with an arbitrary large, supercritical p_0. Standard methods based on the operator version of the variation of constants formula (see [209,336,353]) are not applicable in this case. Nevertheless, such equations possess global attractors, see [55,98,101]; the attractors upper semicontinuously depend on parameters λ, g. Estimates of dimension of global attractors for problems with a supercritical non-linearity are given by Zelik [390].

3.3. *Semilinear equations*

Semilinear equations include non-linearities only in the lower-order terms, they have the form

$$\partial_t u = -Au + \mathcal{F}_0(u),$$

where A is a linear operator which is in some sense positive, $\mathcal{F}(u)$ is the non-linearity which is subordinate in some sense. Many important equations from applications have such structure. Often (in particular in the subcritical growth case) one can apply the theory of linear operators to invert the linear part using the variation of constant formula

$$u(t) = u(0) + \mathcal{F}_1(u)(t), \quad \mathcal{F}_1(u)(t) = \int_0^t e^{-(t-\tau)A}\mathcal{F}_0\big(u(\tau)\big)\,d\tau.$$

When the operator $\mathcal{F}_1(v)$ is Lipschitz continuous in an appropriate space of functions $v(x,t)$ the proofs of basic properties of the semigroups, such as the existence of solutions, compactness properties, smoothing property and differentiability are relatively simple (see [209,336,353]). To provide good properties of $\mathcal{F}_1(u)$ the non-linearity $\mathcal{F}_0(u)$ has to satisfy certain regularity conditions, in particular subcritical growth conditions. Note that even for the semilinear equations in the supercritical growth case this approach is not sufficient and one has to use more specific methods from PDE, see the preceding Subsection 3.2, see also [55,98].

3.3.1. *Parabolic semilinear systems (reaction–diffusion systems).* In a bounded domain Ω we consider the parabolic system

$$\partial_t u = \nu a \Delta u - f(x,u) + \lambda u - g(x) \tag{72}$$

with Neumann boundary condition

$$\frac{\partial u}{\partial l} = 0 \quad \text{on } \partial\Omega. \tag{73}$$

(Equations with Dirichlet boundary conditions may be considered in a similar way.) Here $u = (u_1, \dots, u_m)$, $f = (f_1, \dots, f_m)$, $g = (g_1, \dots, g_m)$, a is a positive diagonal matrix, $\lambda > 0$. The diagonal matrix a is positive, $g \in H^0 = (L_2(\Omega))^m$, $\nu > 0$. Systems that involve a certain dependence on ∇u can be studied in a similar way. Properties of such systems depend strongly on the structure of $f(x,u)$.

We assume that f is continuously differentiable with respect to all arguments and satisfies the sign condition

$$f(x,u)\cdot u \geqslant \mu_0 |u|^{p_0}, \quad \mu_0 > 0, \qquad \sum_{k,i} \frac{\partial f_k}{\partial u_i} \xi_k \xi_i \geqslant 0, \quad \xi_k \in \mathbb{R}^m. \tag{74}$$

This condition implies, in particular, that if $\lambda = 0$ the time-dependent solutions tend to a steady-state solution as $t \to +\infty$, and the difference of any two steady state solutions is a constant. Therefore all non-trivial dynamics in this example is generated by the instability thanks to $\lambda > 0$. We also impose the growth condition

$$\big|f(x,u)\big| \leqslant \mu_1 |u|^{p_0 - 1} + C, \quad p_0 > 2. \tag{75}$$

We consider $\lambda \gg 1$, $\nu \ll 1$. The following typical results are proven in [55].

THEOREM 3.3.1. *The semigroup $\{S_t\}$ generated by* (72), (73) *in $H^0 = (L_2(\Omega))^m$ has a global (H^0, H^1_w) attractor $\mathcal{A}$.*

Note that the above theorem does not contain restrictions on the powers p_0.

Now we give a simple and typical formal derivation of the existence of an absorbing ball. We multiply (72) by u and integrate over Ω. After integration by parts we obtain

$$\partial_t \|u\|^2 + \nu \int_\Omega a \nabla u \cdot \nabla u \, dx \leqslant \int_\Omega \big[-f(x,u)\cdot u + \lambda |u|^2 \big] dx. \tag{76}$$

From (74) we infer using Young's inequality that

$$\begin{aligned} -f(x,u)\cdot u + \lambda |u|^2 + gu &\leqslant -\mu_0 |u|^{p_0} + \left(\lambda + \frac{1}{2}\right)|u|^2 + \frac{1}{2}|g|^2 \\ &\leqslant -\frac{1}{2}|u|^2 + \frac{1}{2}|g|^2 + C(\lambda+1)^{p_0/(p_0-2)} \mu_0^{-2/(p_0-2)} \\ &= -\frac{1}{2}|u|^2 + \frac{1}{2}|g|^2 + C_1. \end{aligned}$$

Using this inequality in the right-hand side of (76) we obtain the following differential inequality

$$\partial_t \|u\|^2 \leqslant -\frac{1}{2}\|u\|^2 + \frac{1}{2}\|g\|^2 + C_1|\Omega|, \tag{77}$$

where $|\Omega|$ is the volume of Ω, which implies the estimate

$$\big\|u(t)\big\|^2 \leqslant \big\|u(0)\big\|^2 e^{-t/2} + \big(\|g\|^2 + 2C_1|\Omega|\big)\big(1 - e^{-t/2}\big).$$

This estimate implies that the set

$$B_0 = \big\{u \in H^0\colon\ \|u\|^2 \leqslant 2\big(\|g\|^2 + 2C_1|\Omega|\big)\big\}$$

is an absorbing ball for S_t.

When the non-linearity satisfies the smoothness condition

$$\big|f(x, u+z) - f(x,u) - \nabla_u f(x,u)z\big| \leqslant C\big(1 + |u| + |z|\big)^{p_1}|z|^{1+\gamma}, \tag{78}$$

where $\gamma > 0$ is sufficiently small and the growth is subcritical, namely

$$(d-2)p_1 < 4 \quad \text{when } d > 2 \tag{79}$$

a stronger result holds.

THEOREM 3.3.2. *The semigroup $\{S_t\}$ generated by* (72), (73) *in $H^0 = (L_2(\Omega))^m$ has a global attractor $\mathcal{A}$. Its Hausdorff dimension satisfies the estimate*

$$\dim_{\mathrm{H}} \mathcal{A} \leqslant C'\lambda^{d/2}\nu^{-d/2}, \tag{80}$$

and for a x-independent g we have a lower bound for $\dim_{\mathrm{H}} \mathcal{A}$:

$$C''\lambda^{d/2}\nu^{-d/2} \leqslant \dim_{\mathrm{H}} \mathcal{A}. \tag{81}$$

The upper estimate in (80) is obtain by methods of Subsection 2.1. The lower estimate in (80) is obtained based on Theorem 2.1.4. Namely, we can find an equilibrium point with a large instability index as a constant vector. See Chapter 10 [55] for details.

REMARK. Consider a large domain Ω_r which is obtained by r-dilation of a fixed domain Ω_1, $\Omega_r = r\Omega_1$, let $\mathcal{A}_r$ be the global attractor of (72), (73) in $(L_2(\Omega_r))^m$. Note that when g is a constant rescaling $x = ry$ reduces Ω_r to Ω_1, but the coefficient ν is replaced by νr^{-2}. Therefore the estimates (80), (81) imply that

$$C_1|\Omega_r| \leqslant \dim_{\mathrm{H}} \mathcal{A}_r \leqslant C_2|\Omega_r|, \tag{82}$$

that is the dimension of the attractor is roughly proportional to the volume $r^d|\Omega_1| = |\Omega_r|$ of the domain Ω_r. This can be interpreted the following way: every degree of freedom on the attractor requires some volume in the physical space.

A semilinear parabolic system (72) has a Lyapunov function when the function $f(x,u)$ is a gradient,

$$f_i(x,u) = \frac{\partial F(x,u)}{\partial u_i}, \quad i = 1,\dots,m.$$

The global Lyapunov function for (72) is given by

$$\mathcal{L}(u) = \int_\Omega \left[\frac{1}{2}a\nabla u\cdot\nabla u + F(x,u) - \frac{\lambda}{2}|u|^2 + g\cdot u\right]dx.$$

Condition (74) implies that $F(x,u) \geqslant 0$ and grows faster than quadratic as $|u| \to \infty$. Therefore $\mathcal{L}(u)$ is bounded from below.

A typical theorem on the attractor $\mathcal{A}$ from Theorem 3.3.2 is the following

THEOREM 3.3.3. *For a generic g the attractor $\mathcal{A}$ is a regular attractor described in Subsection* 2.3. *It has the tracking property described in Theorem* 2.3.7.

REMARK. In many applications (see [276,290,358,115] and references therein) a natural restriction on solutions is the positivity of $u_1,\dots,u_n$. We do not discuss here conditions on f that lead to the positivity of components $u_1(t),\dots,u_n(t)$ when $u_1(0),\dots,u_n(0)$ are positive.

3.3.2. *Semilinear scalar parabolic equations (multidimensional).* The literature on dynamical properties of scalar parabolic equations of second order is extensive. A review on the asymptotic behavior of semilinear equations is given by Poláčik [331], see also Hale [211]. Semilinear parabolic equations have a linear principal part and a lower-order non-linearity, a typical equation has the form

$$\partial_t u = \Delta u + f(x,u,\nabla u), \tag{83}$$

where Δ is the Laplace operator and $f(x,u,p)$ is a twice differentiable function of arguments $x \in \Omega$ and $u \in \mathbb{R}$, $p \in \mathbb{R}^d$. For a more detailed discussion see [331]. We consider here Dirichlet boundary condition

$$u(x) = 0 \quad \text{for } x \in \partial\Omega. \tag{84}$$

Other types of boundary conditions, for example Neumann condition, can be imposed leading to similar dynamics. There are two essentially different situations, namely when $f(u,\nabla u)$ depends on ∇u and when it is independent of ∇u, $f(u,\nabla u) = f(x,u)$. In the latter case Equation (83) takes the form

$$\partial_t u = \Delta u - f(x,u). \tag{85}$$

We assume that the sign condition (the dissipativity condition) holds

$$f(x,u,\nabla u)u > 0 \quad \text{for } |u| \geqslant b > 0. \tag{86}$$

A very important property of a scalar parabolic second-order equation that does not hold for general parabolic systems is the *Maximum Principle*. From the Maximum Principle in particular follows that a natural space for a semigroup generated by a scalar parabolic semilinear equation is the space $\mathcal{C}(\Omega)$ of continuous functions in Ω.

THEOREM 3.3.4. *Equation* (85) *generates a semigroup* $\{S_t\}$ *in the space* $E = \mathcal{C}(\Omega) \cap \{u_{\partial\Omega} = 0\}$. *Operators* S_t *of this semigroup are continuous and differentiable, they are compact for* $t > 0$. *The set* $|u| \leqslant b$ *is absorbing.*

Now we consider dynamical properties of S_t.

Lyapunov function. The following functional $\mathcal{L}(u)$ is a global Lyapunov function for the semigroup $\{S_t\}$ corresponding to (85):

$$\mathcal{L}(u) = \int_\Omega \frac{1}{2}|\nabla u|^2 + F(x,u)\,dx. \tag{87}$$

Here $F(u)$ is the antiderivative of $f(x,u)$,

$$F(u) = \int_0^u f(v)\,dv.$$

An example of an equation that has a global Lyapunov function but the right-hand side is not a gradient but gradient-like is

$$\partial_t u = a(u)\big(\Delta u + f(x,u)\big), \tag{88}$$

where $a(u) \geqslant a_0 > 0$, Δ is the Laplace operator and $a(u)$ and $f(x,u)$ are twice differentiable functions of arguments $x \in \Omega$ and $u \in \mathbb{R}$. We consider here Dirichlet boundary conditions.

When spatial dimension $d = 1$ Equation (83) also has a Lyapunov function, but the construction is more complicated, see [385].

THEOREM 3.3.5 [42]. *Let* $f(x,u)$ *be of class* $\mathcal{C}^{1+\alpha}$, *be independent of* ∇u, *let all equilibria be hyperbolic. Then the semigroup* $\{S_t\}$ *in the space* $\mathcal{C}^0(\Omega)$ *generated by* (88) *has a regular attractor* $\mathcal{A}$ *and the tracking property takes place. In particular,* $\mathcal{A}$ *is a union of finite-dimensional smooth manifolds and dynamics on the attractor is described by movement along connecting orbits of equilibrium points.*

The proof uses the existence of the Lyapunov function (87), boundedness of the attractor in $\mathcal{C}^0(\Omega)$, differentiability of S_t and spectral properties of linearized operators.

REMARK. It is proven by Brunovský and Poláčik [73] that for a generic function $f(x,u)$ stable and unstable manifolds of (85) intersect transversally, that is Morse–Smale property holds.

Gradient-dependent non-linearity. When the non-linearity $f(x,u,\nabla u)$ includes dependence on ∇u, in contrast to the case when the non-linearity $f(x,u)$ does not depend on the gradient, the attractor is not regular. Namely, when the non-linearity depends on ∇u the situation is completely different as it is shown by Poláčik [330] and Poláčik and Rybakowski [332]. They consider Dirichlet problem for Equation (83). They prove that for any given analytic vector field in a neighborhood of zero in a finite-dimensional vector space there exist a domain Ω and a non-linearity $f(x,u,\nabla u)$ such that there exists an equilibrium solution of (83) and dynamics of the parabolic equation restricted to the finite-dimensional center-unstable manifold is arbitrarily close (in a certain sense) to the dynamics described by the vector field.

Asymptotic symmetrization. One of properties specific to scalar parabolic equations is asymptotic symmetrization. We consider a domain Ω with a piecewise smooth boundary which is symmetric with respect to a reflection Q in a plane. Without loss of generality we may assume that the plane has equation $x_1 = 0$. When the equation depends on x_1 and $\partial_1 u$ in an even way, then the change of variables x_1 to $-x_1$ does not change the equation. Certain monotonicity conditions on dependence of coefficients on x_1 have to be imposed, see [16,20,36] for details. We here for simplicity consider the case where the equations do not depend on x explicitly and the equation is of the form

$$\partial_t u = va\Delta u - f(u,\nabla u), \qquad u|_{\partial\Omega} = 0 \tag{89}$$

(see [36] for a strongly non-linear case). First results on the asymptotic symmetrization for generic semilinear equations in smooth domains were obtained by Hess and Poláčik [232], see also [229,329]. The general case was considered in [16,20,36].

THEOREM 3.3.6. *Let Ω and the equation be invariant with respect to reflections Q from a family $\widetilde{Q}$. Under conditions of Theorem* 3.3.4 *the attractor $\mathcal{A}$ consists of Q-symmetric functions $v(x) = v(Qx)$ for every $Q \in \widetilde{Q}$.*

COROLLARY 3.3.7. *If the equation is radially symmetric and Ω is a ball, the attractor $\mathcal{A}$ consists of radially symmetric functions.*

The following theorem of Babin and Sell [36] shows that symmetrization is exponentially fast.

THEOREM 3.3.8. *Let the boundary of Ω be smooth. Let Ω and the equation be invariant with respect to reflections Q from a family $\widetilde{Q}$. Then there exists $\gamma > 0$ and $C > 0$ such that for every solution $S_t u_0 = u(x,t)$*

$$\left\| u(t) - Q^* u(t) \right\| \leqslant Ce^{-\gamma t}$$

uniformly in $Q \in \widetilde{Q}$ *where* $Q^*u(x) = u(Qx)$.

The proofs of the symmetrization are based on the Maximum principle and the method of moving planes.

Comparison principle. Maximum principle implies that if two solutions $u(x,t)$, $v(x,t)$ of (89) satisfy at $t=0$ the inequality $u(x,0) \leqslant v(x,0)$ then $u(x,t) \leqslant v(x,t)$ for $t \geqslant 0$. This property implies non-trivial restrictions on the dynamics. For more details and for references see [331].

REMARK. Dynamics of degenerate parabolic equations was studied by Feireisl and Simondon [153] and Feireisl [154].

3.3.3. *Semilinear one-dimensional scalar parabolic equations.* One-dimensional semilinear scalar parabolic equations are of the form

$$\partial_t u = \alpha \partial_x^2 u - f(u, x, \partial_x u), \quad 0 < x < L, \tag{90}$$

with the Dirichlet boundary conditions

$$u(0) = u(L) = 0$$

at $x = 0$, $x = L$ (the case of Neumann or Robin boundary conditions is similar). If (86) and certain growth conditions with respect to $|\partial_x u|$ hold, there exists a global attractor of $\{S_t\}$ generated by (90) in $\mathcal{C}([0, L])$ (see [331,36] for details). In the 1D case the semigroup generated by (90) has a global Lyapunov function (see [385]), therefore one can apply all results of the section concerning scalar equations with $f(x,u)$ independent on ∇u. But in the 1D case one can get much more detailed information. Solutions with one spatial variable always have a limit as $t \to \infty$; the limit is an equilibrium point (see [385]). The equilibrium points of this semigroup are solutions to the ODE

$$0 = \alpha \partial_x^2 z - f(z, x, \partial_x z) \tag{91}$$

with the same boundary condition.

In the generic case $\mathcal{N}$ is a finite set, $\mathcal{N} = \{z_1, \dots, z_N\}$ and every equilibrium point is hyperbolic (see [42]). According to the results of the previous section there exists a global attractor of this equation and it is regular, namely (53), (54) hold. Since the dynamics on the attractor is invertible, every $M^{\mathrm{un}}(z)$ equals a union of trajectories connecting different points of $\mathcal{N}$, therefore $\mathcal{A}$ is a union of such trajectories $u(t)$

$$\lim_{t \to -\infty} u(t) \to z \in \mathcal{N}, \qquad \lim_{t \to +\infty} u(t) = z'(u) \in \mathcal{N}.$$

There are two important properties of scalar 1D equations that make this case very special. First property concerns equilibrium points z, namely differentials $S_t'(z)$. Since

points ρ of the spectrum of $S_t'(z)$ equals exponents $\rho_j = e^{-\lambda_j t}$ of the eigenvalues λ_j, $A'(z)v = -\lambda_j v$, of the ordinary differential operator

$$A'(z)v = \alpha\partial_x^2 v - f_u'(z, x, \partial_x z)v - f_{\partial_x u}'(z, x, \partial_x z)\partial_x v,$$

the multiplicity of its eigenvalues for the Dirichlet or Neumann boundary conditions is always one.

The second property is the possibility to introduce the number $N(v(t))$ of nodal points for a solution $v(x,t)$ of a linear parabolic equation

$$\partial_t v = a(x,t)\partial_x^2 v - b(x,t)v - c(x,t)\partial_x v$$

and use the nodal number to study dynamics of (90) (see Matano [295], Angenent [5,6]; see [331] for more references). Using Maximum principle one deduces that if the number $N(v(t))$ is finite at $t = t_0$ it stays finite and, moreover, it is a non-increasing function of t. Under natural conditions $N(v(t))$ is finite for $t > 0$.

The nodal property implies fulfillment of Morse–Smale property (Henry [231]):

$$M^{\mathrm{un}}(z_i) \cap M^{\mathrm{s}}(z_j) = \varnothing \quad \text{when } \dim M^{\mathrm{un}}(z_i) \leqslant \dim M^{\mathrm{un}}(z_j),$$

otherwise the intersection is transversal.

There is a remarkably detailed description of the structure of attractors of 1D scalar problems in generic situations. Since the attractor is regular, it is a union of trajectories that connect equilibria $z_1, \ldots, z_N$. A non-trivial problem is to determine which equilibria are connected and which are not. When all equilibria are hyperbolic, a complete solution of this problem is given by Brunovský and Fiedler [72] for Dirichlet boundary conditions and Fiedler and Rocha [156,157] for Neumann boundary conditions. Brunovský and Fiedler [72] gave complete rules determining connected equilibria of the Dirichlet problem in terms of nodal numbers of equilibria and order (magnitude with respect to nodal number) of their slopes at the boundary. Fiedler and Rocha [156] gave complete rules determining connected equilibria of the Neumann problem in terms of monotone ordering of their values at both boundary points.

A detailed review of 1D results is given in [331,211] and [158].

3.3.4. *Navier–Stokes equations and equations from mathematical physics.* Here we give a very brief sketch of the theory of global attractors of the Navier–Stokes equations which in many respects determined the development of the theory of global attractors, see [25] for details. The 2D Navier–Stokes (2DNS) equations for viscous incompressible fluids have the form

$$\partial_t u + u \cdot \nabla u - \nu\Delta u = f + \nabla p, \qquad \nabla \cdot u = 0. \tag{92}$$

Here $u = (u_1, u_2)$ is the velocity field, $u = u(x,t) = u(x_1, x_2, t)$, $\nu > 0$ is the kinematic viscosity, $f(x_1, x_2)$ represent volume forces. The Euler non-linearity for 2D Navier–Stokes equations is given by

$$u \cdot \nabla v = u_1\partial_1 v + u_2\partial_2 v. \tag{93}$$

For simplicity we consider here periodic boundary conditions

$$u(x_1 + 2\pi a_1, x_2) = u(x_1, x_2 + 2\pi a_2) = u(x_1, x_2). \tag{94}$$

We denote by $\mathbf{H}^s$, $s \geqslant 1$, the space of vector fields with a finite Sobolev H^s norm which satisfy the divergence-free condition. We denote by Π the orthogonal Leray projection $\Pi : H \to \mathbf{H}_0$, the projection can be explicitly written in terms of Fourier series. Gradient fields are in the null-space of Leray projection. Applying the projection we can rewrite (92) in the form

$$\partial_t u + B(u, u) + \nu A u = f, \tag{95}$$

where

$$B(u, v) = \Pi(u \cdot \nabla v), \qquad Au = -\Pi \Delta u. \tag{96}$$

We consider the initial value problem

$$u|_{t=0} = u_0 \in \mathbf{H}^0. \tag{97}$$

THEOREM 3.3.9. *Let $f \in H^{-1}$. For any $u_0 \in H^0$ there exists unique solution $u(t)$ of* 2D *NS system* (92) *with initial data* (97). *This solution belongs to H^0 for all $t \geqslant 0$.*

The solution mapping

$$S_T : u|_{t=0} \longmapsto u|_{t=T}$$

determines a family of operators $\{S_T\}$, $T \geqslant 0$. The operators S_T form a semigroup that acts in the space $\mathbf{H}^0$.

Many important properties of NS equations can be formulated in terms of the solution semigroup $\{S_t\}$. These basic properties are described in the following theorem (see [55] for a detailed proof, see also [363]).

THEOREM 3.3.10. *Let $f \in H^0$. Then the semigroup $\{S_t\}$ that corresponds to the* 2D *NS system* (92), (94) *and acts in the space H^0 has a global attractor $\mathcal{A}$.*

THEOREM 3.3.11. *Let $f \in H^0$. The attractor $\mathcal{A}$ is compact in H^2 and $\delta_{H^2}(S_t(B), \mathcal{A}) \to 0$ for any bounded in H^0 set B.*

The following important theorem is proven by Constantin, Foias and Temam [121], see also [94]. We give the formulation in the most important case when the Grasshof number

$$G = \frac{\|f\|_0}{\nu^2 \lambda_1},$$

where λ_1 is the first eigenvalue of the Stokes operator, is not small, $G \geqslant 1$.

THEOREM 3.3.12. *If $d \geqslant c' G^{2/3}(1+\log(G))^{2/3}$ then d-dimensional volumes on $\mathcal{A}$ exponentially decay as $t \to \infty$. When $G \geqslant 1$ Hausdorff and fractal dimension of the attractor $\mathcal{A}$ satisfy the inequality* $\dim_{\mathrm{H}} \mathcal{A} \leqslant \dim_{\mathrm{F}} \mathcal{A} \leqslant 2c' G^{2/3}(1+\log(G))^{2/3}$.

REMARK. Lower bounds of the dimension of $\mathcal{A}$ (which are almost precise as proved by Ziane [396]) are obtained by Babin and Vishik [39] and Liu [282,283] based on Theorem 2.1.4. See [25] for a discussion of related results.

3D Navier–Stokes equations. For the 3D Navier–Stokes equations global regularity of solutions with large initial data and forcing terms is not proven. The global regularity of solutions and existence of a finite-dimensional global attractor for the 3D Navier–Stokes equations with general large initial data and forcing terms and a large Coriolis force is proven by Babin, Mahalov and Nicolaenko, see [25] for details. The questions of existence of a global attractor without assuming uniqueness of solutions were studied by Sell [351], Ball [59], Chepyzhov and Vishik [97]. Existence of global attractors in thin domains was proven by Raugel and Sell [337]. Sell [351] and Chepyzhov and Vishik [98] applied trajectory approach to prove the existence of a global attractor for the 3D Navier–Stokes equations (for a brief discussion of related methods see Subsection 4.1).

Attractors of equations of mathematical physics. The existence and properties of attractors of equations of mathematical physics is the subject of intensive study. One of important issues is to obtain an estimate of the dimension of the attractor in terms of physical parameters of the problem. The literature is extensive, see [363,55,103,106,98,25]. We add here several recent references. For the dynamical theory of compressible Navier–Stokes equations see Feireisl [150] and references therein. Attractors of the Cahn–Hilliard equation are studied by Miranville [314,312,313] and Carrive, Miranville, Piétrus and Rakotoson [77]. Attractors for the generalized Benjamin–Bona–Mahony equation are studied by Celebi, Kalantarov and Polat [82].

3.3.5. *Damped hyperbolic equations and systems.* Semilinear wave equation which we discuss here has the form

$$\partial_t^2 u + \gamma \partial_t u = \Delta u - f(u) - g(x), \qquad u|_{\partial\Omega} = 0, \tag{98}$$

with the damping term $\gamma \partial_t u$ with $\gamma > 0$. Here $g \in L_2(\Omega)$. We denote by

$$F(u) = \int_0^u f(v)\,dv$$

the antiderivative of f. We assume that for some $\varepsilon > 0$ the following sign condition holds:

$$F(u) \geqslant -\left(\frac{\lambda_1}{2} - \varepsilon\right)u^2 - C, \qquad f(u)u \geqslant -c - \left(\frac{\lambda_1}{2} - \varepsilon\right)u^2, \tag{99}$$

where λ_1 is the minimum eigenvalue of $-\Delta$. We also assume the growth condition

$$\big|f(u) - f(v)\big| \leqslant C\big(1 + |u| + |v|\big)^{\rho+1-\alpha}|u - v|^{\alpha}$$

with $0 < \alpha \leqslant 1$; sometimes a similar growth condition is imposed on $f'(u)$. When $\alpha = 1$ and

$$(d-2)\rho \leqslant 2 \tag{100}$$

(the last condition is obviously satisfied when $d = 1, 2$) the solution of (98) with initial data $u(0) = u_0$, $\partial_t u(0) = p_0$ is unique. The value $\rho = 2/(d-2)$ is called critical. For example, when the space dimension $d = 3$ the cubic growth is critical.

Weak solutions of (98) exist when ρ is arbitrary; the uniqueness of weak solutions is not known in the case of a general domain Ω when the growth is supercritical.

To write (98) in the form of a first-order equation (61) we introduce a new unknown function $p = \partial_t u$ and obtain from (98) an equivalent system

$$\partial_t u = p, \qquad \partial_t p = -\gamma p + \Delta u - f(u) - g, \qquad u|_{\partial\Omega} = 0. \tag{101}$$

We introduce the spaces

$$\begin{aligned} E &= \big\{y = (u, p)\colon\ u \in W_2^1(\Omega) \cap \{u|_{\partial\Omega} = 0\},\ p \in L_2(\Omega)\big\}, \\ E_1 &= \big\{y = (u, p)\colon\ u \in W_2^2(\Omega) \cap \{u|_{\partial\Omega} = 0\},\ p \in W_2^1(\Omega) \cap \{u|_{\partial\Omega} = 0\}\big\}. \end{aligned} \tag{102}$$

THEOREM 3.3.13. *If $\alpha = 1$, (100) holds and $y_0 = (u_0, p_0) \in E$ then the problem* (101) *has a unique solution $y(t) = (u(t), p(t))$. Operators $S_t y$ are continuous in y uniformly in $t \leqslant T$ and y, $\|y\|_E \leqslant R$. Solutions $S_t y$ are bounded uniformly for $t < \infty$ and $\|y\| \leqslant R$.*

The semigroup $\{S_t\}$ has a global Lyapunov function

$$\mathcal{L}\big((u, p)\big) = \int \left[\frac{1}{2}|p|^2 + \frac{1}{2}|\nabla u|^2 + F(u) + gu\right] dx. \tag{103}$$

The above Lyapunov function was used by Dafermos [124] to study dynamical behavior of hyperbolic damped equations, in particular convergence of bounded solutions to equilibria.

To obtain the existence of an absorbing set we use another functional that depends on an auxiliary parameter η

$$\mathcal{L}_{1\eta}(u, p) = \int_\Omega \left[\frac{1}{2}|p|^2 + \frac{1}{2}|\nabla u|^2 + F(u) + gu\right] dx + \eta \int_\Omega pu\, dx. \tag{104}$$

For small enough η we obtain (see [225,55]),

$$\partial_t \mathcal{L}_{1\eta}(u, p) \leqslant -\delta_1 \mathcal{L}_{1\eta}(u, p) + C_2$$

with $\delta_1 > 0$. This inequality implies existence of a bounded absorbing set

$$B_0 = \big\{(u, p) \in E\colon\ \mathcal{L}_{1\eta}(u, p) \leqslant 2C_2/\delta_1\big\}.$$

The theory of attractors of hyperbolic equations is more difficult technically compared with the theory of parabolic equations because the operators S_t of the corresponding semigroup are not compact when $t > 0$, for a discussion see [336].

The regular global attractor $\mathcal{A}$ of this Equation (98) was constructed by Babin and Vishik for generic g in [38,42]; it was proven there that $\mathcal{A}$ is (E_1, E)-attractor. Haraux [225] and Hale [208] proved the existence of (E, E)-attractor in the subcritical case $(d-2)\rho < 2$ for general f, g; their results imply in particular that the regular attractor constructed in [38, 42] is the global attractor in E. System (101) that generates the semigroup $\{S_t\}$ involves the solution $y(t) = (u(t), \partial_t u(t))$ that corresponds to the solution $u(t)$ of (98). Haraux [225] and Hale [208] use the splitting of the solution $y(t) = S_t y(0)$ in the form

$$y(t) = y_1(t) + y_2(t),$$

where $y_1(t)$ is a solution of the *linear* equation obtained from (101) by setting $f = g = 0$ and $y_2 = y - y_1$ is a correction.

$$\|y_1(t)\|_E \leqslant Ce^{-\gamma t}, \quad \gamma > 0. \tag{105}$$

When ρ is subcritical, $y_2(t)$ belongs to a compact set in E when $y(0)$ is in B_0. Such a splitting was used also by Webb [378,379] in the study of asymptotical behavior of individual solutions.

There are several works which treat the case of the critical exponent ρ when $(d-2)\rho = 2$, this case requires more refined techniques. Existence of the global attractor in the critical case with the cubic growth $f(u)$ for $d = 3$ was proven by Babin and Vishik in [44], but the attraction was proven in the weak topology of $E = H^1 \times H^0$. The existence of the global attractor $\mathcal{A}$ in the norm-induced topology of E in the critical case was proven in [46,52], see also [55]. In these works the non-linearity was split in the form $f(u) = f_0(u) + f_1(u)$ where $f_0(u)$ is monotone and $f_1(u)$ has a lower rate of growth at infinity. The solution respectively splits $y(t) = y_1(t) + y_2(t)$ where $y_1(t)$ is a solution of a *non-linear* equation (101) with $f(u)$ replaced by $f_0(u)$ and $y_2 = y - y_1$ is a correction. In this case $y_1(t)$ again satisfies (105), and $y_2(t)$ belongs to a compact set in E when $y(0)$ is in B_0. Note that the set $\mathcal{A}$ is the same for the attractor in the weak topology and the norm-induced topology, but the attraction in the norm-induced topology is stronger. It was proven in [55] that the attractor $\mathcal{A}$ is compact in E_1 and for generic g is regular and has the structure described in Theorem 2.3.5.

Remaining technical restrictions of [46,52,55] on general $f(u)$ of critical growth were removed by Arrieta, Carvalho and Hale [9], they improved the method of [52,55]. Time-periodic equations with a critical growth were considered by Ceron and Lopes [83], Lopes [284]. Ladyzhenskaya [264] in the critical case proved existence of the global attractor in the space $E_1 = H^2 \times H^1$.

REMARK. When the growth of $f(u)$ is supercritical, namely with the power $\rho < \frac{4}{d-2}$ (that is $p < 5$ for $d = 3$) it was proven by Kapitanski in [248] that solutions of the damped wave equation exist and are unique when Ω is a compact Riemannian manifold without boundary, and the equation has a global attractor. The proof uses Strichartz inequalities. In

$\mathbb{R}^d$ global attractors were constructed by Feireisl [145,146]. For general bounded domains existence of generalized attractors with supercritical growth of non-linearity with power $p < \frac{4}{d-2} + 1$ was proven by Babin and Vishik [44]. Existence of a connected generalized attractor was proven by Ball [60] where one can find a detailed exposition of the theory of generalized global attractors of damped wave equations and more references.

Dimension of attractors. For a generic g the global attractor $\mathcal{A}$ is regular and its Hausdorff dimension is given by the formula

$$\dim \mathcal{A} = \max_{z_j \in \mathcal{N}} \dim M^{\mathrm{u}}(z_j).$$

The asymptotic upper and lower estimates of $\dim \mathcal{A}$ are given in [39]. For arbitrary subcritical f and g the finite dimensionality of the attractor was proved by Ghidaglia and Temam [199] by a different method.

REMARK. The theory of attractors can be expanded to semilinear hyperbolic equations with the damping more general than caused by the term $\gamma \partial_t u$ in (98). The damping terms can be non-linear, can depend on x, in particular be localized on a set smaller than the entire domain (see Feireisl and Zuazua [155], Feireisl [146]). The damping may include differential operators (see Belleri and Pata [63], Carvalho and Cholewa [78], Zhou [395]). Equations with a non-linear damping were studied by Haraux [226], Raugel [335]. Damping via the boundary dissipation was studied by Chueshov, Eller and Lasiecka [108].

3.4. *Fragmentation complexity of attractors of PDE*

In this subsection we give examples of equations which have global attractors with a large fragmentation number and, therefore, high complexity.

3.4.1. *Scalar parabolic equation with humpy coefficients.* We consider here, following [17] and [24], a scalar parabolic equation of the form

$$\partial_t u = a\Delta u - F'(u, x), \quad t \geqslant 0, \tag{106}$$

where $a > 0$ is a constant, Δ is the Laplacian, $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, $u = u(x, t)$. We impose the Dirichlet boundary condition (84). In (106) $F'(u, x)$ is the derivative of the potential $F(u, x)$ with respect to u, the potential is non-negative

$$F(u, x) \geqslant 0. \tag{107}$$

We assume that $F'(u, x)$ is continuously differentiable with respect to (u, x) and we assume the following absorption condition:

$$\sup_{u \leqslant \beta_1} F'(u, x) \leqslant 0, \qquad \inf_{u \geqslant \beta_2} F'(u, x) \geqslant 0, \tag{108}$$

where $\beta_1 < \beta_2$. Everywhere in this subsection for simplicity we assume $\beta_1 = -\beta_2$. We denote the class of functions $F(u,x)$ that satisfy the above conditions by $\mathcal{NL}(\beta_1,\beta_2)$. We take as Ω the cube $\Omega_N = \{x\colon |x_i| \leqslant N, i = 1,\dots,d\}$ where $N > 0$ is a (large) integer. We consider smaller cubes with side $R \leqslant 1$

$$\Omega(\bar{j},R) = \big\{x \in \mathbb{R}^d\colon |x - x_0(\bar{j})|_\infty \leqslant R/2\big\}, \quad |x|_\infty = \max_i |x_i|,$$

the cubes are centered at the points $x_0(\bar{j}) \in \mathbb{R}^d$:

$$x_0(\bar{j}) = \frac{1}{2}\vec{1} + \bar{j}, \quad \bar{j} \in \mathbb{Z}^d,\ \vec{1} = (1,\dots,1).$$

Obviously, the domain Ω_N is a union of $(2N)^d$ cubes $\Omega(\bar{j},1)$.

From the Maximum Principle using (108) one can easily deduce that the set

$$E(\beta_1,\beta_2) = \big\{u \in C_0(\Omega)\colon\ \beta_1 \leqslant u(x) \leqslant \beta_2\big\}$$

is attracting and invariant.

THEOREM 3.4.1. *When $F \in \mathcal{NL}(\beta_1,\beta_2)$, Equation* (106) *generates a semigroup S_t in $C_0(\Omega) = C(\Omega) \cap \{u|_{\partial\Omega} = 0\}$, this semigroup possesses a global attractor $\mathcal{A}(F,N)$ in $C_0(\Omega)$. This attractor lies in the set $E(\beta_1,\beta_2) \subset C_0(\Omega)$.*

We consider potentials $F(u,x)$ which have a graph with a hump in every box $\Omega(\bar{j},R)$ where $|u|$ is small. As a particular example we consider the function

$$F(u,x) = \psi(x)\big[f(u) + h(u,x)\big], \tag{109}$$

where

$$\psi(x) = \mu \sin^2\left(\frac{\pi x_1}{L}\right)\cdots\sin^2\left(\frac{\pi x_d}{L}\right) + q_0 a, \tag{110}$$

$$f(u) = \frac{1}{4}u^4 - \frac{1}{2}u^2 + \frac{1}{4}$$

with a fixed $\mu > 0$ and a fixed small $q_0 > 0$. The function $h(u,x)$ is a perturbation, it is an arbitrary function that is twice differentiable in u, the derivatives are continuous in x and the following condition holds:

$$\big|\partial_u h(u,x)\big| + \big|h(u,x)\big| < q_0 a, \quad u \in \mathbb{R}, x \in \mathbb{R}^d. \tag{111}$$

We include this function to show that the lower estimates of complexity given below are uniform with respect to perturbations.

The next theorem is a corollary of general results of Babin [24]. It shows that the fragmentation complexity of the attractor (see Definition 1.6.1) tends to infinity as the size

of the domain tends to infinity, the complexity grows proportionally to the volume of the domain Ω_N. We take in (108) $\beta_2 = 1 + q_0 a$, $\beta_1 = -1 - q_0 a$.

THEOREM 3.4.2. *Let $F(u,x)$ be defined by* (109), *let $q_0 > 0$ be sufficiently small. Then for any h satisfying* (111), *for any natural N the global attractor $\mathcal{A}(F, \Omega_N)$ has the fragmentation complexity* $\log_2 \mathsf{Fr}(\mathcal{A}(F, \Omega_N))$ *that satisfies the inequality*

$$\log_2 \mathsf{Fr}\big(\mathcal{A}(F, \Omega_N)\big) \geqslant (2N)^d \tag{112}$$

and the average spatial complexity $\mathsf{cmp}(F)$ *defined by* (25) *satisfies the inequality* $\mathsf{cmp}(F) \geqslant 1$.

Note that the potential $F = F_h$ defined by (109), (110) is a $\mathcal{C}^2$ perturbation of F_0 given by (109), (110) with $h = 0$; according to (111) this perturbation is small in $\mathcal{C}^1$. If $\Theta = \bigcup_h \{F\}$ is the set of potentials defined by (109), (110), (111) then according to (112) the fragmentation complexity of the family of attractors $\mathcal{A}(\Theta, \Omega_N)$ defined by (24) satisfies the estimate

$$\log_2 \mathsf{Fr}\big(\mathcal{A}(\Theta, \Omega_N)\big) \geqslant (2N)^d$$

and $\mathsf{cmp}(\Theta) \geqslant 1$.

We present main ideas of the proof of Theorem 3.4.2. Let $J(N)$ be the set of indices $\bar{j}$ that numerate the domains $\Omega(\bar{j}, R) \subset \Omega_N$

$$\bar{j} \in J(N) = \{-N, \dots, N-1\}^d.$$

We consider binary lattice functions $\zeta(\bar{j})$ on $J(N)$ that take at every point $\bar{j} \in J(N)$ values 0 or 1. For every binary lattice function $\zeta(\bar{j})$ on $J(N)$ following [24] we define an invariant subset $\widetilde{E}(\zeta) \subset E(\beta_1, \beta_2)$ with a non-empty interior. Note that graphs of functions from $E(\beta_1, \beta_2)$ lie in the slab $D = [\beta_1, \beta_2] \times \Omega_N$. Obviously, $D = \bigcup_{\bar{j} \in J(N)} D(\bar{j})$ where $D(\bar{j}) = [\beta_1, \beta_2] \times \Omega(\bar{j}, 1)$. The subset $\widetilde{E}(\zeta)$ is defined as follows. For every cube $\Omega(\bar{j}, 1)$ we find two barrier functions $U^+_{\bar{j}}(x)$ and $U^-_{\bar{j}}(x)$ (solutions of $a\Delta U - F'(U, x) = 0$) such that

$$U^+_{\bar{j}}|_{\partial\Omega(\bar{j},1)} = \beta_1, \qquad U^+_{\bar{j}}(x) < \beta_2 \quad \text{when } x \in \Omega(\bar{j}, 1),$$

$$U^-_{\bar{j}}|_{\partial\Omega(\bar{j},1)} = \beta_2, \qquad U^-_{\bar{j}}(x) > \beta_1 \quad \text{when } x \in \Omega(\bar{j}, 1).$$

The existence of such functions is proven in [24]; it is the most technical part of the construction. We introduce domains

$$D(0, \bar{j}) = \big\{(u, x) \in D(\bar{j})\colon\ u < U^-_{\bar{j}}(x)\big\},$$

$$D(1, \bar{j}) = \big\{(u, x) \in D(\bar{j})\colon\ u > U^+_{\bar{j}}(x)\big\}.$$

For a given binary function $\zeta(\bar{j})$ we define a connected set $D(\zeta) \subset \Omega$ by the formula

$$D(\zeta) = \bigcup_{\bar{j} \in J(N)} D\big(\zeta(\bar{j}), \bar{j}\big).$$

We denote by $\widetilde{E}(\zeta)$ the subset of functions from $E(\beta_1, \beta_2)$ with graphs in $D(\zeta)$. Every set $\widetilde{E}(\zeta)$ is closed, has a non-empty interior $\widetilde{E}^0(\zeta)$ and $\widetilde{E}(\zeta) \cap \widetilde{E}(\zeta') = \emptyset$ when $\zeta \neq \zeta'$. By the maximum principle $\widetilde{E}(\zeta)$ is invariant, the restriction of the semigroup S_t to this set has an attractor $\mathcal{A}_\zeta \subset \widetilde{E}^0(\zeta) \cap \mathcal{A}$. Since the number of binary functions on the lattice $J(N)$ equals $2^{(2N)^d}$ we obtain (112).

Now we give an example of a two-sided estimate of the fragmentation complexity.

Consider one-dimensional equation in the domain $\Omega_N = \{x\colon -N < x < N\}$,

$$\partial_t u = a\partial_x^2 u - F'(u, x), \quad t \geqslant 0,\ -N < x < N, \tag{113}$$

where F is the same as in (109) with $d = 1$, $x = x_1$, $h = 0$ and with the boundary conditions

$$u(-N, t) = 0, \qquad u(N, t) = \xi, \tag{114}$$

$|\xi| \leqslant \frac{1}{2}$. Now the parameter that defines the family Θ is the boundary value ξ. For every ξ we have a global attractor $\mathcal{A}(\xi)$ in the space $\mathcal{C}([-N, N]) \cap \{u(-N) = 0, u(N) = \xi\}$.

THEOREM 3.4.3. *Let $\Theta = \{\xi\colon |\xi| < \epsilon_0\}$, where $\epsilon_0 > 0$. We have for every $\epsilon_0 \leqslant 1/2$ for all $N = 1, 2, \ldots$ the following estimate*

$$1 \leqslant \frac{1}{2N} \log_2 \mathsf{Fr}\big(\mathcal{A}(\Theta, \Omega_N)\big) \leqslant C_0,$$

therefore $C_0 \geqslant \mathsf{cmp}(\Theta) \geqslant 1$.

The lower estimate is similar to (112). The upper estimate of the fragmentation complexity is based on the following observations. First, since the semigroup has a global Lyapunov function, by the invariance principle of La Salle every subattractor contains an equilibrium and the fragmentation number $\mathsf{Fr}(\mathcal{A})$ is not greater than the number of equilibrium points. Second, we use the following lemma.

LEMMA 3.4.4. *Let $f(u, x)$ be a differentiable function which satisfies the following sign condition*

$$f(u, x)u > 0 \quad \text{when } |u| \geqslant 1,\ -\infty < x < \infty,$$

and the growth condition

$$\sup_{|u| \leqslant 1} \big[\big|f(u, x)\big| + \big|\partial_u f(u, x)\big| + \big|\partial_x f(u, x)\big| + \big|\partial_u^2 f(u, x)\big|\big] \leqslant C_1.$$

Then there exist such $\alpha > 0$ and $M_0 > 0$ which depend only on C_1 that for any $\varepsilon > 0$, for any $N \geqslant 1$ there exists a set $\overline{\Xi} \subset [-\frac{1}{2}, \frac{1}{2}]$ with its Lebesgue measure $|\overline{\Xi}| \leqslant \varepsilon/2$ such that if $\xi \notin \overline{\Xi}$ then the equation

$$\partial_x^2 u - f(u, x) = 0 \tag{115}$$

with boundary conditions (114) *has no more than $M_0 e^{\alpha N}/\varepsilon$ solutions.*

Using Lemma 3.4.4 we see that the infimum over Θ of $\frac{1}{2N} \log_2 \mathrm{Fr}(\mathcal{A}(\xi, \Omega_N))$ is not greater than $\alpha + \log_2(M_0/\varepsilon_0) = C_0$.

REMARK. More general parabolic equations of the form

$$a_0(x)\partial_t u = \sum_{i,j=1}^{d} \partial_i \big(a_{ij}(x)\partial_j u\big) - F'(u, x) \tag{116}$$

with a non-negative potential $F(u, x)$ are studied similarly to (106) in [24]. We discuss here for simplicity only the special case (106), we also impose simplifying assumptions; more general cases are considered in [24]. Note, in particular, that the variability of $a_0(x)$, $a_{ij}(x)$ implies non-trivial effects; see [24] for details.

REMARK. Multiple stable, non-constant solutions of parabolic equations were constructed in works [164,192,222,294,383,323,7,79,80,191,223,339,190,189]. The approach of [7, 339,79,80,191,223] is based on singular limit techniques. In the one-dimensional case it is possible to give examples when the long-time dynamics is explicitly described in detail by a reduction to a system of ordinary differential equations (see [7,191,223,339]). Methods of [79,80,191,223] are based on a reduction to a finite-dimensional invariant manifold.

REMARK. A damped hyperbolic equation of the form

$$\partial_t^2 u + \gamma \partial_t u = a\Delta u - F'(u, x), \quad \gamma > 0, \tag{117}$$

with a humpy non-linearity (109), (110) and Dirichlet boundary condition (84) is considered in [26]. The attractor of this equation admits the same lower bound (112) of fragmentation complexity.

Unbounded domain. When the equation is considered in the whole space $\mathbb{R}^d$ we replace (108) by

$$\sup_{u \leqslant \beta_1} F'(u, x) \leqslant -\epsilon, \qquad \inf_{u \geqslant \beta_2} F'(u, x) \geqslant \epsilon, \quad \epsilon > 0, \tag{118}$$

with an arbitrary small ϵ. We denote by $\mathcal{C}_b(\mathbb{R}^d)$ the space of uniformly bounded continuous functions, the norm in $\mathcal{C}_b(\mathbb{R}^d)$ coincides with the norm in $L_\infty(\mathbb{R}^d)$. We use in this space the topology $\mathcal{C}_{\mathrm{loc}}(\mathbb{R}^d)$ of uniform convergence on bounded sets in $\mathbb{R}^d$. This topology is metrizable (one can use a formula similar to (171) to introduce the metric).

THEOREM 3.4.5. *When $F \in \mathcal{NL}(b_1, b_2)$ Equation* (106) *generates a semigroup $\{S_t\}$ in $\mathcal{C}_b(\mathbb{R}^d)$. The operators S_t are continuous in the topology of $\mathcal{C}_{\rm loc}(\mathbb{R}^d)$ and when $t > 0$ they map sets bounded in $\mathcal{C}_b(\mathbb{R}^d)$ into compact sets in $\mathcal{C}_{\rm loc}(\mathbb{R}^d)$. This semigroup possesses a global attractor $\mathcal{A}(F)$ in the topology $\mathcal{C}_{\rm loc}(\mathbb{R}^d)$. This attractor lies in the set $E(b_1, b_2) \subset \mathcal{C}_b(\mathbb{R}^d)$.*

Using binary functions $\zeta(\bar{j})$ defined on $\bar{j} \in \mathbb{Z}^d$ as in the proof of Theorem 3.4.2 we obtain the following theorem.

THEOREM 3.4.6. *Let $F(u, x)$ be defined by* (109) *and $q_0 > 0$ be sufficiently small. Then for any h satisfying* (111) *$\mathcal{A}(F)$ has the infinite fragmentation number*: $\mathsf{Fr}(\mathcal{A}(F)) = \infty$.

3.4.2. *Complexity of attractors of spatially homogeneous systems.* First we show that global attractors of *scalar* spatially homogeneous parabolic equations generically do not have high fragmentation complexity. A spatially homogeneous scalar parabolic equation

$$\partial_t u = \partial_x^2 u - f(u), \quad -L < x < L, \tag{119}$$

with the periodic boundary conditions

$$u(L) = u(-L), \qquad \partial_x u(L) = \partial_x u(-L), \tag{120}$$

where $f(u)$ satisfies the absorption condition

$$f(u)u \geqslant c > 0 \quad \text{for } |u| \geqslant 1$$

generates in the Sobolev space $H^1_{\rm per}([-L, L])$ a semigroup $\{S_t\}$ that possesses a global regular attractor $\mathcal{A}(f, L)$. The equilibria of S_t coincide with $2L$-periodic solutions of the steady-state equation

$$\partial_x^2 z - f(z) = 0.$$

For a generic f this equation has a finite (modulo a spatial shift) number of L-periodic solutions; the number of corresponding curves (with a fixed L) in the phase plane is uniformly bounded as $L \to \infty$. Taking into account that nl-periodic solution with an integer n is l-periodic the number of L-periodic solutions can not grow faster than CL as $L \to \infty$. Therefore the fragmentation number $\mathsf{Fr}(\mathcal{A}(f, L)) \leqslant CL$. When $f \in \Theta$, Θ being a small neighborhood of a given f_0 in $\mathcal{C}^1(\mathbb{R})$, the average spatial complexity of the attractor of a scalar equation

$$\mathsf{cmp}(\Theta) = \lim_{L\to\infty} \inf \frac{\log_2 \mathsf{Fr}(\mathcal{A}(\Theta, L))}{2L} = 0.$$

The simplest equation with positive average spatial complexity of the attractor is the two-component parabolic system of the form

$$\partial_t u = \partial_x^2 u - F'(u) \tag{121}$$

with the boundary condition (120). Here $u = (u_1, u_2) \in \mathbb{R}^2$, $F(u) = F(u_1, u_2)$ is a given smooth non-negative potential function

$$F(u) \geqslant 0 \tag{122}$$

which has the gradient $F' = \nabla F$ which is assumed to be Lipschitzian in u. Here we consider a simple example of $F(u)$, for a more general and detailed treatment see Afraimovich, Babin and Chow [3]. We consider here for simplicity a potential that satisfies for large $|u|$ the absorption condition, namely for some $c > 0$

$$F_0'(u) \cdot u \geqslant c|u|^2 - C, \quad u \in \mathbb{R}^2. \tag{123}$$

The energy functional

$$\mathcal{E}_L(u) = \int_{-L}^{L} \left[\frac{1}{2} |\partial_x u|^2 + F(u) \right] dx \tag{124}$$

is the global Lyapunov function for this equation.

Under the above conditions (121) generates in $H^1_{\text{per}}([-L, L])$ a semigroup $\{S_t\}$ that possesses a global (regular for generic F) attractor $\mathcal{A}(F, L)$. The equilibria of S_t coincide with $2L$-periodic solutions of the steady-state system

$$\partial_x^2 U - F'(U) = 0. \tag{125}$$

Now we introduce conditions which guarantee that the attractor $\mathcal{A}(F, L)$ has a high fragmentation number and complexity for large L. We assume that the graph of the potential $F(u)$ has two humps at two points $P_1 = (-R, 0)$ and $P_2 = (R, 0)$, $R > 0$. Namely, $F(u)$ is large in a disc of a smaller radius $r_2 < R$ near the point P_i

$$F(u) \geqslant M \quad \text{when } |u - P_i| \leqslant r_2, i = 1, 2, \tag{126}$$

and is smaller at the boundary of a larger disc

$$F(u) \leqslant m \quad \text{when } |u - P_i| = R, i = 1, 2, \tag{127}$$

where the constants m and M are chosen to satisfy the following condition

$$\frac{r_2 \sqrt{M}}{R \sqrt{m}} > (\pi + 1). \tag{128}$$

Now we introduce invariant classes of solutions, the classes are determined by how many times, in what sequence and in which direction the graphs of solutions wind around the points P_1 and P_2. More precisely, we consider the domain Ω'' in the u-plane obtained by deletion of the points P_1, P_2 from $\mathbb{R}^2$, $\Omega'' = \mathbb{R}^2 \setminus (P_1 \cup P_2)$. The fundamental homotopy

group $\pi_1(\Omega'', \sigma) = \pi_1$ of the set Ω'' is a free group with two generators g_i, $i = 1, 2$, corresponding to counterclockwise cycles in Ω'' around points P_i starting from and ending at the designated point, namely at the origin. Elements of the group π_1 are uniquely represented by irreducible words of the form $b = \prod_{l=1}^{n} g_{i_l}^{k_l}$ with $k_l = \pm 1$; $n = |b|$ is the length of the word b. Homotopy classes of closed curves without a designated point form the set π_1'; the classes correspond to the words from π_1 with the additional restriction $g_{i_1}^{k_1} \neq g_{i_n}^{-k_n}$, the words obtained by cyclic permutations being equivalent. The set of classes of equivalence with respect to the cyclic permutations of the words from the group π_1 is also denoted by π_1'. For every fixed t a solution $u(x,t)$ that does not pass through the points P_1, P_2 is a closed curve in Ω'' and has a homotopy type $h(u(\cdot,t)) \in \pi_1'$. We describe below sets of initial data u on which $h(S_t u)$ does not depend on t, that is the curve $u(x,t)$ has the same homotopy type for every $t \geqslant 0$. First we have to introduce necessary definitions. We introduce the Jacobian (or Fermat–Maupertuis) functional

$$\mathcal{J}(u) = \int_{-L}^{L} \sqrt{2F\big(v(x)\big)}\,\big|\partial_x v(x)\big|\,dx \leqslant \mathcal{E}(v), \tag{129}$$

the value of $\mathcal{J}(v)$ is determined by the graph of the curve v and does not depend for a given curve on its x-parameterization and on L.

Jacobian distance between two points $Q_1, Q_2 \in \mathbb{R}^2$ is given by $\operatorname{dist}_{\mathcal{J}}(Q_1, Q_2) = \inf_v \mathcal{J}(v)$ where the infimum is over the curves $v(s)$ that connect Q_1, Q_2.

DEFINITION 3.4.7. Let a smooth simple Jordan curve C_i bound a domain Ω_i, $C_i = \partial\Omega_i$ (Ω_i is an open domain and $\overline{\Omega}_i$ is its closure). It is called a minimal cycle if it has the following minimality property. For any closed cycle Γ which lies in an ϵ-neighborhood $O_\epsilon(C_i)$ of the cycle in the u-plane, $\Gamma \subset O_\epsilon(C_i) \cap \Omega_i$, once encircles $P_i \in \Omega_i \setminus \overline{O}_\epsilon(C_i)$ and has a point $Q \in \Gamma$ strictly inside Ω_i at the distance ϵ_1 from C_i the following inequality holds: $\mathcal{J}(\Gamma) \geqslant \mathcal{J}(C_i) + \epsilon_2$ where $\epsilon_2 > 0$ depends on ϵ_1 and does not depend on Γ.

DEFINITION 3.4.8. A cycle C_i encircling a disc Ω_i is called η-stable (with $\eta > 0$) with respect to a cycle $C_i^1 = \partial\Omega_i^1$, $\overline{\Omega}_i^1 \subset \Omega_i$, if for any closed curve Γ^0 which lies in $\overline{\Omega}_i \setminus \Omega_i^1$, once encircles $\overline{\Omega}_i^1$ and intersects with C_i^1 there exists a homotopy Γ^t, $0 \leqslant t \leqslant 1$, of this curve inside the ring $\overline{\Omega}_i \setminus \Omega_i^1$ which does not move points of $\Gamma^0 \cap C_i$ such that $\Gamma^1 \subset C_i$, $\mathcal{J}(\Gamma^0) \geqslant \mathcal{J}(\Gamma^1) + \eta$.

THEOREM 3.4.9. *Let* (123), (126)–(128) *hold. Let* $C_i^1 = \{u\colon |u - P_i| = r_2/(\pi+1)\}$, $\eta = 2\sqrt{2M}r_2\pi/(\pi+1) - 2\pi R\sqrt{2m}$. *Then there exist two non-intersecting minimal cycles* C_i, $i = 1, 2$, *that encircle* P_i, $i = 1, 2$. *The cycles* C_i *are* η*-stable with respect to* C_i^1.

DEFINITION 3.4.10. We denote by Ξ the set of closed continuous curves $u(s)$ (parametrized by $s \in [0,1]$, $u(0) = u(1)$) which do not intersect with both domains Ω_i^1 encircled

by C_i^1 and have a finite Jacobian length. We denote by Ξ_b for a given homotopy class $b \in \pi_1'(\mathbb{R}^2 \setminus \bigcup_1^2 \Omega_i^1)$ the set of curves from Ξ which belong to b and put

$$\ell^*(b) = \inf_{u \in \Xi_b} \mathcal{J}(u).$$

We denote by $\Xi'([-L, L])$ and $\Xi_b'([-L, L])$ the set of functions $u \in H^1_{\text{per}}([-L, L])$ which graphs are curves from Ξ and Ξ_b respectively.

We formulate two theorems that follow from Theorems 3.1 and 4.1 and Proposition 6.1 of [3].

THEOREM 3.4.11. *Let* (123)–(128) *hold. Let C_i, $i = 1, 2$, be the η-stable cycles described above; let $b \in \pi_1'$ be a homotopy class. Then the sets $\Xi_b'([-L, L]) \cap \{u\colon \mathcal{E}_L(u) \leqslant \lambda\}$ where $\lambda < \ell^*(b) + \eta$ are invariant with respect to the flow defined in $H^1_{\text{per}}([-L, L])$ by the parabolic equation* (121) *with the periodic boundary conditions* (120).

THEOREM 3.4.12. *Let* (123), (126)–(128) *hold. For every non-trivial homotopy class $b \in \pi_1'$ there exists L such that $L_*|b| \leqslant L \leqslant L^*|b|$ where L^* and L_* do not depend on b and a steady-state L-periodic solution U of Equation* (125) *such that $U \in \Xi_b'([-L, L])$,*

$$\mathcal{E}_L(U) = \mathcal{J}(U) = \ell^*(b), \tag{130}$$

U is a global minimizer of $\mathcal{E}_L(u)$ in $\Xi_b'([-L, L])$, its graph lies in the domain $|u| \leqslant 2\sqrt{C/c}$ and the following integral holds:

$$\sqrt{2F\big(U(x)\big)} = |\partial_x u|, \quad x \in [-L, L]. \tag{131}$$

To estimate the number of global minimizers with a given L we use the following lemma.

LEMMA 3.4.13. *Let C_i, $i = 1, 2$, be the η-stable cycles described above. Let M_1 be the maximum of $F(u)$ over $|u| \leqslant 2\sqrt{C/c}$ where C, c are defined in* (123), *$\delta_\eta = \eta/(4M_1)$. Let $U \in \Xi_b'([-L, L])$ be a global minimizer of $\mathcal{E}_L(U)$. Let $L \leqslant L' \leqslant L + \delta_\eta$. Then there exists a global minimizer of $\mathcal{E}_{L'}(U)$ in $\Xi_b'([-L', L'])$, $\mathcal{E}_{L'}(U) \leqslant \ell^*(b) + \eta$.*

Note that in Theorem 3.4.12 L depends on b, $L = L(b)$. To estimate the average fragmentation complexity of the attractor we have to find for a fixed L many stable equilibria and then take a sequence $L \to \infty$. According to [2] the number of classes b from π_1' with a given length $|b|$ equals $K_{|b|} = 3^{|b|} + 2 + (-1)^{|b|}$. By Theorem 3.4.12 for every $b \in \pi_1'$ we have a global minimizer of the periodic problem with the half-period $L(b) \in [L_*|b|, L^*|b|]$. Using Lemma 3.4.13 we conclude that we have global minimizers of periodic problems with half periods $L'(b)$ that cover the segment $[L(b), L(b) + \delta_\eta] \subset [L_*|b|, L^*|b|]$. Therefore we have $K_{|b|}$ of intervals with length δ_η that lie in $[L_*|b|, L^*|b|]$. Hence there is a point $L_0 \in [L_*|b|, L^*|b|]$ that belongs to at least $\delta_\eta K_{|b|}/((L^* - L_*)|b|)$ intervals

$[L(b), L(b)+\delta_\eta]$. The attractor $\mathcal{A}(F, L_0)$ contains at least $\delta_\eta K_{|b|}/((L^*-L_*)|b|)$ sub-attractors $\mathcal{A}_b(F, L_0')$ and $\mathsf{Fr}(\mathcal{A}(F, L_0)) \geqslant \delta_\eta K_{|b|}/((L^*-L_*)|b|)$,

$$\log_2 \mathsf{Fr}\big(\mathcal{A}(F, L_0)\big) \geqslant |b| \log_2 3 - C_1.$$

We take $L \to \infty$, it contains a subsequence L_0 we constructed above with $|b| \to \infty$ and we get

$$\mathsf{cmp}\big(\mathcal{A}(F)\big) = \lim_{L\to\infty} \inf \frac{\log_2 \mathsf{Fr}(\mathcal{A}(F, L))}{2L} \geqslant \lim_{|b|\to\infty} \frac{|b| \log_2 3 - C_1}{L^*|b|} = \frac{\log_2 3}{L^*} > 0. \tag{132}$$

COROLLARY 3.4.14. *For every F that satisfy* (123)–(128) *the average spatial fragmentation complexity of the attractor $\mathcal{A}(F)$ is positive*: $\mathsf{cmp}(F) \geqslant c_0 > 0$.

Note that the set of potentials F which satisfy the above conditions is a set with a non-empty interior in $\mathcal{C}^2(\mathbb{R}^2)$ and this set is large.

REMARK. Similar results hold for hyperbolic equations, see [26].

REMARK. Lower bounds of the fragmentation number of attractors of strongly non-linear systems are given in [18].

REMARK. We can consider (119) with F fixed and with the boundary conditions similar to (114)

$$u(-L) = 0, \qquad u(L) = \xi,$$

where ξ lies in a small neighborhood Θ of the origin in the u-plane. The attractor $\mathcal{A}(\xi)$ depends on ξ and similarly to (132) we obtain that its complexity is positive $\mathsf{cmp}(\mathcal{A}(\Theta)) > 0$, we also obtain that for every L for almost all ξ (119) with the above boundary conditions has a finite set of equilibria and its complexity is bounded, $\mathsf{cmp}(\mathcal{A}(\Theta)) \leqslant C$.

3.5. *Equations in unbounded domains*

Compactness properties of sets of functions defined on unbounded domains differ from the properties of functions defined on bounded domains. According to Arzela–Ascoli theorem a sequence of functions in a bounded domain which are uniformly bounded and have uniformly bounded derivatives contains a uniformly convergent subsequence. This is not true in unbounded domains. Consider for example a smooth bounded non-constant function $\varphi(x)$, $-\infty < x < +\infty$ with a bounded derivative which has limits $\varphi(\pm\infty)$ at plus and minus infinity. Translations of this function

$$\varphi(x - ct), \quad -\infty < x < +\infty, \ c \neq 0 \tag{133}$$

when $t \to \infty$ are bounded and derivatives are uniformly bounded in $\mathcal{C}(\mathbb{R})$. If $t \to \infty$ $\varphi(x - ct)$ tends to the same constant $\varphi(-c\infty)$ for every x. Since φ is not constant, the difference $\varphi(x - ct) - \varphi(-c\infty)$ is large on some interval, therefore $\varphi(x - ct)$ does not contain a subsequence with $t_j \to \infty$ which converges on the whole real line in $\mathcal{C}(\mathbb{R})$. Therefore a set bounded in $\mathcal{C}^1(\mathbb{R})$ is not compact in $\mathcal{C}(\mathbb{R})$. Similarly, the embedding of Sobolev spaces $H^s(D) \subset H^\sigma(D)$ with $s > \sigma$ is not compact. Therefore, in unbounded domains additional smoothness does not imply compactness. If $\|u\|$ is any translation invariant norm such that convergence in this norm implies for almost every x pointwise convergence of $u(x)$ the above argument shows that $\varphi(x - ct_j)$, $t_j \to +\infty$, does not contain a convergent subsequence in this norm. These properties of function spaces imply essential differences between the dynamics generated by PDE in bounded and unbounded domains and require introduction of Banach spaces with norms that are not translation invariant, namely weighted spaces. Another possibility is to use weak convergence or convergence on bounded intervals which leads to metric spaces where the metric is not translation invariant.

We consider here the parabolic system of equations

$$\partial_t u = \nu \Delta u - \Lambda_0 u - f(u) - g \tag{134}$$

in an unbounded domain. Here $u = (u_1, \dots, u_n)$, $f = (f_1, \dots, f_n)$, $g = (g_1, \dots, g_n)$. We assume that $f(0) = 0$. Here Λ_0 is a positive matrix,

$$\Lambda_0 u \cdot u \geqslant \lambda_0 |u|^2, \quad \lambda_0 > 0.$$

For simplicity we will consider equations in $\Omega = \mathbb{R}^d$ and we consider for simplicity a second-order operator of the form $\nu\Delta$, though all results of [57] which we cite in the beginning of this section hold for more general systems. Under some natural conditions solutions of such equations exist and are unique in classes of functions that grow at spatial infinity slower than $e^{\varepsilon|x|^2}$, $\varepsilon > 0$ (see [57] for details). Therefore a semigroup $\{S_t\}$ is well-defined in a corresponding space E of functions with a prescribed growth. One can choose different weights and, therefore, different spaces E of initial data for the same equation. The properties of semigroups and their attractors depend on the choice of the space E.

A way to describe the behavior of functions at spatial infinity is via weighted norms. We introduce the norm $\|\cdot\|_{0,\gamma}$ in the space $H^0_\gamma(\mathbb{R}^d) = H^0_\gamma$ by the formula

$$\|u\|_{0,\gamma} = \left[\int \left(1 + |x|^2\right)^\gamma \left|u(x)\right|^2 dx\right]^{1/2}.$$

Similarly,

$$\|u\|^2_{1,\gamma} = \|u\|^2_{0,\gamma} + \|\nabla u\|^2_{0,\gamma}$$

etc. When $\gamma > 0$ the functions from H^0_γ decay at infinity faster than functions from $L_2(\mathbb{R}^d)$. When $\gamma < 0$ they decay slower, and when $\gamma < -d/2$ the weight $(1 + |x|^2)^\gamma$ is integrable

and the space H^0_γ includes functions that do not decay at all, for example every bounded over $\mathbb{R}^d$ measurable function belongs to $H^0_\gamma(\mathbb{R}^d)$. Note that the weighted norms with $\gamma \neq 0$ are not translation invariant, but the function spaces are translation invariant.

As always, we impose the sign conditions

$$f(u)\cdot u \geqslant -C|u|^2, \qquad f'(u) \geqslant -CI \tag{135}$$

and the growth condition

$$\big|f'(u)\big| \leqslant C\big(1+|u|^{p_2}\big), \quad 0 \leqslant p_2 \leqslant p_0,\ p_0 = \min\big(4/d, 2/(d-2)\big) \tag{136}$$

(when $d \leqslant 2$ the only condition on p_2 is $p_2 \geqslant 0$).

3.5.1. *Dynamics in spaces that include non-decaying functions.* The spaces H^0_γ with $\gamma < 0$ are wider than $H^0_0 = L_2(\mathbb{R}^d)$, we discuss this case first (see [57] for more details and variants).

THEOREM 3.5.1 [57]. *Let* $\gamma < 0$, $g \in H_{0,\gamma}$, *let* (135)–(136) *hold. Let* $\gamma_1 = \max((\gamma-1)/(p_2+1), \gamma)$. *Then the semigroup* $\{S_t\}$ *is well-defined on* $H = H^0_{\gamma_1} \cap H^1_\gamma$ *and operators* S_t *are continuous on* H *in the weak topology of the Hilbert space* H.

THEOREM 3.5.2. *Let* $\gamma < -d/2$, *conditions of Theorem* 3.5.1 *hold and*

$$f(u)\cdot u \geqslant -C \quad \textit{for all } u. \tag{137}$$

Then there exists a global attractor $\mathcal{A}$ *of* $\{S_t\}$ *in the weak topology of* H.

The proof is based on a construction of a bounded absorbing ball in H and uses the compactness of a bounded ball in a reflexive space in the weak topology, see a sketch of an analogous proof in the next subsection.

Now we discuss the case $\gamma < -d/2$ in more detail. In this case the attractor $\mathcal{A}$ can be very large (in the next subsection we give examples when it has infinite dimension). Note that equations of the form (134) with $g = const$ in the whole space $\mathbb{R}^d$ are translation invariant and there are examples of *traveling wave solutions* of such equations, that is solutions of the form (133) which in the one-dimensional case solve the equation

$$\nu\partial_x^2\varphi + c\partial_x\varphi - \lambda_0\varphi - f(\varphi) - g = 0 \tag{138}$$

obtained from (134) by plugging in solution $\varphi(x-ct)$ in the form of (133). The traveling wave solutions, in particular their stability were studied starting from Kolmogorov, Petrovski and Piskunov [256], for modern approaches and references see [376,163,160,194,193, 343]. Note that a bounded solution of (138) with $c \neq 0$ has limits $\varphi_\pm$ as $x \to \pm\infty$, they solve equation $\lambda_0\varphi_\pm - f(\varphi_\pm) - g = 0$. Since solutions $\varphi(x-ct)$ form a bounded, strictly invariant set in H they lie in the global attractor $\mathcal{A}$. The discussion of (133) in the beginning of Subsection 3.5 shows that the existence of traveling wave solutions implies that the

attractor cannot be compact in a Banach space which has the norm invariant with respect to translations and contains $L_\infty(\mathbb{R}^1)$.

The results of Theorems 3.5.1 and 3.5.2 can be improved in the following way. Boundedness of the norm $\|g\|_{0,\gamma}$ implies that

$$\int_{|x|\geqslant r} \bigl(1+|x|^2\bigr)^\gamma \bigl|g(x)\bigr|^2 dx \to 0 \quad \text{as } r\to\infty. \tag{139}$$

The rate of decay may be very slow, but it allows to estimate the decay of steady-state solutions and, asymptotically, solutions of parabolic and damped hyperbolic equations as $t\to\infty$. The analysis is more technical than the straightforward weighted norm estimates, but it allows to prove the attraction to the attractor $\mathcal{A}$ and its compactness in the norm-induced topology of H^0_γ. This approach was realized by Feireisl [147], see also [151,305, 303]. In these works the attraction to the attractor and the compactness of the attractor of parabolic and damped hyperbolic equations is proven in the norm-induced topology of H^0_γ.

The weighted norms $\|u\|_{l,\gamma}$ with $\gamma\neq 0$ are not invariant with respect to the translations $T_y u(x) = u(x+y)$. One may introduce a translation-invariant norm based on weighted norms as follows

$$\sup_y \|T_y u\|_{l,\gamma},$$

where $T_y u(x) = u(x+y)$ is a translation. Feireisl [147] applied similar translation invariant spaces for the study of dynamics and global attractors of damped hyperbolic equations. Mielke and Schneider [305] and Mielke [303] study dynamics of Ginzburg–Landau, Swift–Hohenberg and Kuramoto–Shivashinsky equations in unbounded domains in such spaces systematically using properties of spatial translations of solutions.

REMARK. For a review of the dynamical properties of Ginzburg–Landau equation

$$\partial_t u = a\partial_x^2 u + u - b|u|^2 u, \quad -\infty < x < \infty,$$

where a, b are complex numbers with positive real parts see [304].

REMARK. Feireisl [148] has proved convergence as $t\to\infty$ of solution of a scalar parabolic equation in $\mathbb{R}^d$ to steady-state solutions and to soliton-type solutions.

REMARK. Merino [301] studies the semigroup generated by scalar parabolic equations and special systems in the Banach space of the form (134) with the coefficient Λ_0 which depends on x and vanishes at infinity. He studies dynamics in the space of bounded uniformly continuous functions on $\mathbb{R}^d$ and proves the existence of the global attractor in the norm-induced topology, the attractor has finite Hausdorff dimension.

3.5.2. *Dynamics in spaces of decaying functions.* Now we consider the case $\gamma > 0$. We impose the sign conditions (135) and in addition to (136) we assume that $f(u)$ has the second-order zero at the origin, that is $f'(u) = 0$ and the growth condition holds

$$\big|f'(u) - f'(v)\big| \leqslant C|u - v|^{\alpha}\big(1 + |u| + |v|\big)^{q_0}, \tag{140}$$

where $q_0, \alpha > 0$ and for $d > 2$

$$q_0 + \alpha \leqslant p_0, \quad p_0 = \min\big(4/d, 2/(d-2)\big), \tag{141}$$

$p_0 = 4/d$ when $d \leqslant 2$. First we present two theorems from [57].

THEOREM 3.5.3. *Let $\gamma \geqslant 0$, $g \in H_{0,\gamma}$, (135)–(136), (140) hold. Then* (134) *generates a semigroup $\{S_t\}$ in H^0_γ and H^1_γ. Operators S_t on H^0_γ are continuous on H in both the norm-induced and weak topology of $H_{0,\gamma}$ and on $H_{1,\gamma}$ in the weak topology of $H_{1,\gamma}$, they possess the smoothing property, namely they are bounded from H^0_γ to H^2_γ when $t > 0$.*

THEOREM 3.5.4. *Let $\gamma \geqslant 0$. Let conditions of Theorem* 3.5.3 *hold together with the sign condition*

$$f(u) \cdot u \geqslant 0 \quad \textit{for all } u. \tag{142}$$

Then $\{S_t\}$ on H^1_γ has a global attractor in the weak topology of H^1_γ. When $\gamma > 0$ the semigroup $\{S_t\}$ on H^0_γ has a global attractor A in the norm-induced topology of H^0_γ.

REMARK. If the forcing term $g(x)$ decays very fast as $x \to \infty$, it belongs to $H^0_{\gamma_g}$ with large γ_g. One may take initial data in a wider space H^0_γ with $\gamma < \gamma_g$. The attractor $\mathcal{A} = \mathcal{A}(\gamma)$ could depend on γ. But the analysis shows that $\mathcal{A}(\gamma) = \mathcal{A}(\gamma_g)$, the rate of decay of functions on the attractor is determined by the rate of decay of $g(x)$ (see [57]). Note that this property is asymptotical as $t \to \infty$ only, $S_t u_0$ with a finite t belongs to the same space $H_{0,\gamma}$ as u_0, this can be explicitly verified for linear parabolic equations. In addition, the attractor is bounded in H_{2,γ_g}.

Now we give a theorem on existence of the attractor in the case $\gamma = 0$ when the non-linearity is gradient, $f(u) = F'(u)$ where $F(u)$ is a potential function, (134) takes the form of (121) and the global Lyapunov function is given by

$$\mathcal{F}(u) = \int_{\mathbb{R}^d} \left[\frac{\nu}{2}|\nabla u|^2 + F(u)\right] dx. \tag{143}$$

This theorem is similar to results of [18,21]. We also give a sketch of the proof since it is typical (for details and generalizations to a strongly non-linear case and a multidimensional case see [18]).

THEOREM 3.5.5. *Let* $\gamma = 0$, *let* (135)–(136), (140) *hold and* $g = 0$. *Let the potential satisfy the conditions*

$$F(u) > 0 \quad \text{when } |u| > 0, \qquad F(0) = 0, \qquad F(u) \to \infty \quad \text{as } |u| \to \infty \tag{144}$$

and

$$f(u) \cdot u \geqslant \beta |u|^2 \quad \text{when } |u| \leqslant r_0 \text{ or } |u| \geqslant r_1, \tag{145}$$

where $\beta > 0$, $r_0 < r_1$ *are fixed. Then for every* $\mu \geqslant 0$ *the semigroup* $\{S_t\}$ *restricted to the invariant set* $\{\mathcal{F}(u) \leqslant \mu\}$ *has an* $(H_0^0, (H_0^0)_w)$ *attractor* $\mathcal{A}_\mu$ *which is bounded in* H_0^2.

PROOF. Since S_t are continuous in the weak topology (see [57]) and a bounded ball is compact in the weak topology of $H_0^0 = L_2(\mathbb{R}^d)$, it is sufficient to prove the existence of an absorbing ball. Note that (144) implies that

$$F\big(u(x)\big) \geqslant c > 0 \quad \text{when } r_0 \leqslant |u| \leqslant r_1. \tag{146}$$

Multiplying (134) by $\partial_t u$ we obtain the identity

$$\mathcal{F}\big(u(T)\big) + \int_0^T \|\partial_t u\|_{0,0}^2 \, dt = \mathcal{F}\big(u(0)\big). \tag{147}$$

We denote

$$D_1 = \big\{x\colon\ r_0 \leqslant \big|u(x)\big| \leqslant r_1\big\}, \qquad D_2 = \mathbb{R}^d \setminus D_1.$$

Using (146) we conclude that $F(u(x)) \geqslant c$ when $x \in D_1$. Therefore (147) implies that for every T the Lebesgue measure of the set $D_1(T)$ admits the estimate

$$|D_1| \leqslant \frac{1}{c} \mathcal{F}\big(u(0)\big). \tag{148}$$

We multiply (134) by u and obtain after integration by parts

$$\frac{1}{2} \partial_t \|u\|_{0,0}^2 + \nu |\nabla u|^2 + \int_{\mathbb{R}^d} f(u) \cdot u \, dx = 0. \tag{149}$$

From (148) and (145) we deduce that

$$\int_{\mathbb{R}^d} f(u) \cdot u \, dx \geqslant \beta \int_{\mathbb{R}^d} |u|^2 \, dx + \int_{D_1} \big[f(u) \cdot u - |u|^2 \beta\big] dx \geqslant \beta \|u\|_{0,0}^2 - C_1,$$

where C_1 does not depend on u and t. From (149) we obtain the differential inequality

$$\partial_t \|u\|_{0,0}^2 \leqslant 2C_1 - 2\beta \|u\|_{0,0}^2$$

and

$$\|u\|_{0,0}^2(t) \leqslant e^{-2\beta t}\|u\|_{0,0}^2(0) + \frac{C_1}{\beta}\bigl(1 - e^{-2\beta t}\bigr).$$

Therefore the ball $\|u\|_{0,0}^2 \leqslant \frac{2C_1}{\beta}$ is an absorbing ball and we can apply the theorem on existence of a global attractor in the weak topology of H_0^0. □

REMARK. The union $\bigcup_{\mu \geqslant 0} \mathcal{A}_\mu = \mathcal{A}$ is a closed attractor which attracts in the weak topology all bounded sets. The attractor $\mathcal{A}$ is minimal among all such attractors. The attractor $\mathcal{A}$ is, generally speaking, unbounded in H_0^0. Unboundedness of $\mathcal{A}$ for non-linearities similar to considered in the next example can be proven based on results of Afraimovich, Babin and Chow [3].

Below we give an example when the attractor (in the weak topology) constructed in the above theorem cannot be a global attractor (on the invariant set $\{\mathcal{F}(u) \leqslant \mu\}$) in the norm-induced topology.

EXAMPLE 3.5.6. We consider a special case of the above theorem when $d = 1$, the system includes two components, $\nu = 1$ and the potential $F(u)$ has a two-hump structure (126)–(127), and (128) is replaced by the condition

$$\sqrt{M} r_2 > \pi R \sqrt{m}. \tag{150}$$

This example essentially coincides with the example of diverging soliton-like solutions in [18]. We consider system (121) which satisfies (122)–128), (144), (145) and additional condition

$$F(-u) = F(u). \tag{151}$$

Theorems 3.5.4 and 3.5.5 are applicable to this system. We take initial data $u_0 \in H_\gamma^1$ which are odd in x, $u_0(-x) = -u_0(x)$. From (147) we infer that for every solution $u(t)$ there is a sequence $\|\partial_t u(t_j)\|_{0,0}^2 \to 0$ with $t_j \to \infty$. Using standard estimates for solutions of $\partial_t u(t_j) = \partial_x^2 u(t_j) - F'(u(t_j))$ like in Chapter 3.5 of [55] we conclude that a subsequence $u(t_j)$ converges weakly in $L_2(\mathbb{R})$ and strongly in $\mathcal{C}^2$ on every bounded interval to a solution of the steady-state equation (equilibrium)

$$\nu \partial_x^2 z - F'(z) = 0. \tag{152}$$

From (147) $\mathcal{F}(z) \leqslant \mathcal{F}(u_0)$. Since $u(x,t_j)$ are odd in x $z(x)$ is also odd and $z(0)=0$. Multiplication by $\partial_x z$ gives an integral

$$\frac{1}{2}\big|\partial_x z(x)\big|^2 - F\big(z(x)\big) = \frac{1}{2}\big|\partial_x z(0)\big|^2 - F\big(z(0)\big). \tag{153}$$

Since $\mathcal{F}(z) < \infty$ we have a sequence $x_j \to \infty$ such that $F(z(x_j)) \to 0$, $|\partial_x z(x_j)|^2 \to 0$. Therefore

$$\frac{1}{2}\big|\partial_x z(0)\big|^2 - F\big(z(0)\big) = \frac{1}{2}\big|\partial_x z(0)\big|^2 = 0.$$

The only solution of (152) with $\partial_x z(0) = z(0) = 0$ is zero. Therefore every solution $u(t)$ with odd initial data weakly converges to zero. If the attractor $\mathcal{A}$ is in the norm-induced topology, one can choose from any sequence $u(x,t_j)$, $t_j \to \infty$, a subsequence which converges in $L_2(\mathbb{R})$ by the norm to a function on the attractor, this function must be zero and we would have

$$\big\|u(x,t_j)\big\|_{L_2} \to 0, \quad t_j \to \infty. \tag{154}$$

Now we describe a specific way to choose $u_0(x)$ such that $u(x,t)$ does not tend to zero in $L_2(\mathbb{R})$ and gives rise to a pair of solitons slowly moving in opposite directions. To describe the initial data we identify $u=(u_1,u_2)\in\mathbb{R}^2$ with the complex number $u_1+iu_2\in\mathbb{C}$. We set $u_0(x) = R(-1+e^{ixk})$ when $-\frac{2\pi}{k} \leqslant x \leqslant 0$, $u_0(x)=0$ when $x < -\frac{2\pi}{k}$ and for positive x by symmetry $u_0(x) = -u_0(-x)$. We choose $k=\sqrt{2m}/R$ to minimize the integral

$$\frac{1}{2}\mathcal{F}(u_0) \leqslant \int_{-\frac{2\pi}{k}}^{0}\left[\frac{1}{2}|\partial_x u|^2 + m\right]dx = \int_{-\frac{2\pi}{k}}^{0}\left[\frac{1}{2}R^2k^2 + m\right]dx = 2\pi R\sqrt{2m},$$

where m is from (127). The curve $u_0(x)$ in the plane starts (at $x=-\infty$) at the origin, goes around the point $-R$ on the real axis counterclockwise, returns to the origin ant turns around the point R on the real axis clockwise. The curve corresponding to $u(x,t)$ depends on t continuously, it comes arbitrarily close to the origin as $x \to \pm\infty$ and passes through origin at $x=0$. From (129) it follows that the Jacobian length $\mathcal{J}(u)$ of this curve is bounded by $\mathcal{F}(u)$, $\mathcal{J}(u) \leqslant \mathcal{F}(u)$. From results of [2,3] it follows that the curve $u(x,t)$ never gets to the points $\pm R$. In the particular case we consider here the proof is very simple. Namely, if the closed curve connects the origin with the point R it must at least twice intersect the circle $|R-u| = r_2$. The Jacobian length of the arc inside this circle that passes through its center can be estimated as follows:

$$\frac{1}{2}\mathcal{J}(u) \geqslant \int_{x_1}^{x_2}\sqrt{2F(u)}|\partial_x u|\,dx = \int_{s_1}^{s_2}\sqrt{2F(u)}\,ds \geqslant \int_{s_1}^{s_2}\sqrt{2M}\,ds \geqslant 2\sqrt{2M}r_2.$$

Therefore

$$2\sqrt{2M}r_2 \leqslant \frac{1}{2}\mathcal{F}(u_0) \leqslant 2\pi R\sqrt{2m}, \quad \sqrt{M}r_2 \leqslant \pi R\sqrt{m},$$

which contradicts (150). Therefore the homotopy type of the curve is preserved and $\max_x |u(x,t)| \geqslant R$ for all t. Note that

$$\begin{aligned}\left|\left|u(x_1,t)\right|^2 - \left|u(x_2,t)\right|^2\right| &\leqslant \int_{x_1}^{x_2} \left|\partial_x |u|^2\right| dx \leqslant 2\int_{x_1}^{x_2} |u||\partial_x u|\, dx \\ &\leqslant 2\left(\int_{x_1}^{x_2} |\partial_x u|^2\, dx\right)^{1/2}\left(\int_{x_1}^{x_2} |u|^2\, dx\right)^{1/2} \\ &\leqslant 4\mathcal{F}\big(u(t)\big)^{1/2}\big\|u(t)\big\|_{L_2}.\end{aligned}$$

Since the supremum over x_1, x_2 of the left-hand side is greater than R we conclude that $\|u(t)\|_{L_2}$ is separated from zero. This contradicts (154), therefore for this non-linearity the weak L_{2w} attraction to the attractor takes place and the L_2-norm-induced attraction does not.

REMARK. In the one-dimensional scalar case Feireisl [148] proved that if conditions of Theorem 3.5.5 are satisfied then every solution tends to an equilibrium in the $L_2(\mathbb{R})$ norm. Example 3.5.6 shows that the dynamics in the case of two-component system is completely different since the solution we constructed does not tend to an equilibrium even though there exists a global Lyapunov function (143) and the set of equilibria is a single point.

Navier–Stokes equations in an unbounded domain. Attractors of the Navier–Stokes equations in an unbounded channel-like domain were studied by Abergel [1] and Babin [12–14], for a review see [25]. In particular, the existence of a finite-dimensional attractor of the Navier–Stokes equations in a channel with a Poiseuille flow was proven in [12–14] when the flux through the cross-section is not too large; a time-independent asymptotic expansion near infinity of time-dependent solutions on the attractor was obtained in [14]. Note that the Poiseuille flow does not vanish at infinity, creating instability in the infinite part of the channel if the flux is too large. Originally the existence of attractors of the two-dimensional Navier–Stokes system in an unbounded channel-like domain was proven in [1,12,14] in the norm-induced topology of a weighted space with $\gamma > 0$ (it is essentially a condition on the decay of the forcing term) and when $\gamma = 0$ the attraction and compactness was proven in [12,14] in the weak topology. Rosa [340] proved attraction and compactness in the norm-induced topology of H^0_0 when $\gamma = 0$; Ju [247] proved existence of the global attractor in H^1_0; for a discussion of the techniques and more examples see Moise, Rosa and Wang [318].

Note that when $\gamma > 0$ solutions on the attractor admit asymptotic expansion as $|x| \to \infty$ (see [14,25] for details). The question of the existence of such expansion in the case when $\gamma = 0$ is open.

Attractors in the norm-induced topology. In the important case $\gamma = 0$ the existence of the global attractor $\mathcal{A}$ (see Theorem 3.5.5) was proven in [57] in the weak topology. Wang [377] proved that compactness and attraction for $\mathcal{A}$ holds in the strong topology (under condition (142) which is stronger than (145), therefore it does not contradict Example 1). Prizzi [333] proved that the compactness and attraction is in H^1_0 norm, that is under

natural assumptions $\mathcal{A}$ is (H_0^0, H_0^1)-attractor. See also cited above works [340,247,318] on the Navier–Stokes equations.

Damped hyperbolic equations. The theory of attractors of damped hyperbolic equations in unbounded domains combines ideas and methods of the theory of damped hyperbolic equations in bounded domains and the theory of parabolic equations and systems in unbounded domains. For details we refer the reader to the works of Feireisl [147], Guo and Li [204], Karachalios and Stavrakakis [251–253], Belleri and Pata [63], Zelik [389,392].

3.5.3. *Finite and infinite dimension of attractors.* Global attractors of parabolic equations in unbounded domains in contrast to the case of bounded domain may have finite or infinite dimension depending on the non-linearity and the class of solutions.

We give here the result of [57] on dimension of attractors.

THEOREM 3.5.7. *Let $\gamma \geqslant 0$ and conditions of Theorem* 3.5.4 *hold. In addition we make an assumption on the order of vanishing $f'(u)$ at zero, namely we assume that*

$$\left|f'(u)\right| \leqslant |u|^{\alpha_0} C_0(u). \tag{155}$$

Then $\mathcal{A}$ has a finite Hausdorff dimension

$$\dim_{\mathrm{H}} \mathcal{A} \leqslant C\nu^{-d/2}\lambda_0^{-3-2/\alpha_0}\|g\|_{0,0}^2.$$

If instead of (155) *we assume that $-f'(u) \leqslant C|u|^{4/(d+2)}$ then*

$$\dim_{\mathrm{H}} \mathcal{A} \leqslant C\nu^{-d/2}\lambda_0^{-3}\|g\|_{0,0}^2.$$

In a few words, the above results on attractors in unbounded domains can be described as follows. When the equations are linearly stable at the spatial infinity, the only source of instability is the source term $g(x)$ which decays (on average) at infinity. When the solution are sought in a function space which consists of functions that decay at the infinity, the attractors of non-linear equations are finite-dimensional and attract solutions in the norm-induced metric, so they are in a way similar to equations in a bounded domain. When the function spaces contain functions that do not decay at infinity, the non-linearity creates a large perturbation at the infinity. Therefore, even when the linear part is stable, the non-linear terms break the linear stability at the spatial infinity. The dimension of the attractor, generally speaking, is infinite. When the equation is translation invariant, the attractor is translation invariant too.

REMARK. In the situation of Theorem 3.5.5 of the previous subsection the functional space consists of functions that decay at the infinity and the zero solution is stable. Nevertheless our conjecture is that for the non-linearity of the type given in Example 3.5.6 the dimension of the attractor is infinite, more precisely it contains equilibria with arbitrary large index of instability.

REMARK. Efendiev and Miranville [136] prove existence of finite-dimensional global attractors of reaction–diffusion equations with non-linearities that include dependence on the spatial gradient.

When the domain Ω is unbounded, one can prove existence of attractors of semigroups in the weak topology, or in the norm-induced topology of a wider space, but generally speaking the attractors can be infinite-dimensional (see the discussion below, see also [57]). The first example of an infinite-dimensional attractor of a parabolic equation was given in [57]. Since the construction is elementary we present it here.

Let

$$f_0(u) = \begin{cases} -u & \text{for } |u| \leqslant 1, \\ u-2 & \text{for } u \geqslant 1, \\ u+2 & \text{for } u \leqslant -1 \end{cases}$$

and consider the equation

$$\partial_t u = \Delta u - f_0(u).$$

This equation generates a semigroup $\{S_t\}$ in the weighted Sobolev space $H^0_\gamma(\mathbb{R}^d)$, $\gamma < -d/2$, and the semigroup has a global attractor $\mathcal{A}$ (see [57]). The dimension (fractal and Hausdorff) of the attractor is infinite. This fact follows from the following observation. Obviously, $z(x) = 0$ is an equilibrium of $\{S_t\}$. For $|u| \leqslant 1$ the dynamics of $\{S_t\}$ is generated by the linear equation

$$\partial_t u = \Delta u + u. \tag{156}$$

The unstable manifold of z includes bounded solutions $u(t)$, $t \leqslant 0$, of (156) that tend to zero as $t \to -\infty$. Such solutions are given in terms of the Fourier transform $\tilde{u}(\xi, t)$ of $u(x, t)$ by the formula

$$\tilde{u}(\xi, t) = e^{(1-|\xi|^2)t}\tilde{u}_0(\xi), \quad t \leqslant 0, \qquad u(x,t) = (2\pi)^{-d}\int e^{-ix\cdot\xi}\tilde{u}(\xi,t)\,d\xi. \tag{157}$$

Now we describe the set of $\tilde{u}_0$. Let $\widetilde{U}$ be the set of Lebesgue integrable functions $\tilde{v}(\xi)$ which satisfy the following conditions

$$\widetilde{U} = \big\{\tilde{v} \in L_2\big(\mathbb{R}^d\big)\colon\ \tilde{v}(\xi) = 0 \text{ for } |\xi| \geqslant 1/2,\ \big|\tilde{v}(\xi)\big| \leqslant 1 \text{ for } |\xi| \leqslant 1/2\big\}.$$

Note that the set $\widetilde{U}$ is infinite-dimensional, it includes the set of all functions from $L_\infty(B_{1/2})$ with the L_∞ norm less than 1, $B_{1/2} = \{\xi\colon\ |\xi| \leqslant 1/2\}$. We take in (157) $u_0 \in U$ where the set U is defined in terms of Fourier transforms of its elements

$$U = \big\{u_0\colon\ \tilde{u}_0(\xi) \in \widetilde{U}\big\}. \tag{158}$$

The functions from U are analytic, they satisfy, in particular, the inequality $|u(x)| \leqslant 1$ for all $x \in \mathbb{R}^d$. It is proven in [57] that U lies in the attractor $\mathcal{A}$. Since the set U determined by (158) is infinite-dimensional, the attractor $\mathcal{A}$ is infinite-dimensional.

Though the limit (17) is infinite and the fractal dimension is infinite, the massiveness of the attractor can be described in terms of its Kolmogorov entropy $\mathbb{H}_\varepsilon$. We can describe quantitatively the behavior of Kolmogorov entropy as $\varepsilon \to 0$. To take into account the spatial behavior of the functions $u(x)$, $u \in U$, we, following Collet and Eckmann [112], Zelik [388,389,391,394] and Efendiev and Zelik [138] consider the set U_R of restrictions $u(x)$, $|x| < R$, of functions $u \in U$ to a ball B_R of radius R in $\mathbb{R}^d$. The set $U \subset L_2(\mathbb{R}^d)$ given by (158) has the entropy $\mathbb{H}_\varepsilon(U)$ in the space $\mathcal{C}(B_R)$ that admits the estimate

$$\mathbb{H}_\varepsilon(U_R) \leqslant C\left(R + K \ln\frac{1}{\varepsilon}\right)^d \ln\frac{1}{\varepsilon},$$

$$\mathbb{H}_\varepsilon(U_R) \geqslant c_\alpha R^d \left(\ln\frac{1}{\varepsilon}\right)^{d+1-\alpha}, \quad c_\alpha > 0,\ \alpha > 0.$$

Kolmogorov ε-entropy of attractors. Estimates of Kolmogorov ε-entropy of attractors in large (with size tending to infinity) and in unbounded domains were obtained in papers of Collet and Eckmann [112–114], Zelik [388,389,391,394] and Efendiev and Zelik [138]. They consider parabolic reaction–diffusion systems and damped hyperbolic equations. As an example we give here one of results of Efendiev and Zelik [139]. Under natural conditions they proved estimates of the Kolmogorov ε-entropy of restriction of attractors $\mathcal{A}$ of a reaction–diffusion system in an unbounded domain to a ball B_R that are of the form

$$\mathbb{H}_\varepsilon(\mathcal{A}|_{B_R}) \leqslant C R^d \left(\ln\frac{1}{\varepsilon}\right)^{d+1}.$$

They obtain also a lower estimate of $\mathcal{A}$

$$\mathbb{H}_\varepsilon(\mathcal{A}|_{B_R}) \geqslant c_\alpha R^d \left(\ln\frac{1}{\varepsilon}\right)^{d+1-\alpha}, \quad c_\alpha > 0,\ \alpha > 0.$$

The proof of the lower bound is based on a general construction of large infinite-dimensional submanifold of the unstable manifold of an equilibrium which is applicable to reaction–diffusion systems in unbounded domains (see [388,389,391,138]).

4. Generalized attractors

4.1. *Multivalued semigroups and trajectory dynamics*

Multivalued semigroups. There are several situations when it is natural to consider multivalued mappings, that is the mappings S_t which map sets into sets rather than elements into elements. First example is given by equations with the non-linearity which

does not satisfy the Lipschitz condition. When the weak solutions are considered, it is rather common that even when the non-linearity is given by a very regular (analytic) expression, the corresponding operators are not Lipschitz continuous in the functional space in which the existence of solutions is proven. The second example comes from the elliptic equations when several solutions can exist with the same boundary conditions. One of ways to overcome the difficulty of non-uniqueness of weak solutions is to use theory of semigroups of multivalued operators. The theory of attractors of multivalued semigroups generated by PDE was started by Babin and Vishik in [44]. Further works in this direction are by Babin [15], Ball [59,60], Mel'nik [297,298], Valero [370]. Theory of global attractors of differential inclusions is studied by Mel'nik and Valero [299,369], Kapustyan and Mel'nik [249].

Here we give a sketch of the theory developed in [44,15] since it is more elementary.

Let E be a complete metric space. An operator semigroup $\{S_t\}$ acts on subsets of E, $S_t B \subset E$ when $B \subset E$. It is assumed that for every $t \geqslant 0$

$$S_t B = \bigcup_b \{S_t b, b \in B\}. \tag{159}$$

We also assume the semigroup property

$$S_{t+\tau} B \subset S_t S_\tau B. \tag{160}$$

A set B_0 is called absorbing if for any bounded B there exists $T \geqslant 0$ such that $S_t B \subset B_0$ for all $t \geqslant T$. We denote by $[X]_E$ the closure of the set X in the metric of E.

A generalized semigroup of multivalued operators can be constructed as follows. We denote by $\mathcal{U}$ the set of solutions $u(t)$, $0 \leqslant t < \infty$, of the equation $\partial_t u = \mathcal{F}(u(t))$ with initial data in E. The set $S_t b = \{v\colon \bigcup_u v = u(t),\ u(0) = b, u \in \mathcal{U}\}$ is the value of the generalized operator S_t at the one-point set $\{b\}$.

DEFINITION 4.1.1. We call $\mathcal{A}$ a generalized global attractor of a generalized semigroup $\{S_t\}$ if the following three conditions are fulfilled:

(i) $\mathcal{A}$ is compact;
(ii) $\mathcal{A}$ is an attracting set that is $\delta_E(S_t B, \mathcal{A}) \to 0$ as $t \to +\infty$;
(iii) $\mathcal{A}$ is negative invariant, that is $\mathcal{A} \subset S_t \mathcal{A}$ for every $t \geqslant 0$.

A set $\mathcal{A}$ which has the above properties is unique.

THEOREM 4.1.2. *Let* $\{S_t\}$ *satisfy* (159), (160) *and have a compact absorbing set* B_0. *Let for any set* $X \subset E$

$$[S_t X]_E \subset S_t [X]_E. \tag{161}$$

We also assume that for any point $y \in E$ *and* $t \geqslant 0$ *its preimage* $S_t^{-1} y$ *restricted to* B_0 *is closed*

$$S_t^{-1} y \cap B_0 = \left[S_t^{-1} y \cap B_0\right]_E.$$

Then $\{S_t\}$ has a generalized global attractor, it equals the omega-limit set of B_0, $\mathcal{A} = \omega(B_0)$.

This concept can be applied to the following situation. Assume that the equation $\partial_t u = \mathcal{F}(u(t))$ has for every $u_0 \in E$ a non-empty set of solutions $U(u_0)$. Then the action of the semigroup is defined by the formula $S_t u_0 = \{v\colon\ v = u(t), u \in U(u_0)\}$. Examples of generalized attractors for a damped hyperbolic equation when the non-linearity is supercritical or the Lipschitz condition does not hold are given by Babin and Vishik [44]. Ball [59] applied theory of multivalued semigroups to three-dimensional Navier–Stokes equation and proved that if weak solutions are continuous functions of time from $(0, \infty)$ to $L_2(\Omega)$ there is a global attractor in $L_2(\Omega)$. Ball [60] proved existence of a connected generalized attractor of a damped wave equation.

Mel'nik [297,298] and Ball [59] developed variants of the theory of global attractors of multivalued operators, see [59,60] for a detailed discussion of the theory.

Trajectory attractors. Another approach to equations without uniqueness uses the concept of a trajectory attractor. We denote by $\mathcal{U}$ the set of solutions $u(t)$, $0 \leqslant t < \infty$, of the equation $\partial_t u = \mathcal{F}(u(t))$ with initial data in E. The semigroup $\{S_t\}$ acts on a function $u(t)$ as a translation, $S_\tau : u(\cdot) \longmapsto u(\cdot + \tau)$. Under appropriate conditions S_τ has a global attractor in the corresponding metric space (see Section 4.2 for the definitions, see [98] for details of the theory of trajectory attractors). The global attractor $\mathcal{A}_\mathcal{U} \subset \mathcal{U}$ consists of functions of time which are defined for all t, $-\infty < t < \infty$. We set

$$\mathcal{A} = \bigl\{v\colon\ v = u(0), u \in \mathcal{A}_\mathcal{U}\bigr\} \tag{162}$$

and call $\mathcal{A}$ a global attractor. When the topology on the set $\mathcal{U}$ is strong enough, namely when the convergence of functions u in this topology implies the convergence in $\mathcal{C}([0, T]; E)$ for every T, the attractor defined by (162) is a global attractor in the sense of Definition 4.1.1. The trajectory approach to construct global attractors of equations without uniqueness was used by Chepyzhov and Vishik [96–98], Sell [351], Kapustyan and Mel'nik [250], Feireisl [149].

Chepyzhov and Vishik [98] proved the existence of a global attractor for the damped hyperbolic equation with arbitrary power growth of the non-linearity and without Lipschitz condition on it.

4.1.1. *Dynamical approach to elliptic equations.* We consider a cylindrical domain $\Omega = \omega \times \mathbb{R}$ where $\omega \subset \mathbb{R}^d$ is a bounded domain with a smooth boundary, we denote $\vec{x} = (x_0, x_1, \dots, x_d)$ a point in Ω, $(x_1, \dots, x_d) \in \omega$ and $-\infty < x_0 < \infty$. Let $\mathcal{U} = \{u\}$ be a given set of functions $u(\vec{x}) = u(x_0, x)$ on Ω. For a set B of functions $v(x)$, $x \in \omega$, we consider the set $B_\mathcal{U} = \{v \in B\colon\ v = u(0, x), u \in \mathcal{U}\}$ of values taken in B by functions from $\mathcal{U}$ at $x_0 = 0$ (this set may be empty). Starting from correspondence $u(0, x_1, \dots, x_d) \to u(t, x_1, \dots, x_d)$ we can define a generalized operator S_t which for every B maps $B_\mathcal{U}$ into $S_t(B_\mathcal{U})$, $S_t(B_\mathcal{U}) = \{v\colon\ v = u(t, \cdot), u(0, \cdot) \in B_\mathcal{U}\}$. If $\mathcal{U}$ is translation-invariant with respect to shifts in x_0, namely $u(x_0 + t, x) \in \mathcal{U}$ when $u(x_0, x) \in \mathcal{U}$ and $t \geqslant 0$ the operators S_t form a semigroup.

As an example of the set $\mathcal{U}$ we consider the set of solutions of a non-linear elliptic equation. We consider elliptic system

$$\partial_0^2 u + \gamma \partial_0 u + \nu \Delta u - a_0 u - f(u) = g, \tag{163}$$

$u = (u_1, \dots, u_m)$, here γ and a_0 are diagonal matrices, a_0 is positive, elements of γ may have different signs,

$$\gamma_{\min} \leqslant \gamma_i \leqslant \gamma_{\max}.$$

Coefficients $a_0, \nu > 0$. The function $g = g(x)$ does not depend on x_0. We consider solutions that satisfy Neumann boundary condition

$$\left.\frac{\partial u}{\partial l}\right|_{\partial\Omega} = 0 \tag{164}$$

(the case of Dirichlet boundary conditions is similar). We assume that the following growth conditions hold:

$$f(u)u \geqslant \mu_1 |u|^{p_1}, \quad \mu_1 > 0,\ p_1 > 0,$$

$$\bigl|f(u)\bigr| \leqslant C\bigl(1 + |u|\bigr)^{p_2}, \quad \mu_1 > 0,\ p_2 < \max(2, p_1),$$

where

$$p_2 \leqslant 1 + \frac{4}{d-2}$$

when $d > 2$.

Now we define a multivalued semigroup that corresponds to (163) and acts on functions that depend in $x \in \omega$. First we introduce the function space E. A natural choice for the space E is the Sobolev space of fractional order $H^{3/2}(\omega)$. We introduce the space $H^s(\omega)$ by the formula $H_s(\omega) = (1 - \Delta)^{-s/2} H_0$ where $H^0 = L_2(\omega)$, the Laplace operator Δ is taken with Neumann boundary conditions. Since Δ is a negative self-adjoint operator, one can define the above fractional powers using eigenfunction expansions. We take a fractional power because according to the Sobolev trace theorem, if a function $u(x_0, x)$ belongs to $H^2(\mathbb{R} \times \omega)$, its restriction $u(t, x)$ with a fixed $x_0 = t$ belongs to $H^{3/2}(\omega)$.

Let $u_0 \in H^{3/2}(\omega)$. Consider solutions $u(x_0, x)$ of (163), $x_0 \geqslant 0$ such that $u(0, x) = u_0(x)$ which satisfies the uniform boundedness condition

$$\sup_{\tau \geqslant 0} \int_\tau^{\tau+1} \bigl\|u(x_0, \cdot)\bigr\|_{H^0} \, dx_0 < \infty.$$

The set of such solutions we denote by $U_+(u_0)$. It is proven in [15] that this set is not empty for any $u_0 \in H_{3/2}(\omega)$. In the notation of the beginning of the subsection

$$B_{\mathcal{U}} = B \quad \text{for every set } B \subset H^{3/2}(\omega). \tag{165}$$

For any given $t \geqslant 0$ the set $S_t u_0$ is defined as restrictions $u(t,x)$ of all solutions $u \in U_+(u_0)$. For $B \subset E$ we put $S_t B = \bigcup_{b\in B} S_t b$.

THEOREM 4.1.3 [15]. *The operators S_t corresponding to* (163) *map bounded sets into bounded. The semigroup $\{S_t\}$ has an absorbing set B_0 that is compact in $H^{3/2}(\omega)$.*

All conditions of Theorem 4.1.2 can be verified and we obtain the following theorem from [15].

THEOREM 4.1.4. *The semigroup $\{S_t\}$ of multivalued mappings has a global attractor $\mathcal{A}$ which is defined as an omega-limit set by* (6). *The attractor is compact in $H^{3/2}(\omega)$. The attractor consists of values $u(x_0 = 0, \cdot)$ of solutions $u(x_0, x)$ of* (163) *that are defined for all x_0, $-\infty < x_0 < +\infty$, their restrictions $u(t,x)$ are bounded in $H^{3/2}(\omega)$ uniformly in t_0 and they belong to $H^2([\tau, \tau+1]\times\omega)$ for every $\tau \in \mathbb{R}$.*

REMARK. Bounded solutions of (163) of the form $V(x - ct)$ are called traveling wave solutions. Such solutions are defined for all t. Therefore they lie on the attractor $\mathcal{A}$.

REMARK. When there is a wide enough gap in the spectrum of Δ the attractor $\mathcal{A}$ lies in the finite-dimensional inertial manifold $\mathcal{M}$. Therefore it is finite-dimensional itself. Inertial manifolds for elliptic equations were constructed by different methods in [302] and [19]. The approach of [302] uses a reduction to a first-order system. Approach of [19] treats second-order equation (163) directly and gives better estimates of the dimension of manifolds.

Now we consider the case when $f(u)$ is a gradient, $f(u) = \nabla_u F(u)$ and all γ_i are nonzero and have the same sign. In this case the "dynamics" of $\{S_t\}$ on any given full bounded trajectory $u(x_0, x)$, $-\infty < x_0 < +\infty$, on the attractor possesses a Lyapunov function

$$\int \left[-\frac{1}{2}|\partial_0 u|^2 + \frac{1}{2}\nu \nabla u \cdot \nabla u + F(u) - \frac{1}{2} a_0 u \cdot u - g \cdot u \right] dx.$$

Though it is not bounded from above or below on arbitrary functions, it is still very useful. Using this function we obtain the following theorem on the structure of the attractor (see Babin [15], see also Vishik and Zelik [375]. Schulze, Vishik, Witt and Zelik [347] obtained similar results for a cylindrical domain with a piecewise smooth boundary).

THEOREM 4.1.5. *Let $f(u)$ be a gradient, and derivatives of f satisfy the condition $|\nabla f(u)| \leqslant C(1 + |u|^{p_3})$ with $p_3 \leqslant 4/(d-3)$ when $d > 3$. Let the equation*

$$\nu \Delta z - a_0 z - f(z) = g$$

for the equilibria of (163) *have a finite set of solutions $\{z_1(x), \dots, z_N(x)\}$. Then the attractor $\mathcal{A}$ consists of connecting orbits between $z_1, \dots, z_N$.*

REMARK. Fiedler, Scheel and Vishik [159] prove the existence of connecting orbits between the equilibria using Conley index theory. When $f(u)$ is not a gradient they prove existence of non-equilibrium solutions on the attractor which converge to an equilibrium when $x_0 \to +\infty$ or $x_0 \to -\infty$.

Cauchy data approach. Another approach to spatial dynamics generated by an elliptic equation in a cylindrical domain is applied by Calsina, Mora and Solà-Morales [74]. They consider Cauchy problem for (163) prescribing $(u(x,0), \partial_t u(x,0))$ and look for $(u(x,t), \partial_t u(x,t))$ for $t \geqslant 0$. The advantage is that a solution of the Cauchy problem is unique. The difficulty is that a global (and a local) solution does not exist for arbitrary initial data. They construct the semigroup in the following way. Let $\mathcal{U} = \{u\}$ be the set of solutions of (163) which are defined and bounded for all $x_0 \geqslant 0$. For the function $u \in \mathcal{U}$ we consider the mapping $\widetilde{S}_t : (u(0), \partial_0 u(0)) \to (u(t), \partial_0 u(t))$. Since the Cauchy data determine solution uniquely this mapping is one-valued. The semigroup $\widetilde{S}_t$ is defined on the set $\mathcal{U}_0 = \{(u(0, \partial_0 u(0))) \colon u \in \mathcal{U}\}$. On this set the dynamics is well-defined and the attractor exists. Unfortunately an explicit description of $\mathcal{U}_0$ on which the semigroup is defined is not available.

Trajectory approach. Vishik and Zelik [374,375], Schulze, Vishik, Witt and Zelik [347], Fiedler, Scheel and Vishik [159], Zelik [387] applied theory of trajectory attractors to elliptic equations in cylindrical domains. In [374,347] it was proven that there exists a trajectory attractor of the translation semigroup. In [375] it was proven that the attractor is regular.

In the trajectory approach the set $\mathcal{U} = \{u\}$ again is the set of solutions of (163) which are defined and bounded for all $x_0 \geqslant 0$. Now the dynamics is defined on the functions that depend both on x_0 and x by the formula

$$\check{S}_t : u(x_0, x) \to u(x_0 + t, x).$$

The set $\mathcal{U}$ is endowed with the topology of the Fréchet space $H^2_{\mathrm{loc}}(\omega \times [0,\infty])$, which corresponds to convergence (non-uniform in T) in $H^2(\omega \times [0,T])$ for every T. The semigroup $\check{S}_t$ defined on $\mathcal{U}$ has a global attractor $\mathcal{A}_{\mathcal{U}}$. The attractor $\mathcal{A}$ is given by (162). Note that since $\mathcal{C}(\omega \times [0,T]; H^{3/2}(\omega)) \subset H^2(\omega \times [0,T])$ so defined attractor coincides with the attractor of the multivalued semigroup in Theorem 4.1.4.

REMARK. Note that both the Cauchy data approach and trajectory approach allow to avoid multivalued operators. At the same time the domains in which the semigroups are defined are not explicitly given, they are defined in terms of global solutions of non-linear elliptic equations. The advantage of the multivalued approach is that the space in which the semigroup is defined is explicitly given. An important advantage of the trajectory approach is that it allows to endow $\mathcal{U}$ with weaker topologies, for example one can take spaces of functions on $(\omega \times [0,T])$ which do not include $\mathcal{C}(\omega \times [0,T]; E)$ and the restriction for fixed x_0 from such spaces may not belong to E. Therefore the trajectory approach allows to prove existence of attractors in weaker topologies in some situations when the multivalued approach does not work and can be applied to more general non-linearities. The same

observation holds for parabolic and hyperbolic problems without uniqueness, see [98]. For a recent review of the theory of trajectory attractors of elliptic equations in cylindrical domains, in particular for upper and lower bounds for Kolmogorov ε-entropy of these attractors see Mielke and Zelik [306].

Estimates of dimension and invariant manifolds. The dynamical approach to elliptic problems in cylindrical domains was initiated by Kirchgässner [254], he constructed local invariant manifolds for such problems, see [127] for a recent review of the local theory. Mielke [302] proved existence of an invariant essential manifold of the form (163) in the case $d = 1$ when the linear operator $\nu\Delta$ has a wide enough gap. The essential manifold contains all bounded solutions and consequently contains the attractor, but it lacks the exponential attraction property of inertial manifolds, though it has a property of weak normal hyperbolicity which implies proximity to the essential manifold of solutions which are bounded on long intervals. Scheel [345] studied the case $\gamma \to \infty$ and proved existence of invariant manifolds and convergence of the dynamics to the dynamics of the limit parabolic equation. Shapoval [355] proved the existence of an integral manifold for the non-autonomous case when in (163) $f(u)$ depends on x_0, $f(u) = f(u, x_0)$.

Babin [19] proved existence of finite-dimensional invariant inertial manifolds $\mathcal{M}$ corresponding to elliptic systems of the form (163) when the spectrum of $\nu\Delta$ has a wide enough gap. The gap condition is always satisfied when $d = 1$ and is satisfied in the multidimensional case for special domains, see [287]. The gap condition in [19] has the following form. There should exist two consecutive eigenvalues of $\nu\Delta$ such that

$$\lambda_{N+1} - \lambda_N \geqslant 5 \max_{|v| \leqslant R} f'(v), \quad \text{where } R = \max_{u \in \mathcal{A}} \|u\|_{\mathcal{C}(\omega \times (-\infty, \infty))}.$$

When the gap condition holds, the fractal dimension of $\mathcal{A}$ is not greater than N. The method of [19] is based on construction of an extended dynamical system $\mathcal{S}_t$ in the space $E \times E_N$, where E_N is N-dimensional space with has first N eigenfunctions of $\nu\Delta$ as the basis. The orthoprojection in $E = H_{3/2}(\omega)$ onto E_N is denoted by P_N. The extended semigroup is defined in the following way. First, the non-linearity $f'(u)$ is reduced to a globally Lipschitz by a modification for large $|u| \geqslant R$. Then the following semi-Cauchy problem for (163) is considered:

$$u|_{x_0=0} = u_0 \in E, \qquad P_N \partial_{x_0} u|_{x_0=0} = u_+ \in E_N.$$

It has a unique solution for all $(u_0, u_+) \in E \times E_+$ in the class of exponentially growing functions with an appropriate rate of growth. The semigroup $\mathcal{S}_t$ in $E \times E_N$ of one-valued, continuous operators is defined by the formula

$$\mathcal{S}_t : (u_0, u_+) \mapsto (u, P_N \partial_{x_0} u)|_{x_0=t}.$$

This semigroup $\mathcal{S}_t$ has a finite-dimensional, exponentially attracting inertial manifold $\mathcal{M}$ which enjoys the tracking property. This manifold contains trajectories of all bounded and slowly exponentially growing solutions of the extended system defined for $-\infty < x_0 < \infty$. In particular, the projection of this manifold to E contains $\mathcal{A}$. Therefore, the attractor $\mathcal{A}$ has dimension not greater than N.

4.2. *Non-autonomous equations and trajectory attractors*

For a detailed systematic treatment of the theory of attractors of non-autonomous equations and for references see a recent book of Chepyzhov and Vishik [98], here we introduce only basic concepts of the theory. When (61) includes explicit time dependence it becomes non-autonomous and operators S_t do not form a semigroup anymore. A non-autonomous equation has a form

$$\partial_t u = F(u, t). \tag{166}$$

The simplest example is given by the equations of the form

$$\partial_t u = F_0(u) + f(t), \tag{167}$$

where the forcing term f depends on t, $f = f(t)$. We now consider the situation when for a given $u_0 = u(0)$ and f the solution exists and is unique (in appropriate classes). The solution depends on u_0 and f, $u = u(f, u_0)$, where u_0 is a time-independent vector and $f = f(t)$ is time-dependent. Now the solution operators $u(0) \mapsto u(t)$ do not form a semigroup, one has to consider a semiprocess $S(s,t): u(s) \mapsto u(t)$, $t \geqslant s \geqslant 0$. Processes and semiprocesses generated by non-linear PDE were studied in many papers, see [123, 227,98,353].

The *shift operator* $\mathcal{T}_s: f(t) \mapsto f(t+s)$ plays important role in the theory of non-autonomous equations.

If $u(t)$, $t \geqslant 0$, is a solution of (167) the shifted function $(\mathcal{T}_s u)(t) = u(t+s)$, $t \geqslant 0$, is a solution of the shifted equation $\partial_t u = F_0(u) + \mathcal{T}_s f$. We define an operator

$$\check{S}_s: \big(u(0), f(t)\big) \mapsto \big(u\big(f, u(0)\big)(s), \mathcal{T}_s f\big) \tag{168}$$

the operators $\widetilde{S}_s$, $s \geqslant 0$, form a semigroup. This family of operators gives an example of *skew-product dynamics*, see Miller [307], Sell [349], Sacker and Sell [342] and for a detailed discussion Sell and You [353].

Following [98] we call the time-dependent part of Equation (166) the *symbol* σ of the equation, in (167) the symbol is $\sigma = f(t)$; symbols also may include u-dependent functions. Usually the set of symbols σ is endowed with an appropriate topology. For examples of symbols see [98,36].

Similarly to (168) we can consider the dynamics on the set of trajectories

$$\widetilde{S}_s: \big(u(t), f(t)\big) \mapsto \big(u(t+s), f(t+s)\big). \tag{169}$$

One may consider functions $(u(t), f(t))$, $t \geqslant 0$, as elements of a topological space Θ_+ of time-dependent functions. A topology in this space is defined by convergence in an appropriate norm on bounded intervals $[t_1, t_2]$, for example $(\int_{t_1}^{t_2} \|u(\tau)\|_0^2 \, d\tau)^{1/2}$ for all positive t_1, t_2. This convergence can be described similarly to (171) by a metric, therefore results of Section 1 are applicable. The closure of shifts $f(t+s)$ is called the hull of f.

Under appropriate conditions one can prove existence of the global attractor of this semigroup $\{\widetilde{S}_T\}$. The attractor is called trajectory attractor (see [96]). This attractor consists of solutions $u(t)$ defined for $-\infty < t < \infty$ and corresponding to an element $\tilde{f}(t)$ of the hull that is defined for $-\infty < t < \infty$. The construction is directly applicable to 2D NS equations with time-dependent forcing term $f(t)$ and can be easily generalized to more general non-autonomous equations.

To treat the shift operator $\mathcal{T}_s : f(t) \mapsto f(t+s)$ it is convenient to introduce local Banach spaces $\mathcal{B}_{\rm loc}([0,\infty], Y)$ of Y-valued functions.

If Y is a Banach space then $L_{p,\rm loc}([0,\infty], Y)$, $1 \leqslant p \leqslant \infty$, is a space of Y valued functions $u(t)$, $0 \leqslant t < \infty$, with $L_p([0,T], Y)$-convergence on finite intervals:

$$f \to f_0 \quad \text{in } L_{p,\rm loc}\big([0,\infty], Y\big) \text{ when } \|f - f_0\|_{L_p([0,T],Y)} \to 0 \text{ for every } T > 0. \tag{170}$$

One can introduce the metric in $L_{p,\rm loc}([0,\infty], Y)$

$$\varrho_{L_{p,\rm loc}([0,\infty],Y)}(u,v) = \sum_{n=0}^{\infty} 2^{-n} \frac{\|u-v\|_{L_p([n,n+1],Y)}}{1 + \|u-v\|_{L_p([n,n+1],Y)}}. \tag{171}$$

Convergence (170) is equivalent to the convergence generated by (171). The metric space $L_{p,\rm loc}([0,\infty], Y)$ is complete.

Similarly, one introduces $\mathcal{C}^k_{\rm loc}([0,\infty], Y)$ as a space of k times continuously differentiable, Y valued functions $u(t)$ with the convergence

$$u \to u_0 \quad \text{in } \mathcal{C}^k_{\rm loc}\big([0,\infty], Y\big) \text{ when } \|f - f_0\|_{\mathcal{C}^k([0,T],Y)} \to 0 \text{ for every } T > 0, \tag{172}$$

the metric $\varrho_{\mathcal{C}^k_{\rm loc}([0,\infty],Y)}(u,v)$ is introduced similarly to (171). Other spaces $\mathcal{B}_{\rm loc}([0,\infty], Y)$, can be similarly introduced, for example Sobolev spaces $W^1_{p,\rm loc}([0,\infty], Y)$ of Y valued functions etc.

Miller [307] and Sell [348] introduced compactification of time dependence in non-autonomous equations with general time dependences to study their dynamics, an important concept used for this purpose is the hull.

DEFINITION 4.2.1. For an element $f(\tau)$ of a topological functional space $\mathcal{B}_{\rm loc}([0,\infty], Y)$ we take all shifts $f(T+\tau)$, $T \geqslant 0$, and define the *hull* of f

$$H(f) = \text{closure}_{\mathcal{B}_{\rm loc}([0,\infty],Y)} \bigcup_{T \geqslant 0} \big(f(T+\tau)\big).$$

A usual condition for the existence of a global attractor of $\widetilde{S}_t$ defined by (169) is the compactness of the hull $H(f)$ in $\mathcal{B}_{\rm loc}([0,\infty], Y)$ (see [98] for examples and details). Conditions of theorems on attractors of non-autonomous equations include also conditions that

are similar to conditions for the existence of a global attractor of an autonomous equation, for example the existence of a bounded absorbing ball.

Note that the action of $\widetilde{S}_t$ on symbols is explicit and independent on the dynamics of the equation, therefore it is possible to describe separately the u-component of the attractor, see [98] for details.

As a typical example we consider a reaction–diffusion system of the form (72) with $g = g(x,t)$. The non-linearity satisfies conditions of the type of (74) (see Section 6.2 of [98] for optimal conditions and details).

THEOREM 4.2.2. *Let* (74), (78) *hold and the set* $\{\mathcal{T}_s g, s \geqslant 0\}$ *be precompact in* $L_{2,\mathrm{loc}}([0,\infty],(L_2(\Omega))^m)$. *Then* $\widetilde{S}_t$ *possesses a global attractor in* $(L_2(\Omega))^m \times H(f)$.

Further examples include non-autonomous parabolic and damped hyperbolic equations and systems, Navier–Stokes system, see [98].

Note that the global attractor of a non-autonomous equation is invariant with respect to time translations, whereas the non-autonomous equation itself is not. That leads to new interesting phenomena. One such phenomenon concerns the dimension of a global attractor. First we consider an elementary example.

EXAMPLE 4.2.3. Let us consider a linear system of ODE in $\mathbb{C}^N$

$$\partial_t y_j = -y_j + b_j e^{i\omega_j t}, \quad j = 1,\dots,N.$$

The solution of this system has the form

$$y_j = C_j e^{-t} + g_j e^{i\omega_j t}, \quad g_j = \frac{b_j}{1+i\omega_j}.$$

Obviously, every solution tends as $t \to \infty$ to the uniquely defined solution

$$y_j = g_j e^{i\omega_j t}. \tag{173}$$

When the frequencies ω_j are non-commesurable, the closure of the set given by (173) with $t \in \mathbb{R}$ coincides with the N-dimensional torus $\{y\colon |y_j| = |g_j|, j = 1,\dots,N\}$.

The above example shows that the attractor may have a high dimension even when the dynamics is stable. It was shown by Chepyzhov and Vishik that similar effects exist in much more complex situations and significantly change the dimension of attractors. For example, an estimate of dimension of the global attractor of 2D Navier–Stokes system obtained by Chepyzhov and Vishik when the forcing term includes N incommesurable frequencies has the form $\dim_{\mathrm{F}} \mathcal{A} \leqslant d_0(G) + N$ where G is the Grasshof number and $d_0(G)$ is a usual estimate of the dimension for the case when the forcing term is time-independent, see [98] for details.

For a detailed treatment of the theory of trajectory attractors and non-autonomous dynamics and for the bibliography see Chepyzhov and Vishik [96–98], Sell [351], Sell and You [353].

REMARK. When we have an equation with time-periodic coefficients with a given period T_0, a discrete time semigroup $\{S_t\} = \{S_t, t = nT_0, n = 0, 1, \ldots\}$ naturally arises. Many aspects of the theory of global attractors of PDE with periodic coefficients are similar to the theory of autonomous equations. In the scalar case the main differences arise from the absence of a global Lyapunov function; the dynamics is essentially different, see Babin and Chow [28] where the tracking property from Subsection 2.3.3 is extended to the case of non-autonomous periodic and non-periodic slow time dependence.

Acknowledgement

The author's work was supported by AFOSR grant FA9550-04-1-0359. The author expresses his gratitude to Professor E. Titi for useful discussions; he is especially grateful to Professor M.I. Vishik for valuable comments and suggestions.

References

[1] F. Abergel, *Attractor for a Navier–Stokes flow in an unbounded domain*, Attractors, Inertial Manifolds and their Approximation (Marseille-Luminy, 1987), RAIRO Modél. Math. Anal. Numér. **23** (3) (1989), 359–370.

[2] V. Afraimovich, A. Babin and S.-N. Chow, *Spatial chaotic structure of attractors of reaction–diffusion systems*, Trans. Amer. Math. Soc. **348** (12) (1996), 5031–5063.

[3] V. Afraimovich, A. Babin and S.-N. Chow, *Infinitely spatially complex solutions of PDE and their homotopy complexity*, Comm. Anal. Geom. **9** (2) (2001), 281–339.

[4] V.S. Afraimovich and L.A. Bunimovich, *Density of defects and spatial entropy in extended domains*, Physica D **80** (1995), 277–288.

[5] S.B. Angenent, *The Morse–Smale property for a semilinear parabolic equation*, J. Differential Equations **62** (3) (1986), 427–442.

[6] S.B. Angenent, *The zero set of a solution of a parabolic equation*, J. Reine Angew. Math. **390** (1988), 79–96.

[7] S.B. Angenent, J. Mallet-Paret and L.A. Peletier, *Stable transition layers in a semilinear boundary value problem*, J. Differential Equations **67** (2) (1987), 212–242.

[8] V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer (1978).

[9] J. Arrieta, A.N. Carvalho and J.K. Hale, *A damped hyperbolic equation with critical exponent*, Comm. Partial Differential Equations **17** (5–6) (1992), 841–866.

[10] J.M. Arrieta, A.N. Carvalho and A. Rodríguez-Bernal, *Perturbation of the diffusion and upper semicontinuity of attractors*, Appl. Math. Lett. **12** (1999).

[11] J.D. Avrin, *A one-point attractor theory for the Navier–Stokes equation on thin domains with no-slip boundary conditions*, Proc. Amer. Math. Soc. **127** (3) (1999), 725–735.

[12] A.V. Babin, *The asymptotic behavior as $x \to \infty$ of functions lying on the attractor of the two-dimensional Navier–Stokes system in an unbounded plane domain*, Mat. Sbornik **182** (12) (1991), 1683–1709 (Russian); English transl. in Math. USSR-Sb. **74** (2) (1993), 427–453.

[13] A.V. Babin, *The attractor of a Navier–Stokes system in an unbounded channel-like domain*, J. Dynamics Differential Equations **4** (4) (1992), 555–584.

[14] A.V. Babin, *Asymptotic expansion at infinity of a strongly perturbed Poiseuille flow*, Properties of Global Attractors of Partial Differential Equations, Advances in Soviet Mathematics, Vol. 10, Amer. Math. Soc., Providence, RI (1992).

[15] A.V. Babin, *Attractor of the generalized semigroup generated by an elliptic equation in a cylindrical domain*, Izv. Ross. Akad. Nauk Ser. Mat. **58** (2) (1994), 3–18; English transl. in Russian Acad. Sci. Izv. Math. **44** (2) (1995), 207–223.

[16] A.V. Babin, *Symmetrization properties of parabolic equations in symmetric domains*, J. Dynamics Differential Equations **6** (4) (1994), 639–659.
[17] A.V. Babin, *On space-chaotic solutions of scalar parabolic equations with modulated non-linearities*, Russian J. Math. Phys. **3** (3) (1995), 389–392.
[18] A.V. Babin, *Dynamics of spatially chaotic solutions of parabolic equations*, Mat. Sbornik **186** (10) (1995), 1389–1415.
[19] A.V. Babin, *Inertial manifolds for travelling-wave solutions of reaction–diffusion systems*, Comm. Pure Appl. Math. **48** (1995), 167–198.
[20] A.V. Babin, *Symmetry of instabilities for scalar equations in symmetric domains*, J. Differential Equations **123** (1) (1995), 122–152.
[21] A.V. Babin, *Homotopy stable spatially chaotic waves on an infinite interval*, Structure and Dynamics of Nonlinear Waves in Fluids, K. Kirchgässner and A. Mielke, eds, World Scientific (1995), 141–145.
[22] A.V. Babin, *Spatially chaotic solutions of parabolic equations, and the preservation of homotopies*, Dokl. Akad. Nauk **350** (4) (1996), 439–442 (Russian).
[23] A.V. Babin, *Preservation of homotopies, and spatially complex solutions of parabolic equations with several variables*, Funktsional. Anal. i Prilozhen. **30** (3) (1996), 73–76 (Russian); English transl. in Funct. Anal. Appl. **30** (3) (1996), 204–206.
[24] A.V. Babin, *Topological invariants and solutions with high complexity for scalar semilinear PDE*, J. Dynamics Differential Equations **12** (3) (2000), 599–646.
[25] A.V. Babin, *Attractors of Navier–Stokes equations*, Handbook of Mathematical Fluid Dynamics, Vol. II, North-Holland, Amsterdam (2003), 169–222.
[26] A.V. Babin, *Preservation of spatial patterns by a hyperbolic equation*, Discrete Continuous Dynamical Systems **10** (1–2) (2004), 1–19.
[27] A.V. Babin and L. Bunimovich, *Dynamics of stable chaotic waves generated by hyperbolic PDE*, Nonlinearity **9** (1996), 853–875.
[28] A.V. Babin and S.-N. Chow, *Uniform long-time behavior of solutions of parabolic equations depending on slow time*, J. Differential Equations **150** (1998), 264–316.
[29] A. Babin, A. Mahalov and B. Nicolaenko, *Global splitting, integrability and regularity of 3D Euler and Navier–Stokes equations for uniformly rotating fluids*, Eur. J. Mech. B/Fluids **15** (1996), 291–300.
[30] A. Babin, A. Mahalov and B. Nicolaenko, *Global regularity and integrability of 3D Euler and Navier–Stokes equations for uniformly rotating fluids*, Asymptotic Anal. **15** (1997), 103–150.
[31] A. Babin, A. Mahalov and B. Nicolaenko, *Global splitting and regularity of rotating shallow-water equations*, Eur. J. Mech. B/Fluids **16** (1) (1997).
[32] A. Babin, A. Mahalov and B. Nicolaenko, *On the regularity of three-dimensional rotating Euler–Boussinesq equations*, Math. Models Methods Appl. Sci. **9** (7) (1999), 1089–1121.
[33] A. Babin, A. Mahalov and B. Nicolaenko, *Global regularity of 3D rotating Navier–Stokes equations for resonant domains*, Indiana Univ. Math. J. **48** (3) (1999), 1133–1176.
[34] A. Babin and B. Nicolaenko, *Exponential attractors and inertially stable algorithms for Navier–Stokes equations*, Progress in Partial Differential Equations: The Metz Surveys, 3, Pitman Res. Notes Math. Ser., Vol. 314, Longman Sci. Tech., Harlow (1994), 185–198.
[35] A.V. Babin and S.Yu. Pilyugin, *Continuous dependence of attractors on the shape of domain*, Kraev. Zadachi Mat. Fiz. i Smezh. Voprosy Teor. Funktsii, Vol. 26, Zap. Nauchn. Sem. St.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **221** (1995), 58–66, 254 (Russian); English transl. in J. Math. Sci. **87** (2) (1997), 3304–3310.
[36] A.V. Babin and G.R. Sell, *Attractors of non-autonomous parabolic equations and their symmetry properties*, J. Differential Equations **160** (2000), 1–50.
[37] A.V. Babin and M.I. Vishik, *Attractors of quasilinear parabolic equations*, Dokl. Akad. Nauk SSSR **264** (4) (1982), 780–784; English transl. in Soviet Math. Dokl. **25** (3), 703–706.
[38] A.V. Babin and M.I. Vishik, *A regular attractor of a hyperbolic equation*, Uspekhi Mat. Nauk **37** (4) (1982), 89–90.
[39] A.V. Babin and M.I. Vishik, *Attractors of partial differential equations and estimates of their dimension*, Uspekhi Mat. Nauk **38** (4) (1982), 133–187; English transl. in Russian Math. Surveys **38** (4) (1983), 151–213.

[40] A.V. Babin and M.I. Vishik, *Attractors of the Navier–Stokes system and parabolic equations and estimates of their dimension*, Boundary Value Problems of Mathematical Physics and Related Questions in the Theory of Functions, Vol. 14, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **115** (1982), 3–15, 305.

[41] A.V. Babin and M.I. Vishik, *The dimension of attractors of the Navier–Stokes system and other evolution equations*, Dokl. Akad. Nauk SSSR **271** (6) (1983), 1289–1293.

[42] A.V. Babin and M.I. Vishik, *Regular attractors of semigroups and evolution equations*, J. Math. Pures Appl. **62** (4) (1983), 441–491.

[43] A.V. Babin and M.I. Vishik, *Attracteurs maximaux dans les equations aux derivees partielles*, Seminaire Brezis-Lions, 1984, Pitman (1985).

[44] A.V. Babin and M.I. Vishik, *Maximal attractors of semigroups corresponding to evolution differential equations*, Mat. Sbornik **126** (3) (1985), 397–419 (Russian); English transl. in Math. USSR-Sb. **54** (2) (1986), 387–408.

[45] A.V. Babin and M.I. Vishik, *Maximal attractors of semigroups having Lyapunov function*, Differential Equations with Partial Derivatives, Nauka, Novosibirsk (1986), 39–46.

[46] A.V. Babin and M.I. Vishik, *Unstable invariant sets of semigroups of nonlinear operators and their perturbations*, Uspekhi Mat. Nauk **41** (4) (1986), 3–34; English transl. in Russian Math. Surveys **41** (4), 1–41.

[47] A.V. Babin and M.I. Vishik, *Lyapunov function of a perturbed evolution equation*, Uspekhi Mat. Nauk **41** (5) (1986), 210–211.

[48] A.V. Babin and M.I. Vishik, *On unstable sets of evolution equations in the neighbourhood of critical points of a stationary curve*, Izv. Akad. Nauk SSSR, Ser. Mat. **51** (1) (1987), 44–78; English transl. in Math. USSR-Izv. **30**.

[49] A.V. Babin and M.I. Vishik, *Lyapunov stability modulo attractor*, Uspekhi Mat. Nauk **42** (3) (1987), 222–223 (Russian).

[50] A.V. Babin and M.I. Vishik, *Uniform asymptotics of singularly perturbed evolution equations*, Uspekhi Mat. Nauk **42** (5) (1987), 231–232 (Russian).

[51] A.V. Babin and M.I. Vishik, *On the behavior as $t \to \infty$ of solutions of nonlinear evolution equations depending on a parameter*, Dokl. Akad. Nauk SSSR **295** (4) (1987), 786–790; English transl. in Soviet Math. Dokl. **36** (1) (1988), 113–117.

[52] A.V. Babin and M.I. Vishik, *Attractors of parabolic and hyperbolic equations, the character of their compactness and attraction*, Vestnik Moskov. Univ. Ser. I, Mat. Mekh. (3) (1988), 71–73; English transl. in Moscow Univ. Math. Bull. **43** (3) (1988), 70–72.

[53] A.V. Babin and M.I. Vishik, *Spectral and stabilized asymptotical behaviour of solutions of nonlinear evolution equations*, Uspekhi Mat. Nauk **43** (5) (1988), 99–132, 239; English transl. in Russian Math. Surveys **43** (5), 121–164.

[54] A.V. Babin and M.I. Vishik, *Uniform finite-parameter asymptotics of solutions of nonlinear evolutionary equations*, J. Math. Pures Appl. **68** (1989), 399–455.

[55] A.V. Babin and M.I. Vishik, *Attractors of evolution equations*, Nauka, Moscow (1989); English transl. *Attractors of Evolution Equations*, North-Holland, Amsterdam (1992).

[56] A.V. Babin and M.I. Vishik, *Semigroups dependent on a parameter, their attractors and asymptotic behaviour*, Global Analysis and Nonlinear Equations, Voronezh. Gos. Univ., Voronezh (1988), 3–21 (Russian); English transl. in Global Analysis—Studies and Applications, Vol. IV, Lecture Notes in Math., Vol. 1453, Springer, Berlin (1990), 1–19.

[57] A.V. Babin and M.I. Vishik, *Attractors of partial differential evolution equations in an unbounded domain*, Proc. Roy. Soc. Edinburgh A **116** (1990), 221–243.

[58] J.M. Ball, *On the asymptotical behaviour of generalized processes with applications to nonlinear evolutionary equations*, J. Differential Equations **27**, 224–265.

[59] J.M. Ball, *Continuity properties and global attractors of generalized semiflows and the Navier–Stokes equations*, J. Nonlinear Sci. **7** (5) (1997), 475–502.

[60] J.M. Ball, *Global attractors for damped semilinear wave equations*, Discrete Continuous Dynamical Systems **10** (1–2) (2004), 31–52.

[61] P.W. Bates, K. Lu and C. Zeng, *Existence and persistence of invariant manifolds for semiflows in Banach space*, Mem. Amer. Math. Soc. **135** (645) (1998).

[62] P.W. Bates, K. Lu and C. Zeng, *Invariant foliations near normally hyperbolic invariant manifolds for semiflows*, Trans. Amer. Math. Soc. **352** (10) (2000), 4641–4676.
[63] V. Belleri and V. Pata, *Attractors for semilinear strongly damped wave equations on* $\mathbb{R}^3$, Discrete Continuous Dynamical Systems **7** (4) (2001), 719–735.
[64] H. Bellout, F. Bloom and J. Nečas, *Bounds for the dimensions of the attractors of non-linear bipolar viscous fluids*, Asymptotic Anal. **11** (2) (1995), 131–167.
[65] N.P. Bhatia and O. Hajek, *Local Semi-dynamical Systems*, Lecture Notes in Math., Vol. 90, Springer, Berlin (1969).
[66] J.E. Billotti and J.P. La Salle, *Dissipative periodic processes*, Bull. Amer. Math. Soc. **77** (1971), 1082–1088.
[67] F. Bloom and W.G. Hao, *Inertial manifolds of incompressible, nonlinear bipolar viscous fluids*, Quart. Appl. Math. **54** (3) (1996), 501–539.
[68] V.G. Bondarevsky, *Energetic systems and global attractors for the 3D Navier–Stokes equations*, Proceedings of the Second World Congress of Nonlinear Analysts, Part 2 (Athens, 1996), Nonlinear Anal. **30** (2) (1997), 799–810.
[69] B. Brefort, J.-M. Ghidaglia and R. Temam, *Attractors for the penalized Navier–Stokes equations*, SIAM J. Math. Anal. **19** (1) (1988), 1–21.
[70] K.J. Brown and A. Hess, *Stability and uniqueness of positive solutions for a semilinear elliptic boundary value problem*, Differential Integral Equations **3** (1990), 201–207.
[71] R.M. Brown, P.A. Perry and Zh. Shen, *On the dimension of the attractor for the non-homogeneous Navier–Stokes equations in non-smooth domains*, Indiana Univ. Math. J. **49** (1) (2000), 81–112.
[72] P. Brunovský and B. Fiedler, *Connecting orbits in scalar reaction diffusion equations. II. The complete solution*, J. Differential Equations **81** (1) (1989), 106–135.
[73] P. Brunovský and P. Poláčik, *The Morse–Smale structure of a generic reaction–diffusion equation in higher space dimension*, J. Differential Equations **135** (1) (1997), 129–181.
[74] Á. Calsina, X. Mora and J. Solà-Morales, *The dynamical approach to elliptic problems in cylindrical domains, and a study of their parabolic singular limit*, J. Differential Equations **102** (2) (1993), 244–304.
[75] Á. Calsina, J. Solà-Morales and M. València, *Bounded solutions of some nonlinear elliptic equations in cylindrical domains*, J. Dynamics Differential Equations **9** (3) (1997), 343–372.
[76] C. Cao, M.A. Rammaha and E.S. Titi, *The Navier–Stokes equations on the rotating 2-D sphere: Gevrey regularity and asymptotic degrees of freedom*, Z. Angew. Math. Phys. **50** (3) (1999), 341–360.
[77] M. Carrive, A. Miranville, A. Piétrus and J.M. Rakotoson, *The Cahn–Hilliard equation for an isotropic deformable continuum*, Appl. Math. Lett. **12** (2) (1999), 23–28.
[78] A.N. Carvalho and J.W. Cholewa, *Attractors for strongly damped wave equations with critical nonlinearities*, Pacific J. Math. **207** (2) (2002), 287–310.
[79] A.N. Carvalho and J.A. Cuminato, *Reaction–diffusion problems in cell tissues*, ICNMSC-USP, Série Mat. **19** (1994).
[80] A.N. Carvalho and A.L. Perreira, *A scalar parabolic equation whose asymptotical behavior is dictated by a system of ordinary differential equations*, J. Differential Equations **112** (1994), 81–130.
[81] R.C. Casten and C.J. Holland, *Instability results for reaction–diffusion equations with Neumann boundary conditions*, J. Differential Equations **27** (1978), 266–273.
[82] A.O. Celebi, V.K. Kalantarov and M. Polat, *Attractors for the generalized Benjamin–Bona–Mahony equation*, J. Differential Equations **157** (2) (1999), 439–451.
[83] S. Ceron and O. Lopes, *α-contractions and attractors for dissipative semilinear hyperbolic equations and systems*, Ann. Mat. Pura Appl. (4) **160** (1991), 193–206.
[84] L. Cherfils and A. Miranville, *Finite-dimensional attractors for a model of Allen–Cahn equation based on a microforce balance*, C. R. Acad. Sci. Paris Sér. I Math. **329** (12) (1999), 1109–1114.
[85] M. Chipot and J.K. Hale, *Stable equilibria with variable diffusion*, Contemp. Math., Vol. 17 (J.A. Smoller, ed.), Amer. Math. Soc., Providence, RI (1983), 209–213.
[86] M. Capinski and N.J. Cutland, *Measure attractors for stochastic Navier–Stokes equations*, Electron. J. Probab. **3** (8) (1998), 15 (electronic).
[87] M. Capinski and N.J. Cutland, *Attractors for three-dimensional Navier–Stokes equations*, Proc. Roy. Soc. London Ser. A **453** (1966) (1997), 2413–2426.

[88] N. Chafee and E.F. Infante, *Bifurcation and stability for a nonlinear parabolic partial differential equation*, Bull. Amer. Math. Soc. **80** (1974), 49–52.
[89] X.-Y. Chen, J.K. Hale and B. Tan, *Invariant foliations for C^1 semigroups in Banach spaces*, J. Differential Equations **139** (2) (1997), 283–318.
[90] X.-Y. Chen and H. Matano, *Convergence, asymptotic periodicity, and finite-point blow-up in one-dimensional semilinear heat equations*, J. Differential Equations **78** (1) (1989), 160–190.
[91] X.Y. Chen and P. Polacik, *Asymptotic periodicity of positive solutions of reaction diffusion equations on a ball*, J. Reine Angew. Math. **472** (1996), 17–51.
[92] V.V. Chepyzhov and A.Yu. Goritskii, *Unbounded attractors of evolution equations*, Properties of Global Attractors of Partial Differential Equations, Adv. Soviet Math., Vol. 10, Amer. Math. Soc., Providence, RI (1992), 85–128.
[93] V.V. Chepyzhov and M.A. Efendiev, *Hausdorff dimension estimation for attractors of nonautonomous dynamical systems in unbounded domains: An example*, Comm. Pure Appl. Math. **53** (5) (2000), 647–665.
[94] V.V. Chepyzhov and A.A. Ilyin, *A note on the fractal dimension of attractors of dissipative dynamical systems*, Nonlinear Anal. **44** (2001), 811–819.
[95] V.V. Chepyzhov and A.A. Ilyin, *On the fractal dimension of invariant sets; applications to the Navier–Stokes equations*, Discrete Continuous Dynamical Systems **10** (1–2) (2004), 117–136.
[96] V.V. Chepyzhov and M.I. Vishik, *Trajectory attractors for the 2D Navier–Stokes system and some generalizations*, Topol. Methods Nonlinear Anal. **8** (2) (1996), 217–243.
[97] V.V. Chepyzhov and M.I. Vishik, *Evolution equations and their trajectory attractors*, J. Math. Pures Appl. (9) **76** (10) (1997), 913–964.
[98] V.V. Chepyzhov and M.I. Vishik, *Attractors of Equations of Mathematical Physics*, Amer. Math. Soc. (2002).
[99] P. Chossat and G. Iooss, *The Couette–Taylor Problem*, Appl. Math. Sci., Vol. 102, Springer, New York (1994).
[100] S.-N. Chow, K. Lu and G.R. Sell, *Smoothness of inertial manifolds*, J. Math. Anal. Appl. **169** (1) (1992), 283–312.
[101] J.W. Cholewa and T. Dlotko, *Global Attractors in Abstract Parabolic Problems*, London Math. Soc. Lecture Notes Series, Vol. 278, Cambridge University Press (2000).
[102] A. Chorin, J. Marsden and S. Smale, eds, *Turbulence Seminar*, Lecture Notes in Math., Vol. 615, Springer, New York (1977).
[103] I.D. Chueshov, *Global attractors in nonlinear problems of mathematical physics*, Uspekhi Mat. Nauk **48** (3) (291) (1993), 135–162 (Russian); English transl. in Russian Math. Surveys **48** (3) (1993), 133–161.
[104] I.D. Chueshov, *On approximate inertial manifolds for stochastic Navier–Stokes equations*, J. Math. Anal. Appl. **196** (1) (1995), 221–236.
[105] I.D. Chueshov, *Theory of functionals that uniquely determine the asymptotic dynamics of infinite-dimensional dissipative systems*, Russian Math. Surveys **53** (4) (1998), 731–776.
[106] I.D. Chueshov, *Vvedenie v teoriyu beskonechnomernykh dissipativnykh sistem* (Russian) [*Introduction to the Theory of Infinite-Dimensional Dissipative Systems*], Universitetskie Lektsii po Sovremennoi Matematike [University Lectures in Contemporary Mathematics], AKTA, Kharkiv (1999), 436 p.
[107] I.D. Chueshov, *Analyticity of global attractors, and determining nodes for a class of nonlinear wave equations with damping*, Mat. Sbornik **191** (10) (2000), 119–136 (Russian); English translation in Sbornik Math. **191** (9–10) (2000), 1541–1559.
[108] I. Chueshov, M. Eller and I. Lasiecka, *On the attractor for a semilinear wave equation with critical exponent and nonlinear boundary dissipation*, Comm. Partial Differential Equations **27** (9–10) (2002), 1901–1951.
[109] I.D. Chueshov and V.K. Kalantarov, *Determining functionals for nonlinear damped wave equations*, Mat. Fiz. Anal. Geom. **8** (2) (2001), 215–227.
[110] I. Chueshov and I. Lasiecka, *Inertial manifolds for von Kármán plate equations*, Special issue dedicated to the memory of Jacques-Louis Lions, Appl. Math. Optim. **46** (2–3) (2002), 179–206.
[111] B. Cockburn, D.A. Jones and E.S. Titi, *Estimating the number of asymptotic degrees of freedom for nonlinear dissipative systems*, Math. Comp. **66** (219) (1997), 1073–1087.
[112] P. Collet and J.-P. Eckmann, *Extensive properties of the complex Ginzburg–Landau equation*, Comm. Math. Phys. **200** (3) (1999), 699–722.

[113] P. Collet and J.-P. Eckmann, *The definition and measurement of the topological entropy per unit volume in parabolic PDEs*, Nonlinearity **12** (3) (1999), 451–473; Erratum: Nonlinearity **14** (4) (2001), 907.
[114] Collet P. and J.-P. Eckmann, *Topological entropy and ε-entropy for damped hyperbolic equations*, Ann. H. Poincaré **1** (4) (2000), 715–752.
[115] P. Collet and J. Xin, *Global existence and large time asymptotic bounds of L_∞ solutions of thermal diffusive combustion systems on $\mathbb{R}^n$*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **23** (4) (1996), 625–642.
[116] P. Constantin and C. Foias, *Navier–Stokes Equations*, The University of Chicago Press (1988).
[117] P. Constantin and C. Foias, *Global Lyapunov exponents, Kaplan–Yorke formulas and the dimension of the attractors for 2D Navier–Stokes equations*, Comm. Pure Appl. Math. **38** (1) (1985), 1–27.
[118] P. Constantin, C. Foias, O.P. Manley and R. Temam, *Determining modes and fractal dimension of turbulent flows*, J. Fluid Mechanics **150** (1985), 427–440.
[119] P. Constantin, C. Foias, B. Nicolaenko and R. Temam, *Spectral barriers and inertial manifolds for dissipative partial differential equations*, J. Dynam. Differential Equations **1** (1) (1989), 45–73.
[120] P. Constantin, C. Foias and R. Temam, *Attractors representing turbulent flows*, Mem. Amer. Math. Soc. **53** (314) (1985).
[121] P. Constantin, C. Foias and R. Temam, *On the dimension of the attractors in two-dimensional turbulence*, Physica D **30** (3) (1988), 284–296.
[122] V. Coti Zelati and P.H. Rabinowitz, *Homoclinic type solutions for a semilinear elliptic PDE on R^n*, Comm. Pure Appl. Math. **45** (10) (1992).
[123] C. Dafermos, *An invariance principle for compact processes*, J. Differential Equations **9** (1971), 239–252.
[124] C.M. Dafermos, *Asymptotic behavior of solutions of evolution equations*, Nonlinear Evolution Equations, Proc. Sympos., Univ. Wisconsin, Madison, Wis., 1977, Publ. Math. Res. Center Univ. Wisconsin, Vol. 40, Academic Press, New York (1978), 103–123.
[125] Z.D. Dai and D.C. Ma, *Exponential attractors of the nonlinear wave equations*, Chinese Sci. Bull. **43** (16) (1998), 1331–1335.
[126] E.N. Dancer and P. Poláčik, *Realization of vector fields and dynamics of spatially homogeneous parabolic equations*, Mem. Amer. Math. Soc. **140** (668) (1999).
[127] F. Dias and G. Iooss, *Water-waves as a spatial dynamical system*, Handbook of Mathematical Fluid Dynamics, Vol. II, North-Holland, Amsterdam (2003), 443–499.
[128] C.R. Doering and J.D. Gibbon, *Note on the Constantin–Foias–Temam attractor dimension estimate for two-dimensional turbulence*, Physica D **48** (2–3) (1991), 471–480.
[129] C.R. Doering and X. Wang, *Attractor dimension estimates for two-dimensional shear flows*, Nonlinear Waves and Solitons in Physical Systems (Los Alamos, NM, 1997), Physica D **123** (1–4) (1998), 206–222.
[130] A. Douady and J. Oesterlé, *Dimension de Hausdorff des attracteurs*, C. R. Acad. Sci. Paris Sér. A-B **290** (24) (1980), A1135–A1138 (French).
[131] L. Dung and B. Nicolaenko, *Exponential attractors in Banach spaces*, J. Dynamics Differential Equations **13** (4) (2001), 791–806.
[132] A. Eden, C. Foias and B. Nicolaenko, *Exponential attractors of optimal Lyapunov dimension for Navier–Stokes equations*, C. R. Acad. Sci. Paris Sér. I Math. **316** (11) (1993), 1211–1215.
[133] A. Eden, C. Foias and B. Nicolaenko, *Exponential attractors of optimal Lyapunov dimension for Navier–Stokes equations*, J. Dynamics Differential Equations **6** (2) (1994), 301–323.
[134] A. Eden, C. Foias, B. Nicolaenko and R. Temam, *Exponential Attractors for Dissipative Evolution Equations*, RAM: Research in Applied Mathematics, Vol. 37, Masson, Paris; Wiley, Chichester (1994).
[135] A. Eden, V. Kalantarov and A. Miranville, *Finite-dimensional attractors for a general class of nonautonomous wave equations*, Appl. Math. Lett. **13** (5) (2000), 17–22.
[136] M. Efendiev and A. Miranville, *Finite-dimensional attractors for reaction–diffusion equations in $\mathbf{R}^n$ with a strong nonlinearity*, Discrete Continuous Dynamical Systems **5** (2) (1999), 399–424.
[137] M. Efendiev, A. Miranville and S. Zelik, *Exponential attractors for a nonlinear reaction–diffusion system in $\mathbb{R}^3$*, C. R. Acad. Sci. Paris Sér. I Math. **330** (8) (2000), 713–718.
[138] M.A. Efendiev and S.V. Zelik, *The attractor for a nonlinear reaction–diffusion system in an unbounded domain*, Comm. Pure Appl. Math. **54** (6) (2001), 625–688.
[139] M.A. Efendiev and S.V. Zelik, *Upper and lower bounds for the Kolmogorov entropy of the attractor for the RDE in an unbounded domain*, J. Dynamics Differential Equations **14** (2) (2002), 369–403.
[140] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam (1976).

[141] P.F. Embid and A.J. Majda, *Averaging over fast gravity waves for geophysical flows with arbitrary potential vorticity*, Comm. Partial Differential Equations **21** (1996), 619–658.
[142] P. Fabrie, C. Galusinski and A. Miranville, *Uniform inertial sets for damped wave equations*, Discrete Continuous Dynamical Systems **6** (2) (2000), 393–418.
[143] E. Feireisl, *Attractors for wave equations with nonlinear dissipation and critical exponent*, C. R. Acad. Sci. Paris Sér. I Math. **315** (5) (1992), 551–555.
[144] E. Feireisl, *Finite-dimensional asymptotic behavior of some semilinear damped hyperbolic problems*, J. Dynamics Differential Equations **6** (1) (1994), 23–35.
[145] E. Feireisl, *Global attractors for semilinear damped wave equations with supercritical exponent*, J. Differential Equations **116** (2) (1995), 431–447.
[146] E. Feireisl, *Asymptotic behaviour and attractors for a semilinear damped wave equation with supercritical exponent*, Proc. Roy. Soc. Edinburgh Sect. A **125** (5) (1995), 1051–1062.
[147] E. Feireisl, *Bounded, locally compact global attractors for semilinear damped wave equations on* $\mathbf{R}^n$, Differential Integral Equations **9** (5) (1996), 1147–1156.
[148] E. Feireisl, *On the long time behaviour of solutions to nonlinear diffusion equations on* $\mathbf{R}^n$, Nonlinear Differential Equations Appl. **4** (1) (1997), 43–60.
[149] E. Feireisl, *Global attractors for the Navier–Stokes equations of three-dimensional compressible flow*, C. R. Acad. Sci. Paris Sér. I Math. **331** (1) (2000), 35–39.
[150] E. Feireisl, *Viscous and/or heat conducting compressible fluids*, Handbook of Mathematical Fluid Dynamics, Vol. I, North-Holland, Amsterdam (2002), 307–371.
[151] E. Feireisl, Ph. Laurençot and F. Simondon, *Global attractors for degenerate parabolic equations on unbounded domains*, J. Differential Equations **129** (2) (1996), 239–261.
[152] E. Feireisl and P. Poláčik, *Structure of periodic solutions and asymptotic behavior for time-periodic reaction–diffusion equations on* $\boldsymbol{R}$, Adv. Differential Equations **5** (4–6) (2000), 583–622.
[153] E. Feireisl and F. Simondon, *Convergence for degenerate parabolic equations*, J. Differential Equations **152** (2) (1999), 439–466.
[154] E. Feireisl and F. Simondon, *Convergence for semilinear degenerate parabolic equations in several space dimensions*, J. Dynamics Differential Equations **12** (3) (2000), 647–673.
[155] E. Feireisl and E. Zuazua, *Global attractors for semilinear wave equations with locally distributed nonlinear damping and critical exponent*, Comm. Partial Differential Equations **18** (9–10) (1993), 1539–1555.
[156] B. Fiedler and C. Rocha, *Heteroclinic orbits of semilinear parabolic equations*, J. Differential Equations **125** (1) (1996), 239–281.
[157] B. Fiedler and C. Rocha, *Orbit equivalence of global attractors of semilinear parabolic differential equations*, Trans. Amer. Math. Soc. **352** (1) (2000), 257–284.
[158] B. Fiedler and A. Scheel, *Dynamics of reaction–diffusion patterns*, Trends in Nonlinear Analysis, Festschrift dedicated to Willi Jäger for his 60th birthday, M. Kirkilionis, R. Rannacher and F. Tomi, eds, Springer, Heidelberg (2003), 23–152.
[159] B. Fiedler, A. Scheel and M.I. Vishik, *Large patterns of elliptic systems in infinite cylinders*, J. Math. Pures Appl. (9) **77** (9) (1998), 879–907.
[160] P.C. Fife, *Long time behavior of solutions of bistable nonlinear diffusion equations*, Arch. Rational Mech. Anal. **70** (1) (1979), 31–46.
[161] P. Fife, *Diffusive waves in inhomogeneous media*, Proc. Edinburgh Math. Soc. **32** (1989), 291–315.
[162] P. Fife and L. Hsiao, *The generation and propagation of internal layers*, Nonlinear Anal. **12** (1988), 19–41.
[163] P.C. Fife and J.B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Rational Mech. Anal. **65** (4) (1977), 335–361.
[164] P. Fife and L.A. Peletier, *Clines induced by variable selection and migration*, Proc. Roy. Soc. London B **214** (1981), 99–123.
[165] W.E. Fitzgibbon, M. Parrott and Y. You, *Global dynamics of coupled systems modeling nonplanar beam motion*, Evolution Equations (Baton Rouge, LA, 1992), Lecture Notes in Pure and Appl. Math., Vol. 168, Dekker, New York (1995), 187–199.
[166] W.E. Fitzgibbon, M.E. Parrott and Y. You, *Finite dimensionality and upper semicontinuity of the global attractor of singularly perturbed Hodgkin–Huxley systems*, J. Differential Equations **129** (1) (1996), 193–237.

[167] F. Flandoli and B. Schmalfuß, *Weak solutions and attractors for three-dimensional Navier–Stokes equations with nonregular force*, J. Dynamics Differential Equations **11** (2) (1999), 355–398.
[168] F. Flandoli and B. Schmalfuss, *Random attractors for the 3D stochastic Navier–Stokes equation with multiplicative white noise*, Stochastics Stochastics Rep. **59** (1–2) (1996), 21–45.
[169] W.H. Fleming, *A selection migration model in population genetics*, J. Math. Biology **2** (1975), 219–233.
[170] C. Foias, O. Manley and R. Temam, *Attractors for the Bénard problem: Existence and physical bounds on their fractal dimension*, Nonlinear Anal. **11** (8) (1987), 939–967.
[171] C. Foias, O. Manley and R. Temam, *Modelling of the interaction of small and large eddies in two-dimensional turbulent flows*, RAIRO Modél. Math. Anal. Numér. **22** (1) (1988), 93–118.
[172] C. Foias, O. Manley, R. Temam and Y. Trève, *Asymptotic analysis of the Navier–Stokes equations*, Physica D, Nonlinear Phenomena **90** (1983), 157–188.
[173] C. Foias, B. Nicolaenko, G.R. Sell and R. Temam, *Inertial manifolds for the Kuramoto–Sivashinsky equation and an estimate of their lowest dimension*, J. Math. Pures Appl. (9) **67** (3) (1988), 197–226.
[174] C. Foias and E. Olson, *Finite fractal dimension and Hölder–Lipschitz parametrization*, Indiana Univ. Math. J. **45** (3) (1996), 603–616.
[175] C. Foias and G. Prodi, *Sur le comportement global des solutions non-stationnaires des équations de Navier–Stokes en dimension 2*, Rend. Sem. Mat. Univ. Padova **39** (1967), 1–34.
[176] C. Foias and J.-C. Saut, *Linearization and normal form of the Navier–Stokes equations with potential forces*, Ann. Inst. H. Poincaré Anal. Non Linéaire **4** (1) (1987), 1–47.
[177] C. Foias and J.-C. Saut, *Asymptotic integration of Navier–Stokes equations with potential forces. I*, Indiana Univ. Math. J. **40** (1) (1991), 305–320.
[178] C. Foias, G.R. Sell and R. Temam, *Variétés inertielles des équations différentielles dissipatives*, C. R. Acad. Sci. Paris Sér. I Math. **301** (5) (1985), 139–141.
[179] C. Foias, G.R. Sell and R. Temam, *Inertial manifolds for nonlinear evolutionary equations*, J. Differential Equations **73** (2) (1988), 309–353.
[180] C. Foias, G.R. Sell and E.S. Titi, *Exponential tracking and approximation of inertial manifolds for dissipative nonlinear equations*, J. Dynamics Differential Equations **1** (2) (1989), 199–244.
[181] C. Foias and R. Temam, *Some analytic and geometric properties of the solutions of the evolution Navier–Stokes equations*, J. Math. Pures Appl. (9) **58** (3) (1979), 339–368.
[182] C. Foias and R. Temam, *On the Hausdorff dimension of an attractor for the two-dimensional Navier–Stokes equations*, Phys. Lett. A **93** (9) (1983), 451–454.
[183] C. Foias and R. Temam, *Determination of the solutions of the Navier–Stokes equations by a set of nodal values*, Math. Comp. **43** (1) (1984), 117–133.
[184] C. Foias and R. Temam, *The connection between the Navier–Stokes equations, dynamical systems, and turbulence theory*, Directions in Partial Differential Equations (Madison, WI, 1985), Publ. Math. Res. Center Univ. Wisconsin, Vol. 54, Academic Press, Boston, MA (1987), 55–73.
[185] C. Foias and R. Temam, *Approximation of attractors by algebraic or analytic sets*, SIAM J. Math. Anal. **25** (5) (1994), 1269–1302.
[186] C. Foias and E.S. Titi, *Determining nodes, finite difference schemes and inertial manifolds*, Nonlinearity **4** (1) (1991), 135–153.
[187] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall (1964).
[188] P.K. Friz and J.C. Robinson, *Parametrising the attractor of the two-dimensional Navier–Stokes equations with a finite number of nodal values*, Physica D **148** (3–4) (2001), 201–220.
[189] H. Fujii and Y. Nishiura, *On the phenomenon of multiple existence of stable equilibria. Coexistence of stable singularly perturbed solutions in systems of reaction–diffusion equations*, Computing Methods in Applied Sciences and Engineering, VII (Versailles, 1985), North-Holland, Amsterdam (1986), 197–212.
[190] H. Fujii, Y. Nishiura and Y. Hosono, *On the structure of multiple existence of stable stationary solutions in systems of reaction–diffusion equations*, Patterns and Waves, Stud. Math. Appl., Vol. 18, North-Holland, Amsterdam (1986), 157–219.
[191] G. Fusco, *A system of ODE which has the same attractor as a scalar parabolic PDE*, J. Differential Equations **69** (1987), 85–110.
[192] G. Fusco and J.K. Hale, *Stable equilibria in a scalar parabolic equation with variable diffusion*, SIAM J. Math. Anal. **16** (1985), 1152–1164.

[193] Th. Gallay and A. Mielke, *Diffusive mixing of stable states in the Ginzburg–Landau equation*, Comm. Math. Phys. **199** (1) (1998), 71–97.
[194] Th. Gallay and G. Raugel, *Stability of travelling waves for a damped hyperbolic equation*, Z. Angew. Math. Phys. **48** (3) (1997), 451–479.
[195] B. García-Archilla, J. Novo and E.S. Titi, *An approximate inertial manifolds approach to postprocessing the Galerkin method for the Navier–Stokes equations*, Math. Comp. **68** (227) (1999), 893–911.
[196] F. Gazzola, *An attractor for a 3D Navier–Stokes type equation*, Z. Anal. Anwendungen **14** (3) (1995), 509–522.
[197] F. Gazzola and V. Pata, *A uniform attractor for a non-autonomous generalized Navier–Stokes equation*, Z. Anal. Anwendungen **16** (2) (1997), 435–449.
[198] J.-M. Ghidaglia, *On the fractal dimension of attractors for viscous incompressible fluid flows*, SIAM J. Math. Anal. **17** (5) (1986), 1139–1157.
[199] J.-M. Ghidaglia and R. Temam, *Propriétés des attracteurs associés à des équations hyperboliques non linéaires amorties*, C. R. Acad. Sci. Paris Sér. I Math. **300** (7) (1985), 185–188.
[200] J.-M. Ghidaglia and R. Temam, *Regularity of the solutions of second order evolution equations and their attractors*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **14** (3) (1987), 485–511.
[201] J.-M. Ghidaglia and R. Temam, *Long time behavior for partly dissipative equations: The slightly compressible 2D-Navier–Stokes equations*, Asymptotic Anal. **1** (1) (1988), 23–49.
[202] J.-M. Ghidaglia and R. Temam, *Lower bound on the dimension of the attractor for the Navier–Stokes equations in space dimension 3*, Mechanics, Analysis and Geometry: 200 years after Lagrange, North-Holland Delta Ser., North-Holland, Amsterdam (1991), 33–60.
[203] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer, New York (1983).
[204] B. Guo and Y. Li, *Attractor for dissipative Klein–Gordon–Schrödinger equations in* $\mathbb{R}^3$, J. Differential Equations **136** (2) (1997), 356–377.
[205] J.K. Hale, *Asymptotic Behavior of Dissipative Systems*, Math. Surveys Monographs, Vol. 25, Amer. Math. Soc., Providence, RI (1988).
[206] J.K. Hale, *Topics in Dynamic Bifurcation Theory*, CBMS Regional Conference Series in Mathematics, Vol. 47, Conference Board of the Mathematical Sciences, Washington, DC (1981).
[207] J.K. Hale, *Theory of Functional Differential Equations*, 2nd edn, Appl. Math. Sci., Vol. 3, Springer, New York (1977), x+365 pp.
[208] J.K. Hale, *Asymptotic behaviour and dynamics in infinite dimensions*, Nonlinear Differential Equations (Granada, 1984), Res. Notes in Math., Vol. 132, Pitman, Boston, MA (1985), 1–42.
[209] J.K. Hale, *Asymptotic Behavior of Dissipative Systems*, Math. Surveys Monographs, Vol. 25, Amer. Math. Soc. (1988).
[210] J.K. Hale, *Attracting manifolds for evolutionary equations*, Workshop on Differential Equations and Nonlinear Analysis (Águas de Lindóia, 1996), Resenhas **3** (1) (1997), 55–72.
[211] J.K. Hale, *Dynamics of a scalar parabolic equation*, Canad. Appl. Math. Quart. **5** (1997), 209–305.
[212] J.K. Hale, J.P. La Salle and M. Slemrod, *Theory of a general class of dissipative processes*, J. Math. Anal. Appl. **39** (1972), 177–191.
[213] J.K. Hale, X.-B. Lin and G. Raugel, *Upper semicontinuity of attractors for approximations of semigroups and partial differential equations*, Math. Comp. **50** (181) (1988), 89–123.
[214] J.K. Hale, L.T. Magalhães and W.M. Oliva, *An Introduction to Infinite-Dimensional Dynamical Systems—Geometric Theory*, Appl. Math. Sci., Vol. 47, Springer, New York (1984).
[215] J.K. Hale and G. Raugel, *Upper semicontinuity of the attractor for a singularly perturbed hyperbolic equation*, J. Differential Equations **73** (2) (1988), 197–214.
[216] J.K. Hale and G. Raugel, *Lower semicontinuity of attractors of gradient systems and applications*, Ann. Mat. Pura Appl. (4) **154** (1989), 281–326.
[217] J.K. Hale and G. Raugel, *Lower semicontinuity of the attractor for a singularly perturbed hyperbolic equation*, J. Dynamics Differential Equations **2** (1) (1990), 19–67.
[218] J.K. Hale and G. Raugel, *Convergence in gradient-like systems and applications*, Z. Angew. Math. Phys. **43** (1992), 63–124.
[219] J.K. Hale and G. Raugel, *A damped hyperbolic equation on thin domains*, Trans. Amer. Math. Soc. **329** (1) (1992), 185–219.

[220] J.K. Hale and G. Raugel, *Limits of semigroups depending on parameters*, Resenhas **1** (1) (1993), 1–45.
[221] J.K. Hale and G. Raugel, *A reaction–diffusion equation on a thin L-shaped domain*, Proc. Roy. Soc. Edinburgh Sect. A **125** (2) (1995), 283–327.
[222] J.K. Hale and C. Rocha, *Bifurcation in a parabolic equation with variable diffusion*, Nonlinear Anal. **9** (1985), 479–494.
[223] J.K. Hale and K. Sakamoto, *Existence and stability of transition layers*, Japan J. Appl. Math. **5** (1988), 367–405.
[224] J.K. Hale and J. Scheurle, *Smoothness of bounded solutions of nonlinear evolution equations*, J. Differential Equations **56** (1985), 142–163.
[225] A. Haraux, *Two remarks on hyperbolic dissipative problems*, Nonlinear Partial Differential Equations and their Applications, Collège de France seminar, Vol. VII (Paris, 1983–1984), 6, Res. Notes in Math., 122, Pitman, Boston, MA (1985), 161–179.
[226] A. Haraux, *Semi-linear hyperbolic problems in bounded domains*, Math. Rep. **3** (1) (1987), i–xxiv and 1–281.
[227] A. Haraux, *Attractors of asymptotically compact processes and applications to nonlinear partial differential equations*, Comm. Partial Differential Equations **13** (11) (1988), 1383–1414.
[228] A. Haraux, *Systèmes dynamiques dissipatifs et applications* (French) [*Dissipative Dynamical Systems and Applications*], Recherches en Mathématiques Appliquées [Research in Applied Mathematics], Vol. 17, Masson, Paris (1991).
[229] A. Haraux and P. Poláčik, *Convergence to a positive equilibrium for some nonlinear evolution equations in a ball*, Acta Math. Univ. Comenian. (N.S.) **61** (2) (1992), 129–141.
[230] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math., Vol. 840, Springer, Berlin (1981).
[231] D. Henry, *Some infinite dimensional Morse–Smale systems defined by parabolic differential equations*, J. Differential Equations **59** (1985), 165–205.
[232] P. Hess and P. Poláčik, *Symmetry and convergence properties for non-negative solutions of nonautonomous reaction–diffusion problems*, Proc. Roy. Soc. Edinburgh Sect. A **124** (3) (1994), 573–587.
[233] A.T. Hill and E. Süli, *Approximation of the global attractor for the incompressible Navier–Stokes equations*, IMA J. Numer. Anal. **20** (4) (2000), 633–667.
[234] D. Hoff and M. Ziane, *Compact attractors for the Navier–Stokes equations of one-dimensional, compressible flow*, C. R. Acad. Sci. Paris Sér. I Math. **328** (3) (1999), 239–244.
[235] D. Hoff and M. Ziane, *The global attractor and finite determining nodes for the Navier–Stokes equations of compressible flow with singular initial data*, Indiana Univ. Math. J. **49** (3) (2000), 844–889.
[236] B.R. Hunt and V.Y. Kaloshin, *Regularity of embeddings of infinite-dimensional fractal sets into finite-dimensional spaces*, Nonlinearity **12** (5) (1999), 1263–1275.
[237] D. Iftimie and G. Raugel, *Some results on the Navier–Stokes equations in thin 3D domains*, Special issue in celebration of Jack K. Hale's 70th birthday, Part 4 (Atlanta, GA/Lisbon, 1998), J. Differential Equations **169** (2) (2001), 281–331.
[238] Yu.S. Ilyashenko, *Weakly contracting systems and attractors of Galerkin approximations of Navier–Stokes equations on the two-dimensional torus*, Adv. in Mech. **5** (1–2) (1982), 31–63; Selected translations. Selecta Math. Soviet. **11** (3) (1992), 203–239.
[239] A.A. Ilyin, *On the dimension of attractors for Navier–Stokes equations on two-dimensional compact manifolds*, Differential Integral Equations **6** (1) (1993), 183–214.
[240] A.A. Ilyin, *Lieb–Thirring inequalities on the N-sphere and in the plane, and some applications*, Proc. London Math. Soc. (3) **67** (1) (1993), 159–182.
[241] A.A. Ilyin, *Partially dissipative semigroups generated by the Navier–Stokes system on two-dimensional manifolds, and their attractors*, Mat. Sbornik **184** (1) (1993), 55–88 (Russian); English transl. in Russian Acad. Sci. Sb. Math. **78** (1) (1994), 47–76.
[242] A.A. Ilyin, *Navier–Stokes equations on the rotating sphere. A simple proof of the attractor dimension estimate*, Nonlinearity **7** (1) (1994), 31–39.
[243] A.A. Ilyin, *Attractors for Navier–Stokes equations in domains with finite measure*, Nonlinear Anal. **27** (5) (1996), 605–616.
[244] A.A. Ilyin, *Global averaging of dissipative dynamical systems*, Rend. Accad. Naz. Sci. XL Mem. Mat. Appl. (5) **22** (1998), 165–191.

[245] D.A. Jones and E.S. Titi, *On the number of determining nodes for the 2D Navier–Stokes equations*, J. Math. Anal. Appl. **168** (1) (1992), 72–88.

[246] D.A. Jones and E.S. Titi, *Upper bounds on the number of determining modes, nodes, and volume elements for the Navier–Stokes equations*, Indiana Univ. Math. J. **42** (3) (1993), 875–887.

[247] N. Ju, *The H_1-compact global attractor for the solutions to the Navier–Stokes equations in two-dimensional unbounded domains*, Nonlinearity **13** (4) (2000), 1227–1238.

[248] L. Kapitanski, *Minimal compact global attractor for a damped semilinear wave equation*, Comm. Partial Differential Equations **20** (7–8) (1995), 1303–1323.

[249] A.V. Kapustyan and V.S. Mel'nik, *Global attractors of multivalued semidynamical systems and their approximations*, Kibernet. Sistem. Anal. **5** (1998), 102–111, 189 (Russian); English transl. in Cybernet. Systems Anal. **34** (5) (1998), 719–725.

[250] A.V. Kapustyan and V.S. Mel'nik, *On the global attractors of multivalued semidynamical systems and their approximations*, Dokl. Akad. Nauk **366** (4) (1999), 445–448.

[251] N.I. Karachalios and N.M. Stavrakakis, *Existence of a global attractor for semilinear dissipative wave equations on $\mathbb{R}^N$*, J. Differential Equations **157** (1) (1999), 183–205.

[252] N.I. Karachalios and N.M. Stavrakakis, *Global attractor for the weakly damped driven Schrödinger equation in $H^2(\mathbb{R})$*, Nonlinear Differential Equations Appl. **9** (3) (2002), 347–360.

[253] N.I. Karachalios and N.M. Stavrakakis, *Estimates on the dimension of a global attractor for a semilinear dissipative wave equation on $\mathbb{R}^N$*, Discrete Continuous Dynamical Systems **8** (4) (2002), 939–951.

[254] K. Kirchgässner, *Wave solutions of reversible systems and applications*, J. Differential Equations **45** (1982), 113–127.

[255] K. Kishimoto and H.F. Weinberger, *The spatial homogeneity of stable equilibria of some reaction–diffusion systems on convex domains*, J. Differential Equations **58** (1985), 15–21.

[256] A.N. Kolmogorov, I.G. Petrovskii and N.S. Piskunov, *Etude de la di usion avec croissance de la quantit e de mati ere et son application a un probleme biologique*, Moscow Univ. Math. Bull. **1** (1937), 1–25.

[257] A.N. Kolmogorov and V.M. Tihomirov, *ε-entropy and ε-capacity of sets in functional space*, Amer. Math. Soc. Transl. (2) **17** (1961), 277–364.

[258] N.V. Krylov, *Nonlinear Elliptic and Parabolic Equations of Second Order*, Reidel (1987).

[259] I. Kukavica and J.C. Robinson, *Distinguishing smooth functions by a finite number of point values, and a version of the Takens embedding theorem*, Physica D **196** (1–2) (2004), 45–66. 37L30.

[260] S. Kuksin and A. Shirikyan, *Stochastic dissipative PDEs and Gibbs measures*, Comm. Math. Phys. **213** (2) (2000), 291–330.

[261] O.A. Ladyzhenskaya, *A dynamical system that is generated by the Navier–Stokes equations*, Dokl. Akad. Nauk SSSR **205** (1972), 318–320.

[262] O.A. Ladyzhenskaya, *A dynamical system generated by Navier–Stokes equations*, Zap. Nauchn. Sem. LOMI **27** (1972), 91–115; English transl. in J. Soviet Math. **3** (4) (1975), 458–479.

[263] O.A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, 2nd edn, Gordon and Breach, New York (1969).

[264] O.A. Ladyzhenskaya, *Attractors of nonlinear evolution problems with dissipation*, Kraev. Zadachi Mat. Fiz. i Smezhn. Vopr. Teor. Funktsii, Vol. 18, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **152** (1986), 72–85, 182; English transl. in J. Soviet Math. **40** (5) (1988), 632–640.

[265] O.A. Ladyzhenskaya, *Finding minimal global attractors for the Navier–Stokes equations and other partial differential equations*, Uspekhi Mat. Nauk **42** (6) (258) (1987), 25–60; English transl. in Russian Math. Surveys **42** (6) (1987), 27–73.

[266] O.A. Ladyzhenskaya, *Estimates for the fractal dimension and number of deterministic modes for invariant sets of dynamical systems*, Kraev. Zadachi Mat. Fiz. i Smezhn. Vopr. Teor. Funktsii, Vol. 19, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **163** (1987), 105–129, 188; English transl. in J. Soviet Math. **49** (5) (1990), 1186–1201.

[267] O.A. Ladyzhenskaya, *Estimates for the fractal dimension and number of deterministic modes for invariant sets of dynamical systems*, Kraev. Zadachi Mat. Fiz. i Smezhn. Vopr. Teor. Funktsii, Vol. 19, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **163** (1987), 105–129, 188; English transl. in J. Soviet Math. **49** (5) (1990), 1186–1201.

[268] O.A. Ladyzhenskaya, *Attractors for second-order quasilinear parabolic equations of general form*, Kraev. Zadachi Mat. Fiz. i Smezhn. Vopr. Teor. Funktsii, Vol. 20, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **171** (1989), 163–173, 186; English transl. in J. Soviet Math. **56** (2) (1991), 2389–2396.
[269] O.A. Ladyzhenskaya, *Some additions and refinements to my papers on the theory of attractors for abstract semigroups*, Kraev. Zadachi Mat. Fiz. i Smezh. Voprosy Teor. Funktsii, Vol. 21, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **182** (1990), 102–112, 172 (Russian); English transl. in J. Soviet Math. **62** (3) (1992), 2789–2794.
[270] O.A. Ladyzhenskaya, *Attractors for Semigroups and Evolution Equations*, Lezioni Lincee, Cambridge University Press, Cambridge (1991).
[271] O. Ladyzhenskaya, *First boundary value problem for the Navier–Stokes equations in domains with non-smooth boundaries*, C. R. Acad. Sci. Paris Sér. I Math. **314** (4) (1992), 253–258.
[272] O.A. Ladyzhenskaya, *Attractors for the modifications of the three-dimensional Navier–Stokes equations*, Philos. Trans. Roy. Soc. London Ser. A **346** (1679) (1994), 173–190.
[273] O.A. Ladyzhenskaya, *An attractor for a 3D Navier–Stokes type equation*, Z. Anal. Anwendungen **14** (3) (1995), 509–522.
[274] O.A. Ladyzhenskaya and G.A. Seregin, *Smoothness of solutions of equations describing generalized Newtonian flows and estimates for the dimensions of their attractors*, Izv. Math. **62** (1) (1998), 55–113.
[275] O.A. Ladyzhenskaja, V.A. Solonnikov, N.N. Ural'seva, *Linear and Quasi-Linear Equations of Parabolic Type*, Amer. Math. Soc., Providence (1968).
[276] A.W. Leung, *Systems of Nonlinear Partial Differential Equations. Applications to Biology and Engineering*, Math. Appl., Kluwer Academic Publishers, Dordrecht (1989).
[277] J.L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéares*, Dunod, Paris (1969).
[278] J.L. Lions, O.P. Manley, R. Temam and S. Wang, *Physical interpretation of the attractor dimension for the primitive equations of atmospheric circulation*, J. Atmospheric Sci. **54** (9) (1997), 1137–1143.
[279] J.L. Lions, R. Temam and S. Wang, *Geostrophic asymptotics of the primitive equations of the atmosphere*, Topol. Methods Nonlinear Anal. **4** (1994), 253–287, special issue dedicated to J. Leray.
[280] J.-L. Lions, R. Temam and S.H. Wang, *Mathematical theory for the coupled atmosphere–ocean models (CAO III)*, J. Math. Pures Appl. (9) **74** (2) (1995), 105–163.
[281] J.-L. Lions, R. Temam and S. Wang, *A simple global model for the general circulation of the atmosphere*, Comm. Pure Appl. Math. **50** (1997), 707–752.
[282] V.X. Liu, *A sharp lower bound for the Hausdorff dimension of the global attractors of the 2D Navier–Stokes equations*, Comm. Math. Phys. **158** (2) (1993), 327–339.
[283] V.X. Liu, *Remarks on the Navier–Stokes equations on the two- and three-dimensional torus*, Comm. Partial Differential Equations **19** (5–6) (1994), 873–900.
[284] O. Lopes, *Compact attractor for a nonlinear wave equation with critical exponent*, Proc. Roy. Soc. Edinburgh Sect. A **115** (1–2) (1990), 61–64.
[285] A. Mahalov, S. Leibovich and E.S. Titi, *Invariant helical subspaces for the Navier–Stokes equations*, Arch. Rational Mech. Anal. **112** (3) (1990), 193–222.
[286] J. Mallet-Paret, *Negatively invariant sets of compact maps and an extension of a theorem of Cartwright*, J. Differential Equations **22** (1976), 331–348.
[287] J. Mallet-Paret and G.R. Sell, *Inertial manifolds for reaction diffusion equations in higher space dimensions*, J. Amer. Math. Soc. **1** (4) (1988), 805–866.
[288] J. Mallet-Paret, G.R. Sell and Z.D. Shao, *Obstructions to the existence of normally hyperbolic inertial manifolds*, Indiana Univ. Math. J. **42** (3) (1993), 1027–1055.
[289] R. Mañé, *On the dimension of the compact invariant sets of certain nonlinear maps*, Dynamical Systems and Turbulence, Warwick 1980 (Coventry, 1979/1980), Lecture Notes in Math., Vol. 898, Springer, Berlin (1981), 230–242.
[290] O. Manley, M. Marion and R. Temam, *Fully nonlinear multispecies reaction–diffusion equations*, Appl. Math. Lett. **8** (4) (1995), 7–11.
[291] J.E. Marsden and M. McCracken, *The Hopf Bifurcation and its Applications*, Appl. Math. Series, Vol. 19, Springer, New York (1976).
[292] P. Massatt, *Some properties of condensing maps*, Ann. Mat. Pura Appl. (4) **125** (1980), 101–115.
[293] P. Massatt, *Stability and fixed points of point-dissipative systems*, J. Differential Equations **40** (2) (1981), 217–231.

[294] H. Matano, *Asymptotic behavior and stability of solutions of semilinear diffusion equations*, Publ. Res. Inst. Math. Sci. **15** (1979), 401–458.
[295] H. Matano, *Nonincrease of the lap-number of a solution for a one-dimensional semilinear parabolic equation*, J. Fac. Sci. Univ. Tokyo Sect. IA Math. **29** (2) (1982), 401–441.
[296] H. Matano and K.-I. Nakamura, *The global attractor of semilinear parabolic equations on S^1*, Discrete Continuous Dynamical Systems **3** (1) (1997), 1–24.
[297] V.S. Mel'nik, *Multivalued semiflows and their attractors*, Dokl. Akad. Nauk **343** (3) (1995), 302–305 (Russian).
[298] V.S. Mel'nik, *On estimates for the fractal and Hausdorff dimensions of sets that are invariant under multivalued mappings*, Mat. Zametki **63** (2) (1998), 217–224 (Russian); English transl. in Math. Notes **63** (1–2) (1998), 190–196.
[299] V.S. Mel'nik and J. Valero, *On attractors of multivalued semi-flows and differential inclusions*, Set-Valued Anal. **6** (1) (1998), 83–111.
[300] V.S. Mel'nik and J. Valero, *On global attractors of multivalued semiprocesses and nonautonomous evolution inclusions*, Set-Valued Anal. **8** (4) (2000), 375–403.
[301] S. Merino, *On the existence of the compact global attractor for semilinear reaction diffusion systems on $\mathbb{R}^N$*, J. Differential Equations **132** (1) (1996), 87–106.
[302] A. Mielke, *Essential manifolds for an elliptic problem in an infinite strip*, J. Differential Equations **110** (2) (1994), 322–355.
[303] A. Mielke, *The complex Ginzburg–Landau equation on large and unbounded domains: Sharper bounds and attractors*, Nonlinearity **10** (1) (1997), 199–222.
[304] A. Mielke, *The Ginzburg–Landau equation in its role as a modulation equation*, Handbook of Dynamical Systems, Vol. 2, North-Holland, Amsterdam (2002), 759–834.
[305] A. Mielke and G. Schneider, *Attractors for modulation equations on unbounded domains – existence and comparison*, Nonlinearity **8** (5) (1995), 743–768.
[306] A. Mielke and S. Zelik, *Infinite-dimensional trajectory attractors of elliptic boundary value problems in cylindrical domains*, Uspekhi Mat. Nauk **57** (4) (346) (2002), 119–150 (Russian); English transl. in Russian Math. Surveys **57** (4) (2002), 753–784.
[307] R.K. Miller, *Almost periodic differential equations as dynamical systems with applications to existence of a.p. solutions*, J. Differential Equations **1** (1965), 337–345.
[308] J. Milnor, *On the concept of attractors*, Comm. Math. Phys. **99** (1985), 177–195; *On the concept of attractors: Correction and remarks*, Comm. Math. Phys. **102** (1985), 517–519.
[309] A. Miranville, *Exponential attractors for nonautonomous evolution equations*, Appl. Math. Lett. **11** (2) (1998), 19–22.
[310] A. Miranville, *Exponential attractors for a class of evolution equations by a decomposition method*, C. R. Acad. Sci. Paris Sér. I Math. **328** (2) (1999), 145–150.
[311] A. Miranville, *Exponential attractors for a class of evolution equations by a decomposition method. II. The non-autonomous case*, C. R. Acad. Sci. Paris Sér. I Math. **328** (10) (1999), 907–912.
[312] A. Miranville, *A model of Cahn–Hilliard equation based on a microforce balance*, C. R. Acad. Sci. Paris Sér. I Math. **328** (12) (1999), 1247–1252.
[313] A. Miranville, *Some models of Cahn–Hilliard equations in nonisotropic media*, Math. Model. Numer. Anal. **34** (3) (2000), 539–554.
[314] A. Miranville, *Some generalizations of the Cahn–Hilliard equation*, Asymptotic Anal. **22** (3–4) (2000), 235–259.
[315] A. Miranville and X. Wang, *Upper bound on the dimension of the attractor for nonhomogeneous Navier–Stokes equations*, Discrete Continuous Dynamical Systems **2** (1) (1996), 95–110.
[316] A. Miranville and X. Wang, *Attractors for nonautonomous nonhomogeneous Navier–Stokes equations*, Nonlinearity **10** (5) (1997), 1047–1061.
[317] A. Miranville and M. Ziane, *On the dimension of the attractor for the Bénard problem with free surfaces*, Russian J. Math. Phys. **5** (4) (1997), 489–502.
[318] I. Moise, R. Rosa and X. Wang, *Attractors for non-compact semigroups via energy equations*, Nonlinearity **11** (5) (1998), 1369–1393.
[319] I. Moise, R. Temam and M. Ziane, *Asymptotic analysis of the Navier–Stokes equations in thin domains*, Dedicated to Olga Ladyzhenskaya, Topol. Methods Nonlinear Anal. **10** (2) (1997), 249–282.

[320] S. Montgomery-Smith, *Global regularity of the Navier–Stokes equation on thin three-dimensional domains with periodic boundary conditions*, Electron. J. Differential Equations **11** (1999), 19 pp.
[321] X. Mora and J. Solà-Morales, *Existence and nonexistence of finite-dimensional globally attracting invariant manifolds in semilinear damped wave equations*, Dynamics of Infinite-Dimensional Systems (Lisbon, 1986), NATO Adv. Sci. Inst. Ser. F Comput. Systems Sci., Vol. 37, Springer, Berlin (1987), 187–210.
[322] B. Nicolaenko and W.J. Qian, *Inertial manifolds for nonlinear viscoelasticity equations*, Nonlinearity **11** (4) (1998), 1075–1093.
[323] Y. Nishiura, *Coexistence of infinitely many stable solutions to reaction diffusion systems in the singular limit*, Dynamics Reported, New Series, No. 3, Springer (1994), 25–103.
[324] S. Njamkepo, *Existence of a global attractor for the slightly compressible 2-D Navier–Stokes equations in the case of a thermohydraulic problem*, Math. Models Methods Appl. Sci. **6** (1) (1996), 59–75.
[325] S. Njamkepo, *Global existence and long-time behavior for a partially dissipative system modelling a polymerization–crystallization reaction*, Math. Models Methods Appl. Sci. **8** (2) (1998), 219–249.
[326] G. Papanicolaou and X. Xin, *Reaction–diffusion fronts in periodically layered media*, J. Statist. Phys. **63** (5–6) (1991), 915–931.
[327] J. Pauwelussen, *Nerve impulse propagation in a branching nerve system: A simple model*, Physica D **4** (1981), 67–88.
[328] S.Yu. Pilyugin, *Shadowing in Dynamical Systems*, Lecture Notes in Math., Vol. 1706, Springer, Berlin (1999).
[329] P. Poláčik, *Transversal and nontransversal intersections of stable and unstable manifolds in reaction diffusion equations on symmetric domains*, Differential Integral Equations **7** (5–6) (1994), 1527–1545.
[330] P. Poláčik, *High-dimensional ω-limit sets and chaos in scalar parabolic equations*, J. Differential Equations **119** (1) (1995), 24–53.
[331] P. Poláčik, *Parabolic equations: Asymptotic behavior and dynamics on invariant manifolds*, Handbook of Dynamical Systems, Vol. 2, North-Holland, Amsterdam (2002), 835–883.
[332] P. Poláčik and K. Rybakowski, *Imbedding vector fields in scalar parabolic Dirichlet BVPs*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **22** (4) (1995), 737–749.
[333] M. Prizzi, *A remark on reaction–diffusion equations in unbounded domains*, Discrete Continuous Dynamical Systems **9** (2) (2003), 281–286.
[334] M. Protter and H. Weinberger, *Maximum Principles in Differential Equations*, Prentice-Hall (1967).
[335] G. Raugel, *Une équation des ondes avec amortissement non linéaire dans le cas critique en dimension trois*, C. R. Acad. Sci. Paris Sér. I Math. **314** (3) (1992), 177–182.
[336] G. Raugel, *Global attractors in partial differential equations*, Handbook of Dynamical Systems, Vol. 2, North-Holland, Amsterdam (2002), 885–982.
[337] G. Raugel and G. Sell, *Navier–Stokes equations on thin 3D domains. I. Global attractors and global regularity of solutions*, J. Amer. Math. Soc. **6** (3) (1993), 503–568.
[338] J.C. Robinson, *Infinite-Dimensional Dynamical Systems. An Introduction to Dissipative Parabolic PDEs and the Theory of Global Attractors*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge (2001).
[339] C. Rocha, *Examples of attractors in scalar reaction–diffusion equations*, J. Differential Equations **73** (1988), 178–195.
[340] R. Rosa, *The global attractor for the 2D Navier–Stokes flow on some unbounded domains*, Nonlinear Anal. **32** (1) (1998), 71–85.
[341] D. Ruelle and F. Takens, *On the nature of turbulence*, Comm. Math. Phys. **20** (1971), 167; **23** (1971), 343–344.
[342] R.J. Sacker and G.R. Sell, *Lifting properties in skew-product flows with applications to differential equations*, Mem. Amer. Math. Soc. **190** (1977).
[343] B. Sandstede, *Stability of travelling waves*, Handbook of Dynamical Systems, Vol. 2, North-Holland, Amsterdam (2002), 983–1055.
[344] D.H. Sattinger, *Topics in Stability and Bifurcation Theory*, Lecture Notes in Math., Vol. 309, Springer, Berlin (1973).
[345] A. Scheel, *Existence of fast traveling waves for some parabolic equations: A dynamical systems approach*, J. Dynamics Differential Equations **8** (4) (1996), 469–547.

[346] B. Schmalfuss, *Stochastic dissipative PDE's and Gibbs measures*, Stochastic Anal. Appl. **17** (6) (1999) 1075–1101.
[347] B.-W. Schulze, M.I. Vishik, I. Witt and S.V. Zelik, *The trajectory attractor for a nonlinear elliptic system in a cylindrical domain with piecewise smooth boundary*, Rend. Accad. Naz. Sci. XL Mem. Mat. Appl. (5) **23** (1999), 125–166.
[348] G.R. Sell, *Nonautonomous differential equations and topological dynamics. I. The basic theory*, Trans. Amer. Math. Soc. **127** (1967), 241–262; *II. Limiting equations*, Trans. Amer. Math. Soc. **127** (1967), 263–283.
[349] G.R. Sell, *Topological Dynamics and Ordinary Differential Equations*, Van Nostrand, New York (1971).
[350] G.R. Sell, *Differential equations without uniqueness and classical topological dynamics*, J. Differential Equations **14** (1973), 42–56.
[351] G.R. Sell, *Global attractors for the three-dimensional Navier–Stokes equations*, J. Dynamics Differential Equations **8** (1) (1996), 1–33.
[352] G.R. Sell and Y.C. You, *Inertial manifolds: The nonselfadjoint case*, J. Differential Equations **96** (2) (1992), 203–255.
[353] G.R. Sell and Y. You, *Dynamics of Evolutionary Equations*, Appl. Math. Sci., Vol. 143, Springer, New York (2002).
[354] Z.D. Shao and E.S. Titi, *Parameterizing the global attractor of the Navier–Stokes equations by nodal values*, Numer. Funct. Anal. Optim. **16** (3–4) (1995), 547–563.
[355] A.B. Shapoval, *The integral manifold of a nonlinear elliptic equation in a cylinder*, Math. Notes **61** (3–4) (1997), 391–395.
[356] V.Yu. Skvortsov, *On the attractor of a quasilinear elliptic equation in a strip*, Vestnik Moskov. Univ. Ser. I Mat. Mekh., No. 4 (1993), 74–76 (Russian); English transl. in Moscow Univ. Math. Bull. **48** (4) (1993), 49–51.
[357] V.Yu. Skvortsov and M.I. Vishik, *The asymptotics of solutions of reaction–diffusion equations with small parameter*, Properties of Global Attractors of Partial Differential Equations, Adv. Soviet Math., Vol. 10, Amer. Math. Soc., Providence, RI (1992), 149–172.
[358] H. Smith, *Monotone Dynamical Systems*, Amer. Math. Soc. (1995).
[359] E. Stein, *Harmonic Analysis. Real-Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton University Press (1993).
[360] J.T. Stuart, *Bifurcation theory in non-linear hydrodynamic stability*, Applications of Bifurcation Theory (Proc. Advanced Sem., Univ. Wisconsin, Madison, Wis., 1976), Publ. Math. Res. Center, No. 38, Academic Press, New York (1977), 127–147.
[361] Temam R., *Navier–Stokes Equations: Theory and Numerical Analysis*, North-Holland, Amsterdam (1984).
[362] R. Temam, *Attractors for Navier–Stokes equations*, Nonlinear Partial Differential Equations and their Applications, Collège de France seminar, Vol. VII (Paris, 1983–1984), 10, Res. Notes in Math., Vol. 122, Pitman, Boston, MA (1985), 272–292.
[363] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci., Vol. 68, Springer (1988).
[364] R. Temam, *Attractors for the Navier–Stokes equations: Localization and approximation*, J. Fac. Sci. Univ. Tokyo Sect. IA Math. **36** (3) (1989), 629–647.
[365] R. Temam, *Approximation of attractors, large eddy simulations and multiscale methods*, Turbulence and Stochastic Processes: Kolmogorov's Ideas 50 Years on, Proc. Roy. Soc. London Ser. A **434** (1890) (1991), 23–39.
[366] R. Temam and M. Ziane, *Navier–Stokes equations in three-dimensional thin domains with various boundary conditions*, Adv. Differential Equations **1** (4) (1996), 499–546. Approx. inert manifolds.
[367] L.X. Tian, Y.R. Liu and Z.R. Liu, *Local attractors for weakly damped forced KdV equation in thin 2D domains*, Appl. Math. Mech. **21** (10) (2000), 1131–1138.
[368] E.S. Titi, *Une variété approximante de l'attracteur universel des équations de Navier–Stokes, non linéaire, de dimension finie*, C. R. Acad. Sci. Paris Sér. I Math. **307** (8) (1988), 383–385.
[369] Kh. Valero and V.S. Mel'nik, *On global attractors of nonautonomous evolution inclusions in Banach spaces*, Dokl. Akad. Nauk **375** (6) (2000), 730–733 (Russian).
[370] J. Valero, *Attractors of parabolic equations without uniqueness*, J. Dynamics Differential Equations **13** (4) (2001), 711–744.

[371] M.I. Vishik, *Asymptotic Behaviour of Solutions of Evolutionary Equations*, Lezioni Lincee [Lincei Lectures], Cambridge University Press, Cambridge (1992).
[372] M.I. Vishik and M.Yu. Skvortsov, *Asymptotic behavior of elements of attractors corresponding to singularly perturbed parabolic equations*, Mat. Sbornik **182** (12) (1991), 1769–1785 (Russian); English transl. in Math. USSR-Sb. **74** (2) (1993), 513–529.
[373] M.I. Vishik and M.Yu. Skvortsov, *Asymptotics of trajectories lying on the attractor of a singularly perturbed parabolic equation*, Vestnik Moskov. Univ. Ser. I Mat. Mekh., No. 6 (1991), 11–16, 102 (Russian); English transl. in Moscow Univ. Math. Bull. **46** (6) (1991), 12–16.
[374] M.I. Vishik and S.V. Zelik, *A trajectory attractor of a nonlinear elliptic system in a cylindrical domain*, Mat. Sbornik **187** (12) (1996), 21–56 (Russian); English transl. in Sb. Math. **187** (12) (1996), 1755–1789.
[375] M.I. Vishik and S.V. Zelik, *A regular attractor of a nonlinear elliptic system in a cylindrical domain*, Mat. Sbornik **190** (6) (1999), 23–58 (Russian); English transl. in Sb. Math. **190** (5–6) (1999), 803–834.
[376] A.I. Volpert, V.A. Volpert and V.A. Volpert, *Traveling Wave Solutions of Parabolic Systems*, Amer. Math. Soc. (1994).
[377] B.X. Wang, *Attractors for reaction–diffusion equations in unbounded domains*, Physica D **128** (1) (1999), 41–52.
[378] G.F. Webb, *Compactness of bounded trajectories of dynamical systems in infinite-dimensional spaces*, Proc. Roy. Soc. Edinburgh Sect. A **84** (1–2) (1979), 19–33.
[379] G.F. Webb, *A bifurcation problem for a nonlinear hyperbolic partial differential equation*, SIAM J. Math. Anal. **10** (5) (1979), 922–932.
[380] G.F. Webb, *Existence and asymptotic behavior for a strongly damped nonlinear wave equation*, Canad. J. Math. **32** (3) (1980), 631–643.
[381] X. Xin, *Existence and stability of travelling waves in periodic media governed by a bistable nonlinearity*, J. Dynamics Differential Equations **3** (4) (1991), 541–573.
[382] Y. Yan, *Dimensions of attractors for discretizations for Navier–Stokes equations*, J. Dynamics Differential Equations **4** (2) (1992), 275–340.
[383] E. Yanagida, *Stability of stationary distribution in space-dependent population growth processes*, J. Math. Biol. **15** (1982), 401–441.
[384] Y. You, *Global dynamics of dissipative generalized Korteweg–de Vries equations*, A Chinese summary appears in Chinese Ann. Math. Ser. A **17** (4) (1996), 651; Chinese Ann. Math. Ser. B **17** (4) (1996), 389–402.
[385] T.I. Zelenjak, *Stabilization of solutions of boundary value problems for a second-order parabolic equation with one space variable*, Differencial'nye Uravneniya **4** (1968), 34–45.
[386] S.V. Zelik, *Boundedness of solutions of a nonlinear elliptic system in a cylindrical domain*, Mat. Zametki **61** (3) (1997), 447–450 (Russian); English transl. in Math. Notes **61** (3–4) (1997), 365–369.
[387] S.V. Zelik, *Trajectory attractor of a nonlinear elliptic system in an unbounded domain*, Mat. Zametki **63** (1) (1998), 135–138 (Russian); English transl. in Math. Notes **63** (1–2) (1998), 120–123.
[388] S.V. Zelik, *An attractor of a nonlinear system of reaction–diffusion equations in* $\mathbf{R}^n$ *and estimates for its* ε*-entropy*, Mat. Zametki **65** (6) (1999), 941–944 (Russian); English transl. in Math. Notes **65** (5–6) (1999), 790–792.
[389] S.V. Zelik, *The attractor of a quasilinear hyperbolic equation with dissipation in* $\mathbf{R}^n$*: Dimension and* ε*-entropy*, Mat. Zametki **67** (2) (2000), 304–308 (Russian); English transl. in Math. Notes **67** (1–2) (2000), 248–251.
[390] S. Zelik, *The attractor for a nonlinear reaction–diffusion system with a supercritical nonlinearity and its dimension*, Rend. Accad. Naz. Sci. XL Mem. Mat. Appl. (5) **24** (2000), 1–25.
[391] S.V. Zelik, *The attractor for a nonlinear reaction–diffusion system in the unbounded domain and Kolmogorov's* ε*-entropy*, Math. Nachr. **232** (2001), 129–179.
[392] S.V. Zelik, *The attractor for a nonlinear hyperbolic equation in the unbounded domain*, Discrete Continuous Dynamical Systems **7** (3) (2001), 593–641.
[393] S.V. Zelik, *The attractor for a nonlinear hyperbolic equation in the unbounded domain*, Discrete Continuous Dynamical Systems **7** (3) (2001), 593–641.
[394] S.V. Zelik, *Attractors of reaction–diffusion systems in unbounded domains and their spatial complexity*, Comm. Pure Appl. Math. **56** (5) (2003), 584–637.

[395] S. Zhou, *Dimension of the global attractor for strongly damped nonlinear wave equation*, J. Math. Anal. Appl. **233** (1) (1999), 102–115.

[396] M. Ziane, *Optimal bounds on the dimension of the attractor of the Navier–Stokes equations*, Physica D **105** (1–3) (1997), 1–19.

[397] M. Ziane, *On the two-dimensional Navier–Stokes equations with the free boundary condition*, Appl. Math. Optim. **38** (1) (1998), 1–19.

This page intentionally left blank

CHAPTER 15

Hamiltonian PDEs

Sergei B. Kuksin

Department of Mathematics, Heriot-Watt University, Edinburgh EH14 4AS, Scotland, UK
and Steklov Institute of Mathematics, 8 Gubkina St., 117966 Moscow, Russia
E-mail: *kuksin@ma.hw.ac.uk*

With an appendix by Dario Bambusi

Dipartimento di Mathematica, Politecnico di Milano, Via Saldini 50, 20133 Milano, Italy
E-mail: *dario.bambusi@unimi.it*

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

1. Introduction

In this work we discuss qualitative properties of solutions for Hamiltonian partial differential equations in the finite volume case. That is, when the space-variable x belongs to a finite domain and appropriate boundary conditions are specified on the domain's boundary (or x belongs to the whole space, but the equation contains a potential term, where the potential growths to infinity as $|x| \to \infty$, cf. below Example 5.5 in Section 5.2). Most of these properties have analogies in the classical finite-dimensional Hamiltonian mechanics. In the infinite-volume case properties of the equations become rather different due to the phenomenon of radiation, and we do not touch them here.

Our bibliography is by no means complete.

NOTATION. By $\mathbb{T}^n$ we denote the torus $\mathbb{T}^n = \mathbb{R}^n/2\pi\mathbb{Z}^n$ and write $\mathbb{T}^1 = S^1$; by $\mathbb{R}^n_+$—the open positive octant in $\mathbb{R}^n$; by $\mathbb{Z}_0$—the set of non-zero integers. By $B_\delta(x; X)$ we denote an open δ-ball in a space X, centred at $x \in X$. Abusing notation, we denote by x both the space-variable and an element of an abstract Banach space X. For an invertible linear operator J we set $\bar{J} = -J^{-1}$. The Lipschitz norm of a map f from a metric space M to a Banach space is defined as $\sup_{m\in M} \|f(m)\| + \sup_{m_1\neq m_2} \frac{\|f(m_1)-f(m_2)\|}{\operatorname{dist}(m_1,m_2)}$.

2. Symplectic Hilbert scales and Hamiltonian equations

2.1. *Hilbert scales and their morphisms*

Let X be a real Hilbert space with a scalar product $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_X$ and a Hilbert basis $\{\varphi_k \mid k \in \widetilde{\mathbb{Z}}\}$, where $\widetilde{\mathbb{Z}}$ is a countable subset of some $\mathbb{Z}^n$. Let us take a positive sequence $\{\theta_k \mid k \in \widetilde{\mathbb{Z}}\}$ which goes to infinity with k. For any s we define X_s as a Hilbert space with the Hilbert basis $\{\varphi_k \theta_k^{-s} \mid k \in \widetilde{\mathbb{Z}}\}$. By $\| \cdot \|_s$ and $\langle \cdot, \cdot \rangle_s$ we denote the norm and the scalar product in X_s (in particular, $X_0 = X$ and $\langle \cdot, \cdot \rangle_0 = \langle \cdot, \cdot \rangle$). The totality $\{X_s\}$ is called a *Hilbert scale*, the basis $\{\varphi_k\}$—the *basis of the scale* and the scalar product $\langle \cdot, \cdot \rangle$—the *basic scalar product of the scale*.

A Hilbert scale may be continuous or discrete, depending on whether $s \in \mathbb{R}$ or $s \in \mathbb{Z}$. The objects we define below and the theorems we discuss are valid in both cases.

A Hilbert scale $\{X_s\}$ possesses the following properties:

(1) X_s is compactly embedded in X_r if $s > r$ and is dense there;

(2) the spaces X_s and X_{-s} are conjugated with respect to the scalar product $\langle \cdot, \cdot \rangle$. That is, for any $u \in X_s \cap X_0$ we have

$$\|u\|_s = \sup\bigl\{\langle u, u' \rangle \mid u' \in X_{-s} \cap X_0, \|u'\|_{-s} = 1\bigr\};$$

(3) the norms $\| \cdot \|_s$ satisfy the interpolation inequality; linear operators in the spaces X_s satisfy the interpolation theorem.

Concerning these and other properties of the scales see [77] and [59].

For a scale $\{X_s\}$ we denote by $X_{-\infty}$ and X_∞ the linear spaces $X_{-\infty} = \bigcup X_s$ and $X_\infty = \bigcap X_s$.

Scales of Sobolev functions are the most important for this work:

EXAMPLE 2.1. Basic for us is the Sobolev scale of functions on the d-dimensional torus $\{H^s(\mathbb{T}^d;\mathbb{R}) = H^s(\mathbb{T}^d)\}$. A space $H^s(\mathbb{T}^d)$ is formed by functions $u:\mathbb{T}^d \to \mathbb{R}$ such that

$$u = \sum_{l\in\mathbb{Z}^d} u_l e^{il\cdot x}, \quad \mathbb{C} \ni u_l = \bar{u}_{-l},\ \|u\|_s^2 = \sum_l \big(1+|l|\big)^{2s}|u_l|^2 < \infty.$$

The basis $\{\varphi_k\}$ is formed by all distinct properly normalised functions $\operatorname{Re} e^{il\cdot x}$ and $\operatorname{Im} e^{il\cdot x}$, $l \in \mathbb{Z}^d$.

We shall also use the sub-scale $\{H^s(\mathbb{T}^d)_0\}$, where a space $H^s(\mathbb{T}^d)_0$ consists of functions from $H^s(\mathbb{T}^d)$ with zero mean-value.

EXAMPLE 2.2. Consider the scale $\{H_0^s(0,\pi)\}$, where a space $H_0^s = H_0^s(0,\pi)$ is formed by the odd 2π-periodic functions $u = \sum_{k=1}^\infty u_k \sin kx$ such that $\|u\|_s^2 = \sum |k|^{2s}|u_k|^2 < \infty$. Since $\{\sin nx\}$ is a complete system of eigenfunctions of the operator $-\Delta$ in $L_2(0,\pi)$ with the domain of definition $\{u \in H^2(0,\pi) \mid u(0) = u(\pi) = 0\}$, then an equivalent definition of these spaces is that $H_0^s = \mathcal{D}(-\Delta)^{s/2}$ (see [77]). In particular,

$$\begin{aligned} &H_0^1 = \big\{u \in H^1(0,\pi) \mid u(0) = u(\pi) = 0\big\}, \qquad H_0^2 = H^2(0,\pi) \cap H_0^1, \\ &H_0^3 = \big\{u \in H^3(0,\pi) \mid u(0) = u_{xx}(0) = u(\pi) = u_{xx}(\pi) = 0\big\}. \end{aligned} \tag{2.1}$$

Given two scales $\{X_s\}$, $\{Y_s\}$ and a linear map $L: X_\infty \to Y_{-\infty}$, we denote by $\|L\|_{s_1,s_2} \leqslant \infty$ its norm as a map $X_{s_1} \to Y_{s_2}$. We say that L defines a (*linear*) *morphism of order* d of the two scales for $s \in [s_0, s_1]$, $s_0 \leqslant s_1$,[1] if $\|L\|_{s,s-d} < \infty$ for every $s \in [s_0, s_1]$. If in addition the inverse map L^{-1} exists and defines a morphism of order $-d$ of the scales $\{Y_s\}$ and $\{X_s\}$ for $s \in [s_0 + d, s_1 + d]$, we say that L defines an *isomorphism of order* d for $s \in [s_0, s_1]$. If $\{X_s\} = \{Y_s\}$, then an isomorphism is called an *automorphism*.

EXAMPLE 2.3. Multiplication by a non-vanishing C^r-smooth function defines a zero-order automorphism of the Sobolev scale $\{H^s(\mathbb{T}^n)\}$ for $-r \leqslant s \leqslant r$.

If L is a morphism of scales $\{X_s\}$, $\{Y_s\}$ of order d for $s \in [s_0, s_1]$, then adjoint maps L^* form a morphism of the scales $\{Y_s\}$ and $\{X_s\}$ of the same order d for $s \in [-s_1+d, -s_0+d]$. It is called the *adjoint morphism*.

If $L = L^*$ ($L = -L^*$) on the space X_∞, then the morphism L is called symmetric (antisymmetric).

If L is a symmetric morphism of $\{X_s\}$ of order d for $s \in [s_0, d - s_0]$, where $s_0 \geqslant d/2$, then the adjoint morphism L^* is defined for $s \in [s_0, d - s_0]$ and coincide with L on X_∞; hence, $L^* = L$. We call L a *selfadjoint morphism*. Anti-selfadjoint morphisms are defined similarly.

EXAMPLE 2.4. The operator Δ defines a selfadjoint morphism of order 2 of the Sobolev scale $\{H^s(\mathbb{T}^n)\}$ for $-\infty < s < \infty$. The operators $\partial/\partial x_j$, $1 \leqslant j \leqslant n$, define anti-selfadjoint morphisms of order one. The automorphism in Example 1.1 is selfadjoint.

[1] Or $s \in (s_0, s_1)$, etc.

Let $\{Y_s\}$, $\{Y_s\}$ be two scales and $O_s \subset X_s$, $s \in [a, b]$, be a system of (open) domains, compatible in the following sense:

$$O_{s_1} \cap O_{s_2} = O_{s_2} \quad \text{if } a \leqslant s_1 \leqslant s_2 \leqslant b.$$

Let $F : O_a \to Y_{-\infty}$ be a map such that for every $s \in [a, b]$ its restriction to O_s defines an analytic (C^k-smooth) map $F : O_s \to Y_{s-d}$. Then F is called an analytic (C^k-smooth) morphism of order d for $s \in [a, b]$.

EXAMPLE 2.5. Let $\{X_s\}$ be the Sobolev scale $\{H^s(\mathbb{T}^d)\}$ and $f(u, x)$ be a smooth function. Then the map $F : u(x) \mapsto f(u(x), x)$, $X_a \to X_a$, is smooth if $a > \frac{d}{2}$, so on these spaces ord $F = 0$. If f is analytic, then so is F.

Now let us assume that $d = 1$, f is analytic, $f(0, x) \equiv 0$ and consider F as a map in the scale $\{H_0^s = H_0^s(0, \pi), s \in \mathbb{Z}\}$. For $s \geqslant 1$ the map $F : H_0^s \to H^s(0, \pi)$ is analytic. Since $Fu(0) = Fu(\pi) = 0$, then due to (2.1) for $s = 1$ and $s = 2$ $F(H_0^s) \subset H_0^s$. So on the spaces H_0^1 and H_0^2 we have ord $F = 0$. Since in general for $u \in H_0^\infty$, $F(u) \in H_0^2$ but $\notin H_0^3$ (see (2.1)), then on the spaces H_0^s, $s \geqslant 3$, we have ord $F > 0$.

If $f(u, x)$ is odd in u and even in x (e.g., is x-independent), or vice versa, then $F(H_0^s) \subset H_0^s$ for $s \geqslant 1$, so ord $F = 0$ for any $s \geqslant 1$.

Given a C^k-smooth function $H : X_d \supset O_d \to \mathbb{R}$, $k \geqslant 1$, we consider its *gradient map* with respect to the paring $\langle \cdot, \cdot \rangle$:

$$\nabla H : O_d \to X_{-d}, \quad \big\langle \nabla H(u), v \big\rangle = dH(u)v \ \forall v \in X_d.$$

The map ∇H is C^{k-1}-smooth.

If O_d belongs to a system of compatible domains O_s, $a \leqslant s \leqslant b$, and the gradient map ∇H defines a C^{k-1}-smooth morphism of order d_H for $a \leqslant s \leqslant b$, we write that ord $\nabla H = d_H$.

2.2. *Symplectic structures*

For simplicity we restrict ourselves to constant-coefficient symplectic structures. For the general case see [59].

Let $\{X_s\}$ be a Hilbert scale and J be its anti-selfadjoint automorphism of order d for $-\infty < s < \infty$. Then the operator $\bar{J} = -J^{-1}$ defines an anti-selfadjoint automorphism of order $-d$. We define a two-form α_2 as

$$\alpha_2 = \bar{J}\, dx \wedge dx,$$

where by definition $\bar{J}\, dx \wedge dx\, [\xi, \eta] = \langle \bar{J}\xi, \eta \rangle$. Clearly, $\bar{J}\, dx \wedge dx$ defines a continuous skew-symmetric bilinear form on $X_r \times X_r$ if $r \geqslant -d/2$. Therefore any space X_r, $r \geqslant -d/2$, becomes a *symplectic* (*Hilbert*) *space* and we shall write it as a pair (X_r, α_2).

The pair $(\{X_s\}, \alpha_2)$ is called a *symplectic* (*Hilbert*) *scale*.

EXAMPLE 2.6. Let us take the index-set $\mathcal{Z}$ to be the union of non-intersecting subsets $\mathcal{Z}_+$ and $\mathcal{Z}_-$, provided with an involution $\mathcal{Z} \to \mathcal{Z}$ which will be denoted $j \mapsto -j$, such that $-\mathcal{Z}_\pm = \mathcal{Z}_\mp$. Let us consider a Hilbert scale $\{X_s\}$ with a basis $\{\phi_k, k \in \mathcal{Z}\}$ and a sequence $\{\theta_k, k \in \mathcal{Z}\}$, such that $\theta_{-j} \equiv \theta_j$. Take J to be the linear operator, defined by the relations

$$J\phi_k = \phi_{-k} \quad \forall k \in \mathcal{Z}_+, \qquad J\phi_k = -\phi_{-k} \quad \forall k \in \mathcal{Z}_-.$$

It defines an anti-selfadjoint automorphism of the scale of zero order, and $\bar{J} = J$. The symplectic scale $(\{X_s\}, \alpha_2 = \bar{J}\, dx \wedge dx = J\, dx \wedge dx)$ will be called a *Darboux scale*.

Let $(\{X_s\}, \alpha_2 = \bar{J}\, dx \wedge dx)$ and $(\{Y_s\}, \beta_2 = \overline{\Upsilon}\, dy \wedge dy)$ be two symplectic Hilbert scales and $O_s \subset X_s$, $a \leqslant s \leqslant b$, be a system of compatible domains. A C^1-smooth morphism of order d_1

$$F : O_s \to Y_{s-d_1}, \quad a \leqslant s \leqslant b,$$

is *symplectic* if $F^*\beta_2 = \alpha_2$. That is, if $\langle \overline{\Upsilon} F_*(x)\xi, F_*(x)\eta \rangle_Y \equiv \langle \bar{J}\xi, \eta \rangle_X$, or

$$F^*(x)\overline{\Upsilon} F_*(x) = \bar{J} \quad \forall x.$$

A symplectic morphism F as above is called a *symplectomorphism* if it is a diffeomorphism.

2.3. *Hamiltonian equations*

To a C^1-smooth function h on a domain $O_d \subset X_d$, the symplectic form α_2 as above corresponds the *Hamiltonian vector field* V_h, defined by the usual relation (cf. [2,43]):

$$\alpha_2\big[V_h(x), \xi\big] = -dh(x)\xi \quad \forall \xi.$$

That is, $\langle \bar{J} V_h(x), \xi \rangle \equiv -\langle \nabla h(x), \xi \rangle$ and

$$V_h(x) = J\nabla h(x).$$

The vector field V_h defines a continuous map $O_d \to X_{-d-d_J}$. Usually we shall assume that V_h is smoother than that and defines a smooth morphism of order $d_1 \leqslant 2d + d_J$ for all s from some segment.

For any C^1-smooth function h on $O_d \times \mathbb{R}$ we denote by V_h the non-autonomous vector field $V_h(x,t) = J\nabla_x h(x,t)$, where ∇_x is the gradient in x, and consider the corresponding *Hamiltonian equation* (or *Hamiltonian system*)

$$\dot{x} = J\nabla_x h(x,t) = V_h(x,t). \tag{2.2}$$

A partial differential equation, supplemented by some boundary conditions, is called a *Hamiltonian partial differential equation*, or an *HPDE*, if under a suitable choice of a

symplectic Hilbert scale $(\{X_s\}, \alpha_2)$, a domain $O_d \subset X_d$ and a Hamiltonian h, it can be written in the form (2.2). In this case the vector field V_h is unbounded, $\operatorname{ord} V_h = d_1 > 0$. That is,

$$V_h : O_d \times \mathbb{R} \to X_{d-d_1}.$$

Usually O_d belongs to a system of compatible domains O_s, $s \geqslant d_0$, and V_h (as a function of x) defines an analytic morphism of order d_1 for $s \geqslant d_0$.

A continuous curve $x : [t_0, t_1] \to O_d$ is called a *solution of* (2.2) *in the space* X_d if it defines a C^1-smooth map $x : [t_0, t_1] \to X_{d-d_1}$ and both parts of (2.2) coincide as curves in X_{d-d_1}. A solution x is called *smooth* if it defines a smooth curve in each space X_s.

If a solution $x(t)$, $t \geqslant t_0$, of (2.2) such that $x(t_0) = x_0$ exists and is unique, we write $x(t_1) = S_{t_0}^{t_1} x_0$, or $x(t_1) = S^{t_1 - t_0} x_0$ if the equation is autonomous (we do not assume that $t_1 \geqslant t_0$). The operators $S_{t_0}^{t_1}$ and S^t are called *flow-maps* of the equation. Clearly, $S_{t_0}^{t_1}$ equals $(S_{t_1}^{t_0})^{-1}$ on a joint domain of definition of the two operators.

A non-linear PDE is called *strongly non-linear* if its non-linear part contains as many derivatives as the linear part. Strongly non-linear Hamiltonian PDEs may possess rather unpleasant properties. In particular, for some of them, every non-zero solution develops a singularity in finite time, see an example in Section 1.4 of [59].

We shall call a non-linear PDE *quasilinear* if its non-linear part contains less derivatives then the linear one. A quasilinear equation can be written in the form (2.2) with

$$h(x, t) = \frac{1}{2}\langle Ax, x\rangle + h_0(x, t), \tag{2.3}$$

where A is a linear operator which defines a selfadjoint morphism of the scale (so $\nabla h(x, t) = Ax + \nabla h_0(x, t)$) and $\operatorname{ord} \nabla h_0 < \operatorname{ord} A$.

The class of Hamiltonian PDEs contains many important equations of mathematical physics, some of them are discussed below. The first difficulty one comes across when studies this class is absence of a general theorem which would guarantee that (locally in time) an equation has a unique solution.[2] Such a theorem exists for semilinear equations, where Equation (2.2) will be called *semilinear* if its Hamiltonian has the form (2.3) and $\operatorname{ord} J\nabla h_0 \leqslant 0$ (see [69] and Section 1.4 of [59]).

EXAMPLE 2.7 (*Equations of the Korteweg–de Vries type*). Let us take for $\{X_s\}$ the scale of zero mean-value Sobolev spaces $H^s(S^1)_0$ as in Example 2.1 and choose $J = \partial/\partial x$, so $d_J = 1$. For a Hamiltonian h we take $h(u) = \int_0^{2\pi} (-\frac{1}{8}u'(x)^2 + f(u))\,dx$ with some analytic function $f(u)$. Then $\nabla h(u) = \frac{1}{4}u'' + f'(u)$ and the equation takes the form

$$\dot{u}(t, x) = \frac{1}{4}u''' + \frac{\partial}{\partial x} f'(u).$$

For $f(u) = \frac{1}{4}u^3$ we get the classical Korteweg–de Vries (KdV) equation. The map V_h defines an analytic morphism of order 3 of the scale $\{X_s\}$, for $s > 1/2$. The equation

[2] Still, see [47] for a theory which applies to some classes of quasilinear HPDEs.

has the form (2.2), (2.3), where $\operatorname{ord} JA = 3$ and $\operatorname{ord} J\nabla h_0 = 1$. It is quasilinear, but not semilinear.

EXAMPLE 2.8 (*NLS—non-linear Schrödinger equation*). Let $X_s = H^s(\mathbb{T}^n; \mathbb{C})$, where this Sobolev space is treated as a real Hilbert space, and the basic scalar product of the scale is $\langle u, v\rangle = \operatorname{Re}\int u\bar{v}\,dx$. For J we take the operator $Ju(x) = iu(x)$, so $\operatorname{ord} J = 0$ and $(\{X_s\}, \bar{J}\,du \wedge du)$ is a Darboux scale. We choose

$$h(u) = \frac{1}{2}\int_{\mathbb{T}^n} \big(|\nabla u|^2 + V(x)|u|^2 + g(x, u, \bar{u})\big)\,dx,$$

where V is a smooth real function and $g(x, u, v)$ is a smooth function, real if $v = \bar{u}$. Then $\nabla h(u) = -\Delta u + V(x)u + \frac{\partial}{\partial \bar{u}} g$ and (2.2) takes the form

$$\dot{u} = i\left(-\Delta u + V(x)u + \frac{\partial}{\partial \bar{u}}\, g(x, u, \bar{u})\right), \quad u = u(t, x),\ x \in \mathbb{T}^n. \tag{2.4}$$

This is a semilinear Hamiltonian equation in any space X_{d_0}, $d_0 > n/2$, with $\operatorname{ord} A = 2$ and $\operatorname{ord} \nabla h_0 = 0$.

Non-linear Schrödinger equation (2.4) with $n = 1$, $V(x) = \text{const}$ and $g = \gamma|u|^4$, $\gamma \neq 0$, is called the *Zakharov–Shabat equation*. The equation with $\gamma > 0$ is called *defocusing* and with $\gamma < 0$—*focusing*.

EXAMPLE 2.9 (*1D NLS with Dirichlet boundary conditions*). Let us choose for X_s the space $H_0^s(0, \pi; \mathbb{C})$ (see Example 2.2), $Ju(x) = iu(x)$ and

$$h(u) = \frac{1}{2}\int_0^\pi \big(|u_x|^2 + V(x)|u|^2 + g\big(x, |u|^2\big)\big)\,dx,$$

where g is smooth and 2π-periodic in x. Now $\nabla h(u) = -u_{xx} + V(x)u + f(x, |u|^2)u$, where $f = \frac{\partial g}{\partial |u|^2}$, and (2.2) becomes

$$\dot{u} = i\big(-u_{xx} + V(x)u + f\big(x, |u|^2\big)u\big), \qquad u(0) = u(\pi) = 0. \tag{2.5}$$

For $s = 1$ and 2 the non-linear term defines a smooth map $X_s \to X_s$ (see Example 2.5), so in these spaces this is a semilinear equation with $\operatorname{ord} A = 2$ and $\operatorname{ord} \nabla h_0 = 0$. If in addition f is even in x, then the non-linear term defines a smooth map for every $s \geqslant 1$. This map is analytic if f is.

EXAMPLE 2.10 (*Non-linear wave equations*). Now let $X_s = H^s(\mathbb{T}^n) \times H^s(\mathbb{T}^n)$ and $\alpha_2 = \bar{J}\,d\eta \wedge d\eta$, where $\eta = (u, v)$ and $J(u, v) = \bar{J}(u, v) = (-v, u)$. Let

$$h(u, v) = \int_{\mathbb{T}^n} \left(\frac{1}{2}v^2 + \frac{1}{2}|\nabla u|^2 - f(x, u)\right) dx. \tag{2.6}$$

The corresponding Hamiltonian equation is

$$\dot{u} = -v, \qquad \dot{v} = -\Delta u - f'_u(x, u). \tag{2.7}$$

Or

$$\ddot{u} = \Delta u + f'_u(x, u), \quad u = u(t, x),\ x \in \mathbb{T}^n. \tag{2.8}$$

This is a *non-linear wave equation* (with the periodic boundary conditions). We have seen that this equation can be re-written as the system (2.7) which is an HPDE. This Hamiltonian form of the equation is inconvenient since the quadratic part of the Hamiltonian (2.6) corresponds to the linear operator $(u, v) \to \frac{1}{2}(-\Delta u, v)$ which does not define an isomorphism of the scale $\{X_s\}$ (of some order m). It turns out that the non-linear wave equation (2.8) admits another Hamiltonian representation (2.2), where the Hamiltonian h has the form (2.3), the operator A defines an isomorphism of the scale and $\operatorname{ord} A < \operatorname{ord} \nabla h_0$ (so the equation is quasilinear). We note that the corresponding linear operator JA *is not* differential. See [52] and [59], also see below Section 4.3, where the non-linear wave equation $\ddot{u} = u_{xx} - \sin u$ (the Sine-Gordon equation) is considered in details.

3. Basic theorems on Hamiltonian systems

Basic theorems from the classical Hamiltonian formalism (see [2,43]) remain true for Hamiltonian equations (2.2) in Hilbert scales, provided that the theorems are properly formulated. In this section we present three corresponding results. Their proofs can be found in [52,59].

Let $(\{X_s\}, \alpha_2 = \bar{J}\, dx \wedge dx)$ and $(\{Y_s\}, \beta_2 = \overline{\Upsilon}\, dy \wedge dy)$ be two symplectic scales and (for simplicity) $\operatorname{ord} J = \operatorname{ord} \Upsilon = d_J \geqslant 0$. Let $\Phi : Q \to O$ be a C^1-smooth symplectic map, where Q and O are domains in Y_d and X_d, $d \geqslant 0$. If $d_J > 0$, we have to assume that

(H1) for any $|s| \leqslant d$ linearised maps $\Phi_*(y)$, $y \in Q$, define linear maps $Y_s \to X_s$ which continuously depend on y.

The first theorem states that symplectic maps transform Hamiltonian equations to Hamiltonian:

THEOREM 3.1. *Let $\Phi : Q \to O$ be a symplectic map as above* (so (H1) *holds if $d_J > 0$*). *Let us assume that the vector field V_h of Equation* (2.2) *defines a C^1-smooth map $V_h : O \times \mathbb{R} \to X_{d-d_1}$ of order $d_1 \leqslant 2d$ and that this vector field is tangent to the map Φ* (*i.e., for every $y \in Q$ and every t the vector $V_h(\Phi(y), t)$ belong to the range of the linearised map $\Phi_*(y)$*). *Then Φ transforms solutions of the Hamiltonian equation $\dot{y} = \Upsilon \nabla_y H(y, t)$, where $H = h \circ \Phi$, to solutions of* (2.2).

COROLLARY 3.2. *If under the assumptions of Theorem* 3.1 $\{X_s\} = \{Y_s\}$ *and $h \circ \Phi = h$, $\Phi^* \alpha_2 = \alpha_2$, then Φ preserves the class of solutions for* (2.2).

For Hamiltonian PDEs (and for Hamiltonian equations (2.2)) Theorem 2.1 plays the same role as its classical finite-dimensional counterpart plays for usual Hamiltonian equations: it is used to transform an equation to a normal form, usually in the vicinity of an invariant set (e.g., of an equilibrium).

To apply Theorem 3.1 one needs regular ways to construct symplectic transformations. For classical finite-dimensional systems symplectic transformations usually are obtained either via generating functions, or as Lie transformations (i.e., as flow-maps of additional Hamiltonians), see [2,43,40]. For infinite-dimensional symplectic spaces generating functions play negligible role, while the Lie transformations remain an important tool. An easy but important corresponding result is stated in the theorem below.

Let $(\{X_s\}, \alpha_2)$ be a symplectic Hilbert scale as above and O be a domain in X_d.

THEOREM 3.3. *Let f be a C^1-smooth function on $O \times \mathbb{R}$ such that the map $V_f : O \times \mathbb{R} \to X_d$ is Lipschitz in (x, t) and C^1-smooth in x. Let O_1 be a subdomain of O. Then the flow-maps $S_t^\tau : (O_1, \alpha_2) \to (O, \alpha_2)$ are symplectomorphisms* (*provided that they map O_1 to O*). *If the map V_f is C^k-smooth or analytic, then the flow-maps are C^k-smooth or analytic as well.*

The assumption that the map V_f is Lipschitz can be replaced by the much weaker assumption that for a solution $x(t)$ of the equation $\dot{x} = V_f(x)$, the linearised equation $\dot{\xi} = V_{f*}(x(t))\xi$ is such that its flow maps are bounded linear transformations of the space X_d. See [59].

Usually Theorem 3.3 is applied in the situation when $|f| \ll 1$, or $|t - \tau| \ll 1$. In these cases the flow-maps are close to the identity and the corresponding transformations of the space of C^1-smooth functions on O, $H \mapsto H \circ S_t^\tau$, can be written as Lie series (cf. [40]). In particular, the following simple result holds:

THEOREM 3.4. *Under the assumptions of Theorem* 3.3, *let H be a C^1-smooth function on O. Then*

$$\frac{d}{d\tau} H\big(S_t^\tau(x)\big) = \{f, H\}\big(S_t^\tau(x)\big), \quad x \in O_1. \tag{3.1}$$

In this theorem $\{f, H\}$ denotes the *Poisson bracket* of the two functions:

$$\{f, H\}(x) = \big\langle J\nabla f(x), \nabla H(x)\big\rangle.$$

It is well defined since $J\nabla f = V_f \in X_d$ by assumptions.

Theorem 3.3 and formula (3.1) make from symplectic flow-maps S_t^τ a tool which is well suited to prove KAM-theorems for Hamiltonian PDEs, see the scheme of the proof of Theorem 5.1 in Section 5.1 below.

An immediate consequence of Theorem 3.4 is that for an autonomous Hamiltonian equation $\dot{x} = J\nabla f(x)$ such that $\operatorname{ord} J\nabla f = 0$, a C^1-smooth function H is an integral of motion[3] if and only if $\{f, H\} \equiv 0$.

[3] That is, $H(x(t))$ is a time-independent quantity for any solution $x(t)$.

If $d' = \operatorname{ord} J\nabla f > 0$ and $O = O_d$ belongs to a system of compatible domains $O_s \subset X_s$, $s \in [d_0, d]$, where $d_0 = d - d'$, then H such that $\{f, H\} \equiv 0$ is an integral of motion for the equation $\dot{x} = J\nabla f(x)$, provided that

$$\operatorname{ord} J\nabla f = d' \quad \text{and} \quad \operatorname{ord} \nabla H = d_H \quad \text{for } s \in [d_0, d],$$

where $d' + d_H \leqslant 2d$. Indeed, since $d_0 - d_H \geqslant -d_0$, then H is a C^1-smooth function on O_{d_0}. Since any solution $x(t)$ is a C^1-smooth curve in O_{d_0} by the definition of a solution, then

$$\frac{d}{dt} H(x) = \langle \nabla H(x), \dot{x} \rangle = \langle \nabla H(x), J\nabla f(x) \rangle = \{f, H\}(x) = 0.$$

In particular, f is an integral of motion for the equation $\dot{x} = J\nabla f(x)$ in O_d if we have $\operatorname{ord} J = d_J$ and $\operatorname{ord} \nabla f = d_f$ for $s = d$ and for $s \in [d, d - d_f - d_J]$, where $d \geqslant d_f + d_J/2$. That is, if the equation is being considered in sufficiently smooth spaces.

EXAMPLE 3.5. Let us consider a non-linear Schrödinger equation (2.5) such that $g(u, \bar{u}) = g_0(|u|^2)$, and take $H(u) = \|u\|_0^2 = |u|_{L_2}^2$. Now $d' := \operatorname{ord} J\nabla f = 2$ for $s \in (n/2, \infty)$, and $\operatorname{ord} \nabla H = 0$. Elementary calculations show that $\{f, H\} \equiv 0$. So L_2-norm is an integral of motion for solutions of (2.5) in X_s if $s > n/2 + 2$. (In fact this result remains true for solutions of much lower smoothness, see [15].)

4. Lax-integrable equations

4.1. *General discussion*

Let us take a Hamiltonian PDE and write it as a Hamiltonian equation in a suitable symplectic Hilbert scale $(\{X_s\}, \alpha_2 = \bar{J}\, du \wedge du)$:

$$\dot{u} = J\nabla H(u). \tag{4.1}$$

This equation is called Lax-integrable if there exists an additional Hilbert scale $\{Z_s\}$ (real or complex), and finite order linear morphisms $\mathcal{L}_u$ and $\mathcal{A}_u$ of this scale which depend on the parameter $u \in X_\infty$, such that a curve $u(t)$ is a smooth solution for (4.1) if and only if

$$\frac{d}{dt} \mathcal{L}_{u(t)} = [\mathcal{A}_{u(t)}, \mathcal{L}_{u(t)}]. \tag{4.2}$$

The operators $\mathcal{A}_u$ and $\mathcal{L}_u$, treated as morphisms of the scale $\{Z_s\}$, are assumed to depend smoothly on $u \in X_d$ where d is sufficiently large, so the left-hand side of (4.2) is well defined (for details see [59]). The pair of operators $\mathcal{L}$, $\mathcal{A}$ is called the *Lax pair*.[4]

[4]Due to a deep result by Krichever and Phong [48], any Lax-integrable PDE is a Hamiltonian system. The corresponding symplectic structure belongs to a bigger class than that described in Section 2.2, so to apply our techniques we need to assume a priori that the equation has the form (4.1).

In most known examples of Lax-integrable equations the relation between the scales $\{X_s\}$ and $\{Z_s\}$ is the following: spaces X_s are formed by T-periodic Sobolev vector-functions, while $\mathcal{A}$ and $\mathcal{L}$ are differential or integro-differential operators with u-dependent coefficients, acting in a scale $\{Z_s\}$ of TL-periodic Sobolev vector-functions. Here L is some fixed integer.

Let $u(t)$ be a smooth solution for (4.1). We set $\mathcal{L}_t = \mathcal{L}_{u(t)}$ and $\mathcal{A}_t = \mathcal{A}_{u(t)}$.

LEMMA 4.1. *Let $\chi_0 \in Z_\infty$ be a smooth eigenvector of $\mathcal{L}_0$, i.e., $\mathcal{L}_0\chi_0 = \lambda\chi_0$. Let us assume that the initial-value problem*

$$\dot{\chi} = \mathcal{A}_t\chi, \qquad \chi(0) = \chi_0, \tag{4.3}$$

has a unique smooth solution $\chi(t)$. Then

$$\mathcal{L}_t\chi(t) = \lambda\chi(t) \quad \forall t. \tag{4.4}$$

PROOF. Let us denote the left-hand side of (4.4) by $\xi(t)$, the right-hand side—by $\eta(t)$ and calculate their derivatives. We have:

$$\frac{d}{dt}\xi = \frac{d}{dt}\mathcal{L}\chi = [\mathcal{A}, \mathcal{L}]\chi + \mathcal{L}\mathcal{A}\chi = \mathcal{A}\mathcal{L}\chi = \mathcal{A}\xi$$

and

$$\frac{d}{dt}\eta = \frac{d}{dt}\lambda\chi = \lambda\mathcal{A}\chi = A\eta.$$

Thus, both $\xi(t)$ and $\eta(t)$ solve the problem (4.3) with χ_0 replaced by $\lambda\chi_0$ and coincide by the uniqueness assumption. □

Due to this lemma the discrete spectrum of the operator $\mathcal{L}_u$ is an integral of motion for Equation (4.1). In particular, a set $\mathcal{T}$ formed by all smooth vectors $u \in X_\infty$ such that the eigenvalues of the operator $\mathcal{L}_u$ belong to a fixed subset of $\mathbb{C} \times \mathbb{C} \times \cdots$, is invariant for the flow of Equation (4.1). A remarkable discovery, made by Novikov [68] and Lax [61], is that for integrable Hamiltonian PDEs, considered on finite space-intervals with suitable boundary conditions, some sets $\mathcal{T}$ as above are finite-dimensional symplectic submanifolds $\mathcal{T}^{2n}$ of all symplectic spaces (X_s, α_2), and restriction of Equation (4.1) to any $\mathcal{T}^{2n}$ is an integrable Hamiltonian system. Moreover, for some integrable equations it is known that the union of all these manifolds $\mathcal{T}^{2n}$ is dense in every space X_s. Solutions that fill a manifold $\mathcal{T}^{2n}$ are called *finite-gap solutions*, and the manifold itself—a *finite-gap manifold*. See, e.g., [32,83,8,59].

4.2. *Korteweg–de Vries equation*

The KdV equation

$$\dot{u} = \frac{1}{4}\frac{\partial}{\partial x}\big(u_{xx} + 3u^2\big), \quad u(t,x) \equiv u(t,x+2\pi), \ \int_0^{2\pi} u\,dx \equiv 0, \tag{4.5}$$

takes the form (4.1) in the symplectic Hilbert scale $(\{X_s\}, \alpha_2 = \bar{J}\,du \wedge du)$, where X_s is the Sobolev space $H^s(S^1)_0$ and $Ju = (\partial/\partial x)u$, see Example 2.7. Due to Lax himself, this equation is Lax-integrable and the corresponding Lax pair is

$$\mathcal{L}_u = -\frac{\partial^2}{\partial x^2} - u, \qquad \mathcal{A}_u = \frac{\partial^3}{\partial x^3} + \frac{3}{2}u\frac{\partial}{\partial x} + \frac{3}{4}u_x.$$

Taking for $\{Z_s\}$ the Sobolev scale of 4π-periodic functions and applying Lemma 4.1 we obtain that smooth 4π-periodic spectrum of the operator $\mathcal{L}_u$ is an integral of motion. It is well known that the spectrum of $\mathcal{L}_u$ is formed by eigenvalues

$$\lambda_0 < \lambda_1 \leqslant \lambda_2 < \lambda_3 \leqslant \lambda_4 < \cdots \nearrow \infty,$$

and that the corresponding eigenfunctions are smooth, provided that the potential u is. Let us take any integer n-vector $\mathbf{V}$,

$$\mathbf{V} = (V_1, \ldots, V_n) \in \mathbb{N}^n, \quad V_1 < \cdots < V_n.$$

Denoting $\Delta_j = \lambda_{2j} - \lambda_{2j-1} \geqslant 0$, $j = 1, 2, \ldots$, we define the set $\mathcal{T}^{2n}_{\mathbf{V}}$ as

$$\mathcal{T}^{2n}_{\mathbf{V}} = \big\{u(x) \mid \Delta_j \neq 0 \text{ iff } j \in \{V_1, \ldots, V_n\}\big\}.$$

Clearly $\mathcal{T}^{2n}_{\mathbf{V}}$ equals to the union $\mathcal{T}^{2n}_{\mathbf{V}} = \bigcup_{r\in\mathbb{R}^n_+} T^n_{\mathbf{V}}(r)$, where $\mathbb{R}^n_+ = \{r \mid r_j > 0\ \forall j\}$ and

$$T^n_{\mathbf{V}}(r) = \big\{u(x) \in \mathcal{T}^{2n}_{\mathbf{V}} \mid \Delta_j = r_j\ \forall j\big\}.$$

Since the 4π-periodic spectrum $\{\lambda_j\}$ is an integral of motion for (KdV), then the sets $T^n_{\mathbf{V}}(r)$ are invariant for the KdV-flow. Due to the classical theory of the Sturm–Liouville operator $\mathcal{L}_u$, the set $\mathcal{T}^{2n}_{\mathbf{V}}$ is a smooth submanifold of any space X_s, foliated to the smooth n-tori $T^n_{\mathbf{V}}(r)$. For all these results see, e.g., [46].

Due to Novikov and Lax, there exist an analytic map $\Phi = \Phi_{\mathbf{V}} : \{(r,\xi)\} = \mathbb{R}^n_+ \times \mathbb{T}^n \to X_s$ (s is any integer), and an analytic function $h = h^n(r)$ such that $T^n_{\mathbf{V}}(r) = \Phi(\{r\} \times \mathbb{T}^n)$, and for any $\xi_0 \in \mathbb{T}^n$ the curve $u(t) = \Phi(r, \xi_0 + t\nabla h(r))$ is a smooth solution for (4.5). As a function of t, this solution is quasiperiodic.[5] The celebrated Its–Matveev formula explicitly represents Φ in terms of theta-functions, see in [32,31,8,59].

[5] A continuous curve $u : \mathbb{R} \to X$ is quasiperiodic if there exist $n \in \mathbb{N}$, $\phi \in \mathbb{T}^n$, $\omega \in \mathbb{R}^n$ and a continuous map $U : \mathbb{T}^n \to X$ such that $u(t) = U(\phi + t\omega)$.

4.3. *Other examples*

Sine-Gordon. The Sine-Gordon (SG) equation on the circle

$$\ddot{u} = u_{xx}(t,x) - \sin u(t,x), \quad x \in S^1 = \mathbb{R}/2\pi\mathbb{Z},$$

is another example of a Lax-integrable HPDE.

First the equation has to be written in a Hamiltonian form. The most straightforward way to do this is to write (SG) as the system

$$\dot{u} = -v, \qquad \dot{v} = -u_{xx} + \sin u(t,x).$$

One immediately sees that this system is a semilinear Hamiltonian equation in the symplectic scale $(\{X_s = H^s(S) \times H^s(S)\}, \alpha_2 = \bar{J}\, d\eta \wedge d\eta)$, where $\eta = (u, v)$ and $J(u,v) = (-v, u)$.

Now we derive another Hamiltonian form of (SG), more convenient for its analysis. To do this we consider the shifted Sobolev scale $\{X_s = H^{s+1}(S^1) \times H^{s+1}(S^1)\}$, where the space X_0 is given the scalar product

$$\langle \xi_1, \xi_2 \rangle = \int_{S^1} \left(\xi_{1x}' \cdot \xi_{2x}' + \xi_1 \cdot \xi_2\right) dx,$$

and any space X_s—the product $\langle \xi_1, \xi_2 \rangle_s = \langle A^s \xi_1, \xi_2 \rangle$. Here A is the operator $A = -\partial^2/\partial x^2 + 1$. Obviously, A defines a selfadjoint automorphism of the scale of order one. The operator $J(u,w) = (-\sqrt{A}\, w, \sqrt{A}\, u)$ defines an anti-selfadjoint automorphism of the same order. We provide the scale with the symplectic form $\beta_2 = \bar{J}\, d\xi \wedge d\xi$. We note that (SG) can be written as the system

$$\dot{u} = -\sqrt{A}\, w, \qquad \dot{w} = \sqrt{A}\big(u + A^{-1} f'\big(u(x)\big)\big), \tag{4.6}$$

where $f(u) = -\cos u - \frac{1}{2}u^2$, and that (4.6) is a semilinear Hamiltonian equation in the symplectic scale as above with the Hamiltonian $H(\xi) = \frac{1}{2}\langle \xi, \xi \rangle + \int f(u(x))\, dx$, $\xi = (u, w)$.

Let us denote by X_s^o (X_s^e) subspaces of X_s formed by odd (even) vector functions $\xi(x)$. Then $(\{X_s^o\}, \beta_2)$ and $(\{X_s^e\}, \beta_2)$ are symplectic sub-scales of the scale above. The space X_s^o and X_s^e (with $s \geqslant 0$) are invariant for the flow of Equation (4.6). The restricted flows correspond to the SG equation under the odd periodic and even periodic boundary conditions, respectively.

The SG equation is Lax-integrable under periodic, odd periodic and even periodic boundary conditions. That is, Equation (4.6) is Lax-integrable in the all three symplectic scales defined above. See [8,59].

Zakharov–Shabat equation. Let us take the symplectic Hilbert scale ($X_s = H^s(S^1, \mathbb{C})$, $\bar{J}\,du \wedge du$) as in the Example 2.8. The defocusing and focusing Zakharov–Shabat equations

$$\dot{u} = i\big(-u_{xx} + mu \pm \gamma|u|^2 u\big), \quad \gamma > 0, \tag{4.7}$$

both are Lax-integrable, see [83,8].

5. KAM for PDEs

In this section we discuss the 'KAM for PDEs' theory. Here we cover all relevant topics, except the theory of time-periodic solutions of Hamiltonian PDEs. The latter is reviewed in the Appendix, written by Dario Bambusi. We avoid completely the classical finite-dimensional KAM-theory which deals with time-quasiperiodic solutions of finite-dimensional Hamiltonian systems and instead refer the reader to the recent survey [78].

5.1. *An abstract KAM-theorem*

Let $(\{X_s\}, \alpha_2 = \bar{J}\,du \wedge du)$ be a symplectic Hilbert scale, $-d_J = \operatorname{ord}\bar{J} \leqslant 0$; A be an operator which defines a selfadjoint automorphism of the scale of order $d_A \geqslant -d_J$ and H be a Fréchet–analytic functional on X_{d_0}, $d_0 \geqslant 0$, such that $\operatorname{ord}\nabla H = d_H < d_A$:

$$\nabla H : X_{d_0} \to X_{d_0 - d_H}.$$

We assume that $d_A \leqslant 2d_0$, so the quadratic form $\frac{1}{2}\langle Au, u\rangle$ is well defined on the space X_{d_0}.

In this section we consider the quasilinear Hamiltonian equation with the Hamiltonian $H_\varepsilon(u) = \frac{1}{2}\langle Au, u\rangle + \varepsilon H(u)$:

$$\dot{u}(t) = J\big(Au(t) + \varepsilon\nabla Hu(t)\big). \tag{5.1}$$

We assume that the scale $\{X_s\}$ admits a basis $\{\varphi_k, k \in \mathbb{Z}_0 = \mathbb{Z}\backslash\{0\}\}$ such that

$$A\varphi_j^\pm = \lambda_j^A \varphi_j^\pm, \qquad J\varphi_j^\pm = \mp\lambda_j^J \varphi_j^\pm \quad \forall j \geqslant 1, \tag{5.2}$$

with some real numbers λ_j^J, λ_j^A. In particular, the spectrum of the operator JA is $\{\pm i\lambda_j \mid \lambda_j = \lambda_j^J\lambda_j^A\}$. The numbers λ_j are called the *frequencies* of the linear system

$$\dot{u} = JAu. \tag{5.3}$$

Let us fix any $n \geqslant 1$. Then the $2n$-dimensional linear space

$$\operatorname{span}\big\{\varphi_j^\pm \mid 1 \leqslant j \leqslant n\big\} \tag{5.4}$$

is invariant for Equation (5.3) and is foliated to the invariant tori

$$T^n = T^n(I) = \left\{ \sum_{j=1}^{n} u_j^{\pm} \varphi_j^{\pm} \mid {u_j}^{+2} + {u_j}^{-2} = 2I_j \ \forall j \right\}. \tag{5.5}$$

If $I \in \mathbb{R}^n_+$, then $T^n(I)$ is an n-torus. Providing it with the coordinates $q = (q_1, \dots, q_n)$, where $q_j = \operatorname{Arg}(u_j^+ + iu_j^-)$, we see that Equation (5.3) defines on $T^n(I)$ the motion

$$\dot{q} = (\lambda_1, \dots, \lambda_n) =: \omega. \tag{5.6}$$

So all solutions for the linear equation in $T^n(I)$ are quasiperiodic curves with the frequency-vector ω. Our goal in this section is to present and discuss a KAM-theorem which implies that under certain conditions 'most of' trajectories of Equation (5.6) on the torus $T^n(I)$ persist as time-quasiperiodic solutions of the perturbed equation (5.1), if $\varepsilon > 0$ is sufficiently small.

To state the result we assume that the operator A and the function H analytically depend on an additional n-dimensional parameter $a \in \mathcal{A}$, where $\mathcal{A}$ is a connected bounded open domain in $\mathbb{R}^n$. Then $\lambda_j = \lambda_j(a)$. We assume that the first n frequencies $\lambda_l = \omega_l$ depend on a in the non-degenerate way:

(H1) $\det\{\partial\omega_l/\partial a_k \mid 1 \leqslant k, l \leqslant n\} \not\equiv 0$;

and that the following spectral asymptotic holds:

(H2) $|\lambda_j(a) - K_1 j^{d_1} - K_1^1 j^{d_1^1} - K_1^2 j^{d_1^2} - \cdots| \leqslant K j^{\tilde{d}}$, $\operatorname{Lip} \lambda_j \leqslant j^{\tilde{d}}$,

where $d_1 := d_A + d_J \geqslant 1$, $K_1 > 0$, $\tilde{d} < d_1 - 1$ and the dots stand for a finite sum with exponents $d_1 > d_1^1 > d_1^2 > \cdots$.

Let us denote by X_s^c the complexification of a space X_s and assume that Equation (5.1) is quasilinear and analytic:

(H3) the set $X_{d_0} \times \mathcal{A}$ admits in $X^c_{d_0} \times \mathbb{C}^n$ a complex neighbourhood Q such that the map $\nabla_x H : Q \to X^c_{d_0 - d_H}$ is complex-analytic and bounded uniformly on bounded subsets of Q. Moreover, $d_H + d_J \leqslant \tilde{d}$.

Finally, we shall need the following non-resonance condition:

(H4) For all integer n-vectors s and $(M_2 - n)$-vectors l such that $|s| \leqslant M_1$, $1 \leqslant |l| \leqslant 2$ we have,

$$s \cdot \omega(a) + l_{n+1}\lambda_{n+1}(a) + \cdots + l_{M_2}\lambda_{M_2}(a) \not\equiv 0, \tag{5.7}$$

where the integers $M_1 > 0$ and $M_2 > n$ are to be specified.

Relations (5.7) with $|l| = 1$ and $|l| = 2$ are called, respectively, the first and the second *Melnikov condition*.

Let us fix any $I_0 \in \mathbb{R}^n_+$ and denote by Σ_0 the map $\mathbb{T}^n \times \mathcal{A} \to X_{d_0}$ which sends (q, a) to the point of the torus $T^n(I_0)$ with the coordinate q.

THEOREM 5.1. *Suppose the assumptions* (H1)–(H3) *hold. Then there exist integers* $M_1 > 0$ *and* $M_2 > n$ *such that if* (H4) *is fulfilled, then for arbitrary* $\gamma > 0$ *and for sufficiently small* $\varepsilon < \bar{\varepsilon}(\gamma)$, *a Borel subset* $\mathcal{A}_\varepsilon \subset \mathcal{A}$ *and a Lipschitz map* $\Sigma_\varepsilon : \mathbb{T}^n \times \mathcal{A}_\varepsilon \to X_{d_0}$, *analytic in* $q \in \mathbb{T}^n$, *can be found with the following properties*:

(a) $\mathrm{mes}(\mathcal{A}\backslash\mathcal{A}_\varepsilon) \leqslant \gamma$;
(b) *the map* Σ_ε *is* $C\varepsilon$*-close to* $\Sigma_0|_{\mathbb{T}^n\times\mathcal{A}_\varepsilon}$ *in the Lipschitz norm*;
(c) *each torus* $\Sigma_\varepsilon(\mathbb{T}^n \times \{a\})$, $a \in \mathcal{A}_\varepsilon$, *is invariant for the flow of Equation* (5.1) *and is filled with its time-quasiperiodic solutions of the form* $u_\varepsilon(t; q) = \Sigma_\varepsilon(q + \omega' t, a)$, $q \in \mathbb{T}^n$, *where the frequency vector* $\omega'(a)$ *is* $C\varepsilon$*-close to* $\omega(a)$ *in the Lipschitz norm*;
(d) *the solutions* u_ε *are linearly stable*.[6]

If ∇H defines an analytic map of order d_H on every space X_d, $d \geqslant d_0$, then the solutions u_ε, constructed in the theorem, are smooth. Indeed, if $u_\varepsilon(t)$ is a solution, then due to the equation $JAu_\varepsilon(t)$ is a smooth curve in $X_{d_0-d_H-d_J}$. Since JA is an automorphism of the scale of order d_1, then $u_\varepsilon(t)$ is a smooth curve in $X_{d_0-d_H-d_J+d_1} \subset X_{d_0+1}$. Iterating this arguments we see that u_ε is a smooth curve in each space X_s.

In the semilinear case (i.e., when $d_H + d_J \leqslant \tilde{d} < d_1 - 1$ and $\tilde{d} \leqslant 0$) the theorem is proved in [49,50] (see also [52,73]). The semilinearity restriction $\tilde{d} \leqslant 0$ was removed in [57] (see also [59] and [46]). Simultaneously with [49,50] a related KAM-theorem for infinite-dimensional *Hamiltonian systems with short interactions* was proved by Pöschel [71] (following Eliasson's work [33] on lower-dimensional invariant tori for finite-dimensional systems). The systems (5.1), defined by HPDEs, are not short-interacted, but results of [71] apply to some equations from non-equilibrium statistical physics. For systems with short interaction a KAM-theory for infinite-dimensional invariant tori also is available, see [39,72] and references in [72]. We note that [39] was the first work where the KAM theory was applied to infinite-dimensional Hamiltonian systems.

For some specific HPDEs (5.1) the assertions of Theorem 5.1 can be proven for any $n \geqslant 1$ even if the parameter a is only one-dimensional. In particular, this can be done for the non-linear wave equation as in Example 5.3 below, where $V(x) \equiv a$ and the constant a is the one-dimensional parameter. See [16] and [4].

The proof of Theorem 5.1 is rather technical. For its well-written outline in the semilinear case see [28]. Below we present the proof's scheme in the form which suits our further purposes.

THE SCHEME OF THE PROOF OF THEOREM 5.1. We start with the semilinear case and assume for simplicity that $\lambda_j^J \equiv 1$. Then $I = (I_1, \ldots, I_n)$ and $q = (q_1, \ldots, q_n)$ form a symplectic coordinate system in the space (2.3). We set $Y = \mathrm{span}\{\varphi_j^\pm,\ j > n\} \subset X$, and denote by $y_j^\pm$, $j > n$, the coordinates in Y with respect to the basis $\{\varphi_j^\pm\}$. To study the vicinity of a torus $T^n(I_0)$, we make the substitution $I = I_0 + p$. Then $\bar{J}\, du \wedge du = dp \wedge dq + dy^+ \wedge dy^-$, and $T^n(I_0) = \{p = 0,\ y = 0\}$. In the new variables Equation (2.1) takes the form

$$\dot{q} = \nabla_p \mathcal{H}_\varepsilon, \qquad \dot{p} = -\nabla_q \mathcal{H}_\varepsilon, \qquad \dot{y} = J\nabla_y \mathcal{H}_\varepsilon,$$

with the Hamiltonian

$$\mathcal{H}_\varepsilon = H_0(p, y) + \varepsilon H_1(p, q, y), \quad H_0 = \omega \cdot p + \frac{1}{2}\langle Ay, y\rangle. \tag{5.8}$$

[6] If Equation (5.1) is not semilinear (i.e., if $d_J + d_H > 0$), then this assertion is proved provided that the equation satisfies some mild regularity condition, see Theorem 8.4 in [59].

The vector ω and the operator A depend on the parameter a; the function H_1 depends on a and I_0. We call H_0 the *integrable part of the Hamiltonian* $\mathcal{H}_\varepsilon$.

Retaining the terms of H_1 which are affine in p and quadratic in y, we write H_1 as

$$H_1 = H_1^1 + H_1^3, \quad H_1^1 = h(q) + h^p(q)\cdot p + \langle h^y(q), y\rangle + \langle h^{yy}(q)y, y\rangle,$$
$$H_1^3 = \mathrm{O}\big(|p|^2 + \|y\|^3 + |p|\,\|y\|\big) =: \mathcal{O}(p, q, y).$$

Next in the vicinity of the torus $T^n = \{p = 0,\ y = 0\}$ we make a symplectic change of variable to kill the part εH_1^1 of the perturbation εH_1. This change of variable is a transformation S_1 which is the time-ε shift along trajectories of an additional Hamiltonian F. Here the recipe is that to kill H_1^1, F should be of the same structure, so $F = f(q) + f^p(q)\cdot p + \langle f^y(q), y\rangle + \langle f^{yy}(q)y, y\rangle$. Due to Theorem 3.4 we can write the transformed Hamiltonian $\mathcal{H}_\varepsilon \circ S_1$ as

$$\mathcal{H}_\varepsilon \circ S_1 = H_0 + \varepsilon H_1 + \varepsilon\langle J\nabla_y F,\ \nabla_y H_0\rangle + \varepsilon\nabla_p F\cdot\nabla_q H_0 - \varepsilon\nabla_q F\cdot\nabla_p H_0$$
$$+ \mathrm{O}\big(\varepsilon^2\big) + \mathcal{O}.$$

Since $\nabla_p H_0 = \omega$, $\nabla_q H_0 = 0$ and $\nabla_y H_0 = Ay$, then the linear in ε term vanishes if the following relations hold:

$$(\omega\cdot\nabla) f = h, \qquad (\omega\cdot\nabla) f^p = h^p,$$
$$(\omega\cdot\nabla) f^y - JAf^y = h^y, \qquad (\omega\cdot\nabla) f^{yy} + [f^{yy},\ JA] = h^{yy}.$$

We take these relations as equations for f, f^p, f^y and f^{yy} (called *'the homological equations'*) and try to solve them.

Since the equations have constant coefficients, then decomposing f, f^p, ... in Fourier series in q, we find for their components (and for matrix components of the operator f^{yy}) explicit formulae. Certain terms in these formulae contain small divisors, which vanish for some values of the vector $\omega = \omega(a)$. Careful analysis of these divisors show that all of them are bounded away from zero if $a \notin \mathcal{A}_1$, where $\mathcal{A}_1$ is a Borel subset of $\mathcal{A}$ of small measure. When the equations are solved, we get a symplectic transformation which in a sufficiently small neighbourhood of T^n transforms the Hamiltonian $\mathcal{H}_\varepsilon$ to a Hamiltonian which differs from its integrable part by $\mathrm{O}(\varepsilon^2)$.

The explanation above has some flows. The most important one is that the first and the second homological equations can be solved only if the mean values of h and h^p vanish. To fulfil the first condition we change the Hamiltonian εH_1 by a constant (this change is irrelevant since it does not affect the equations of motion), while to fulfil the second we subtract from εH_1 the average $\varepsilon\langle h^p\rangle\cdot p$ and add it to the integrable part H_0, thus changing the term $\omega\cdot p$ to $\omega^2\cdot p$, where $\omega^2 = \omega + \varepsilon\langle h^p\rangle$. Similar, to solve the last homological equation we subtract from the operator h^{yy} the average of its diagonal part and add the corresponding quadratic form to H_0. Thus, the transformed Hamiltonian becomes

$$\mathcal{H}_2 := \mathcal{H}_\varepsilon \circ S_1 = \omega_2\cdot p + \frac{1}{2}\langle A_2 y, y\rangle + \varepsilon^2 H_2(p, q, y) + \mathcal{O}(p, q, y).$$

This transformation is called *the KAM-step*.

Next we perform the second KAM-step. Under the condition that $a \notin \mathcal{A}_2$ we find a transformation S_2 which sends the Hamiltonian $\mathcal{H}_2$ to $\mathcal{H}_3 = \mathcal{H}_2 \circ S_2 = \omega_3 \cdot p + \frac{1}{2}\langle A_3 y, y\rangle + (\varepsilon^2)^2 H_2 + \mathcal{O}(p, q, y)$, etc. After m steps we find transformations $S_1, \dots, S_m$ such that

$$\mathcal{H}_\varepsilon \circ S_1 \circ \cdots \circ S_m = \omega_m \cdot p + \frac{1}{2}\langle A_m y,\ y\rangle + \varepsilon^{2^m} H_m + \mathcal{O}(p, q, y) =: \mathcal{H}_m.$$

The torus $T^n = \{p = 0, y = 0\}$ is 'almost invariant' for the equation with the Hamiltonian $\mathcal{H}_m$. Hence, the torus $S_1 \circ \cdots \circ S_m(T^n)$ is 'almost invariant' for the original one. Since the sequence ε^{2^m} converges to zero super-exponentially fast, we can choose the sets $\mathcal{A}_1, \mathcal{A}_2, \dots$ in such a way that $\operatorname{mes}(\mathcal{A}_\infty = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots) < \gamma$, for any $a \notin \mathcal{A}_\infty$ the vectors $\omega_m(a)$ converge to a limiting vector $\omega'(a)$, and the transformations $S_1 \circ \cdots \circ S_m$ converge to a limiting map $\Sigma_\varepsilon(\cdot, a)$, defined on T^n. Then the torus $\Sigma_\varepsilon(T^n,\ a)$ is invariant for Equation (5.1) and is filled with its quasiperiodic solutions $t \to \Sigma_\varepsilon(q + \omega' t, a)$. □

If the equation is not semilinear, then the situation is more complicated since to solve the forth homological equation we have to remove from the operator h^{yy} the whole of its diagonal part (not only its average). Because of that the operator A in the integrable part of the Hamiltonian gets terms which form a small q-dependent diagonal operator of a positive order. Accordingly, the forth homological equation becomes more difficult and cannot be solved by the direct Fourier method. Its resolution follows from a non-trivial lemma, based on properties of fast-oscillating Fourier integrals, proved in [57] (see also [59,46]).

5.2. *Applications to 1D HPDEs*

Theorem 1 well applies to parameter-depending quasilinear HPDEs with one-dimensional space variable in a finite interval, supplemented by boundary conditions such that spectrum of the linear operator JA is not multiple. Indeed, for such equations assumption (H2) follows from usual spectral asymptotics, (H3) is obvious if the non-linearity is analytic, while (H1) and (H4) hold if the equation depends on the additional parameter in a non-degenerate way. More explicitly it means the following. In the examples which we consider below, the equations depend on a potential $V(x;\ a)$, which is analytic in a and smooth in x. The non-degeneracy means that in a functional space, formed by functions of x and a of the required smoothness, the potential V should not belong to some analytic subset of infinite codimension.

Below we just list the examples. In each case application of Theorem 5.1 is straightforward. The theorem applies if dimension of the parameter a is $\geqslant n$ and dependence of the potential V on a is non-degenerate as it was explained above. In the first three examples the potential $V(x; a)$ is real, smooth in x and analytic in a. The function $f(x, v; a)$ is real, smooth in x and analytic in v and a. Details can be found in [52,53,59,57].

EXAMPLE 5.2. Non-linear Schrödinger equation (NLS), cf. Example 2.8:

$$\dot{u} = i\big(-u_{xx} + V(x; a)u + \varepsilon f\big(x, |u|^2; a\big)u\big), \quad u = u(t, x), x \in [0, \pi]; \tag{5.9}$$

$$u(t, 0) \equiv u(t,\ \pi) \equiv 0. \tag{5.10}$$

Now $d_J = 0$, $d_A = 2$, $\tilde{d} = d_H = 0$ and we view the Dirichlet boundary conditions as the odd periodic ones (cf. Example 2.9). The theorem applies in the scale of odd periodic functions with $d_0 = 1$ or 2. If f is even and 2π-periodic in x, then the theorem applies with any $d_0 \geqslant 1$ and the constructed quasiperiodic solutions are smooth.

EXAMPLE 5.3. Non-linear string equation: $w(t, x)$ satisfies (5.10) and

$$\ddot{w} = w_{xx} - V(x; a)w + \varepsilon f(x, w; a),$$

where now $V > 0$ and $f(x, w) = 0$ if $w = 0$ or $x = 0$. Let us denote $U = (u, -(-\Delta)^{-1/2}\dot{u})$. It is a matter of direct verification that U satisfies a semilinear Hamiltonian equation (5.1) in a suitable symplectic Hilbert scale, formed by odd periodic Sobolev vector-functions (cf. Equation (4.6)). Now $d_A = 1$, $d_J = 0$, $\tilde{d} = d_H = -1$. Cf. [79] and [16,4].

EXAMPLE 5.4 (*KdV-type equations*). KdV-type equation

$$\dot{u} = \frac{\partial}{\partial x}\big(-u_{xx} + V(x; a)u + \varepsilon f(x, u; a)\big); \quad x \in S^1, \int_{S^1} u\,dx \equiv 0, \tag{5.11}$$

cf. Example 2.7. Now $d_J = 1$, $d_A = 2$, $\tilde{d} = d_H = 0$.

Theorem 5.1 also applies if $x \in \mathbb{R}^1$ and the potential $V(x;\ a)$ grows sufficiently fast when $x \to \infty$.

EXAMPLE 5.5. Non-linear Schrödinger equation on the line:

$$\dot{u} = i\big(-u_{xx} + \big(x^2 + \mu x^4 + V(x; a)\big)u + \varepsilon f\big(|u|^2; a\big)u\big), \quad \mu > 0,$$
$$u = u(t,\ x),\ x \in \mathbb{R},\ u \to 0 \text{ as } |x| \to \infty.$$

Here the potential V is smooth, analytic in a and vanishes as $|x| \to \infty$. The real-valued function f is analytic. Now $d_J = 0$, $d_A = 4/3$, $d_H = 0$. Another example of this sort see in [52], Section 2.5.

The time-quasiperiodic solutions, constructed in Examples 5.2–5.5, are linearly stable. Therefore they should be observable in numerical models for the corresponding equations. Indeed, quasiperiodic behaviour of solutions for 1D HPDEs with small non-linearity was observed in many experiments, starting from the famous numerics of Fermi, Pasta and Ulam [36]; e.g., see [82].

5.3. *Multiple spectrum*

In Examples 5.2, 5.3 the equations are considered under the Dirichlet boundary conditions. If we replace them by the periodic ones

$$u(t, x) \equiv u(t, x + 2\pi),$$

then Theorem 5.1 would not apply since now the frequencies of the corresponding linear equations are asymptotically double: they have the form $\{\lambda_j^{\pm},\ j \geqslant 1\}$, where $|\lambda_j^+ - \lambda_j^-| \to 0$ as $j \to \infty$. It is clear that the numbers $\{\lambda_j^{\pm}\}$ cannot be re-ordered to meet the spectral asymptotic condition (H2). Still, for some semilinear equations (5.1) assertions of the theorem remain true if the frequencies λ_j are not single, but asymptotically they have the same multiplicity $m \geqslant 2$ and behave regularly. A corresponding result is proved by Chierchia and You in [27], using the scheme, explained in Section 5.1. We do not give precise statement of their theorem, but note that it applies to the non-linear string equation in Examples 5.3 under the periodic boundary conditions. The result is the same: if the non-degeneracy condition holds, then for ε small enough and for most (in the sense of measure) values of the n-dimensional parameter a, solutions of the linear equation (5.3) which fill in a torus $T^n(I)$, $I \in \mathbb{R}^n_+$, persist as linearly stable time-quasiperiodic solutions of the corresponding non-linear equation (5.1).

We note that this persistence result was proved earlier by Bourgain [16], who used another KAM-scheme, discussed in the next section.

5.4. *Space-multidimensional problems*

The abstract Theorem 5.1 is a flexible tool to study 1D HPDEs, but it *never* applies to space-multidimensional equations since the spectral assumption (H2) never holds in dimensions > 1. The first KAM-theorem which applies to higher-dimensional HPDEs, is due to Bourgain [19]. In that work the 2D NLS equation as in Example 2.8 is considered. For technical reasons the potential term Vu is replaced there by the convolution $V * u$:

$$\dot{u} = i\left(-\Delta u + V(x;a) * u + \varepsilon \frac{\partial}{\partial \bar{u}} g(u, \bar{u})\right), \quad u = u(t,x),\ x \in \mathbb{T}^2. \tag{5.12}$$

The potential $V(x;a)$ is real analytic and $g(u,\bar{u})$ is a real-valued polynomial of u and $\bar{u}$. This equation has the form (5.1), where $Au = -\Delta u + V * u$ and $Ju = iu$. The basis $\{\varphi_k\}$ as in (5.2) is formed by normalised exponents $\{e^{is\cdot x}$ and $ie^{is\cdot x},\ s \in \mathbb{Z}^2\}$, re-numerated properly, and

$$\lambda_s^J \equiv 1, \qquad \lambda_s^A = |s|^2 + \widehat{V}(s;a),$$

where $\{\widehat{V}(s;a)\}$ are the Fourier coefficients of V. For any n, the linear equation $(5.12)|_{\varepsilon=0}$ has quasiperiodic solutions

$$u = \sum_{j=1}^{n} z_{s_j} e^{i\lambda_{s_j}^A t} \varphi_{s_j}(x) \tag{5.13}$$

(these are trajectories of Equation (5.6) on the n-torus (5.5), where $I_j = \frac{1}{2}|z_{s_j}|^2$ and $I_s = 0$ if s differs from all s_j). For simplicity let us assume that $a_j = \widehat{V}(s_j;\ a),\ j = 1, \ldots, n$. Then

the result of [19] is that for most values of the parameter a (in the same sense as in Theorem 5.1), the solution (5.13) persists as a time-quasiperiodic solution of Equation (5.12). In contrast to the 1D case it is unknown if the new solutions are linearly stable.

The proof in [19] is based on a KAM-scheme, different from that described in Section 5.1. Originally this scheme is due to Craig and Wayne [29] who used it to construct periodic solutions of non-linear wave equations, using certain techniques due to Fröhlich–Spencer [38]. Also see [16].

Now we briefly describe the scheme, using the notations from Section 5.1. When the perturbation εH_1 is decomposed as in (5.8), we extract the term $\varepsilon\langle h^{yy}(q)y,\ y\rangle$ from εH_1^1 and add it to the integrable part H_0. After this the Hamiltonian to be killed is the sum of the three terms $h(q) + h^p(q) + \langle h^y(q), y\rangle$; accordingly the Hamiltonian F is a sum of three terms as well. We have to find them from the first three homological equations. The first two are not difficult, but the third one is a real problem since the operator A no longer has constant coefficients but equals $A_0 + \hat{A}(q)$, where $\hat{A}$ is a bounded operator of order ε (it changes from one KAM-step to another). The resolution of this equation for high KAM steps is the most difficult part of implementation the Craig–Wayne–Bourgain KAM-scheme.

Recently Bourgain managed to develop this scheme father and applied it to high-dimensional equations. We are not ready to discuss this and related results, and instead refer the reader to the original publications [23]. Also see [34].

5.5. *Perturbations of integrable equations*

Let us consider a quasilinear HPDE on a finite space-interval, which is an integrable Hamiltonian equation (4.1) in some symplectic Hilbert scale $(\{X_s\}, \alpha_2 = \bar{J}\,dx \wedge dx)$. As we explained in Section 4.1, this equation has invariant finite-gap symplectic manifolds $\mathcal{T}^{2n}$ such that restriction of (4.1) to any of them is integrable. In this section we discuss the results on persistence of quasiperiodic solutions that fill in these manifolds, provided by the KAM for PDEs theory. We shall see that they are very similar to the celebrated Kolmogorov theorem, which states that *most of quasiperiodic solutions of a non-degenerate analytic integrable* (*finite-dimensional*) *Hamiltonian system persist under small perturbations of the Hamiltonian*; see [1,65,78] and Addendum in [59]. We state the main result as a

THEOREM 5.6 (Metatheorem). *Most of quasiperiodic solutions that fill in any finite-gap manifold $\mathcal{T}^{2n}$ as above persist under small Hamiltonian quasilinear analytic perturbations of the integrable equation. If the finite-gap solutions in $\mathcal{T}^{2n}$ are linearly stable, then the new solutions are linearly stable as well.*

In the assertion above the statement 'most of quasiperiodic solutions persist' means the following. Due to the Liouville–Arnold theorem [2,43], the manifold $\mathcal{T}^{2n}$ can be covered by charts, diffeomorphic to $B \times \mathbb{T}^n = \{p, q\}$ (B is a ball in $\mathbb{R}^n$), with chart-maps $\Phi_0 : B \times \mathbb{T}^n \to \mathcal{T}^{2n}$ such that $\Phi_0^*\alpha_2 = dp \wedge dq$, and the curves $\Phi_0(p, q + t\nabla h(p))$ are solutions of the integrable equation, where $h(p) = H \circ \Phi_0(p, q)$. Let us denote by ε the small coefficient in front of the perturbation. Then for every chart there exists a Borel subset $B_\varepsilon \subset B$ and a map $\Phi_\varepsilon : B_\varepsilon \times \mathbb{T}^n \to X_d$ (d is fixed), with the following properties:

(i) $\mathrm{mes}(B\setminus B_\varepsilon)\to 0$ as $\varepsilon\to 0$;
(ii) the map $\Phi_\varepsilon : B_\varepsilon\times\mathbb{T}^n\to X_d$ is $C\sqrt{\varepsilon}$-close to Φ_0 in the Lipschitz norm and is analytic in $q\in\mathbb{T}^n$;
(iii) there exists a map $\omega_\varepsilon : B_\varepsilon\to\mathbb{R}^n$, $C\varepsilon$-close to the gradient map ∇h in the Lipschitz norm, such that the curves $t\mapsto\Phi_\varepsilon(p, q+t\omega_\varepsilon(p))$, $p\in B_\varepsilon$, $q\in\mathbb{T}^n$, are solutions for the perturbed equation.

The statement of Theorem 5.6 is proven under a number of assumptions (see [59,35]). These assumptions are checked for such basic integrable HPDEs as KdV, Sine- and Sinh-Gordon equations. There are no doubts that they also hold for the Zakharov–Shabat equations[7] (but the theorem in [59,35] does not apply to the Kadomtsev–Petviashvili equation). Below we present a scheme of the proof and discuss the restrictions on the integrable HPDE which allow to implement it.

We view (4.1) as an equation in the Hilbert space X_d, and denote the quasilinear Hamiltonian of the perturbed equation as

$$H_\varepsilon=\frac{1}{2}\langle Ax,\, x\rangle+h_0(x)+\varepsilon h_1(x).$$

Accordingly, $H_0=\frac{1}{2}\langle Ax,\, x\rangle+h_0$ is the Hamiltonian H of the unperturbed equation (4.1).

Step 1. Let us consider any finite-gap solution $u_0(t)=\Phi_0(p_0,\, q_0+t\nabla h(p_0))$ and linearise (4.1) about it:

$$\dot v=J\big(\nabla H\big(u_0(t)\big)\big)_* v. \tag{5.14}$$

The theory of integrable equations provides tools to reduce this equation to constant coefficients by means of a time-quasiperiodic substitution $v(t)=G(p_0, q_0+t\nabla h(p_0))\tilde v(t)$, where $G(p,q)$, $(p,q)\in B\times\mathbb{T}^n$, is a symplectic linear map $G(p,q): Y_d\to Z_d$ (see [59, Sections 5, 6]). Here Y_d is a fixed symplectic subspace of Z_d of codimension $2n$. The restriction, which we impose at this step, is that the operator $G(p,q)$ is a compact perturbation of the embedding $Y_d\to Z_d$, which analytically depends on (p,q).

Step 2. The map G from the Step 1 defines an analytic map

$$B\times\mathbb{T}^n\times Y_d\to X_d,$$

linear and symplectic in $y\in Y_d$. This map defines a symplectomorphism

$$B\times\mathbb{T}^n\times B_\delta(Y_d)\to X_d,\quad B_\delta(Y_d)=\big\{\|y\|_d<\delta\big\}, \tag{5.15}$$

such that linearisation in y at $y=0$ of the latter equals the former ([59, Section 7]).

Step 3. We use the map (5.15) to pass in the Hamiltonian H_ε to the variables (p,q,y). Retaining linear and quadratic in y terms we get

$$H_\varepsilon(p,\, q,\, y)=h(p)+\frac{1}{2}\big\langle\mathcal{A}(p)y,\, y\big\rangle+h_3(p,q,y)+\varepsilon h_1(p,q,y), \tag{5.16}$$

[7] See [41] for an *ad hoc* KAM-theorem for the defocusing equation.

where $h_3 = \mathrm{O}(\|y\|_d^3)$. Calculations show that $h_3(p, q, y)$ contains terms such that their gradient maps have the same order as the operator $\mathcal{A}(p)$. If this really was the case, then the Hamiltonian equation would not be quasilinear, which would complicate its study a lot. Fortunately, this does not happen due to a cancellation of a very general nature (see Lemma 7.5 in [59]), and we have

$$\operatorname{ord} \nabla h_3 < \operatorname{ord} \mathcal{A}(p) - 1. \tag{5.17}$$

Step 4. Invariant tori of the unperturbed system with the Hamiltonian $H_0(p, q, y)$ have the form $\{p = \text{const}, y = 0\}$. Let us scale the variables near a torus $\{p = a, y = 0\}$: $p = a + \varepsilon^{2/3}\tilde{p}$, $q = \tilde{q}$, $y = \varepsilon^{1/3}\tilde{y}$. In the scaled variables the perturbed equation has the Hamiltonian

$$\text{const} + \omega(a) \cdot \tilde{p} + \frac{1}{2}\langle \mathcal{A}(a)\tilde{y}, \tilde{y}\rangle + \mathrm{O}\big(\varepsilon^{1/3}\big), \quad \omega(a) = \nabla h(a). \tag{5.18}$$

So we have got the system (5.1), written in the form (5.8), with ε replaced by $\varepsilon^{1/3}$. If Theorem 5.1 applies, then most of the finite-gap tori $\{p = \text{const}\}$ persist in the perturbed equation, as states the Metatheorem. To be able to use the theorem we have to check the assumptions (H1)–(H4).

The condition (H2) holds if the integrable equation is 1D (if the spectrum is asymptotically double, e.g., if the unperturbed equation is the Sine-Gordon equation under the periodic boundary conditions, then one should use a version of the Metatheorem, based on the Chierchia–You result). The quasilinearity condition (H3) holds due to (5.17). The assumption (H1) now takes the form

$$\operatorname{Hess} h(p) \not\equiv 0. \tag{5.19}$$

This is exactly Kolmogorov's non-degeneracy condition for the integrable system on $\mathcal{T}^{2n}$. The assumption (H4) with $\omega = \nabla h(a)$ is the second non-degeneracy condition, which needs verification.

Summing up what was said above, we see that Theorem 5.1 implies the Metatheorem if the unperturbed integrable equation is 1D quasilinear, the linear operator $G(p, q)$ from Step 1 possesses the required regularity properly and the non-degeneracy assumptions (5.19) and (5.7) hold true.

The scheme we have just explained was suggested in [51], where it was used to prove an abstract KAM-theorem, which next was applied to Birkhoff-integrable infinite-dimensional systems and to perturbed KdV equations. See [59,35] for a more general abstract theorem, based on the same scheme.

Steps 1–2 are not the only way to reduce an integrable equation to the normal form (5.16). Another approach to get it had been initiated by Kappeler [44]. It was developed further in a number of publications and finally in [45] it was proved that the KdV equation is Birkhoff-integrable. It means the following. Let us take the Darboux scale $(\{X_s\}, \alpha_2)$ with the index-set $\mathcal{Z} = \mathbb{Z}_0$, and $\theta_k = |k|$ (see Example 2.6). Then there exists

a map $\Phi : X_\infty \to H^\infty(S^1)_0$ which extends to analytic maps $X_s \to H^s(S^1)_0$, $s \geqslant 0$, such that

$$h \circ \Phi(u) = \sum_{j=1}^{\infty} j^3\big(u_j^2 + u_{-j}^2\big) + \big\langle \text{a function of } u_l^2 + u_{-l}^2,\ l = 1, 2, \ldots \big\rangle. \quad (5.20)$$

Here $\{u_k, k \in \mathbb{Z}_0\}$ are coefficients of decomposition of $u \in X_s$ in the basis $\{\varphi_k\}$ and h is the KdV-Hamiltonian (see Example 2.7). Moreover, the Hamiltonian (5.20) defines an analytic Hamiltonian vector field of order three in each space X_d, $d \geqslant 1$. In the transformed variables the N-gap tori of the KdV equation take the form (5.5), where $n \geqslant N$ and exactly N numbers I_j are non-zero. Now let us take a torus (5.5), where $I \in \mathbb{R}^n_+$. Making a change of variables as in Section 5.1, we arrive at the Hamiltonian (5.18). Detailed and readable derivation of the normal form (5.20) see in [46].

Reduction to the Birkhoff normal form (5.20) uses essentially specifics of the KdV's L-operator. Still, similar arguments apply as well to the defocusing Zakharov–Shabat equation, see [41]. Presumably, the Birkhoff normal forms exist for some other integrable equations with selfadjoint L-operators, but not for equations with non-selfadjoint operators. In particular, the focusing Zakharov–Shabat equation cannot be reduced to the form (5.20) since for this equation some finite-gap tori are linearly unstable [26], while all invariant tori of the form (5.5) for the Hamiltonian (5.20) are linearly stable.

EXAMPLE 5.7 (*Perturbed KdV equation*). Consider the equation

$$\dot{u}(t, x) = \frac{1}{4}\frac{\partial}{\partial x}\big(u'' + 3u^2 + \varepsilon f(x, u)\big), \quad x \in S^1; \ \int_{S^1} u\, dx \equiv 0, \quad (5.21)$$

where f is smooth in x, u and analytic in u. The Metatheorem applies and implies that most of finite-gap KdV-solutions persist as time-quasiperiodic solutions of (5.21). Moreover, these solutions are smooth and linearly stable.

This result was first stated in [51]. The proof contains some gaps. Two the most serious of them are that Theorem 5.1, proved then only for semilinear equations, was used in a quasilinear case, and that the non-degeneracy assumptions (5.19) and (5.7) were taken for granted. These gaps were filled in later. The quasilinear version of Theorem 5.1 was proved in [57] (preprint of this paper appeared in 1995), and the non-degeneracy conditions were verified in [12]. Also see [59, Section 6.2.1]. The arguments in [12,59] are general and applies to other equations.

For a complete proof of 'KAM for KdV' see [59,35] and [46].

The Metatheorem (in its rigorous form as in [59,35] and [46]), applies to quasilinear Hamiltonian perturbations of any higher equation from the KdV-hierarchy, provided that the non-degeneracy relations are checked for this equation. It can be done in the same way as in Example 5.7. See [46], where the non-degeneracy of the second KdV equation is verified.

EXAMPLE 5.8 (*Perturbed SG equation*). Consider the equation

$$\ddot{u} = u_{xx} - \sin u + \varepsilon f(u, x), \qquad u(t, 0) = u(t, \pi) = 0, \tag{5.22}$$

where $f(0, x) \equiv 0$ (and $f \in C^\infty$ is analytic in u). The Metatheorem applies to prove persistence most of finite-gap solutions of the SG-equation, see [11,59,35]. In general, due to the phenomenon explained in Example 2.9, the persisted solutions are only H^2-smooth in x. But if f is x-independent and odd in u, then they are smooth.

In difference with the KdV-case, large amplitude finite-gap SG-solutions, as well as the corresponding persisted solutions of (5.22), in general are not linearly stable.

To end this section we note that since the persisted solutions $u_\varepsilon(t)$ have the form

$$u_\varepsilon(t) = \Phi_\varepsilon\big(p, q + t\omega_\varepsilon(p)\big) = \Phi_0\big(p, q + t\omega_\varepsilon(p)\big) + \mathrm{O}\big(\sqrt{\varepsilon}\big),$$

then to calculate them with the accuracy $\sqrt{\varepsilon}$ for all values of time t, we can use the "finite gap map" Φ_0 with the corrected frequency vector. Moreover, $\omega_\varepsilon(p) = \nabla h(p) + \varepsilon W_1(p) + \mathrm{O}(\varepsilon^2)$, where the vector $W_1(p)$ can be obtained by averaging over the corresponding finite-gap torus of some explicit quantity, see [59, p. 147].

5.6. *Small amplitude solutions of HPDEs*

Let us consider the non-linear string equation

$$u_{tt} = u_{xx} - mu + f(u), \quad u = u(t, x),\ 0 \leqslant x \leqslant \pi; \qquad u(t, 0) = u(t, \pi) = 0. \tag{5.23}$$

Here $m > 0$ and f is an odd analytic function of the form

$$f(u) = \kappa u^3 + \mathrm{O}\big(u^5\big), \quad \kappa > 0.$$

Since $m,\ \kappa > 0$, then constants $a,\ b > 0$ can be found such that $-mu + f(u) = -a \sin bu$. Hence, Equation (5.23) can be written as

$$u_{tt} = u_{xx} - a \sin bu + \mathrm{O}\big(u^5\big).$$

After the scaling $u = \varepsilon w$, $\varepsilon \ll 1$, the higher-order perturbation transforms to a small one, and we can apply the Metatheorem (cf. Example 5.8) to prove that small-amplitude parts of the finite-gap manifolds $\mathcal{T}^{2n}$, $n = 1, 2, \ldots$, for the SG equation $u_{tt} = u_{xx} - a \sin bu$ with the Dirichlet boundary conditions mostly persist in (5.23). To put this scheme through, the small-amplitude parts

$$\mathcal{T}^{2n}_\delta = \big\{(u, \dot{u}) \in \mathcal{T}^{2n} \mid \|u\| + \|\dot{u}\| < \delta\big\}, \quad 0 < \delta \ll 1,$$

of the manifolds $\mathcal{T}^{2n}$ have to be studied in details. This task was accomplished in [14], where the following results were proved:

(i) the sets $\overline{\mathcal{T}_\delta^{2n}}$ are smooth manifolds which contain the origin,
(ii) they are in one-to-one correspondence with their tangent spaces at the origin,
(iii) these tangent spaces are invariant spaces for the Klein–Gordon equation $u_{tt} = u_{xx} - (ab)u$.

Another proof of (i)–(iii) was suggested in [59]. It is based on some ideas from [44] and applies to other integrable equations. After (i)–(iii) are obtained, a version of the Metatheorem (or a version of Theorem 5.1) applies to prove that most of finite-gap solutions from a manifold $\mathcal{T}_\delta^{2n}$ persist in (5.23) in the following sense: the $2n$-dimensional Hausdorff measure of the persisted part of the manifold, divided by a similar measure of $\mathcal{T}_\delta^{2n}$, converges to one as $\delta \to 0$. See [13] for a proof and [53] for discussion.

Similar results hold for the NLS equation

$$i\dot{u} = u_{xx} + mu + f\big(|u|^2\big)u, \qquad f(0) = 0, \qquad f'(0) = \gamma \neq 0, \tag{5.24}$$

where f is analytic, since it is a higher-order perturbation of the Zakharov–Shabat equation (4.7). But it turns out that it is easier to approximate (5.24) near the origin by its partial Birkhoff normal form. The latter is an integrable infinite-dimensional Hamiltonian system (which is not an HPDE), and a sibling of the Metatheorem applies to prove that most of its time-quasiperiodic solutions persist in (5.24), see [60]. More on the techniques of Birkhoff normal forms in HPDE see in [74] and [46]. The classical reference for finite-dimensional Birkhoff normal forms is the book [65].

6. Around the Nekhoroshev theorem

The classical Nekhoroshev theorem [66] deals with nearly-integrable Hamiltonian systems with analytic Hamiltonians $H_\varepsilon(p,q) = h(p) + \varepsilon H(p,q)$ on the phase-space $P \times \mathbb{T}^n$, $P \subset \mathbb{R}^n$, given the usual symplectic structure $dp \wedge dq$. Under the assumption that the Hamiltonian $h(p)$ satisfies a mild non-degeneracy condition called *the steepness*, the theorem states that the action variables change exponentially slow along trajectories of the system. Namely, there exist constants $a, b \in (0,1)$ such that for any trajectory $(p(t), q(t))$ of the system we have

$$\big|p(t) - p(0)\big| \leqslant C\varepsilon^a \quad \text{if } |t| \leqslant \exp\big(\varepsilon^{-b}\big). \tag{6.1}$$

Strictly convex functions $h(p)$ form an important class of the steep Hamiltonians. An alternative proof of the theorem which applies in the convex case was suggested by Lochak [63]. It is based on clever approximation of a trajectory $(p(t), q(t))$ by a time-periodic solution of the equation which is a high-order normal form for H_ε. So rational frequency-vectors play for the Lochak approach very important role.

Original Nekhoroshev's proof contains two parts, analytical and geometrical. The techniques, developed in the analytical part of the proof, allow to get the following result, which we call below the quasi-Nekhoroshev theorem: Let us consider the Hamiltonian H_ε,

depending on an additional vector-parameter $\omega \in \Omega \Subset \mathbb{R}^n$, $H_\varepsilon = p \cdot \omega + \varepsilon H(p,q)$. Then for any $\gamma > 0$ there exists a Borel subset $\Omega_\gamma \subset \Omega$ ('the Diophantine subset') such that $\mathrm{mes}(\Omega \setminus \Omega_\gamma) < \gamma$, and (6.1) with $C = C_\gamma$ holds if $\omega \in \Omega_\gamma$. Note that in the Cartesian coordinates (x, y), corresponding to the action-angle variables (p, q) (i.e., $x_j = \sqrt{2p_j}\cos q_j$, $y_j = \sqrt{2p_j}\sin q_j$), the Hamiltonian H_ε reeds as

$$H_\varepsilon = \frac{1}{2}\sum \omega_j \left(x_j^2 + y_j^2\right) + \varepsilon H(x, y).$$

That is, H_ε is a perturbation of the quadratic Hamiltonian H_0. So the quasi-Nekhoroshev theorem implies long-time stability of the zero equilibrium for an analytical Hamiltonian

$$H(x, y) = H_0 + h, \quad h = \mathrm{O}\left(\left|(x, y)\right|^3\right), \tag{6.2}$$

provided that the vector ω belongs to the Diophantine set. In [67] Niederman used the Lochak approach to get a stronger theorem on stability for (6.2). Namely, he proved that the equilibrium is stable during the exponentially long time if the vector ω does not satisfies resonant relations up to order four, and h is convex in a certain sense.[8]

To get a corresponding theorem which applies to all small initial data is a non-trivial task, resolved by Niederman [67] by means of the Lochak approach.

No analogy of the Nekhoroshev theorem for HPDEs is known yet, but a number of *ad hoc* quasi-Nekhoroshev theorems for HPDEs were proved, mostly by Bourgain and Bambusi, see [3,4,22] and references therein. These works discuss stability of the equilibrium for HPDEs (mostly 1D) with Hamiltonians of the form (6.2). Under some restrictions on the quadratic part H_0 and on the higher-order part h, it is proved that if the initial data u_0 is an ε-small and 'very' smooth function, then a solution stays very close to the corresponding invariant torus of the linear system with the Hamiltonian H_0, during the time which is polynomially large in ε^{-1}, or even exponentially large. This result is obtained either under the 'quasi-Nekhoroshev' condition that the spectrum of the operator A is 'highly non-resonant', or under the opposite assumption (needed to apply the Lochak–Niederman technique) that the spectrum is 'very resonant'. In particular, the following result is proved in [3] (also see [75,22]): Let us consider the NLS equation (5.24) in the scale $\{H_0^s(0, \pi)\}$ of odd 2π-periodic functions. Assume that $u_0(x) = \sum_{k=1}^N u_{k0}\sin kx$, denote $\varepsilon = |u_0(x)|_{L_2} \ll 1$ and write the solution $u(t, x)$ of (5.24) as $u = \sum u_k(t)\sin kx$. Then there exist $\varepsilon_* > 0$ and constants $C_1, C_2 > 0$ such that for $\varepsilon < \varepsilon_*$ and $|t| \leqslant C_1 \exp(\varepsilon_*/\varepsilon)^{1/N} =: T_\varepsilon$ we have

$$\sum_{k=1}^{\infty} \left(\left|u_k(t)\right|^2 - |u_{k0}|^2\right)^2 \leqslant C_2 \varepsilon^{4+1/N}. \tag{6.3}$$

Let us set $T^N = \{u(x) = \sum_{k=1}^N u_k \sin kx \mid |u_k| = |u_{k_0}|\}$. This is an n-torus of diameter $\sim \varepsilon$ and (6.3) implies that

$$\mathrm{dist}_{H_0^s}\left(u(t),\, T^n\right) \leqslant C_s \varepsilon^{1+1/N} \quad \forall |t| \leqslant T_\varepsilon,$$

[8] Independently this result was obtained in [9] by means of the Nekhoroshev's techniques.

if $s < -1/4$. Thus, during the time T_ε the trajectory $u(t)$ remains very close to its projection to T^N. The latter is a trajectory of an N-dimensional dynamical system, so the time of its return to a $\rho\varepsilon$-neighbourhood ($\rho \ll 1$) of the initial point 'should' be of order ρ^{-N}. Same is true for the trajectory $u(t)$, if ε is small in terms of ρ. The phenomenon of the pathologically good recurrence properties of small-amplitude trajectories of some non-integrable 1D HPDEs is well known from numerics (e.g., see [82]). We have seen that the quasi-Nekhoroshev theorems as above explain it up to some extend.

7. Invariant Gibbs measures

If Equation (4.1) is a finite-dimensional Hamiltonian system with $u = (p, q) \in (\mathbb{R}^{2n}, dp \wedge dq)$, then any measure $f(H(p,q))\,dp\,dq$ such that the function $f \circ H$ is Lebesgue-integrable, is invariant for the equation. The most important among these measures is the Gibbs measure $e^{-H}\,dp\,dq$ (the Hamiltonian H is assumed to grow to infinity with $|(p, q)|$). Now let us consider an HPDE (4.1). Say, the zero-mass ϕ^4-equation

$$\ddot{u} = u_{xx} - u^3, \quad u = u(t, x),\ x \in S^1.$$

This equation is equivalent to the system

$$\begin{aligned} \dot{u} &= -Bv, \\ \dot{v} &= Bu + B^{-1}\big(u^3 - u\big), \end{aligned} \tag{7.1}$$

where $B = \sqrt{1 - \Delta}$. Denoting $\xi = (u, v)$ we can see that this is a Hamiltonian system in the symplectic scale $(\{Z_s = H^{s+1/2}(\mathbb{T}^2; \mathbb{R}^2)\}, \alpha_2 = \bar{J}\,d\xi \wedge d\xi)$, where $J(u, v) = (-v, u)$, with the Hamiltonian

$$H(\xi) = \frac{1}{2}\|\xi\|_0^2 + \int \left(\frac{1}{4}|u|^4 - \frac{1}{2}|u|^2\right) dx, \quad \xi = (u, v).$$

Here $\|\cdot\|_0$ is the norm in the space $H^{1/2}(S^1; \mathbb{R}^2)$ (cf. Section 8.3). The natural question is if the formal expression

$$\mu = e^{-H(\xi)}\,d\xi \tag{7.2}$$

defines a measure in a suitable function space $\Xi = \{\xi(x)\}$, invariant for flow-maps of Equation (7.1). Since the Lebesgue measure $d\xi$ does not exist in an infinite-dimensional function space, then to make the right-hand side of (7.2) meaningful we write it as

$$\mu = e^{-\int (\frac{1}{4}|u|^4 - \frac{1}{2}|u|^2)\,dx} e^{-\frac{1}{2}\|\xi\|_0^2}\,d\xi.$$

Now $\exp -\frac{1}{2}\|\xi\|_0^2\,d\xi$ is a well-defined Gaussian measure, supported by a suitable space Ξ, formed by functions of low smoothness, and $0 < p(\xi) \leqslant C$, where $p(\xi) = e^{-\int (\frac{1}{4}|u|^4 - \frac{1}{2}|u|^2)\,dx}$. Therefore if

(i) $p(\xi)$ is a Borel function on Ξ, then μ is a well-defined Borel measure on Ξ.

To check that it is invariant for Equation (7.1) we have to verify that

(ii) the flow-maps of (7.1) are well-defined on $\operatorname{supp}\mu$ and preserve the measure.

The corresponding result was first stated by Friedlander [37]. Unfortunately, his arguments contain serious flaws. Complete proofs appeared later in works of Zhidkov, McKean and Vaninsky and Bourgain, see the books [20,84] and references therein. Similar arguments apply to the 1D NLS equation (2.4), where the non-quadratic term q satisfies certain restrictions.

For higher-dimensional HPDEs the task of constructing the Gibbs measures becomes much more difficult. The only known result is due to Bourgain who proved that for the defocusing 2D NLS equation

$$i\dot{u} = \Delta u - |u|^2 u, \quad x \in \mathbb{T}^2,$$

the Gibbs measure (7.2) exists and is invariant. The main difficulty here is the step (ii) which is now based on highly non-trivial results on regularity of corresponding flow-maps in Sobolev spaces of low smoothness; see in [20].

8. The non-squeezing phenomenon and symplectic capacity

8.1. *The Gromov theorem*

Let $(\mathbb{R}^{2n}, \beta_2)$ be the space $\mathbb{R}^{2n} = \{x_1, x_{-1}, \dots, x_{-n}\}$ with the Darboux symplectic form $\beta_2 = \sum dx_j \wedge dx_{-j}$. By $B_r(x) = B_r(x; \mathbb{R}^{2n})$ and $C^j_\rho = C^j_\rho(\mathbb{R}^{2n})$, $1 \leqslant j \leqslant n$, we denote the following balls and cylinders in $\mathbb{R}^{2n}$:

$$B_r(x) = \big\{y \mid |y - x| < r\big\}, \qquad C^j_\rho = \big\{y = (y_1, \dots, y_{-n}) \mid y_j^2 + y_{-j}^2 < \rho^2\big\}.$$

The famous (*non*-)*squeezing theorem* by M. Gromov [42] states that if f is a symplectomorphism $f : B_r(x) \to \mathbb{R}^{2n}$ such that its range belongs to some cylinder $x_1 + C^j_\rho$, $x_1 \in \mathbb{R}^{2n}$, then $\rho \geqslant r$. For an alternative proof, references and discussions see [43].

8.2. *Infinite-dimensional case*

Let us consider a symplectic Hilbert scale $(\{Z_s\}, \alpha_2)$ with a basis $\{\varphi_j \mid j \in \mathbb{Z}_0\}$. We assume that this is a shifted Darboux scale (cf. Example 2.4 in Section 2.2). It means that the basis can be renormalised to a basis $\{\tilde{\varphi}_j \mid j \in \mathbb{Z}_0\}$ (each $\tilde{\varphi}_j$ is proportional to φ_j) which is a Darboux basis for the form α_2 and a Hilbert basis of some space Z_d:

$$\langle \tilde{\varphi}_j, \tilde{\varphi}_k \rangle_d = \delta_{j,k}, \qquad \alpha_2[\tilde{\varphi}_j, \tilde{\varphi}_{-k}] = \operatorname{sgn} j\, \delta_{j,k} \quad \forall j, k. \tag{8.1}$$

These relations imply that

$$\alpha_2[\xi, \eta] = \langle \bar{J}\xi, \eta \rangle_d, \qquad \bar{J}\tilde{\varphi}_j = \operatorname{sgn} j\, \tilde{\varphi}_{-j} \quad \forall j. \tag{8.2}$$

In particular, $\bar{J} = J$.

Below we skip the tildes and re-denote the new basis back to $\{\varphi_j\}$.

In this scale we consider a semilinear Hamiltonian equation with the Hamiltonian $H(u) = \frac{1}{2}\langle Au, u\rangle_d + h(u,t)$. Due to (8.2) it can be written as

$$\dot{u} = JAu + J\nabla^d h(u,t), \tag{8.3}$$

where ∇^d signifies the gradient in u with respect to the scalar product of Z_d.

If a Hamiltonian PDE is written in the form (8.3), then the symplectic space (Z_d, α_2) is called the (*Hilbert*) *Darboux phase space* for this PDE. Below we study properties of flow-maps of Equation (8.3) in its Darboux phase space.

Let us assume that the operator A has the form

(H1) $Au = \sum_{j=1}^{\infty} \lambda_j(u_j\varphi_j + u_{-j}\varphi_{-j})$ $\forall u = \sum u_j\varphi_j$, where λ_j's are some real numbers.

Then $JAu = \sum_{j=1}^{\infty} \lambda_j(u_{-j}\varphi_{-j} - u_j\varphi_j)$, so the linear operators e^{tJA} are direct sums of rotations in the planes $\mathbb{R}\varphi_j + \mathbb{R}\varphi_{-j} \subset Z_d$, $j = 1, 2, \ldots$.

We also assume that the gradient map $\nabla^d h$ is smoothing:

(H2) there exists $\gamma > 0$ such that $\operatorname{ord} \nabla^d h = -\gamma$ for $s \in [d-\gamma, d+\gamma]$. Moreover, the maps

$$\nabla^d h : Z_s \times \mathbb{R} \to Z_{s+\gamma}, \quad s \in [d-\gamma, d+\gamma],$$

are C^1-smooth and bounded.[9]

For any t and T we denote by O_t^T any open subset of the domain of definition of the flow-map S_t^T in Z_d, such that for each bounded subset $Q \subset O_t^T$ the set $\bigcup_{\tau\in[t,T]} S_t^\tau(Q)$ is bounded in Z_d.[10]

In the theorem below the balls B_r and the cylinders C_ρ^j, $j \geqslant 1$, are defined in the same way as in Section 8.1.

THEOREM 8.1. *Assume that* (H1) *and* (H2) *hold and that a ball* $B_r = B_r(u_0; Z_d) := \{\|y - u_0\|_d < r\}$ *belongs to* O_t^T *together with some* ε*-neighbourhood,* $\varepsilon > 0$. *Then the relation*

$$S_t^T(B_r) \subset v_0 + C_\rho^j(Z_d) \tag{8.4}$$

with some $v_0 \in Z_d$ *and* $j \geqslant 1$ *implies that* $\rho \geqslant r$.

PROOF. Without lost of generality we may assume that

$$v_0 = 0, \qquad j = 1.$$

Arguing by contradiction we assume that (8.4) holds with $\rho < r$ and choose any $\rho_1 \in (\rho, r)$.

[9] I.e., they send bounded sets to bounded.

[10] This set should be treated as a 'regular part of the domain of definition'.

For $n \geqslant 1$ we denote by E^{2n} the subspace of Z_d, spanned by the vectors $\{\varphi_j, |j| \leqslant n\}$, and provide it with the usual Darboux symplectic structure (it is given by the form $\alpha_2|_{E^{2n}}$). By Π_n we denote the orthogonal projection $\Pi_n : Z_d \to E^{2n}$. We set

$$H^n = \frac{1}{2}\langle Au, u\rangle_d + h\big(\Pi_n(u), t\big)$$

and denote by $S^T_{(n)t}$ flow-maps of the Hamiltonian vector filed V_{H^n}. Any map $S^T_{(n)t}$ decomposes to the direct sum of a symplectomorphism of E^{2n} and of a linear symplectomorphism of $Z_d \ominus E^{2n}$. So the theorem's assertion with the map S^T_t replaced by $S^T_{(n)t}$ follows from the Gromov theorem, applied to the symplectomorphism

$$E^{2n} \to E^{2n}, \qquad x \mapsto \Pi_n S^T_{(n)t}\big(i(x) + u_0\big),$$

where i stands for the embedding of E^{2n} to Z_d.

Proofs of the two easy lemmas below can be found in [54].

LEMMA 8.2. *Under the theorem's assumptions the maps $S^T_{(n)t}$ are defined on B_r for $n \geqslant n'$ with some sufficiently large n', and there exists a sequence $\varepsilon_n \underset{n\to\infty}{\longrightarrow} 0$ such that*

$$\big\| S^T_t(u) - S^T_{(n)t}(u) \big\| \leqslant \varepsilon_n \tag{8.5}$$

for $n \geqslant n'$ and for every $u \in B_r$.

LEMMA 8.3. *For any $u \in B_r$ we have $S^T_t(u) = e^{(T-t)JA}u + \widetilde{S}^T_t(u)$, where $\widetilde{S}^T_t$ is a C^1-smooth map in the scale $\{Z_s\}$ and* ord $\widetilde{S}^T_t = -\gamma$ *for $s \in [d-\gamma, d+\gamma]$.*

Now we continue the proof of the theorem. Since its assertion holds for any map $S^T_{(n)t}$ ($n \geqslant n'$) and since the ball B_r belongs to this map's domain of definition (see Lemma 8.2), then for each $n \geqslant n'$ there exists a point $u_n \in B_r$ such that $S^T_{(n)t}(u_n) \notin C^1_{\rho_1}(0)$. That is,

$$\big|\Pi_1 S^T_{(n)t}(u_n)\big| \geqslant \rho_1. \tag{8.6}$$

By the weak compactness of a Hilbert ball, we can find a weakly converging subsequence

$$u_{n_j} \rightharpoonup u \in B_r, \tag{8.7}$$

so

$$u_{n_j} \to u \quad \text{strongly in } Z_{d-\gamma}.$$

Due to Lemma 8.3 this implies that $\widetilde{S}^T_t(u_{n_j}) \to \widetilde{S}^T_t(u)$ in Z_d, and using (8.7) we obtain the convergence:

$$S^T_t(u_{n_j}) \rightharpoonup S^T_t(u). \tag{8.8}$$

Noting that $|\Pi_1 S_t^T(u_n)| = |\Pi_1 S_{(n)t}^T u_n + \Pi_1(S_t^T - S_{(n)t}^T)u_n|$ and using (8.6), (8.5) we get:

$$\left|\Pi_1 S_t^T(u_n)\right| \geqslant \rho_1 - \varepsilon_n, \quad n \geqslant n'. \tag{8.9}$$

Since by (8.8) $\Pi_1 S_t^T(u_{n_j}) \to \Pi_1 S_t^T(u)$ in E^2, then due to (8.9) we have $|\Pi_1 S_t^T(u)| \geqslant \rho_1$. This contradicts (8.4) because $\rho_1 > \rho$. The obtained contradiction proves the theorem. □

8.3. *Examples*

EXAMPLE 8.4. Let us consider the non-linear wave equation

$$\ddot{u} = \Delta u - \tilde{f}(u; t, x), \tag{8.10}$$

where $u = u(t, x)$, $x \in \mathbb{T}^n$. The function $\tilde{f}$ is a polynomial in u of a degree D such that its coefficients are smooth functions of t and x. We set $f = \tilde{f} - u$, denote by B the linear operator $B = \sqrt{1 - \Delta}$ and write (8.10) as the system of two equations:

$$\begin{aligned} \dot{u} &= -Bv, \\ \dot{v} &= Bu + B^{-1} f(u; t, x). \end{aligned} \tag{8.11}$$

Let us take for $\{Z_s\}$ the shifted Sobolev scale $Z_s = H^{s+1/2}(\mathbb{T}^n; \mathbb{R}^2)$, where $\langle \xi, \eta \rangle_s = \int_{\mathbb{T}^n} B^{2s+1}\xi \cdot \eta \, dx$ (its basic scalar product is the scalar product in $H^{1/2}$). We set $\alpha_2 = \bar{J}\, d\xi \wedge d\xi$, where $J\xi = (-v, u)$ for $\xi = (u, v)$. Choosing for $\{\psi_j, j \in \mathbb{N}\}$ a Hilbert basis of the space $H^{1/2}(\mathbb{T}^n)$, formed by properly normalised and enumerated non-zero functions $\sin s \cdot x$ and $\cos s \cdot x$ ($s \in \mathbb{Z}^n$), we set

$$\tilde{\varphi}_j = (\psi_j, 0), \qquad \tilde{\varphi}_{-j} = (0, \psi_j), \quad j \in \mathbb{N}.$$

The obtained symplectic scale $(\{Z_s\}, \alpha_2)$ is a Darboux scale. It is easy to see that (8.11) is a Hamiltonian equation with the Hamiltonian

$$H(u, v) = \frac{1}{2}\big\langle B(u, v), (u, v)\big\rangle_0 + \int F(u; t, x)\, dx,$$

where $F'_u = f$. So $Z_0 = H^{1/2}(\mathbb{T}^n, \mathbb{R}^2)$ is the Darboux phase space for the non-linear wave equation, written in the form (8.11).

To apply Theorem 8.1 we have to check the conditions (H1) and (H2). The first one (with $A = B$) holds trivially since $\tilde{\varphi}_j$'s are eigenfunctions of the Laplacian. The condition (H2) holds in the following three cases:

(a) $n = 1$,
(b) $n = 2$, $D \leqslant 4$,
(c) $n = 3$, $D \leqslant 2$.

The case (a) and the case (b) with $D \leqslant 2$ can be checked using elementary tools, see [54]. Arguments in the case (b) with $3 \leqslant D \leqslant 4$ and in the case (c) are based on a Strichartz-type inequality, see [17].

In the cases (a)–(c), Theorem 8.1 applies to Equation (8.10) in the form (8.11) and shows that the flow maps cannot squeeze $H^{1/2}$-balls to narrow cylinders. This result can be interpreted as impossibility of 'locally uniform' energy transition to high modes, see in [54].

EXAMPLE 8.5. For a non-linear Schrödinger equation

$$\dot u = i\Delta u + i f_u'\big(|u|^2\big)u, \quad x \in \mathbb{T}^n \tag{8.12}$$

(cf. Example 2.7), the Darboux phase space is the L_2-space $L_2(\mathbb{T}^n;\mathbb{C})$ with the basis, formed by normalised exponents $\{e^{is\cdot x}, ie^{is\cdot x}\}$. Now the assumption (H2) fails (and it is very unlikely that the flow-maps of (8.12) satisfy the assertions of Lemmas 8.2 and 8.3). So we smooth out the Hamiltonian of (8.12) and replace it by

$$H_\xi = \frac{1}{2}\int \big(|\nabla u|^2 + f\big(|U|^2\big)\big)\,dx, \quad U = u * \xi,$$

where $u * \xi$ is the convolution of u with a function $\xi \in C^\infty(\mathbb{T}^n, \mathbb{R})$. The corresponding Hamiltonian equation is

$$\dot u = i\Delta u + i\big(f'\big(|U|^2\big)U\big) * \xi. \tag{8.13}$$

This smoothed equation satisfies (H1), (H2), and Theorem 8.1 applies to its flow-maps.

8.4. *Symplectic capacity*

Another way to prove Theorem 8.1 uses a new object—symplectic capacity —which is interesting on its own.

Symplectic capacity in a Hilbert Darboux space (Z_d, α_2) as in Section 8.2 (below we abbreviate Z_d to Z), is a map c which associates to any open subset $O \subset Z$ a number $c(O) \in [0, \infty]$ and satisfies the following properties:

(1) *Translational invariance*: $c(O) = c(O + \xi)$ for any $\xi \in Z$;
(2) *Monotonicity*: if $O_1 \supset O_2$, then $c(O_1) \geqslant c(O_2)$;
(3) 2*-homogeneity*: $c(\tau O) = \tau^2 c(O)$;
(4) *Normalisation*: for any ball $B_r = B_r(x; Z)$ and any cylinder $C_r^j = C_r^j(Z)$ we have $c(B_r) = c(C_r^j) = \pi r^2$.

(We note that for $x = 0$ the cylinder contains the ball and is 'much bigger', but both sets have the same capacity.)

(5) *Symplectic invariance*: for any symplectomorphism $\Phi : Z \to Z$ and any domain O, $c(\Phi(O)) = c(O)$.

If (Z, α_2) is a finite-dimensional Darboux space, then existence of a capacity with properties (1)–(5) is equivalent to the Gromov theorem. Indeed, if a capacity exists, then the squeezing (8.4) with $\rho < r$ is impossible due to (2), (4) and (5). On the opposite, the quantity

$$\tilde{c}(O) = \sup\{\pi r^2 \mid \text{there exists a symplectomorphism which sends } B_r \text{ in } O\}$$

obviously satisfies (1)–(3) and (5). Using the Gromov theorem we see that $\tilde{c}$ satisfies (4) as well.

If (Z, α_2) is a Hilbert Darboux space, then the finite-dimensional symplectic capacity, obtained in [43], can be used to construct a capacity c which meets (1)–(4). This capacity turns out to be invariant under symplectomorphisms, which are flow-maps S_t^T as in Theorem 8.1, see [54]. This result also implies Theorem 8.1.

9. The squeezing phenomenon and the essential part of the phase-space

Example 8.4 shows that flow-maps of the non-linear wave equation (8.11) satisfy the Gromov property. This means (more or less) that *flow of generalised solutions for a non-linear wave equation cannot squeeze a ball in a narrow cylinder*. On the contrary, behaviour of the flow formed by *classical* solutions for the non-linear wave equation in sufficiently smooth Sobolev spaces exhibits 'a lot of squeezing', at least if we put a small parameter δ in front of the Laplacian. Corresponding results apply to a bigger class of equations. Below we discuss them for non-linear Schrödinger equations; concerning the non-linear wave equation (8.10) see the author's paper in GAFA 5:4.

Let us consider the non-linear Schrödinger equation:

$$\dot{u} = -i\delta\Delta u + i|u|^{2p}u, \tag{9.1}$$

where $\delta > 0$ and $p \in \mathbb{N}$, supplemented by the odd periodic boundary conditions:

$$\begin{aligned} u(t,x) &= u(t, x_1, \dots, x_j + 2\pi, \dots, x_n) \\ &= -u(t, x_1, \dots, -x_j, \dots, x_n), \quad j = 1, \dots, n, \end{aligned} \tag{9.2}$$

where $n \leqslant 3$. Clearly, any function which satisfies (9.2) vanishes at the boundary of the cube K^n of half-periods, $K^n = \{0 \leqslant x_j \leqslant \pi\}$. The problem (9.1), (9.2) can be written in the Hamiltonian form (2.2) if for the symplectic Hilbert scale $(\{X_s\}, \alpha_2)$ one takes the scale formed by odd periodic complex Sobolev functions, $X_s = H^s_{\text{odd}}(\mathbb{R}^n/2\pi\mathbb{Z}^n; \mathbb{C})$, and $\alpha_2 = i\,du \wedge du$ (cf. Example 2.8).

Due to a non-trivial result of Bourgain (which can be extracted from [15]), flow-maps S^t for (9.1), (9.2) are well defined in the spaces X_s, $s \geqslant 1$. In particular, they are well defined in the space C^∞ of smooth odd periodic functions. Denoting by $|\cdot|_m$ the

C^m-norm, $|u|_m = \sup_{|\alpha|=m} \sup_x |\partial_x^\alpha u(x)|$, we define below the set $\mathfrak{A}_m \subset C^\infty$ which we call the essential part of the smooth phase-space for the problem (9.1), (9.2) with respect to the C^m-norm, or just the *essential part of the phase-space*:

$$\mathfrak{A}_m = \big\{u \in C^\infty \mid u \text{ satisfies (9.2) and the condition (9.3)}\big\},$$

where

$$|u|_0 \leqslant K_m \delta^\mu |u|_m^{1/(2pm\varkappa+1)}, \tag{9.3}$$

with a suitable $K_m = K_m(\varkappa)$ and $\mu = m\varkappa/(2pm\varkappa + 1)$. Here $\varkappa$ is any fixed constant $\varkappa \in (0, 1/3)$.

Intersection of the set $\mathfrak{A}_m$ with the R-sphere in the C^m-norm (i.e., with the set $\{|u|_m = R\}$) has the C^0-diameter $\leqslant 2K_m \delta^\mu R^{1/(2pm\varkappa+1)}$. Asymptotically (as $\delta \to 0$ or $R \to \infty$) this is much smaller than the C^0-diameter of the sphere, which equals $C_m R$. Thus, $\mathfrak{A}_m$ is an 'asymptotically narrow' subset of the smooth phase space.

The theorem below states that for any $m \geqslant 2$ the set $\mathfrak{A}_m$ is a recursion subset for the dynamical system, and gives a control for the recursion time:

THEOREM 9.1. *Let $u(t) = u(t, \cdot)$ be a smooth solution for* (9.1), (9.2) *and $|u(t_0)|_0 = U$. Then there exists $T \leqslant t_0 + \delta^{-1/3} U^{-4p/3}$ such that $u(T) \in \mathfrak{A}_m$ and $\frac{1}{2}U \leqslant |u(T)|_0 \leqslant \frac{3}{2}U$.*

Since L_2-norm of a solution is an integral of motion (see Example 3.5) and $|u(t)|_0 \geqslant |u(t)|_{L_2(K^n)}$, then we obtain the following

COROLLARY 9.2. *Let $u(t)$ be a smooth solution for* (9.1), (9.2) *and $|u(t)|_{L_2(K^n)} \equiv W$. Then for any $m \geqslant 2$ this solution cannot stay outside $\mathfrak{A}_m$ longer than the time $\delta^{-1/3} W^{-4p/3}$.*

For the theorem's proof we refer the reader to Appendix 3 in [58]. Here we explain why 'something like this result' should be true. Presenting the arguments it is more convenient to operate with the Sobolev norms $\| \cdot \|_m$. Let us denote $\|u(t_0)\|_0 = A$. Arguing by contradiction, we assume that for all $t \in [t_0, t_1] = L$, where $t_1 = t_0 + \delta^{-1/3} U^{-4p/3}$, we have

$$C\delta^a \|u\|_m^b < \|u\|_0, \tag{9.4}$$

where $m \geqslant 3$ is a fixed number. Since $\|u(t)\|_0 \equiv A$, then (9.4) and the interpolation inequality imply the upper bounds

$$\big\|u(t)\big\|_l \leqslant C_l A^{1-\frac{l}{m}+\frac{l}{mb}} \delta^{-\frac{la}{mb}}, \quad 0 \leqslant l \leqslant m,\ t \in L. \tag{9.5}$$

In particular, $\delta \|\Delta u\|_1 \leqslant C_3 A^{1-\frac{3}{m}+\frac{3}{mb}} \delta^{1-\frac{3a}{mb}}$. Therefore if $mb > 3a$, then for $t \in L$ Equation (9.1), treated as a dynamical system in H^1_{odd}, is a perturbation of the trivial equation

$$\dot u = i|u|^{2p}u. \tag{9.6}$$

Elementary arguments show that the H^1-norm of each non-zero solution for (9.6) grows linearly with time. This implies a lower bound for $\sup_{t\in L}\|u(t)\|_1$, where $u(t)$ is the solution for (9.1), (9.2) which we discuss. It turns out that one can choose a and b in such a way that $mb > 3a$ and the lower bound we have just obtained contradicts (9.5) with $l = 1$. This contradiction shows that (9.4) cannot be true for all $t \in L$. In other words, $\|u(\tau)\|_0 \leqslant C\delta^a \|u(\tau)\|_m^b$ for some $\tau \in L$. At this moment τ the solution enters a domain, similar to the essential part $\mathfrak{A}_m$.

Let us consider any trajectory $u(t)$ for (9.1), (9.2) such that $|u(t)|_{L_2(K^n)} \equiv W \sim 1$, and discuss the time-averages $\langle |u|_m \rangle$ and $\langle \|u\|_m^2 \rangle^{1/2}$ of its C^m-norm $|u|_m$ and its Sobolev norm $\|u\|_m$, where we set

$$\langle |u|_m \rangle = \frac{1}{T}\int_0^T |u|_m \, dt, \qquad \langle \|u\|_m^2 \rangle^{1/2} = \left(\frac{1}{T}\int_0^T \|u\|^2 \, dt\right)^{1/2},$$

and the time T of averaging is specified below. While the trajectory stays in $\mathfrak{A}_m$, we have

$$|u|_m \geqslant \left(W K_m^{-1} \delta^{-\mu}\right)^{1/(1-2p\mu)}.$$

One can show that this inequality implies that each visit to $\mathfrak{A}_m$ increases the integral $\int |u|_m \, dt$ by a term bigger than δ to a negative degree. Since these visits are sufficiently frequent by the corollary, then we obtain a lower estimate for the quantity $\langle |u|_m \rangle$. Details can be found in [55]. Here we present a better result which estimates the time-averaged Sobolev norms. For a proof see Section 4.1 of [58].

THEOREM 9.3. *Let $u(t)$ be a smooth solution for Equation* (9.1), (9.2) *such that $|u(t)|_{L_2(K^n)} \geqslant 1$. Then there exists a sequence $k_m \nearrow 1/3$ and constants $C_m > 0$, $\delta_m > 0$ such that $\langle \|u\|_m^2 \rangle^{1/2} \geqslant C_m \delta^{-2mk_m}$, provided that $m \geqslant 4$, $\delta \leqslant \delta_m$ and $T \geqslant \delta^{-1/3}$.*

The results stated in Theorems 9.1, 9.3 remain true for Equations (9.1) with dissipation. I.e., for the equations with δ replaced by $\delta\nu$, where ν is a unit complex number such that $\operatorname{Re}\nu \geqslant 0$ and $\operatorname{Im}\nu \geqslant 0$.[11] If $\operatorname{Im}\nu > 0$, then smooth solutions for (9.1), (9.2) converge to zero in any C^m-norm. Since the essential part $\mathfrak{A}_m$ clearly contains a sufficiently small C^m-neighbourhood of zero, then eventually any smooth solution enter $\mathfrak{A}_m$ and stays there forever. Theorem 9.3 states that the solution will visit the essential part much earlier, before its norm decays. Moreover, results, similar to Theorem 9.3, are true for solutions of the damped-driven equation $\dot{u} + \delta\Delta u - i|u|^2 u = \eta(t,x)$, where the force η is a random field, smooth in x, and stationary mixing in t. See [56] and [58].

Acknowledgements

I thank for the hospitality FIM (ETH, Zürich), where this paper was completed. The research was supported by EPSRC, grants GR/N63055/01 and GR/S68712/01.

[11] The only correction is that if $\operatorname{Im}\nu > 0$, then in Theorem 9.3 one should take $T = \delta^{-1/3}$.

Appendix. Families of periodic orbits in reversible PDEs, by D. Bambusi

A.1. *Introduction*

Some families of periodic solutions of PDEs can be constructed using KAM theory; however a different approach leading to stronger results and simpler proofs is available. It is based on the Lyapunov–Schmidt decomposition combined with a suitable analysis of small denominators. The main advantage of this approach is elimination of the second Melnikov condition (see (5.7)). As a consequence it is applicable to problems with periodic boundary conditions and to some equations in more than one space dimension. Most of the general theory has been developed for equations that are of second order in time and we will mainly deal with this case. Moreover, we will concentrate on problems involving small denominators and only briefly report on results of a different kind.

A.2. *An abstract theorem for non-resonant PDEs*

Let $\{X_s\}$ be a scale of Hilbert spaces with norms $\|\cdot\|_s$ and scalar product $\langle\cdot;\cdot\rangle_s$. Let A be a (linear) morphism of the scale, and assume that there exists a Hilbert basis $\{\varphi_j\}_{j=1}^{\infty}$ of X_0 such that

$$A\varphi_j = \omega_j^2\varphi_j, \quad \omega_j > 0.$$

Let us fix s, consider a neighbourhood $\mathcal{U}$ of the origin in X_s and a smooth map $g:\mathcal{U}\to X_s$, having at the origin a zero of second order. We are interested in families of small amplitude periodic solutions of the equation

$$\ddot{x} + Ax = g(x). \tag{A.1}$$

EXAMPLE A.1. The non-linear wave equation with periodic boundary conditions:

$$w_{tt} - w_{xx} + V(x)w = f(x,w), \tag{A.2}$$

$$w(x,t) = w(x+2\pi,t), \qquad w_x(x,t) = w_x(x+2\pi,t), \tag{A.3}$$

where the potential V and the non-linearity f are smooth periodic of period 2π in x, and $f(x,w) = \mathrm{O}(|w|^2)$. Let λ_j be the periodic eigenvalues of the Sturm–Liouville operator $-\partial_{xx} + V(x)$ and assume $\lambda_j > 0\ \forall j$. Then the frequencies are $\omega_j := \sqrt{\lambda_j}$. In this case $X_s = H^s(\mathbb{T})$, and f induces a smooth operator from X_s to itself, provided that $s > 1/2$.

EXAMPLE A.2. The non-linear plate equation in the d-dimensional cube:

$$w_{tt} + \Delta\Delta w + aw = f(w), \quad x \in \mathcal{Q}, \tag{A.4}$$

$$w|_{\partial\mathcal{Q}} = \Delta w|_{\partial\mathcal{Q}} = 0, \tag{A.5}$$

where $a > 0$, $f(w) = \mathrm{O}(|w|^3)$ and

$$\mathcal{Q} := \big\{x = (x_1, \dots, x_d) \in \mathbb{R}^d\colon\ 0 < x_i < \pi\big\}.$$

Then the eigenfunctions of the linearised system are given by

$$\varphi_n = \sin(n_1 x_1)\sin(n_2 x_2)\cdots\sin(n_d x_d)$$

and the corresponding frequencies are $\omega_n = \sqrt{(n_1^2 + \cdots + n_d^2)^2 + a}$, where $n \in \mathbb{Z}^d$ and $n_i \geqslant 1$. To fit the abstract scheme we order the basis in such a way that the frequencies are in non-decreasing order. Now $X_0 = L^2(\mathcal{Q})$, and $X_s = D((\Delta\Delta)^s) \subset H^{4s}$ endowed with the graph norm. If the non-linearity f is smooth and odd (i.e. $f(-w) = -f(w)$), then it defines a smooth map from X_s to itself for any $s > [d/2]/4$ (see Example 2.5).

In the linear approximation ($g \equiv 0$) the general solution of (A.1) is the superposition of the linear normal modes, i.e. of the families of periodic solutions

$$x^{(j)}(t) = \big(a_j \cos(\omega_j t) + b_j \sin(\omega_j t)\big)\varphi_j. \tag{A.6}$$

Fix one of the families, say $x^{(1)}$. To ensure its persistence in the non-linear problem we make the following assumptions:

(H1) (Non-resonance) For small enough $\gamma > 0$ there exists a closed set $W_\gamma \subset \mathbb{R}^+$ having ω_1 as an accumulation point both from the right and from the left, and such that for any $\omega \in W_\gamma$ one has

$$|\omega l - \omega_j| \geqslant \frac{\gamma}{l}, \quad \forall l \geqslant 1,\ \forall j \geqslant 2. \tag{A.7}$$

(H2) (Non-degeneracy) Let $g_r(x)$ be the first non-vanishing (homogeneous) Taylor polynomial of g. Assume that $r \geqslant 3$ and $\beta_0 \neq 0$, where

$$\beta_0 := \begin{cases} \langle g_r(\varphi_1), \varphi_1\rangle_0 & \text{if } r \text{ is odd,} \\ \langle g_{r+1}(\varphi_1), \varphi_1\rangle_0 & \text{if } r \text{ is even.} \end{cases} \tag{A.8}$$

Denoting $\xi_1(\omega_1 t) = \cos(\omega_1 t)\varphi_1$ one has

THEOREM A.3. *Suppose that assumptions* (H1), (H2) *hold. Then there exist a set $\mathcal{E} \subset \mathbb{R}$ having zero as an accumulation point, a positive ω_*, and a family of periodic solutions $\{x_\varepsilon(t)\}_{\varepsilon\in\mathcal{E}}$ of* (A.1) *with frequencies $\{\omega^\varepsilon\}_{\varepsilon\in\mathcal{E}}$ fulfilling*

$$\sup_t \big\|x_\varepsilon(t) - \varepsilon\xi_1(t\omega^\varepsilon)\big\|_s \leqslant C\varepsilon^r, \quad \big|\omega^\varepsilon - \omega_1\big| \leqslant C\varepsilon^{r-1}. \tag{A.9}$$

Moreover, the set $\mathcal{E}$ is in one to one correspondence either with $W_\gamma \cap [\omega_1, \omega_1 + \omega_*)$ if $\beta_0 < 0$, or with $W_\gamma \cap (\omega_1 - \omega_*, \omega_1]$ if $\beta_0 > 0$.

PROOF. We consider only the case of odd r, the general case can be obtained by a slightly different treatment of the forthcoming equation ω. We are looking for an X_s-valued function $q(t)$ which is 2π-periodic and reversible (i.e. $q(t) = q(-t)$), and for a positive ω, close to ω_1, such that $q(\omega t)$ is a solution of (A.1). They must satisfy the equation

$$L_\omega q = g(q), \quad L_\omega := \omega^2 \frac{d^2}{dt^2} + A, \tag{A.10}$$

which will be considered as an ω-dependent functional equation in the space $\mathcal{H} \subset H^1(\mathbb{T}, X_s)$, formed by the reversible periodic functions. Equation (A.10) is studied using the Lyapunov–Schmidt decomposition, namely by decomposing it into an equation on $\operatorname{Ker} L_{\omega_1} \equiv \operatorname{span}(\xi_1)$ and an equation on its orthogonal complement R. Precisely, denote by Q the projector on ξ_1 and by P the projector on R and make the Ansatz $q = \varepsilon\xi_1 + \varepsilon^r u$, where $u \in R$. Then (A.10) is equivalent to the system

$$\omega^2 = \omega_1^2 + \beta\varepsilon^{r-1}, \tag{A.11}$$

$$L_\omega u = P g_r(\xi_1) + PG(\varepsilon, u), \tag{A.12}$$

$$-\beta\xi_1 = Q g_r(\xi_1) + QG(\varepsilon, u) \tag{A.13}$$

for the unknowns (ε, u, β). Here G contains all higher-order corrections and $\omega \in W_\gamma$ is a parameter. Equations (A.11), (A.12) and (A.13) are called the ω, the P and the Q equation, respectively.

First one solves the P equation (A.12). To this end one has to invert the linear operator $L_\omega|_R$. Its eigenfunctions are $\cos(lt)\varphi_j$, and the corresponding eigenvalues are

$$\lambda_{jl} = -l^2\omega^2 + \omega_j^2 = (l\omega + \omega_j)(\omega_j - l\omega), \quad j \geqslant 2,\ l \geqslant 1.$$

By (A.7), $|\lambda_{jl}| > C\gamma$. So $(L_\omega|_R)^{-1}$ exists and is bounded. Applying this operator to the P equation and using the implicit function theorem one obtains a smooth function $u(\varepsilon)$ that depends parametrically on $\omega \in W_\gamma$ and solves the P equation.

Inserting $u(\varepsilon)$ in the Q equation one determines the parameter β as a function of ε. In particular one has $\beta(\varepsilon) = C\beta_0 +$ higher-order corrections, where $C > 0$. Inserting $\beta(\varepsilon)$ in the ω equation one gets an equation for ε (remember that ω is fixed), which is a perturbation of the equation $\omega^2 - \omega_1^2 = C\beta_0\varepsilon^{r-1}$. By the non-degeneracy this can be reduced to a fixed point equation for ε^{r-1} which is solvable by the contraction mapping principle. □

REMARK A.4. The theorem holds also in the case $r = 2$, but in this case the non-degeneracy condition takes a more complicated form.

Theorem A.3 was proved in [5]. The technique of the Lyapunov–Schmidt decomposition was used for the first time to construct families of periodic solutions in PDEs by Craig and Wayne [29] who considered the model problem of the wave equation with periodic boundary conditions (see Example A.1); we will report on this work in Section A.4.

EXAMPLE A.5. Consider the non-linear wave equation with periodic boundary conditions (see Example A.1). Let ω_1 be such that $\omega_1 \neq \omega_j$ for each $j \neq 1$. Decompose V into its average a and a part $\widetilde{V}$ of zero average, then condition (H1) is satisfied if a belongs to an uncountable set which is dense in a neighbourhood of the origin (for the proof see Lemma 3.1 of [7]). Condition (H2) can be expressed in terms of the eigenfunctions of the Sturm–Liouville operator. If it holds, then Theorem A.3 applies and ensures persistence of the corresponding family of periodic orbits. Note that, in a difference with the case of Dirichlet boundary conditions (see Example 5.3), the non-linearity does not need to have some particular parity.

EXAMPLE A.6. Consider the non-linear plate equation (see Example A.2). In the case $d = 1$ all the frequencies are simple and the assumption (H1) is satisfied if a is chosen in a subset of $\mathbb{R}^+$ having full measure. In the case $d > 1$, all the frequencies are multiple except the smallest one. Taking for ω_1 the smallest frequency, (H1) is fulfilled if a belongs to a dense uncountable subset of $[0, 1/4]$. (H2) holds trivially provided the Taylor expansion of f at zero does not vanish identically (remember that $f(-w) = f(w)$). Then Theorem A.3 ensures persistence of the corresponding family of periodic orbits (for details see [7]).

A.3. *The resonant case*

It is possible to generalise the above theorem to the case when the frequencies satisfy some resonance relations. We will consider only the Lagrangian case, when $g = -\nabla H$.

Fix a frequency ω_1 of the linearised system. We replace the assumption (H1) by the following one:

(H1R) For any small enough γ there exists a closed set $W_\gamma \subset \mathbb{R}^+$ having ω_1 as an accumulation point both from the right and from the left, and such that for any $\omega \in W_\gamma$ one has

$$\text{either} \quad |\omega l - \omega_j| \geqslant \frac{\gamma}{l}, \quad \text{or} \quad l\omega_1 - \omega_j = 0. \tag{A.14}$$

To pass to the non-degeneracy assumption, we define the resonant set as

$$\mathcal{I}_R := \{k \geqslant 1\colon\ \exists l \geqslant 1\colon\ l\omega_1 - \omega_k = 0\}, \tag{A.15}$$

consider the linear space generated by $\{\varphi_k\}_{k\in\mathcal{I}_R}$, and denote by $\mathcal{N}$ its closure in the graph norm of $D(A)$. Note that all solutions of the linearised system with initial datum in $\mathcal{N}$ and vanishing initial velocity are periodic of period $2\pi/\omega_1$. Let H_r be the first non-vanishing Taylor coefficient of H. For $x \in \mathcal{N}$ define the average of H_r by

$$\langle H_r\rangle(x) := \frac{\omega_1}{2\pi}\int_0^{2\pi/\omega_1} H_r\big(\cos(At)x\big)\,dt.$$

Consider the hypersurface $\mathcal{S} \subset \mathcal{N}$ of the points $x \in \mathcal{N}$ such that $\langle x; Ax\rangle_0 = 1$.

(H2R) There exists a non-degenerate critical point x_0 of the functional $\langle H_r\rangle|_{\mathcal{S}}$. The corresponding Lagrange multiplier β_0 does not vanish.

Denote by $\xi_0(\omega_1 t)$ the solution of the linearised system with initial datum x_0 and vanishing initial velocity.

THEOREM A.7 [6]. *Suppose the assumptions* (H1R), (H2R) *hold. Then there exists a family of periodic solutions* $\{x_\varepsilon(t)\}_{\varepsilon\in\mathcal{E}}$ *of* (A.1) *with frequencies* ω^ε, *satisfying*

$$\sup_t \left\|x_\varepsilon(t) - \varepsilon\xi_0(t\omega^\varepsilon)\right\|_s \leqslant C\varepsilon^r, \quad \left|\omega^\varepsilon - \omega_1\right| \leqslant C\varepsilon^{r-1}. \tag{A.16}$$

The set $\mathcal{E}$ *has the same properties as in the non-resonant case.*

The proof is obtained by proceeding as in the non-resonant case. The only difference is that in this case the kernel of L_{ω_1} is no longer one-dimensional, but is isomorphic to $\mathcal{N}$ (the isomorphism being given by the map $x \mapsto \cos(At/\omega_1)x$). So the Q equation can be transformed into an equation in $\mathcal{N}$. The latter turns out to be a perturbation of the equation for the critical points of $\langle H_r\rangle|_{\mathcal{S}}$, and the non-degeneracy condition (H2R) allows to solve it by the implicit function theorem.

Applying the above theorem, one can construct countably many families of periodic solutions of the ϕ^4-model

$$w_{tt} - w_{xx} = \pm w^3 + \text{higher-order terms}$$

with Dirichlet boundary conditions, and also higher frequency periodic solutions of the non-linear plate equation of Example A.2 (see [6,7], see also [62,21]).

In general it is difficult to check condition (H2R). In the case of Hamiltonian systems with $n < \infty$ degrees of freedom, topological arguments allow to avoid it. Indeed, the Weinstein–Moser theorem (see [80,64]) ensures that close to a minimum of the energy, on each surface of a constant energy there exist at least n periodic orbit. In general they do not form regular families. A corresponding result for PDEs is not available at present. However there exists an *ad hoc* variational result for the wave equation

$$w_{tt} - w_{xx} = \pm w^p + \text{higher-order terms}, \quad p \geqslant 2, \tag{A.17}$$

which ensures that, having fixed $j \geqslant 1$, there exists a sequence of periodic orbits accumulating at zero, whose frequencies accumulate at j (which plays here the role of the jth linear frequency). The corresponding theorem is due to Berti and Bolle [10].

Periodic solutions in the non-linear wave equation

$$w_{tt} - w_{xx} + f(x, w) = 0, \qquad u(0, t) = u(\pi, t) = 0, \tag{A.18}$$

where constructed for the first time by Rabinowitz [76] using global variational methods and a Lyapunov–Schmidt decomposition. Rabinowitz proved that, under suitable assumptions on f, Equation (A.18) has at least one periodic solution with period $T = 2\pi p/q$, for any choice of the integers p and q. Note that, when the period T is commensurable with

2π, the operator $L_\omega|_R$ has a compact inverse, i.e. there are no small denominators. The work [76] was followed by a series of papers, simplifying the proof and sharpening the result (see [24] and references therein). In particular, we mention the paper [25] by Brezis, Coron and Nirenberg, where existence of periodic orbits is proved by a particularly simple method: the authors write a variational principle, dual to the usual one, and look for its critical points, using the mountain pass lemma. It is remarkable that in this approach the Q equation becomes trivial.

A.4. *Weakening the non-resonance condition*

The main limitation of the results presented in Sections A.2 and A.3 rests in the non-resonance conditions (H1) and (H1R). Indeed, such conditions are fulfilled with large probability (in a suitable parameter space) when $\omega_j \sim j^\nu$ with $\nu > 1$; when $\nu = 1$ the non-resonance conditions are satisfied typically on uncountable sets of zero measure, but when $\nu < 1$ they are satisfied only exceptionally (as in the plate equation). As a consequence the results of Sections A.2 and A.3 are not applicable to general equations in more than one space dimensions. Furthermore, the method of Lyapunov–Schmidt decomposition can be extended to the case of reversible systems of first order in time, but the approach of Section A.2 is no more applicable.

In order to avoid such limitations one would like to be able to work with the weaker non-resonance condition "there exists a $\tau > 0$ such that $|l\omega - \omega_j| \geqslant \gamma/l^\tau$". This was done by Craig and Wayne [29] who used the Nash–Moser theorem to solve the P equation. The application of the Nash–Moser theorem requires to construct and estimate the inverse of the linear operator describing the linearisation of the P equation at an approximate solution. This is the main difficulty of Craig–Wayne's approach. To overcome it they use the techniques by Fröhlich and Spencer [38], performing a careful analysis of small denominators (cf. Section 5.3). The method by Craig and Wayne was extended by Bourgain in order to construct periodic (and also quasiperiodic) solutions in higher-dimensional equations. The resulting method seems very general, but at present a theorem "ready for application" is not available. We present here the result obtained by Bourgain by applying this method to the non-linear wave equation

$$w_{tt} - \Delta w + aw + w^3 = 0 \tag{A.19}$$

on $\mathbb{T}^d$. Fix a multiindex $n \in \mathbb{Z}^d$ different from zero, and let

$$\xi_n(\omega_n t, x) := \cos(n \cdot x + \omega_n t), \quad \omega_n := \sqrt{n_1^2 + \cdots + n_d^2 + a},$$

be the corresponding symmetric reversible solution.

THEOREM A.8 [18]. *If a belongs to a certain subset of $\mathbb{R}^+$ of full measure, then there exists a Cantor set $\mathcal{E}$ of positive measure, accumulating at zero, and a family of periodic solutions $\{w_\varepsilon(t,x)\}_{\varepsilon\in\mathcal{E}}$ of* (A.19) *with frequencies ω^ε, satisfying*

$$\left|\varepsilon\xi_n(\omega^\varepsilon t, x) - w_\varepsilon(t,x)\right| \leqslant C\varepsilon^3, \quad \left|\omega_n - \omega^\varepsilon\right| \leqslant C\varepsilon^2.$$

In the case $d = 1$, the result was proved in [29]; subsequently, still in the case $d = 1$, Kuksin introduced a simpler technique to find the "large measure result" of Theorem A.8 (see in [20, pp. 90–94]).

The Craig–Wayne–Bourgain method also allows to deal with first order in time equations. For example, it was applied to the Schrödinger equation in one [30] or two space dimensions [19] (see Section 5.4).

A.5. *The water wave problem*

A particular problem that has attracted the attention of many researchers since the very beginning of the theory of PDEs is that of existence of standing water waves. The first rigorous proof of their existence was obtained only recently by Plotnikov and Toland [70]; we present here their result.

Consider a perfect fluid lying above a horizontal bottom, and confined between two parallel vertical walls. The fluid is subject to gravity, and atmospheric pressure acts at the free surface. This is a dynamical system governed by the Euler equations supplemented by appropriate boundary conditions. It was pointed out by Zakharov that this system is Hamiltonian (see [81]). The corresponding Hamiltonian function is the energy of the fluid, and conjugated variables are given by the wave profile and the velocity potential at the free surface.

In the linear approximation the general solution is given by the superposition of the normal modes. The problem is to continue the normal modes to families of periodic solutions of the non-linear system (the standing waves). Fix one of the normal modes, and denote by $\eta(t, x_1)$ the corresponding profile of the free surface (x_1 being the horizontal variable). Then it is possible to choose the depth h, the width l of the region occupied by the fluid and the gravitational constant g in such a way that the period of the solution is normalised to 2π and the linear frequencies fulfil a suitable non-resonance condition. Denote by (g_0, l_0, h_0) a choice of the parameters realising such conditions, then one has

THEOREM A.9 [70]. *There exists an infinite set $\mathcal{E} \subset \mathbb{R}$ having zero as an accumulation point and, for any $\varepsilon \in \mathcal{E}$, there exist g_ε, l_ε and a standing wave solution of the water wave problem with gravity g_ε in a box of width l_ε. Moreover, denoting by η_ε the corresponding profile of the free surface, one has*

$$\left|\eta_\varepsilon(t, x_1) - \varepsilon^2\eta(t, x_1)\right| < C\varepsilon^3, \qquad |g_\varepsilon - g_0| + |l_\varepsilon - l_0| \leqslant C\varepsilon.$$

The main difficulties in proving this result are as follows: firstly, the linear frequencies behave as $\omega_n \sim n^{1/2}$, so the non-resonance conditions that can be satisfied are quite weak. Secondly, the mathematical formulation of the problem involves an unbounded non-linear and non-local operator. To overcome these difficulties, Plotnikov and Toland use the Lagrangian description of the fluid motion and apply the Lyapunov–Schmidt approach to handle the resulting non-linear problem. The P equation now is solved by means of the Nash–Moser theorem. The required invertibility of the linearised operator is obtained in two steps: first it is reduced to a suitable canonical form, and next this canonical form (which is essentially a perturbation of an operator involving derivatives and Hilbert transform) is studied in detail.

References

[1] V.I. Arnold, *Proof of a theorem of A.N. Kolmogorov on the conservation of quasiperiodic motions under a small change of the Hamiltonian function*, Russian Math. Surveys **18** (5) (1963), 9–36.

[2] V.I. Arnold, *Mathematical Methods in Classical Mechanics*, 3rd edn, Springer, Berlin (1989).

[3] D. Bambusi, *Nekhoroshev theorem for small amplitude solutions in nonlinear Schrödinger equation*, Math. Z. **130** (1999), 345–387.

[4] D. Bambusi, *On long time stability in Hamiltonian perturbations of non-resonant linear PDEs*, Nonlinearity **12** (1999), 823–850.

[5] D. Bambusi, *Lyapunov center theorem for some nonlinear PDEs: A simple proof*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. **29** (2000), 823–837.

[6] D. Bambusi and S. Paleari, *Families of periodic orbits for resonant PDE's*, J. Nonlinear Sci. **11** (2001), 69–87.

[7] D. Bambusi and S. Paleari, *Families of periodic orbits for some PDE's in higher dimensions*, Comm. Pure Appl. Anal. **1** (2002), 269–279.

[8] E.D. Belokolos, A.I. Bobenko, V.Z. Enolskii, A.R. Its and V.B. Matveev, *Algebro-Geometric Approach to Nonlinear Integrable Equations*, Springer, Berlin (1994).

[9] G. Benettin, F. Fasso and M. Guzzo, *Nekhoroshev stability of elliptic equilibra of Hamiltonian systems*, Comm. Math. Phys. **197** (1998), 347–360.

[10] M. Berti and P. Bolle, *Periodic solutions of nonlinear wave equations with general nonlinearities*, Comm. Math. Phys. **243** (2003), 315–328.

[11] R.F. Bikbaev and S.B. Kuksin, *A periodic boundary-value problem for the Sine-Gordon equation, small Hamiltonian perturbations of it, and KAM-deformations of finite-gap tori*, St.-Petersburg Math. J. **4** (1993), 439–468.

[12] A.I. Bobenko and S.B. Kuksin, *Finite-gap periodic solutions of the KdV equation are nondegenerate*, Phys. Lett. A **161** (3) (1991), 274–276.

[13] A.I. Bobenko and S.B. Kuksin, *The nonlinear Klein–Gordon equation on an interval as a perturbed Sine-Gordon equation*, Comment. Math. Helv. **70** (1995), 63–112.

[14] A.I. Bobenko and S.B. Kuksin, *Small-amplitude solutions of the Sine-Gordon equation on an interval under Dirichlet or Neumann boundary conditions*, J. Nonlinear Sci. **5** (1995), 207–232.

[15] J. Bourgain, *Fourier transform restriction phenomenon for certain lattice subsets and applications to nonlinear evolution equations*, Geom. Funct. Anal. **3** (1993), 107–156 and 209–262.

[16] J. Bourgain, *Construction of quasi-periodic solutions for Hamiltonian perturbations of linear equations and applications to nonlinear PDE*, Internat. Math. Res. Notices (1994), 475–497.

[17] J. Bourgain, *Aspects of long time behaviour of solutions of nonlinear Hamiltonian evolution equations*, Geom. Funct. Anal. **5** (1995), 105–140.

[18] J. Bourgain, *Construction of periodic solutions of nonlinear wave equations in higher dimension*, Geom. Funct. Anal. **5** (1995), 629–639.

[19] J. Bourgain, *Quasi-periodic solutions of Hamiltonian perturbations of 2D linear Shrödinger equation*, Ann. Math. **148** (1998), 363–439.

[20] J. Bourgain, *Nonlinear Schrödinger Equations*, Hyperbolic Equations and Frequency Interactions, Amer. Math. Soc. (1999).

[21] J. Bourgain, *Periodic solutions of nonlinear wave equations*, Harmonic Analysis and Partial Differential Equations, Chicago University Press (1999), 69–97.

[22] J. Bourgain, *On diffusion in high-dimensional Hamiltonian systems and PDE*, J. Anal. Math. **80** (2000), 1–35.

[23] J. Bourgain, *Green's Function Estimates for Lattice Schrödinger Operators and Applications*, Ann. Math. Studies, Princeton University Press, Princeton (2004).

[24] H. Brezis, *Periodic solutions of nonlinear vibrating string and duality principle*, Bull. Amer. Math. Soc. **8** (1983), 409–426.

[25] H. Brezis, J. Coron and L. Nirenberg, *Free vibrations for a nonlinear wave equation and a theorem by P. Rabinowitz*, Comm. Pure Appl. Math. **33** (1980), 667–689.

[26] D. Cai, D.W. McLaughlin and K.T.R. McLaughlin, *The nonlinear Schrödinger equation as both a PDE and a dynamical system*, Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., Elsevier, Amsterdam (2002).

[27] L. Chierchia and J. You, *KAM tori for 1D nonlinear wave equations with periodic boundary conditions*, Comm. Math. Phys. **211** (2000), 497–525.
[28] W. Craig, *Problèmes de petitts diviseurs dans les équations aux dérivées partielles*, Panoramas et Synthéses, no. 9, Société Mathématique de France (2000).
[29] W. Craig and C.E. Wayne, *Newton's method and periodic solutions of nonlinear wave equations*, Comm. Pure Appl. Math. **46** (1993), 1409–1498.
[30] W. Craig and C.E. Wayne, *Periodic solutions of nonlinear Schrödinger equations and the Nash–Moser method*, Hamiltonian Mechanics. Integrability and Chaotic Behavior, NATO ASI, Vol. B331, Plenum Press (1994), 103–122.
[31] B.A. Dubrovin, *Theta-functions and nonlinear equations*, Russian Math. Surveys **36** (2) (1981), 11–80.
[32] B.A. Dubrovin, V.B. Matveev and S.P. Novikov, *Nonlinear equations of Korteweg–de Vries type, finite zone linear operators, and Abelian varieties*, Russian Math. Surveys **31** (1) (1976), 55–135.
[33] L.H. Eliasson, *Perturbations of stable invariant tori*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. IV Ser. **15** (1988), 115–147.
[34] L.H. Eliasson and S.B. Kuksin, *KAM for NLS* (2005), Preprint.
[35] L.H. Eliasson, S.B. Kuksin, S.Marmi and J.-C. Yoccoz, *Dynamical systems and small divisors*, Lecture Notes in Math., Vol. 1784, ch. KAM-persistence of finite-gap solutions, Springer, Berlin (2002).
[36] E. Fermi, J.R. Pasta and S.M. Ulam, *Studies of nonlinear problems*, Collected works of E. Fermi, Vol. 2, Chicago University Press, Chicago (1965).
[37] L. Friedlander, *An invariant measure for the equation* $u_{tt} - u_{xx} + u^3 = 0$, Comm. Math. Phys. **98** (1985), 1–16.
[38] J. Fröhlich and T. Spencer, *Absence of diffusion in Anderson tight binding model for large disorder or low energy*, Comm. Math. Phys. **88** (1983), 151–184.
[39] J. Fröhlich, T. Spencer and C.E. Wayne, *Localization in disordered nonlinear dynamical systems*, J. Statist. Phys. **42** (1986), 247–274.
[40] G.E. Giacaglia, *Perturbation Methods in Non-Linear Systems*, Springer, Berlin (1972).
[41] B. Grébert and T. Kappeler, *Perturbation of the defocusing nonlinear Schrödinger equation*, Milan J. Math. **71** (2003), 141–174.
[42] M. Gromov, *Pseudoholomorphic curves in symplectic manifolds*, Invent. Math. **82** (1985), 307–347.
[43] H. Hofer and E. Zehnder, *Symplectic Invariants and Hamiltonian Dynamics*, Birkhäuser, Basel (1994).
[44] T. Kappeler, *Fibration of the phase-space for the Korteweg–de Vries equation*, Ann. Inst. Fourier **41** (1991), 539–575.
[45] T. Kappeler and M. Makarov, *On Birkhoff coordinates for KdV*, Ann. H. Poincaré **2** (2001), 807–856.
[46] T. Kappeler and J. Pöschel, *KAM & KdV*, Springer (2003).
[47] T. Kato, *Quasi-linear equations of evolutions, with applications to partial differential equations*, Lecture Notes in Math., Vol. 448, Springer, Berlin (1975), 25–70.
[48] I.M. Krichever and D.H. Phong, *Symplectic forms in the theory of solitons*, Surv. Differ. Geom., Vol. IV, Int. Press, Boston (1998), 239–313.
[49] S.B. Kuksin, *Hamiltonian perturbations of infinite-dimensional linear systems with an imaginary spectrum*, Funct. Anal. Appl. **21** (1987), 192–205.
[50] S.B. Kuksin, *Perturbations of quasiperiodic solutions of infinite-dimensional Hamiltonian Systems*, Izv. Akad. Nauk SSSR Ser. Mat. **52** (1988), 41–63: English transl. in Math. USSR-Izv. **32** (1) (1989).
[51] S.B. Kuksin, *The perturbation theory for the quasiperiodic solutions of infinite-dimensional Hamiltonian systems and its applications to the Korteweg–de Vries equation*, Math. USSR-Sb. **64** (1989), 397–413.
[52] S.B. Kuksin, *Nearly Integrable Infinite-Dimensional Hamiltonian Systems*, Springer, Berlin (1993).
[53] S.B. Kuksin, *KAM-theory for partial differential equations*, Proceedings of the First European Congress of Mathematics, Vol. 2, Birkhäuser (1994), 123–157.
[54] S.B. Kuksin, *Infinite-dimensional symplectic capacities and a squeezing theorem for Hamiltonian PDEs*, Comm. Math. Phys. **167** (1995), 531–552.
[55] S.B. Kuksin, *Growth and oscillations of solutions of nonlinear Schrödinger equation*, Comm. Math. Phys. **178** (1996), 265–280.
[56] S.B. Kuksin, *Oscillations in space-periodic nonlinear Schrödinger equations*, Geom. Funct. Anal. **7** (1997), 338–363.

[57] S.B. Kuksin, *A KAM-theorem for equations of the Korteweg–de Vries type*, Rev. Math. Math. Phys. **10** (3) (1998), 1–64.
[58] S.B. Kuksin, *Spectral properties of solutions for nonlinear PDEs in the turbulent regime*, Geom. Funct. Anal. **9** (1999), 141–184.
[59] S.B. Kuksin, *Analysis of Hamiltonian PDEs*, Oxford University Press, Oxford (2000).
[60] S.B. Kuksin and J. Pöschel, *Invariant Cantor manifolds of quasi-periodic oscillations for a nonlinear Schrödinger equation*, Ann. Math. **143** (1996), 149–179.
[61] P.D. Lax, *Periodic solutions of the KdV equations*, Comm. Pure Appl. Math. **28** (1975), 141–188.
[62] B.V. Lidskij and E.I. Shulman, *Periodic solutions of the equation* $u_{tt} - u_{xx} + u^3 = 0$, Funct. Anal. Appl. **22** (1988), 332–333.
[63] P. Lochak, *Canonical perturbation theory via simultaneous approximation*, Russian Math. Surveys **47** (6) (1992), 57–133.
[64] J. Moser, *Periodic orbits near an equilibrium and a theorem by Alan Weinstein*, Comm. Pure Appl. Math. **29** (1976), 724–747.
[65] J. Moser and C.L. Siegel, *Lectures on Celestial Mechanics*, Springer, Berlin (1971).
[66] N.N. Nekhoroshev, *Exponential estimate of the stability of near integrable Hamiltonian systems*, Russian Math. Surveys **32** (6) (1977), 1–65.
[67] L. Niederman, *Nonlinear stability around an elliptic equilibrium point in an Hamiltonian system*, Nonlinearity **11** (1998), 1465–1479.
[68] S.P. Novikov, *A periodic problem for the Korteweg–de Vries equation, I*, Funct. Anal. Appl. **8** (1974), 236–246.
[69] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, Berlin (1983).
[70] P. Plotnikov and J. Toland, *Nash–Moser theory for standing water waves*, Arch. Rational Mech. Anal. **159** (2001), 1–83.
[71] J. Pöschel, *On elliptic lower dimensional tori in Hamiltonian systems*, Math. Z. **202** (1989), 559–608.
[72] J. Pöschel, *Small divisors with spatial structure in infinite dimensional Hamiltonian systems*, Comm. Math. Phys. **127** (1990), 351–393.
[73] J. Pöschel, *A KAM-theorem for some nonlinear PDEs*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. IV Ser. 15 **23** (1996), 119–148.
[74] J. Pöschel, *Quasi-periodic solutions for a nonlinear wave equation*, Comment. Math. Helv. **71** (1996), 269–296.
[75] J. Pöschel, *On Nekhoroshev estimates for a nonlinear Schrödinger equation and a theorem by Bambusi*, Nonlinearity **12** (1999), 1587–1600.
[76] P. Rabinowitz, *Free vibrations for a semilinear wave equation*, Comm. Pure Appl. Math. **31** (1978), 31–68.
[77] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vol. 2, Academic Press, New York (1975).
[78] M.B. Sevryuk, *The classical KAM theory at the dawn of the twenty-first century*, Moscow Math. J. **3** (3) (2003), 1113–1144.
[79] C.E. Wayne, *Periodic and quasi-periodic solutions of nonlinear wave equations via KAM theory*, Comm. Math. Phys. **127** (1990), 479–528.
[80] A. Weinstein, *Normal modes for nonlinear Hamiltonian systems*, Invent. Math. **20** (1973), 47–57.
[81] V.E. Zakharov, *Stability of periodic waves of finite amplitude on the surface of a deep fluid*, Appl. Mech. Tech. Phys. **2** (1968), 190–194.
[82] V.E. Zakharov, M.F. Ivanov and L.N. Shur, *On the abnormally slow stochastisation in some two-dimensional field theory models*, JETP Lett. **30** (1) (1979), 39–44.
[83] V.E. Zakharov, S.V. Manakov, S.P. Novikov and L.P. Pitaevskij, *Theory of Solitons*, Plenum Press, New York (1984).
[84] P.E. Zhidkov, *Korteweg–de Vries and Nonlinear Schrödinger Equations: Qualitative Theory*, Springer, Berlin (2001).

This page intentionally left blank

CHAPTER 16

Extended Hamiltonian Systems

M.I. Weinstein

Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027
E-mail: miw2103@columbia.edu

Contents

HANDBOOK OF DYNAMICAL SYSTEMS, VOL. 1B
Edited by B. Hasselblatt and A. Katok

1. Introduction

In this chapter we discuss Hamiltonian partial differential wave equations which are defined on unbounded spatial domains, a class of so-called *extended Hamiltonian systems*. The examples we consider are the nonlinear Schrödinger and Klein–Gordon equations defined on $\mathbb{R}^3$. These may be viewed as infinite-dimensional Hamiltonian systems, which have coherent solutions, e.g., spatially uniform equilibria, spatially nonuniform solitary standing waves.... Questions of interest include the dynamics in a neighborhood of these states (stability to small perturbations), stability under small Hamiltonian perturbations of the dynamical system, the behavior of solutions on short, intermediate and infinite time scales and the manner in which these coherent states participate in the structure of solutions on these time scales.

The contrast in dynamics between Hamiltonian systems of extended type and those of compact type is striking. Compact Hamiltonian systems arising, for example, from finite-dimensional Hamiltonian systems or Hamiltonian partial differential equations (PDEs) governing an evolutionary process defined on a bounded spatial domain, are systems governed by finite or infinite systems of ordinary differential equations (ODEs) with a *discrete* set of frequencies. Many fundamental phenomena and questions here involve the persistence or breakdown of regular (e.g., time periodic or quasiperiodic) solutions and their dynamical stability relative to small perturbations. A stable state of the system is one around which neighboring trajectories oscillate. KAM theory implies states persist in the presence of small Hamiltonian perturbations (structural stability) provided certain arithmetic *nonresonance* conditions on the set of frequencies of the unperturbed state hold [1,27,11,3].

In contrast, extended Hamiltonian systems arising from Hamiltonian PDEs are systems involving continuous as well as discrete spectra of frequencies. Stable states are expected to be *asymptotically stable*; states initially nearby the unperturbed state remain close and even converge to it in an appropriate metric. Since the flow is in an infinite-dimensional space, this does not contradict the Hamiltonian character of the phase flow, which in finite-dimensional spaces preserves volume. Convergence to an asymptotic state occurs through a mechanism of radiating energy to infinity. It is also possible that some states of the system are long-lived *metastable states*. These are states which persist on long time scales, but decay as $t \to \infty$. This structural instability due to Hamiltonian perturbations occurs due to nonlinearity induce resonances of states associated with discrete and continuous spectra, precisely that which is precluded in the setting of KAM theory.

2. Overview

We consider partial differential equations for which the linear part (the small amplitude limit) has spatially localized and time-periodic "bound state" solutions, which are dynamically stable. Such solutions of the linear dynamical system are associated with the discrete spectrum of linear self-adjoint operator generating the flow. Also associated with this operator, due to the unboundedness of the spatial domain, is continuous spectrum with corresponding spatially extended (nondecaying) radiation states. These bound and radiation states are central to the linear dynamics. Arbitrary finite energy initial conditions can,

by the spectral theorem for self-adjoint operators, be decomposed into a superposition of discrete and continuous spectral states. Their amplitudes evolve with time according to an infinite system of decoupled linear ordinary differential equations. The discrete component of the solution is quasiperiodic in time and localized in space, while the continuous spectral component disperses to zero (e.g., in a local L^2 sense) as time advances. We seek to understand the dynamics in the *weakly nonlinear regime*, the regime where nonlinearity is present and where the initial data are small in an appropriately chosen norm.

Except in very special cases of integrable systems where, in an appropriate "nonlinear basis", bound (soliton) and radiation states evolve decoupled from one another, nonlinearity induces coupling and exchange of energy among bound and radiation states. It is this situation which interests us and we consider as examples the following two nonlinear wave equations of Hamiltonian type: the nonlinear Schrödinger equation (NLS) and the nonlinear Klein–Gordon equation (NLKG)[1]

$$\text{NLS} \qquad i\partial_t \Phi = \big(-\Delta + V(x)\big)\Phi + g|\Phi|^2\Phi,$$
$$\Phi(t,x) \in \mathbb{C},\ (t,x) \in \mathbb{R}^1 \times \mathbb{R}^3, \tag{2.1}$$
$$\text{NLKG} \quad \big(\partial_t^2 - \Delta + m^2 + V(x)\big)u = gu^3,$$
$$u = u(t,x) \in \mathbb{R},\ (t,x) \in \mathbb{R}^1 \times \mathbb{R}^3. \tag{2.2}$$

The nonlinear coupling coefficient, g, is real and taken to be either zero or of order unity. The particular nonlinear Schrödinger equation, (2.1), with a nontrivial potential is also called the Gross–Pitaevskii equation (G–P). Applications, especially of the nonlinear Schrödinger equation, abound. These range from the fundamental physics of Bose–Einstein condensation [29,16] to nonlinear optics, e.g., nonlinear optical pulse propagation in inhomogeneous media [30,19,18].

The remainder of the chapter is outlined as follows:

- In Section 3 we shall introduce solitary wave solutions of the nonlinear Schrödinger equation, (2.1).
- We then discuss a variational approach to H^1 orbital Lyapunov stability of solitary waves in Section 4.
- The detailed behavior of solutions containing solitary wave components requires a detailed understanding of spectral properties of the solitary wave. The linearization about a stable solitary wave may have (a) discrete spectrum consisting of eigenstates with zero frequency associated with the equation's symmetries, and neutral ("internal") oscillatory eigenstates and (b) continuous spectrum associated with spatially extended radiation states. The presence or absence of these neutral oscillatory states, a property which is not derivable from the variational characterization the solitary wave, has an important effect on the dynamics on all time scales. Section 5 contains a discussion of asymptotic stability of solitary waves in the simple case where there are no neutral oscillations.

[1] Two other examples are cited at the end of this section. We have not attempted a comprehensive survey in this chapter.

- The case where there are neutral oscillatory eigenstates is considerably more rich in phenomena and requires deeper mathematical study. We shall see that these latter states typically decay to zero on very large time scales. The mechanism for decay is nonlinearity induced resonance of discrete and continuum states; the continuous spectral modes act as an effective dissipative heat bath with computable dissipation rate. In Section 6 the ideas and methods of analysis are introduced in the simpler context of the nonlinear Klein–Gordon equation. Then, in Section 7 we return to NLS/G–P to study the weakly nonlinear regime of multimode nonlinear Schrödinger equations.

A theme throughout this article is that we view each PDE as a Hamiltonian system comprised of two subsystems: (a) a finite-dimensional subsystem describing the evolution of coherent spatially localized states and (b) an infinite-dimensional part, governing the radiation of energy to spatial infinity. For a general small norm initial condition, the solution has different behaviors on different time scales: "initial phase", "large but finite time" and "infinite time". This behavior is elucidated by derivation of an appropriate *normal form*, which makes explicit the key mechanism for energy transfer among bound and radiation states. *The direction of energy flow is an emerging property, a consequence of the initial condition being localized, resonant coupling of bound to dispersive waves due to nonlinearity and the property of local decay of dispersive waves.*

There are close connections between these phenomena and their analysis with the computation of lifetimes of quantum states (transition theory), the perturbation theory of embedded eigenvalues in continuous spectra, and parametrically forced Hamiltonian systems; see, for example, [42–44,24,25,38,39].

In conclusion we remark that the asymptotic stability and scattering of coherent structures for infinite volume Hamiltonian systems has been considered in other contexts as well. Two important other studies are (i) the long time dynamics resulting from a classical particle interacting with a scalar wave field; see, for example, [26] and (ii) the stability of the Minkowski metric for the Einstein equations of the gravitational field [10].

3. Linear and nonlinear bound states

In this section we introduce bound states of the linear ($g = 0$) and nonlinear ($g \neq 0$) Schrödinger equation (2.1).

Bound states of the unperturbed problem

Let $H = -\Delta + V(x)$. We assume that $V(x)$ is smooth, real-valued and sufficiently rapidly decaying, so that H defines a self-adjoint operator in L^2. Additionally, we assume that the spectrum of H consists of continuous spectrum extending from zero to positive infinity and two discrete negative eigenvalues, each of multiplicity one[2]:

$$\sigma(H) = \{E_{0*}, E_{1*}\} \cup [0, \infty).$$

[2]The general case of any finite number of bound states can be considered as well. The case of $m \leqslant 2$ bound states captures all key phenomena we wish to discuss and keeps the presentation as simple as possible.

Therefore, there exist eigenvalues E_{j*}, with smooth, square-integrable normalized eigenfunctions, $\psi_{j*}(x)$, $j = 0, 1$, such that

$$H\psi_{j*} = E_{j*}\psi_{j*}, \quad \langle \psi_{j*}, \psi_{k*} \rangle = \delta_{jk}. \tag{3.1}$$

We also introduce spectral projections onto the discrete eigenstates and continuous spectral part of H, respectively:

$$P_{j*}f \equiv \langle \psi_{j*}, f \rangle \psi_{j*}, \quad j = 0, 1,$$
$$P_{c*} \equiv I - P_{0*} - P_{1*}.$$

Nonlinear bound states

We seek solutions of (2.1) of the form

$$\phi = e^{-iEt}\Psi_E.$$

Substitution into (2.1) yields the following elliptic problem for the bound states of NLS

$$H\Psi_E + g|\Psi_E|^2\Psi_E = E\psi_E, \quad \psi_E \in H^1. \tag{3.2}$$

Note that if Ψ_E is any solution of (3.2) then for any $\theta \in \mathbb{R}$, $\Psi_E e^{i\theta}$ is a solution.

THEOREM 3.1 [36]. *For each $j = 0, 1$ we have a one-parameter family, bound states depending on the complex parameter α_j and defined for $|\alpha_j|$ sufficiently small*:

$$\Psi_{\alpha_j}(x) \equiv \alpha_j\big(\psi_{j*}(x) + \mathcal{O}(|\alpha_j|^2)\big),$$
$$E_j = E_{j*} + \mathcal{O}(|\alpha_j|^2).$$

For α_j complex and $|\alpha_j|$ small, the set $\{\Psi_{\alpha_0}\}$ is called the nonlinear ground state family and $\{\Psi_{\alpha_j}\colon\ j \geqslant 1\}$, the family of nonlinear excited states. Below, we shall also use the notation Ψ_{E_j} to denote a real-valued nonlinear bound state and parametrize the family of states by $\Psi_{E_j}(x)e^{i\theta}$, $\theta \in \mathbb{R}$.

The proof uses standard bifurcation theory [32], which is based on the implicit function theorem. The analysis extends to the case of nonlocal nonlinearities.

In what follows we shall "time-modulate" these bound states. For convenience, we shall use the notation: $\Psi_j(t, x) = \Psi_{\alpha_j(t)}$ and $E_j(t) = E_j(|\alpha_j(t)|^2)$.

An alternative approach to the construction of nonlinear bound states is by *variational methods*. The variational characterization is of particular interest in the case of the ground state, due to its role in establishing its *dynamic* stability. Our point of departure for the

variational approach is the observation that NLS has the following two conserved integrals, which are constant in time on solutions of NLS:

$$\mathcal{H}[\Phi] \equiv \int |\nabla \Phi|^2 + V(x)|\Phi|^2 + \frac{1}{2} g|\Phi|^4 \, dx, \tag{3.3}$$

$$\mathcal{N}[\Phi] = \int |\Phi|^2 \, dx, \tag{3.4}$$

$\mathcal{H}$ is a Hamiltonian, which generates NLS;

$$i\partial_t \Phi = \frac{\delta \mathcal{H}[\Phi, \Phi^*]}{\delta \Phi^*}.$$

Its time-invariance for the NLS flow is associated with time-translation symmetry, while the time-invariance of $\mathcal{N}$ is related to the phase symmetry $\Phi \mapsto \Phi e^{i\gamma}$, $\gamma \in \mathbb{R}$.

Nonlinear bound states of NLS are H^1 solutions, $F_\star$, of the elliptic equation

$$HF_\star + g|F_\star|^2 F_\star = E_\star F_\star, \tag{3.5}$$

for some choice of E and can also be viewed as critical points of the functional

$$J_E[f] \equiv \mathcal{H}[F] + E\mathcal{N}[F]. \tag{3.6}$$

That is, we have that the first variation of $J_{E_\star}$ vanishes at $E_\star$, i.e. $\delta J_{E_\star}[F_\star] = 0$.

The nonlinear ground state has a characterization as a constrained minimizer

$$I_\theta = \inf \left\{ \mathcal{H}[F] \colon\ F \in H^1,\ \mathcal{N}[F] = \theta \right\}. \tag{3.7}$$

For θ small, the minimum in (3.7) is attained at the ground state obtained in Theorem 3.1. The value of the frequency parameter, E, of a ground state Ψ_E depends on θ.

4. Orbital stability of ground states

In this section we discuss the orbital Lyapunov stability of ground states of NLS. Note that by the phase invariance of NLS, we have that the *orbit* of the ground state

$$\mathcal{O}_{\text{gs}} = \left\{ \Psi_{E_0}(x) e^{i\gamma} \colon\ \gamma \in [0, 2\pi) \right\} \tag{4.1}$$

is a one-parameter family of ground states. The ground state is stable in the following sense. If initially $\Phi(x, t = 0)$ is H^1 close to some phase-translate of Ψ_{E_0} then, for all $t \neq 0$, $\Phi(x, t)$ is H^1 close to some (typically t dependent) phase-translate of Ψ_{E_0}. In order to make this precise, we introduce a metric which measures the distance from an arbitrary H^1 function to the ground state orbit:

$$\text{dist}(u, \mathcal{O}_{\text{gs}}) = \inf_\gamma \left\| u - \Psi_{E_0} e^{i\gamma} \right\|_{H^1}. \tag{4.2}$$

Thus a more precise statement of stability is as follows. For any $\varepsilon > 0$ there is a $\delta > 0$ such that if

$$\operatorname{dist}\big(\Phi(\cdot,0), \mathcal{O}_{\mathrm{gs}}\big) < \delta \tag{4.3}$$

then for all $t \neq 0$

$$\operatorname{dist}\big(\Phi(\cdot,t), \mathcal{O}_{\mathrm{gs}}\big) < \varepsilon.$$

The proof of stability is now sketched. Let ε be an arbitrary positive number. We have for $t \neq 0$, by choosing δ in (4.3) sufficiently small

$$\begin{aligned}
\varepsilon^2 &\sim J_{E_0}\big[\Phi(\cdot,0)\big] - J_{E_0}[\Psi_0] \\
&= J_{E_0}\big[\Phi(\cdot,t)\big] - J_{E_0}[\Psi_0] \quad \text{by conservation laws} \\
&= J_{E_0}\big[\Phi(\cdot,t)e^{i\gamma}\big] - J_{E_0}[\Psi_0] \quad \text{by phase invariance} \\
&= J_{E_0}\big[\Psi_0 + u(\cdot,t) + iv(\cdot,t)\big] - J_{E_0}[\Psi_0] \\
&\quad \text{(definition of the perturbation } u + iv,\ u, v \in \mathbb{R}) \\
&\sim \big(L_+u(t), u(t)\big) + \big(L_-v(t), v(t)\big) \\
&\quad \text{(by Taylor expansion and } \delta J_{E_0}[\Psi_0] = 0).
\end{aligned} \tag{4.4}$$

The operators L_+ and L_- are, respectively, the real and imaginary parts of the second variational derivative of J_{E_0}, the linearized operator about the ground state. If L_+ and L_- were positive definite operators, implying the existence of positive constants C_+ and C_- such that

$$(L_+u, u) \geqslant C_+\|u\|_{H^1}^2, \tag{4.5}$$

$$(L_-v, v) \geqslant C_-\|v\|_{H^1}^2 \tag{4.6}$$

for all $u, v \in H^1$, then it would follow from (4.4) that the perturbation about the ground state, $u(x,t) + iv(x,t)$, would remain of order ε in H^1 for all time $t \neq 0$. The situation is however considerably more complicated. The relevant facts to note are as follows.

(1) $L_-\Psi_{E_0} = 0$, with $\Psi_{E_0} > 0$. Hence, Ψ_{E_0} is the ground state of L_-, $0 \in \sigma(L_-)$, and L_- is nonnegative with continuous spectrum $[|E_0|, \infty)$.

(2) For small L^2 nonlinear ground states, L_+ has exactly one strictly negative eigenvalue and continuous spectrum $[|E_0|, \infty)$.

The zero eigenvalue of L_- and the negative eigenvalue of L_+ constitute two *bad* directions, which are treated as follows, noting that $u(\cdot,t)$ and $v(\cdot,t)$ are not arbitrary H^1 functions but are rather constrained by the dynamics of NLS.

To control L_-, we choose $\gamma(t)$ so as to minimize the distance of the solution to the ground state orbit, (4.2). This yields the codimension one constraint on v: $(v(\cdot,t), \Psi_{E_0}) = 0$, subject to which (4.6) holds with $C_- > 0$.

To control L_+, we observe that since L^2 is invariant on solutions, we have the codimension constraint on u: $(u(\cdot,t), \Psi_{E_0}) = 0$. Although Ψ_{E_0} is not the ground state of L_+, it can be shown by constrained variational analysis that for small amplitude nonlinear ground states, $C_+ > 0$ in (4.5).

Thus, positivity (coercivity) estimates (4.5) and (4.6) hold and J_{E_0} serves as a Lyapunov functional which controls the distance of the solution to the ground state orbit. The argument presented here appears in greater detail and greater generality in [55], where it is proved that

$$\partial_E \|\psi_E\|_2^2 > 0 \tag{4.7}$$

implies orbital stability of any constrained energy *local* minimizers. The approach is inspired by the seminal article [2], who refers to origins in the work of Boussinesq. A general functional analytic setting is given in [20], where it is shown that (4.7) is necessary and sufficient for stability. See also the related compactness-based variational approach to stability in [9] for a proof of orbital stability of constrained *global* minimizers.

5. Asymptotic stability of ground states I. No neutral oscillations

The type of stability discussed in the previous section is that encountered in the setting of finite-dimensional Hamiltonian systems; if the initial conditions are close to the group orbit of the ground state, then the solution remains close for all time. *Asymptotic stability*, in which the solution asymptotically converges to the state of interest, cannot apply in finite-dimensional Hamiltonian systems as this would violate the volume preserving constraint on the phase flow. However, in an infinite-dimensional setting not all norms are equivalent and there are processes, namely radiation of energy to infinity, which facilitate asymptotic convergence to a preferred state.

We now discuss such a result for small amplitude ground states of NLS. The notion of stability can be seen as a natural refinement of the ideas of Section 4. Instead of freezing the "energy" E of the individual ground state, whose stability is under study, and allowing the phase, θ to evolve in order to *approximately* track the solution, we instead construct $E(t)$ and $\theta(t)$ to evolve in time in such a way that the deviation of the solution, $\phi(\cdot,t)$ and the *modulated* ground state $\Psi_{E(t)} e^{-i(\int_0^t E(s)\,ds - \theta(t))}$ tends to zero in an appropriate norm.

Before stating a result along these lines we need to briefly discuss some spectral properties of the ground state. Recall that by stability of the ground state (in the Lyapunov sense) it is necessary that all spectrum of the generator of the linearized flow, $-i\mathcal{H}_0$ about the ground state lie on the imaginary axis. Zero is an isolated eigenvalue, arising from symmetries of the equation and the continuous spectrum consists of vertical semi-infinite lines $[i|E_0|, i\infty)$ and $(-i\infty, -i|E_0|]$. The key hypotheses are: (i) that $\mathcal{H}_0$ has no nonzero eigenvalues in the gap between $-i|E_0|$ and $i|E_0|$, and thus solutions of the linearized evolution with periodic or quasiperiodic oscillations about the ground state are precluded (see $(h3)$ below), and (ii) that H has neither an eigenvalue nor a "resonance" at zero energy [21], a hypothesis on the behavior of $(H - zI)^{-1}$ as $z \to 0$, which holds for generic $V(x)$, and ensures sufficiently strong dispersive time-decay estimates of the linearized evolution.

THEOREM 5.1 [40,41,33]. *Consider NLS in spatial dimension $n = 3$. Assume the following*

(h1) *The multiplication operator $f \mapsto \langle x\rangle^{\sigma} V(x) f$, where $\sigma > 3$ is bounded on $H^2(\mathbb{R}^3)$.*
(h2) *The Fourier transform of V, $\widehat{V} \in L^1(\mathbb{R}^3)$.*
(h3) *Zero is neither an eigenvalue nor a resonance of the operator $H = -\Delta + V$.*
(h4) *H_0 acting on L^2 has exactly one negative eigenvalue $E_{0*} < 0$, with $H_0\psi_{0*} = E_{0*}\psi_{0*}$, $\|\psi_{0*}\|_2 = 1$.*

Let the initial condition ϕ_0 be sufficiently small in $H^1 \cap L^2(\langle x\rangle^2\, dx)$. Then, there exist smooth functions $E(t)$ and $\theta(t)$, such that $\lim_{t\to\pm\infty} E(t) = E^{\pm}$ and $\lim_{t\to\pm\infty}\theta(t) = \theta^{\pm}$ exist and

$$\lim_{t\to\pm\infty} \left\| \phi(\cdot,t) - e^{-i(\int_0^t E(s)\,ds - \theta(t))}\Psi_{E(t)} \right\|_{L^4(\mathbb{R}^3)} = 0. \tag{5.1}$$

To prove Theorem 5.1 we seek a solution in the form of a modulated nonlinear ground state and a dispersive correction:

$$\phi(x,t) = \Psi_{E(t)} e^{-i(\int_0^t E(s)\,ds - \theta(t))} + \eta(t). \tag{5.2}$$

Substitution into NLS and projection onto the subspaces associated with the discrete and continuum modes of the linearized flow yields a coupled system of equations for $E(t), \theta(t)$ and $\eta(t)$. Asymptotic convergence of $E(t)$ and $\theta(t)$, and decay of $\eta(\cdot,t)$ as $t \to \pm\infty$ are proved using local decay [21] and dispersive L^p, $p > 2$, estimates [22].

We postpone further discussion of this analysis to our discussion of the case where the linearized dynamics has neutral oscillations about the ground state, e.g., which may result from H possessing two or more eigenvalues; see Sections 6 and 7. The analogous coupled ODE–PDE system requires considerably deeper study. In the next section, Section 6, we discuss the metastability and decay of neutral oscillations in the context of the nonlinear Klein–Gordon equation. Then, in Section 7 we turn to the case of NLS, where the same mechanisms are at work.

6. Resonance and radiation damping of neutral oscillations—metastability of bound states of the nonlinear Klein–Gordon equation

Consider the nonlinear Klein–Gordon equation (NLKG) with a potential $V(x)$, assumed to be smooth and sufficiently rapidly decaying as $|x| \to \infty$ ($x \in \mathbb{R}^3$):

$$\left(\partial_t^2 + B^2\right)u = gu^3. \tag{6.1}$$

Here, $B^2 = -\Delta + m^2 + V(x)$, is a strictly positive operator with a single eigenvalue, Ω^2, satisfying $0 < \Omega^2 < m^2$, with corresponding L^2-normalized eigenfunction $\varphi(x)$, which satisfies $B^2\varphi = \Omega^2\varphi$. We also assume that the essential spectrum of B^2 is absolutely continuous and is given by the semi-infinite interval $[m^2, \infty)$. The parameter, g, is taken to be real and either zero or of order unity.

Corresponding to the discrete spectral part of B^2 is a family of time-periodic and spatially localized solutions:

$$u_b(t,x;R,\theta) = R\cos(\Omega t+\theta)\varphi(x) \tag{6.2}$$

of the linear Klein–Gordon equation

$$\left(\partial_t^2 + B^2\right)u = 0. \tag{6.3}$$

For any sufficiently smooth and localized (finite energy) initial conditions, the solution to (6.3) has the decomposition:

$$u(t,x) = u_b(t,x;R_0,\theta_0) + \eta(t,x), \tag{6.4}$$

where R_0 and θ_0 are constants determined by the initial conditions and $\eta(t,x)$ disperses to zero as t tends to infinity.

QUESTION. What is the character of solutions to the *nonlinear* problem $g \neq 0$ for initial data which are small in an appropriate norm?

REMARK 6.1. This question is of independent interest for the nonlinear Klein–Gordon equation. We wish, however, to also point out the relation of this question to the large time asymptotics of NLS. Recall our assumption in Theorem 5.1 that H have only one eigenvalue, E_{0*}, which by Theorem 3.1, gives rise to a branch of nonlinear ground states. If H has two eigenvalues, then there is an additional branch of nonlinear excited states. For NLKG, the role of the nonlinear ground state is played by the zero solution and the dynamics of the nonlinear excited state can be understood by our analysis of how the unperturbed time-periodic bound state of the Klein–Gordon equation decays due to resonant energy transfer to radiation modes under a nonlinear Hamiltonian perturbation.

The following result [45] gives a detailed description of solutions.

THEOREM 6.1. *Consider the nonlinear Klein–Gordon equation* (6.1), *with* $V(x)$, *real-valued and satisfying*

(h1) *There exists* $\delta > 5$ *such that for all* $|\alpha| \leqslant 2$, $|\partial^\alpha V(x)| \leqslant C_\alpha \langle x\rangle^{-\delta}$.

(h2) $(-\Delta+1)^{-1}((x\cdot\nabla)^l V(x))(-\Delta+1)^{-1}$ *is bounded on* L^2 *for* $|l| \leqslant 10$.

(h3) *Zero is not a resonance of the operator* $-\Delta + V$, [21].

(h4) *Nonlinear analogue of the Fermi Golden Rule resonance condition*; *see, for example*, [43].

$$\Gamma \equiv \frac{\pi}{3\Omega}\left(P_c\varphi^3, \delta(B-3\Omega)P_c\varphi^3\right) \equiv \frac{\pi}{3\Omega}\left|\left(\mathcal{F}_c\varphi^3\right)(3\Omega)\right|^2 > 0.$$

Here, P_c *denotes the projection onto the continuous spectral subspace of* B *and* $\mathcal{F}_c$ *denotes the Fourier transform relative to the continuous spectral part of* B.

Assume that the initial data $u(x,0)$ and $\partial_t u(x,0)$ are such that their norms are sufficiently small in $W^{2,2}\cap W^{2,1}$ and $W^{1,2}\cap W^{1,1}$, respectively. Then, the solution of the initial value problem with $g\neq 0$ decays to zero as $t\to\pm\infty$. In particular,

$$u(x,t)=R(t)\cos\bigl(\Omega t+\theta(t)\bigr)\varphi(x)+\eta(x,t),$$
$$\bigl|R(t)\bigr|\leqslant C|t|^{-1/4},\ \bigl\|\eta(\cdot,t)\bigr\|_{L^8}\leqslant C|t|^{-3/4}. \tag{6.5}$$

To prove this result, it is natural to first decompose the solution into its discrete and continuous spectral components:

$$u(x,t)=a(t)\varphi(x)+\eta(x,t),\quad \bigl\langle\varphi,\eta(\cdot,t)\bigr\rangle=0. \tag{6.6}$$

Then, a and η satisfy a coupled system equations, which in the zero amplitude limit is the decoupled linear system:

$$\bigl(\partial_t^2+\Omega\bigr)a(t)=0,\qquad \bigl(\partial_t^2+B^2\bigr)\eta(t,x)=0. \tag{6.7}$$

The latter has time-periodic and spatially localized solution $a(t)=R\cos(\Omega t+\theta),\ \eta\equiv 0$, corresponding to (6.2). For small norm solutions, the equations for a and η are coupled, and can be analyzed by a variant of the arguments outlined in Section 7. We point out that the slow decay of the solution, $u(x,t)$, quantified in the estimates (6.5), is governed by the effective oscillator equation

$$\partial_t^2 a+\bigl(\Omega^2+\mathcal{O}\bigl(|a|^2\bigr)\bigr)a\sim-\Gamma a^4\partial_t a,\quad \Gamma>0. \tag{6.8}$$

Equation (6.8) is a *damped* equation, which governs the transfer of energy from the oscillator to the dispersive wave-field. Γ is the derived nonlinear friction coefficient.

7. Asymptotic stability II. Multiple bound states and selection of the ground state in NLS

We now return to the nonlinear Schrödinger equation (NLS)

$$i\partial_t\Phi=H\Phi+g|\Phi|^2\Phi,\quad H=-\Delta+V,\ x\in\mathbb{R}^3. \tag{7.1}$$

In Section 5 we saw, in the case where H has exactly one bound state, that solutions with small initial conditions asymptotically, as $t\to\pm\infty$, approach an asymptotic ground state. In this section we consider the case where the Schrödinger operator $H=-\Delta+V(x)$ has multiple bound states.

For the linear Schrödinger equation ($g=0$), the solution can be expressed as

$$e^{-iHt}\phi_0=\sum_j\langle\psi_{j*},\phi_0\rangle\psi_{j*}e^{-iE_{j*}t}+e^{-iHt}P_{c*}\phi_0, \tag{7.2}$$

where $e^{-iHt}P_{c*}\phi_0$ decays to zero as $t \to \pm\infty$. The time decay of the continuous spectral part of the solution can be expressed, under suitable smoothness, decay and genericity assumptions on $V(x)$, in terms of *local decay estimates* [21,31]:

$$\left\| \langle x \rangle^{-\sigma} e^{-iHt} P_{c*}\phi_0 \right\|_{L^2(R^3)} \leqslant C \langle t \rangle^{-3/2} \left\| \langle x \rangle^{\sigma} \phi_0 \right\|_{L^2(R^3)}, \tag{7.3}$$

$\sigma \geqslant \sigma_0 > 0$, and $L^1 \to L^\infty$ *decay estimates* [22,56]

$$\left\| e^{-iHt} P_{c*}\phi_0 \right\|_{L^\infty(R^3)} \leqslant C|t|^{-3/2} \left\| \phi_0 \right\|_{L^1(R^3)}. \tag{7.4}$$

Therefore, the large time behavior of typical solutions of the linear Schrödinger equation ($g = 0$) is quasiperiodic.

Consider now the case of NLS with $g \neq 0$ and V is such that the Schrödinger operator H has exactly two bound states: $\psi_{0*}e^{-iE_{0*}t}$ and $\psi_{1*}e^{-iE_{1*}t}$, with $H\psi_{j*} = E_{j*}\psi_{j*},\ \psi_{j*} \in L^2$. By Theorem 3.1 NLS has ground state and excited state branches of nonlinear bound states $\Psi_{\alpha_0}e^{-iE_0t}$ and $\Psi_{\alpha_1}e^{-iE_1t}$, with $\Psi_{\alpha_j} \in L^2$ satisfying

$$H\Psi_{\alpha_j} + g|\Psi_{\alpha_j}|^2\Psi_{\alpha_j} = E_j\Psi_{\alpha_j}. \tag{7.5}$$

Here, α_j denotes a coordinate along the jth nonlinear bound state branch and

$$E_j = E_{j*} + \mathcal{O}\big(|\alpha_j|^2\big).$$

We are interested in the behavior of solutions to NLS with initial conditions of small norm. In contrast to asymptotic quasiperiodic behavior (7.2)–(7.3), we find that the generic long time behavior is a ground state plus dispersive radiation [47]:

THEOREM 7.1. *Consider NLS with a $V(x)$ a smooth and short range* (*sufficiently decaying*) *potential supporting two bound states as described above. Furthermore, assume that the linear Schrödinger operator, H, has no zero energy resonance* [21]. *Assume the* (*generically satisfied*) *nonlinear Fermi golden rule resonance condition*[3]

$$\Gamma_{\omega_*} \equiv g^2\pi\left\langle \psi_{0*}\psi_{1*}^2, \delta(H - \omega_*)\psi_{0*}\psi_{1*}^2 \right\rangle > 0 \tag{7.6}$$

holds, where

$$\omega_* = 2E_{1*} - E_{0*} > 0. \tag{7.7}$$

Then, there exist constants $k_0 \geqslant 3$ and $\sigma_0 \geqslant 2$ such that for any $\sigma \geqslant \sigma_0$ and $k \geqslant k_0$, if $\|\langle x \rangle^{\sigma}\phi(0)\|_{H^k}$ is sufficiently small, we have the following characterization of the large time dynamics of the solution $\phi(t)$ of the initial value problem for NLS with initial data $\phi(0)$.

[3]The operator $f \mapsto \delta(H - \omega_*)f$ projects f onto the generalized eigenfunction of H with generalized eigenvalue ω_*. The expression in (7.6) is finite by local decay estimates (7.3); see, e.g., [43].

As $t \to \infty$

$$\phi(t) \to e^{-i\omega_j(t)}\Psi_{\alpha_j(\infty)} + e^{i\Delta t}\phi_+, \tag{7.8}$$

in L^2, *where* either $j = 0$ *or* $j = 1$. *The phase* ω_j *satisfies*

$$\omega_j(t) = \omega_j^\infty t + \mathcal{O}(\log t). \tag{7.9}$$

Here, $\Psi_{\alpha_j(\infty)}$ *is a nonlinear bound state* (*Section* 3), *with frequency* $E_j(\infty)$ *near* E_{j*}. *When* $j = 0$, *the solution is asymptotic to a nonlinear ground state, while in the case* $j = 1$ *the solution is asymptotic to a nonlinear excited state. Generically,* $j = 0$.

See also the related results on [4–6,13,14,49,50]. Nongeneric solutions which converge asymptotically to an excited state were constructed in [51].

We give a sketch of the analysis. In analogy with the approach discussed in Section 5 for the one bound state case, we represent the solution in terms of the dynamics of the bound state part, described through the evolution of the *collective coordinates* $\alpha_0(t)$ and $\alpha_1(t)$, and a remainder ϕ_2, whose dynamics is controlled by a dispersive equation. In particular we have

$$\phi(t,x) = e^{-i\int_0^t E_0(s)\,ds - i\widetilde{\Theta}(t)}\big(\Psi_{\alpha_0(t)} + \Psi_{\alpha_1(t)} + \phi_2(t,x)\big). \tag{7.10}$$

We substitute (7.10) into NLS and use the nonlinear equations (7.5) for Ψ_{α_j} to simplify. Anticipating the decay of the excited state, we center the dynamics about the ground state. We therefore obtain for $\Phi_2 \equiv (\phi_2, \overline{\phi_2})^{\mathrm{T}}$ the equation:

$$i\,\partial_t\Phi_2 = \mathcal{H}_0(t)\Phi_2 + \mathcal{G}\big(t, x, \Phi_2; \partial_t\vec{\alpha}(t), \partial_t\vec{\alpha}, \partial_t\widetilde{\Theta}(t)\big), \tag{7.11}$$

where $\mathcal{H}_0(t)$ denotes the matrix operator which is the linearization about the time-dependent nonlinear ground state $\Psi_{\alpha_0(t)}$. The idea is that in order for $\phi_2(t,x)$ to decay dispersively to zero we must choose $\alpha_0(t)$ and $\alpha_1(t)$ to evolve in such a way as to remove all secular resonance terms from $\mathcal{G}$. Thus we require,

$$P_b\big(\mathcal{H}_0(t)\big)\Phi_2(t) = 0, \tag{7.12}$$

where $P_b(\mathcal{H}_0)$ and $P_c = I - P_b(\mathcal{H}_0)$ denote the discrete and continuous spectral projections of $\mathcal{H}_0$. Since the discrete subspace of $\mathcal{H}_0(t)$ is four-dimensional (consisting of a generalized null space of dimension two plus two oscillating neutral modes), (7.12) is equivalent to four orthogonality conditions implying four differential equations for α_0, α_1 and their complex conjugates. These equations are coupled to the dispersive partial differential equation for Φ_2. At this stage we have that NLS is equivalent to a dynamical system consisting of a finite-dimensional part governing $\vec{\alpha}_j = (\alpha_j, \overline{\alpha_j})$, $j = 0, 1$, coupled to an infinite-dimensional dispersive part governing Φ_2:

$$\begin{aligned} i\,\partial_t\vec{\alpha} &= \mathcal{A}(t)\vec{\alpha} + \vec{F}_\alpha, \\ i\,\partial_t\Phi_2 &= \mathcal{H}_0(t)\Phi_2 + \vec{F}_\phi. \end{aligned} \tag{7.13}$$

We expect $\mathcal{A}(t)$ and $\mathcal{H}_0(t)$ to have limits as $t \to \pm\infty$. We fix $T > 0$ arbitrarily large, and to study the dynamics on the interval $[0, T]$. In this we follow the strategy of [5,14]. We shall rewrite (7.13) as:

$$\begin{aligned} i\,\partial_t \vec{\alpha} &= \mathcal{A}(T)\vec{\alpha} + \big(\mathcal{A}(t) - \mathcal{A}(T)\big)\vec{\alpha} + \vec{F}_\alpha, \\ i\,\partial_t \Phi_2 &= \mathcal{H}_0(T)\Phi_2 + \big(\mathcal{H}_0(t) - \mathcal{H}_0(T)\big)\Phi_2 + \vec{F}_\phi \end{aligned} \tag{7.14}$$

and implement a perturbative analysis about the *time-independent* reference linear, respectively, matrix and differential, operators $\mathcal{A}(T)$ and $\mathcal{H}_0(T)$.

More specifically, we analyze the dynamics of (7.14) by using (1) the eigenvalues of $\mathcal{A}(T)$ to calculate the key resonant terms and (2) the dispersive estimates of $e^{-i\mathcal{H}_0(T)t}P_c(T)$ [13,17,34,35].

Next we explicitly factor out the rapid oscillations from α_1 and show that, after a near identity change of variables $(\alpha_0, \alpha_1) \mapsto (\tilde{\alpha}_0, \tilde{\beta}_1)$, that the modified ground and excited state amplitudes satisfy the perturbed *dispersive normal form*:

$$\begin{aligned} i\,\partial_t \tilde{\alpha}_0 &= (c_{1022} + i\,\Gamma_\omega)|\tilde{\beta}_1|^4\tilde{\alpha}_0 + F_\alpha[\tilde{\alpha}_0, \tilde{\beta}_1, \eta, t], \\ i\,\partial_t \tilde{\beta}_1 &= (c_{1121} - 2i\,\Gamma_\omega)|\tilde{\alpha}_0|^2|\tilde{\beta}_1|^2\tilde{\beta}_1 + \ F_\beta[\tilde{\alpha}_0, \tilde{\beta}_1, \eta, t]. \end{aligned} \tag{7.15}$$

REMARK 7.1. For finite-dimensional Hamiltonian systems the normal form coefficients are real. That they are complex here, with imaginary part $\sim \Gamma_\omega$, is to due NLS being an infinite-dimensional Hamiltonian system with discrete spectral states resonating with continuum spectral states. The positivity of Γ_ω reflects the energy flow from the excited state to the ground state and continuum states, and the resulting damping of the nonlinear excited state.

It follows from (7.15) that a *Nonlinear Master Equation* governs $P_j = |\tilde{\alpha}_j|^2$, the power in the jth mode:

$$\begin{aligned} \frac{dP_0}{dt} &= 2\Gamma P_1^2 P_0 + R_0(t), \\ \frac{dP_1}{dt} &= -4\Gamma P_1^2 P_0 + R_1(t). \end{aligned} \tag{7.16}$$

Coupling to the dispersive part, Φ_2, is through the source terms R_0 and R_1. The expression "master equation" is used since the role played by (7.16) is analogous to the role of master equations in the quantum theory of open systems [15].

REMARK 7.2. An interesting phenomenon is anticipated by the system obtained from (7.16), by dropping the decaying correction terms $R_j(t)$:

$$\begin{aligned} \frac{dp_0}{dt} &= 2\Gamma p_1^2 p_0, \\ \frac{dp_1}{dt} &= -4\Gamma p_1^2 p_0. \end{aligned} \tag{7.17}$$

First, it is easy to see from (7.17) that $p_1(t)$ decays to zero as $t \to \infty$ unless $p_0(0) = 0$. Furthermore, note that the resulting equation has the conservation law $2p_0(t) + p_1(t) = 2p_0(0) + p_1(0)$, the "total energy". Therefore, since $p_1(t) \to 0$ as $t \to \infty$, we have

$$p_0(\infty) = p_0(0) + \frac{1}{2}p_1(0).$$

Thus we expect that half the energy in decaying excited state is transferred to the ground state and half to continuum radiation.

The detailed behavior of the system (7.16) coupled to the dispersive part can be characterized on short, intermediate and long time scales. We consider the system (7.16) on three time intervals: $I_0 = [0, t_0]$ (initial phase) $I_1 = [t_0, t_1]$ (embryonic phase) and $I_2 = [t_1, \infty)$ (selection of the ground state). A careful analysis reveals an effective finite-dimensional reduction to a system of equations for the "effective mode powers": $Q_0(t)$ and $Q_1(t)$, closely related to $P_0(t)$ and $P_1(t)$, whose character on different time scales dictates the full infinite-dimensional dynamics, in a manner analogous to role of a center manifold reduction of a dissipative system [7].

Initial phase—$t \in I_0 = [0, t_0]$. Here, I_0 is the maximal interval on which $Q_0(t) \leqslant 0$. If $t_0 = \infty$, then $P_0(t) = \mathcal{O}(\langle t \rangle^{-2})$ and the ground state decays to zero. In this case, we show that the excited state amplitude has a limit as well (which may or may not be zero). This case is nongeneric.

Embryonic phase—$t \in I_1 = [t_0, t_1]$. If $t_0 < \infty$, then for $t > t_0$:

$$\begin{aligned} \frac{dQ_0}{dt} &\geqslant 2\Gamma' Q_0 Q_1^2, \\ \frac{dQ_1}{dt} &\leqslant -4\Gamma' Q_0 Q_1^2 + \mathcal{O}\big(\sqrt{Q_0} Q_1^m\big), \quad m \geqslant 4. \end{aligned} \tag{7.18}$$

Therefore, Q_0 is monotonically increasing; *the ground state grows*. Furthermore, if Q_0 is small relative to Q_1, then

$$\frac{Q_0}{Q_1} \quad \text{is montonically increasing,}$$

in fact exponentially increasing; *the ground state grows rapidly relative to the excited state.*

Selection of the ground state $t \in I_2 = [t_1, \infty)$. There exists a time $t = t_1$, $t_0 \leqslant t_1 < \infty$, at which the $\mathcal{O}(\sqrt{Q_0} Q_1^m)$ term in (7.18) is dominated by the leading ("dissipative") term. For $t \geqslant t_1$ we have

$$\begin{aligned} \frac{dQ_0}{dt} &\geqslant 2\Gamma' Q_0 Q_1^2, \\ \frac{dQ_1}{dt} &\leqslant -4\Gamma' Q_0 Q_1^2. \end{aligned} \tag{7.19}$$

It follows that $Q_0(t) \to Q_0(\infty) > 0$ and $Q_1(t) \to 0$ as $t \to \infty$; *the ground state is selected.*

Acknowledgements

We thank S.B. Kuksin for helpful remarks on the manuscript. This work was supported in part by a grant from the U.S. National Science Foundation.

Appendix. Notation

H^s denotes the Sobolev space of functions obtain via the closure of C_0^∞ in a norm: $\|f\|_{H^s}^2 = \sum_{|\alpha| \leqslant s} \|\partial^\alpha f\|_{L^2}^2$.
$P_{b*} = P_{b*}(A)$ denotes the projection onto the discrete spectral subspace of bound states (L^2 eigenstates) of an operator A. $P_{c*} = I - P_{b*}$ denotes the projection onto the continuous spectral subspace.
$H = -\Delta + V$, self-adjoint Schrödinger operator on L^2, with smooth, sufficiently decaying potential, $V(x)$.
$\mathcal{H}_0$ matrix linearization of NLS about the ground state.

References

[1] V.I. Arnol'd, *Geometric Methods in the Theory of Ordinary Differential Equations*, Springer, New York (1983).
[2] T.B. Benjamin, *The stability of solitary waves*, Proc. Roy. Soc. London A **328** (1972), 153–183.
[3] J. Bourgain, *Quasi-periodic solutions of Hamiltonian perturbations of 2D linear Schrödinger equations*, Ann. of Math. **148** (2) (1998), 363–439.
[4] V.S. Buslaev and G.S. Perel'man, *Scattering for the nonlinear Schrödinger equation: States close to a soliton*, St. Petersburg Math. J. **4** (1993), 1111–1142.
[5] V.S. Buslaev and G.S. Perel'man, *On the stability of solitary waves or nonlinear Schrödinger equation*, Nonlinear Evolution Equations, Amer. Math. Soc. Transl. Ser. 2, Vol. 164, Amer. Math. Soc., Providence, RI (1995), 75–98.
[6] V.S. Buslaev and C. Sulem, *On asymptotic stability of solitary waves for nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire **20** (2003), 419–475.
[7] J. Carr, *Applications of Centre Manifold Theory*, Springer, New York (1981).
[8] T. Cazenave, *An Introduction to the Nonlinear Schrödinger Equation*, Textos de Métodos Matemáticos, Vol. 26, Instituto de Matemática, UFRJ, Rio De Janeiro (1989).
[9] T. Cazenave and P.-L. Lions, *Orbital stability of standing waves for some nonlinear Schrödinger equations*, Comm. Math. Phys. **85** (1982), 549–561.
[10] D. Christodoulou and S. Klainerman, *The Global Nonlinear Stability of the Minkowski Space*, Princeton Mathematical Series, Vol. 41, Princeton University Press, Princeton, NJ (1993).
[11] W. Craig and C.E. Wayne, *Newton's method and periodic solutions of nonlinear wave equations*, Comm. Pure Appl. Math. **46** (1993), 1409–1498.
[12] H.L. Cycon, R.G. Froese, W. Kirsch and B. Simon, *Schrödinger Operators*, Springer, Berlin (1987).
[13] S. Cuccagna, *Stabilization of solutions to nonlinear Schrödinger equations*, Comm. Pure Appl. Math. **54** (9) (2001), 1110–1145; Erratum: **58** (1) (2005), 147.

[14] S. Cuccagna, *On asymptotic stability of ground states of nonlinear Schrödinger equations*, Rev. Math. Phys. **15** (8) (2003), 877–903.
[15] E.B. Davies, *Quantum Theory of Open Systems*, Academic Press (1976).
[16] L. Erdös and H.T. Yau, *Derivation of the nonlinear Schrödinger equation from a many particle Coulomb system*, Adv. Theor. Math. Phys. **71** (1999), 463–512.
[17] M. Goldberg and W. Schlag, *Dispersive estimates for Schrödinger operators in dimensions one and three*, Comm. Math. Phys. **251** (1) (2004), 157–178.
[18] R.H. Goodman, P.J. Holmes and M.I. Weinstein, *Strong NLS-soliton defect interactions*, Physica D **192** (2004), 215–248.
[19] R.H. Goodman, R.E. Slusher and M.I. Weinstein, *Stopping light on a defect*, J. Opt. Soc. Am. B **19** (2002), 1635–1652.
[20] M. Grillakis, J. Shatah and W. Strauss, *Stability theory of solitary waves in the presence of symmetry. I*, J. Funct. Anal. **74** (1) (1987), 160–197.
[21] A. Jensen and T. Kato, *Spectral properties of Schrödinger operators and time-decay of wave functions*, Duke Math. J. **46** (1979), 583–611.
[22] J.-L. Journé, A. Soffer and C.D. Sogge, *Decay estimates for Schrödinger operators*, Comm. Pure Appl. Math. **44** (1991), 573–604.
[23] T. Kato, *On nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Phys. Théor. **46**, 113–129.
[24] E. Kirr and M.I. Weinstein, *Parametrically excited Hamiltonian partial differential equations*, SIAM J. Math. Anal. **33** (2001), 16–52.
[25] E. Kirr and M.I. Weinstein, *Metastable states in parametrically excited multimode Hamiltonian partial differential equations*, Comm. Math. Phys. **236** (2003), 335–372.
[26] A. Komech, M. Kunze and H. Spohn, *Effective dynamics for a mechanical particle coupled to a wave field*, Comm. Math. Phys. **203** (1999), 1–19.
[27] S.B. Kuksin, *Nearly Integrable Infinite-Dimensional Hamiltonian Systems*, Lecture Notes in Math., Vol. 1556, Springer, Berlin (1993).
[28] M. Kwong, *Uniqueness of positive solutions of* $\Delta u - u + u^p = 0$, Arch. Rational Mech. Anal. **105** (1989), 243–266.
[29] E.H. Lieb, R. Seiringer and J. Yngvason, *A rigorous derivation of the Gross–Pitaevskii energy functional for a two-dimensional Bose gas*, Comm. Math. Phys. **224** (2001), 17–31.
[30] J.V. Moloney and A.C. Newell, *Nonlinear Optics*, Westview Press (2004).
[31] M. Murata, *Rate of decay of local energy and spectral properties of elliptic operators*, Japan J. Math. **6** (1980), 77–127.
[32] L. Nirenberg, *Topics in Nonlinear Functional Analysis*, Courant Institute Lecture Notes (1974).
[33] C.-A. Pillet and C.E. Wayne, *Invariant manifolds for a class of dispersive, Hamiltonian, partial differential equations*, J. Differential Equations **141** (1997), 310–326.
[34] I. Rodnianski, W. Schlag and A. Soffer, *Dispersive analysis of the charge transfer models*, Comm. Pure Appl. Math. **58** (2) (2005), 149–216.
[35] I. Rodnianski and W. Schlag, *Time decay for solutions of Schrödinger equations with rough and time-dependent potentials*, Invent. Math. **155** (3) (2004), 451–513.
[36] H.A. Rose and M.I. Weinstein, *On the bound states of the nonlinear Schrödinger equation with a linear potential*, Physica D **30** (1988), 207–218.
[37] W. Schlag, *Stable manifolds for orbitally unstable NLS*, http://xxx.lanl.gov/pdf/math.AP/0405435.
[38] I.M. Sigal, *Nonlinear wave and Schrödinger equations I. Instability of time-periodic and quasiperiodic solutions*, Comm. Math. Phys. **153** (1993), 297.
[39] I.M. Sigal, *General characteristics of nonlinear dynamics*, Spectral and Scattering Theory; Proceedings of the Taniguchi international workshop, M. Ikawa, ed., Marcel Dekker, Inc., New York (1994).
[40] A. Soffer and M.I. Weinstein, *Multichannel nonlinear scattering theory for nonintegrable equations*, Integrable Systems and Applications (Ile d'Oleron, 1988), T. Balaban, C. Sulem and P. Lochak, eds, Springer Lecture Notes in Phys., Vol. 342, Springer, Berlin (1989), 312–327.
[41] A. Soffer and M.I. Weinstein, *Multichannel nonlinear scattering theory for nonintegrable equations I & II*, Comm. Math. Phys. **133** (1990), 119–146; J. Differential Equations **98** (1992), 376–390.

[42] A. Soffer and M.I. Weinstein, *Dynamic theory of quantum resonances and perturbation theory of embedded eigenvalues*, Proceedings of Conference on Partial Differential Equations and Applications (University of Toronto, June 1995), P. Greiner, V. Ivrii, L. Seco and C. Sulem, eds, CRM Lecture Notes.
[43] A. Soffer and M.I. Weinstein, *Time dependent resonance theory*, Geom. Funct. Anal. **8** (1998), 1086–1128.
[44] A. Soffer and M.I. Weinstein, *Nonautonomous Hamiltonians*, J. Statist. Phys. **93** (1998), 359–391.
[45] A. Soffer and M.I. Weinstein, *Resonances and radiation damping in Hamiltonian partial differential equations*, Invent. Math. **136** (1999), 9–74.
[46] A. Soffer and M.I. Weinstein, *Ionization and scattering for short lived potentials*, Lett. Math. Phys. **48** (1999), 339–352.
[47] A. Soffer and M.I. Weinstein, *Selection of the ground state in nonlinear Schrödinger equations*, Rev. Math. Phys. **16** (8) (2004), 977–1071; see also http://arxiv.org/abs/nlin/0308020.
[48] C. Sulem and P.-L. Sulem, *The Nonlinear Schrödinger Equation, Self-focusing and Wave Collapse*, Springer (1999).
[49] T.-P. Tsai and H.-T. Yau, *Asymptotic dynamics of nonlinear Schrödinger equations: Resonance dominated and radiation dominated solutions*, Comm. Pure Appl. Math. **55** (2002), 153–216.
[50] T.-P. Tsai and H.-T. Yau, *Relaxation of excited states in nonlinear Schrödinger equations*, Int. Math. Res. Notices **31** (2002), 1629–1673.
[51] T.-P. Tsai and H.-T. Yau, *Stable directions for excited states of nonlinear Schrödinger equations*, Comm. Partial Differential Equations **27** (2002), 2363–2402.
[52] A. Vanderbauwhede and G. Iooss, *Center manifold theory in infinite dimensions*, Dynamics Reported **2** (1990).
[53] R. Weder, *Center manifold for nonintegrable nonlinear Schrödinger equations on the line*, Comm. Math. Phys. **215** (2) (2000), 343–356.
[54] M.I. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal. **16** (1985), 472–491.
[55] M.I. Weinstein, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math. **39** (1986), 51–68.
[56] K. Yajima, *$W^{k,p}$-continuity of wave operators for Schrödinger operators*, J. Math. Soc. Japan **47** (1995), 551–581.

This page intentionally left blank

Author Index of Volume 1A

Roman numbers refer to pages on which the author (or his/her work) is mentioned. Italic numbers refer to reference pages. Numbers between brackets are the reference numbers. No distinction is made between first and co-author(s).

Subject Index of Volume 1A

Author Index

Roman numbers refer to pages on which the author (or his/her work) is mentioned. Italic numbers refer to reference pages. Numbers between brackets are the reference numbers. No distinction is made between first and coauthor(s).

Subject Index