Handbook of Dynamical Systems
Volume 1

Survey I
PRINCIPAL STRUCTURES

Boris Hasselblatt

Anatole Katok

DEPARTMENT OF MATHEMATICS, TUFTS UNIVERSITY, MEDFORD, MA 02155-5597

E-mail address: bhasselb@tufts.edu

DEPARTMENT OF MATHEMATICS, THE PENNSYLVANIA STATE UNIVERSITY, UNIVERSITY PARK, PA 16802-6401

E-mail address: katok_a@math.psu.edu

Contents

Chapter 1. Introduction 13
1. Purpose and structure 13
2. The basic objects of dynamics 14
   a. What is a dynamical system 14
   b. Asymptotic behavior 16
   c. Dynamics without time 16
   d. Orbit properties 16
   e. Transverse behavior and time change 17
3. Equivalence and functorial constructions 17
   a. Isomorphism and invariants 17
   b. Orbit equivalence 18
   c. Classification 18
   d. Functorial constructions 19
   e. Products 19
   f. Restrictions and inducing 19
   g. Irreducibility and decomposition into irreducible components 20
   h. Factors and extensions 20
   i. Inverse limits 20
   j. Suspension 20
   k. Cocycles 21
   l. Skew products and cocycles 21
   m. Orbit equivalence and cocycles 22
   n. Induced action and Mackey range 22
   o. Special flow, integral map, induced map 22
4. Asymptotic behavior and averaging 23
   a. Dissipative and conservative behavior 23
   b. Averaging 23
   c. Amenability 24
   d. Characterizations of amenability 24

Chapter 2. Topological dynamics 27
1. Setting and examples 27
   a. Topological dynamical systems 27
   b. Homogeneous dynamics 28
   c. Group automorphisms and endomorphisms 28
   d. Shifts and symbolic systems 28
2. Basic concepts and constructions 29

a. Topological conjugacy and orbit equivalence 29
b. Invariant sets, inducing 29
c. Topological transitivity and minimality 30
d. Examples of transitivity and minimality 30
e. Isolated sets and attractors 31
f. Factors and almost isomorphism 31
g. Inverse limits 32
h. Natural extension 32
i. Isometric extensions 32
j. Suspensions 33
k. Cocycles and skew products 33
l. Induced action 33
m. Principal classes of asymptotic properties and invariants 33
3. Recurrence 33
a. Limit points 33
b. Recurrence 34
c. Minimality and uniform recurrence 34
d. Nonwandering points, regional recurrence and the center 35
e. Topological transitivity and topological mixing 35
f. Homological and homotopical recurrence, asymptotic cycles 36
g. Rotation number 37
4. Relative behavior of orbits 38
a. Proximality and distality 38
b. Examples of proximal actions 38
c. Classification of distal systems 39
d. Expansiveness 39
5. Orbit growth properties 40
a. Periodic orbits 40
b. The ζ-function for discrete time systems 41
c. Index and algebraic ζ-function 41
d. The ζ-function for flows 42
e. Entropy 43
f. Basic properties of entropy 45
g. Finiteness of entropy 45
h. Growth of separated and spanning sets 45
i. Slow entropy and the Hamming metric 46
j. Weighted zeta-functions 46
k. Pressure 46
l. Higher rank abelian actions 47
m. Complexity of families of orbits 47
6. Symbolic dynamical systems 48
a. Metrics and functions of exponential type 48
b. Shifts 49
c. Topological Markov chains and subshifts of finite type 49
d. Properties of topological Markov chains 50
e. Some subshifts of infinite type 51

f. Complexity of symbolic systems 53
g. The Furstenberg Reduction Principle and multiple recurrence 53
h. Topological Markov chains and subshifts of finite type for other groups 54
7. Low-dimensional topological dynamical systems 54
a. One-dimensional dynamics 54
b. Flows and homeomorphisms on surfaces 55

Chapter 3. Ergodic theory 57
1. Introduction 57
a. Invariant measures and asymptotic distribution 57
b. Quantitative recurrence properties 58
c. The classification problem versus applications 58
d. Dynamical systems and random processes 59
e. Entropy 59
2. Measure spaces, maps, and Lebesgue spaces 60
a. Measure spaces and maps 60
b. Lebesgue spaces 60
c. Lebesgue points 61
d. Measurable partitions 61
e. Conditional measures 62
f. The Radon–Nikodym cocycle 63
g. Relative products 63
3. Setting and examples 63
a. Measurable actions 63
b. Quasi-invariant measures 64
c. Homogeneous dynamics 64
d. Group automorphisms 65
e. Bernoulli shifts 65
f. Markov measure 65
4. Basic concepts and constructions 66
a. Isomorphism and orbit equivalence 66
b. Joinings 66
c. Poincaré Recurrence and induced maps 66
d. Ergodicity 67
e. Kakutani (monotone) equivalence 67
f. Ergodic decomposition 67
g. Factors 67
h. Generators 68
i. Inverse limits 68
j. Natural extensions 68
k. Cocycles 68
l. Isometric extensions 69
m. Suspensions 69
n. Mackey range 69
o. Sections, special representations for flows and cocycle representations for $\mathbb{R}^k$ actions 69

p. Kakutani equivalence for flows 70
q. Induced action on L^p 70
r. Rokhlin Lemma 71
5. Ergodic theorems 71
a. The von Neumann mean ergodic theorem 72
b. The Birkhoff pointwise Ergodic Theorem 72
c. Typical points and recurrence 73
d. Orbit equivalence 73
6. Quantitative recurrence and principal spectral properties 74
a. Ergodicity 74
b. Speed of convergence in ergodic theorems 74
c. Correlation coefficients and spectral measures 74
d. Eigenfunctions 75
e. Rigidity and good periodic approximation 76
f. Weak mixing 76
g. Mild mixing 77
h. Mixing 77
i. Multiple mixing 78
j. Absolutely continuous spectrum 78
k. The K-property 78
l. Decay of correlations 79
7. Entropy 79
a. Entropy and conditional entropy of partitions 79
b. Basic properties of entropy and conditional entropy of a partition 80
c. Entropy of a transformation relative to a partition 81
d. Properties of entropy with respect to a partition 81
e. The Shannon–McMillan–Breiman Theorem 82
f. Entropy of a measure-preserving transformation 82
g. Examples 82
h. Calculation of entropy 83
i. Properties of entropy 84
j. Pinsker algebra, K-property and entropy 84
k. Noninvertible maps 85
l. Slow metric entropy 85
m. Entropy for amenable groups 86
n. Entropy for continuous groups 86
o. Entropy function 86

Chapter 4. Invariant measures in topological dynamics 89
1. Introduction 89
a. Existence of invariant measures 89
b. Topological versus measure-theoretic properties 89
c. Smooth measures 89
2. Existence of invariant measures 89
a. The Kryloff–Bogoliouboff Theorem 89
b. Nonamenability 90

c. Ergodicity 90
d. Ergodic decomposition 90
e. Continuous representation 91
f. The Furstenberg Correspondence Principle; the Szemerédi Theorem 92
3. Unique ergodicity 93
a. Definition and uniform convergence 93
b. Unique ergodicity with trivial recurrence 94
c. Minimal translations of compact abelian groups 94
d. Isometries 94
e. Unipotent affine maps of the torus 95
f. Horocycle flows on surfaces of negative curvature 95
g. Interval exchanges 95
h. Uniquely ergodic realization 96
i. Minimal systems with many invariant measures 96
4. Metric and topological entropy 97
a. Averaging versus maximizing 97
b. Slow entropy 97
c. Measures of high complexity 98
d. The Variational Principle 98
e. Existence of a maximizing measure 99
f. Specification 99
g. Uniqueness of maximal measures 100

Chapter 5. Smooth, Hamiltonian and Lagrangian dynamics 103
1. Differentiable dynamics 103
a. Differentiable dynamical systems 103
b. Linearization 104
c. Semilocal analysis 104
d. Local analysis 105
e. Foliations and holonomy 105
f. Derivative extension and other bundle extensions 106
g. Elliptic, parabolic, hyperbolic and partially hyperbolic behavior 106
h. Prototype examples 107
i. Low-dimensional and conformal dymanics 108
j. Degree of differentiability of smooth dynamical systems 108
2. Basic concepts and constructions 109
a. Conjugacy 109
b. Equivalence of measures 110
c. Local conjugacy and normal forms 110
d. Invariants 111
e. Periodic eigenvalue data 111
f. Stability 112
g. Invariant manifolds and normal forms 112
h. Sections 113
i. Inverse limits 113
j. Suspensions 113

k. Cocycles and extensions 114
l. Isometric extensions 114
m. Smooth invariant measures 114
n. Invariant distributions 115
o. Transversality and Kupka–Smale theorem 116
p. Persistence of recurrence and closing lemma 117
3. Hamiltonian dynamics 119
a. Linear symplectic geometry 119
b. Symplectic geometry 119
c. Examples of symplectic manifolds 120
d. Hamiltonian vector fields and flows 120
e. Symplectic invariants 121
f. Poisson brackets 122
g. The Noether Theorem 122
h. Completely integrable systems, the Liouville–Arnold Theorem 122
4. Lagrangian systems 123
a. The Euler–Lagrange equation 123
b. The Legendre transform 124
c. Geodesic flows 124
5. Contact systems 125
a. Contact forms and contact structures 125
b. Hamiltonian systems preserving a 1-form 126
c. Geodesic flows as contact systems 127
6. Variational methods in dynamics 127
a. Variational description of orbits 127
b. The least action principle in Lagrangian mechanics 128
c. The action principle in Hamiltonian dynamics 129
7. Holomorphic dynamics 130
a. Conformal dynamics 130
b. Holomorphic maps in higher dimension 131

Chapter 6. Hyperbolic dynamics: Orbit instability and structural stability 133
1. Introduction 133
a. The hyperbolic paradigm 133
b. Hyperbolic linear maps 133
2. Main features of hyperbolic behavior 134
a. Growth of the orbit complexity 134
b. Relative behavior of orbits 134
c. Recurrence 134
d. Invariant measures 135
e. Stability 135
f. Prevalence of semilocal phenomena 135
3. Stable manifolds 136
4. Definitions 137
a. Hyperbolic sets 137
b. Nonuniform hyperbolicity 138

c. Partial hyperbolicity 138
d. Flows 138
e. Hölder regularity 138
5. Examples 139
a. Toral automorphisms 139
b. The Smale horseshoe 139
c. The Smale attractor 140
d. Suspensions 140
e. Geodesic flows 140
6. The core theory 141
a. Applications of fixed point results 141
b. The Anosov Closing Lemma 142
c. The Shadowing Lemma 142
d. The Hartman–Grobman Theorem 142
e. Structural stability of hyperbolic sets 143
f. Invariant laminations 143
7. Developments of the theory 145
a. Spectral decomposition 145
b. The Livschitz Theorem 146
c. Specification and equilibrium states 146
d. Sinai–Ruelle–Bowen measure 147
e. Absolutely continuous invariant measures for Anosov systems 147
f. Ergodicity of volume 147
g. Local product structure, Markov partitions 148
h. Stability, moduli and smooth classification 148
i. The stability theorem 149
8. The theory of nonuniformly hyperbolic systems 150
a. Contrast with the uniform case 150
b. Lyapunov exponents and tempering 150
c. Hyperbolic measures and Pesin sets 150
d. Stable manifolds 151
e. Structural theory 151
f. Sinai–Ruelle–Bowen measure 151
g. Comparison 152
9. Partial hyperbolicity 152
a. Structural results 152
b. Invariant foliations 153
c. Stable ergodicity 153

Chapter 7. Elliptic dynamics: Stable recurrent behavior 155
1. Introduction 155
a. Main features 155
b. Linear elliptic maps 155
c. Isometries 156
d. Distinction between different classes of isometries 157
e. Completely integrable systems 157

2. The setting for elliptic dynamics 158
a. Perturbation problem and Diophantine conditions 158
b. Circle diffeomorphisms and twist maps 160
c. Diophantine and Liouvillian behavior 161
d. The Newton method 162
e. Fast periodic approximation in dynamics 163
f. The conjugation-approximation method 164
3. Diophantine phenomena with a single frequency 164
a. Linear stability of Diophantine behavior 164
b. Smooth linearization of circle diffeomorphisms 165
c. The invariant curve theorem 166
d. Twist maps; nondegenerate case 166
e. Neighborhood of an elliptic fixed point 168
f. Caustics in convex billiards and related problems 168
g. Preservation of Diophantine circles without twist 169
4. Diophantine phenomena with several frequencies 170
a. Perturbation of linear maps and vector fields on the torus 170
b. Stability problem in celestial mechanics 170
c. The Kolmogorov theorem in the nondegenerate Hamiltonian case 171
d. Degenerate case and stability of the solar system 171
e. Frequency locking for special symplectic structures 171
f. Preservation of tori in the volume-preserving category 172
5. Liouvillian phenomena 172
a. Linear instability of Liouvillian behavior 172
b. Circle diffeomorphisms with Liouvillian rotation numbers 174
c. Destruction and preservation of Liouvillian circles for twist maps 175
d. Perturbations of isometries in higher dimension 175

Chapter 8. Parabolic dynamics: A special case of intermediate orbit growth 177
1. Introduction 177
a. Systems with intermediate orbit growth 177
b. Parabolic linear paradigm: Jordan blocks and polynomial growth 177
c. Nonlinear systems with parabolic linear part: Local shear 177
d. Parabolic systems with singularities 178
2. Main features of parabolic behavior 178
a. Growth of the orbit complexity 178
b. Relative behavior of orbits 179
c. Recurrence 179
d. Invariant measures 179
e. Mixing properties 180
f. Decay of correlations 180
g. Invariant distributions 181
h. Speed of convergence of ergodic averages 181
i. Rigidity of the measurable orbit structure 181
3. Parabolic systems with uniform structure 181
a. Affine maps on the torus 181

b. Homogeneous dynamics for nonabelian groups 182
c. Extensions 183
d. Time changes 183
4. Flows on surfaces 183
a. Section maps 183
b. Measured foliations 184
c. Topological properties 184
d. Invariant measures and smooth orbit classification 185
e. Mixing and return-time singularities 187
f. Invariant distributions and smooth classification 188
5. Billiards in polygons and polyhedra and related systems 188
a. Billiard flow and the section map 188
b. Integrable billiards 189
c. Rational billiards and quadratic differentials 190
d. Prevalence of minimality and unique ergodicity in rational billiards 191
e. Topologically transitive and ergodic irrational billiards 191
f. Subexponential behavior in polygonal billiards 192
g. Geodesic flows on locally flat surfaces 192
h. Billiards in polyhedra 193

Bibliography 195

CHAPTER 1

Introduction

1. Purpose and structure

Dynamical systems has grown from various roots into a field of great diversity that interacts with many branches of mathematics as well as with the sciences. The purpose of this survey is to describe the general framework for several principal areas of the theory of dynamical systems. We are aware that this is an ambitious goal and that the presentation is bound to be both brief and in many respects superficial.

Our primary aim is to set the stage for the surveys collected in this and the subsequent volume by establishing the unity of the various specialties within dynamics. The range of surveys in these volumes therefore has a strong effect on the presentation given here. Certain topics, which appear in a number of surveys and which we consider as basic for several branches of dynamics, are presented in some detail. Examples are recurrence in topological dynamics, ergodicity, topological and metric entropy, variational principle for entropy, invariant stable and unstable manifolds, cocycles over dynamical systems. Even such topics are usually discussed with only few complete proofs. Topics central to any of the subsequent surveys are often discussed just enough to place them in the greater context, deferring to the corresponding survey for exact statements and further detail. Examples of these are dynamical ζ-functions, variational methods in Lagrangian and Hamiltonian dynamics, KAM theory, dynamics of unipotent homogeneous systems, dynamical methods in combinatorial number theory. Nevertheless, some topics are surveyed here because they play an essential role in the overall picture even though they are not given much attention in subsequent surveys. Bifurcations and applications *per se* are virtually absent here, because they are in the purview of other volumes in the series.

A possible use of this survey is as an introduction to mathematicians unfamiliar with dynamics, and it may be interesting to experts as an overview of a diverse field. With this in mind we pay attention to examples, motivations, informal explanations and discussion of key special cases or simplified versions of general results. Nevertheless, they may often be too brief and may sometimes look cryptic to a nonexpert reader. Expanding the pedagogical aspects of the survey substantially would interfere with its primary goal and expand its size beyond a reasonable limit. Hopefully, a compromise between comprehensiveness and accessibility has been achieved.

A limited number of key results is proved in the text, when the importance of the result, the insights provided by the proof, or its brevity suggested doing so. Other results are provided with sketches or outlines of proofs, many more are only formulated or just mentioned.

The structure of this survey is intended to reflect a coherent framework. Accordingly, this chapter introduces a collection of important notions in generic terms, *i.e.,* without relying on any specific structure of the dynamical system (topological, measure-theoretic, smooth, *etc.*). Although examples are therefore deferred, this serves to provide a structure that organizes the notions and techniques in such a way that later chapters can present large subareas of dynamics in a coherent fashion. Starting from Chapter 2 we introduce basic examples as close to the beginning of each chapter as practicable and then intersperse further examples, as well as comments on previously introduced ones, throughout the chapters. The central structural elements are presented in the following order: the notions of equivalence, principal constructions, recurrence, and orbit growth. The chapters on topological dynamics and ergodic theory follow this pattern closely. The succeeding chapters on smooth dynamics, which are based on the earlier ones, fit into the same framework as well, although the starting point and emphasis is of necessity slightly different. Some background material is incorporated into the text. Examples of these are the treatment of Lebesgue spaces, symplectic manifolds, and Hamiltonian formalism.

When specific results are given without complete proof, we usually provide references to accessible sources, where these can be found. If the original source is mentioned, this is usually done for information only rather than to oblige the reader to consult it. We choose our accessible sources in the following order of preference:

(1) Other surveys in this and the subsequent volume. References to these are distinguished by a format such as [**S-H**], where the S stands for "Survey".
(2) Our book [**KH**], where a variety of topics is presented in settings similar to those of this survey.
(3) Other books from the section "major sources" in the bibliography.
(4) Further books and articles in major journals available at most university libraries.

At the beginning of (sub)sections that introduce a new subject, we occasionally give references to the places in the survey, where that subject is treated in more detail, as well as to other surveys in this volume dedicated to it.

We do not claim to present a comprehensive or even fully representative bibliography on any of the topics. The bibliographies in subsequent surveys and in our major sources are better suited for that purpose. Furthermore, we are aware of the bias in the references toward works that fit to our own point of view on the subject as well as the omission of some important sources with which we are not sufficiently familiar.

2. The basic objects of dynamics

a. What is a dynamical system. The setting for the study of dynamical systems involves a *space, time,* and a *time evolution*.

1. *Phase space.* This is a set with some additional structure, whose elements or points represent possible states of the system. The most basic structures are a measure, a topology or a finite-dimensional differentiable structure. In this survey we also include some more specialized smooth structures, namely symplectic, contact, Lagrangian and holomorphic, as well as homogeneous structures.

Taken together, these cover most of the general aspects of the theory of dynamical systems as well as traditional applications (celestial mechanics, thermodynamics) and more modern ones (diophantine approximations, Riemannian geometry), but, for example, not

infinite-dimensional differentiable dynamics which, within the conceptual framework developed in these volumes, can be treated only partially and with various qualifications.

To summarize, the methods of dynamical systems apply to spaces that are not too big in an appropriate sense (such as (locally) compact, (σ-)finite measure, finite-dimensional).

2. *Time.* Time may be discrete or continuous and may be reversible or irreversible, *i.e.,* parametrized by a group or a semigroup. Again, it is important that this (semi-) group is not too large. Local compactness and second countability are typically required for the methods to apply. On the other hand, it is essential that time be noncompact, in order to allow a notion of behavior asymptotic in time.

We make a point of providing the framework in appropriate generality, but when discussing specific notions and results, we are usually concerned with the classical setting of time given by a one-parameter (semi-) group. In this case, integer time parametrizes a reversible discrete-time process, the natural numbers an irreversible one, real numbers a reversible continuous-time process, and nonnegative real numbers an irreversible one (of which we present no examples). Thus, time is parametrized by $\mathbb{Z}$, $\mathbb{N}_0$, $\mathbb{R}$, or $\mathbb{R}_0^+$.

In the discrete-time case the action is defined by iterates of a single generator, so this map itself is usually referred to as the dynamical system. In the continuous-time case the dynamical system is called a *flow* or *semiflow*, respectively. As a unifying term for these four classical possibilities we use the term *cyclic dynamical system*.

Actions of larger groups are an important area of study in dynamical systems and some fundamental results naturally hold in such generality. The appropriate groups are those that are not too large locally or globally. Specifically, this means local compactness (local), second countability (both local and global) and often amenability (global). Specific reasons for precisely these requirements will be supplied in due course.

On the other hand, there are results and paradigms that do not hold for cyclic dynamical systems but are specific to actions of certain classes of locally compact second countable nonamenable groups such as semisimple Lie groups, lattices in such groups, or groups with property T or for some noncyclic amenable groups, *e.g.,* $\mathbb{Z}^k$ or $\mathbb{R}^k$, $k \geq 2$. Such facts appear in the survey [**S-FK**]. Noncyclic dynamical systems also arise in other surveys in this volume, [**S-B, S-KSS, S-LS, S-T**].

3. *Time evolution.* The time-evolution law is represented by the action of time, given by the (semi-) group G, on the phase space X, *i.e.,* a map $\Phi\colon G \times X \to X$, $(g, x) \mapsto \Phi(g, x) =: \Phi^g(x)$ such that

$$\Phi^e = \mathrm{Id} \text{ and } \Phi^{g_1 g_2} = \Phi^{g_2} \circ \Phi^{g_1} \text{ for all } g_1, g_2 \in G. \tag{1.1}$$

We usually consider left actions, but occasions arise when right actions need to be discussed (*e.g.,* suspensions, see Section 1.3j). We will be explicit at those times.

Dynamics deals with actions that preserve or in some other way respect the structure on X, *i.e.,* continuous or smooth (or piecewise continuous or smooth), measure-preserving, or at least nonsingular actions, *etc.* An important additional aspect needs to be made explicit in the case of continuous time: We require continuous dependence (in an appropriate topology) on the group element. This is, of course, vacuous in the discrete-time case.

Most of the main dynamical phenomena are apparent already in the discrete-time case, with only some layers of technicality added in the corresponding continuous-time setting. In some applications, however, these technical issues are rather central (partial differential equations, statistical mechanics). To illustrate the most basic such issue note that a smooth

flow ($\mathbb{R}$-action) is determined by a single infinitesimal generator, *i.e.,* a vector field, but that the flow appears by way of solving a differential equation rather than by iterating a map as in the discrete-time case. See [**S-FK**] for a more general discussion of that kind, still confined to a finite-dimensional situation. In the infinite-dimensional situation the mere existence of continuous time dynamics becomes a major issue, both in connection with dynamical systems arising from partial differential equations and from continuous-time models in statistical mechanics. This is another reason, along with the large "size" of the phase space, for difficulties in applying the methods from the theory of dynamical systems to these natural infinite-dimensional situations.

b. Asymptotic behavior. The characteristic feature of dynamical theories, which distinguishes them from other areas of mathematics dealing with groups of automorphisms of various mathematical structures, is the emphasis on asymptotic behavior, especially in the presence of nontrivial recurrence, *i.e.,* properties related to the behavior as time goes to infinity (in the sense of leaving any given compact subset of G). Specifically, this suggests the following convenient notation: If $(g_n)_{n\in\mathbb{N}}$ is a sequence in the (semi-) group then "$g_n \to \infty$" as $n \to \infty$ means that for any compact set K there exists an $N \in \mathbb{N}$ such that $g_n \notin K$ for $n \geq N$.

The specific aspects of asymptotic behavior that one can examine depend largely on the structure of the phase space (such as measurable versus topological), and accordingly appear later on, first in Section 1.4 and then when the discussion becomes specific to the respective settings (Section 2.3, Section 2.4, Section 2.5, Section 3.1a, Section 3.5, Section 3.6, Section 3.7, Section 5.1g, Section 6.4).

c. Dynamics without time. In several settings one can use ideas and concepts of a dynamical nature even when there is no action of the kind we consider here. Instead of orbits one may have other equivalence classes with some structure that are "large" enough to give meaning to some ideas of asymptotic behavior.

Instances of such areas of study are

(1) foliations of compact manifolds by noncompact leaves (Section 5.1e, Section 8.4b), and
(2) discrete measurable equivalence relations of measure spaces.

The latter turns out to be the natural setting for the study of orbit equivalence of group actions with an invariant or quasi-invariant measure (see Section 3.4a and Section 3.5d).

In these situations one can define certain actions naturally associated with the structure, such as local holonomy maps in case 1. and the full group in case 2., but these cannot usually be organized into sufficiently manageable group or semigroup structure. Nevertheless, one can still introduce the concept of asymptotic behavior similarly to the previous subsection, which in case 1. amounts to going along a leaf away from any compact subset, and in case 2. to leaving any finite set in an equivalence class.

We do not systematically include any treatment of such situations in this survey, although occasionally specific instances arise in connection with the discussion of group actions [**FM**].

d. Orbit properties. For a point $x \in X$ its *orbit* or *trajectory* is $\mathcal{O}(x) := \Phi(G, x) \subset X$. A point x is said to be *fixed* or *stationary* if $\mathcal{O}(x) = \{x\}$. The action is said to be

transitive if X is an orbit. (This is distinct from the notion of topological transitivity we introduce later.)

For the remainder of this paragraph assume that G is a group. The *stationary subgroup* of x is

$$G(x) := \{g \in G \mid \Phi^g(x) = x\}.$$

Under our standing continuity assumption this is always a closed subgroup.

The orbit of x is said to be *compact* if the factor $G/G(x)$ is compact in the induced topology; in particular a fixed point has a compact orbit. For $G = \mathbb{Z}$ one can then define the *period* of a point x with compact orbit as the positive generator of $G(x)$. For $G = \mathbb{R}$ the same can be done for nonfixed periodic points. Accordingly, such orbits are also said to be *periodic*. For noninvertible cyclic systems one can also define periodic orbits by $\Phi^t(x) = x$ for some $t > 0$; the smallest such t is the period.

A *locally free* point is one for which $G(x)$ is discrete. An action is said to be *locally free* if there is a neighborhood U of the identity Id in G such that $G(x) \cap U = \{\mathrm{Id}\}$ for all $x \in X$. The action is said to be *effective* if $\Phi^g \neq \mathrm{Id}$ for all $g \in G$.

Notice that in the measurable setting, where the notion of a single point is not well defined, most of the above notions are not directly applicable and have to be properly modified. The notion of *Lebesgue point* (Section 3.2b) often provides an appropriate substitute.

e. Transverse behavior and time change. Dynamics of actions of continuous groups includes an important aspect of transverse behavior, which deals with relative behavior of orbits and thus involves more robust properties independent of time changes. Transverse behavior can often be understood by considering sections and the corresponding holonomy maps. The notion of transversality and hence that of a transversal section is quite straightforward in the differentiable case and presents only moderate difficulties in the case of topological dynamics, if both the phase space and the action group are sufficiently nice. In the measurable situation, a section is a set of measure zero and as such does not make direct sense. Nevertheless, sections still can be constructed and provide a useful tool for studying flows and measurable actions of other continuous groups. Transverse behavior is the central feature of "dynamics without time" although in some cases there is an additional geometric structure on leaves, which replaces the homogeneous structure appearing in the case of a group action.

3. Equivalence and functorial constructions

a. Isomorphism and invariants. An *isomorphism* between dynamical systems $\Phi\colon G\times X \to X$ and $\Psi\colon G \times Y \to Y$ is a bijection $h\colon X \to Y$ that preserves or respects the particular structure (diffeomorphism, homeomorphism, measure-preserving, nonsingular map, *etc.*), such that

$$h(\Phi^g(x)) = \Psi^g(h(x)) \text{ for all } g \in G, x \in X.$$

Let us remark parenthetically that while this notion is natural from the categorical point of view, in some particular settings a weaker structure should be preserved for a meaningful working notion. A characteristic example is that for smooth dynamical systems a topological classification is in many situations more natural and tractable than a smooth one (Section 5.2f).

An obvious purpose of introducing the notion of isomorphism is to provide a reasonable equivalence relation, *i.e.*, a working term describing when two systems are to be considered structurally identical. It naturally prompts a search for *invariants, i.e.,* properties preserved by isomorphism. However, many of these invariants were not developed with this question in mind, but arose early on as properties of the orbit structure pertinent to concrete and important qualitative questions. Specific invariants are discussed later in the appropriate contexts.

b. Orbit equivalence. Some of these invariants relate to the orbit structure or transverse behavior and their invariance depends merely on the fact that isomorphisms preserve orbits. This motivates a weaker notion of equivalence:

An *orbit equivalence* between dynamical systems $\Phi\colon G \times X \to X$ and $\Psi\colon G' \times Y \to Y$ is a bijection $h\colon X \to Y$ preserving or respecting the particular structure (diffeomorphism, homeomorphism, measure-preserving, nonsingular map, *etc.*) that sends orbits onto orbits. We say that Ψ is a *time change* of Φ if $h = \mathrm{Id}$.

As it turns out, the essential meaning of orbit equivalence is quite different for the measurable category and the categories that include topology as a part of the structure in the phase space. In fact, the latter case is more similar to *Kakutani equivalence* (Section 3.4e, Section 3.4p, [**S-T**]) in the measurable category.

c. Classification. A natural concept related to the functorial notions of isomorphism and orbit equivalence would be the classification of dynamical systems within a given category up to either of these two fundamental equivalence relations A classification of cyclic dynamical systems in any of the major branches of dynamics is infeasible in full generality because the sets of equivalence classes are usually both huge and lack any convenient structure. Among these branches ergodic theory is the only one where the general classification problems have been seriously posed and investigated. One reason for that is that in ergodic theory at least the phase spaces are standard (Lebesgue spaces, see Section 3.2b) However, the following more limited classification problems are sometimes tractable.

1. *Restricted phase space.* The structure of the phase space may put substantial limitations on the dynamics. The classical examples occur in low-dimensional topological and differentiable dynamics: homeomorphisms and diffeomorphisms of the circle and flows on compact surfaces.

2. *Restrictions of the type of dynamics.* Examples of such *a priori* conditions of a dynamical nature are distality in topological dynamics, discrete spectrum in ergodic theory, hyperbolicity in differentiable dynamics and complete integrability in Hamiltonian dynamics.

3. *Classification up to a weaker type of equivalence.* This is a very characteristic phenomenon in situations with dynamical restrictions. The classical examples are *topological* classification of circle diffeomorphisms and of various classes of hyperbolic differentiable dynamical systems.

4. *Local classification.* Sometimes one can classify perturbations of certain dynamical systems within a natural space of systems. This of course depends on the topology being sufficiently fine. Examples are structural stability in differentiable dynamics and classification up to differentiable cojugacy via local moduli.

5. *Classification on a part of the phase space.* This is an even more general phenomenon which appears in the special situations described above. The orbit structure may be

robust with respect to small or large perturbations on certain invariant sets. These phenomena are central both in the nonuniformly hyperbolic situation and in KAM theory.

6. *Noncyclic dynamical systems.* Actions of certain groups such as semisimple Lie groups of higher rank or lattices in such groups exhibit strong rigidity properties. These put various classification problems for actions of such groups into a different and more accessible category. Such problems can at least be seriously discussed. See [**S-FK**].

d. Functorial constructions. The remainder of this section is dedicated to a description of general constructions that produce new dynamical systems from present ones. Several of these "enlarge" a given dynamical system by an extension process. Of these, some combine several dynamical systems into a single new one by product-like constructions. Conversely, there are also various reductive operations associated with subsets of the phase space. In some cases, these may lead to a decomposition. While there is a certain universality to these constructions, the implementation of several of them depends strongly on the structure (topological, measurable, smooth) of the setting.

The last few subsections (beginning with Section 1.3k) introduce cocycles, among whose applications are several further constructions of this nature.

e. Products. The *(direct) product* of two dynamical systems $\Phi\colon G \times X \to X$ and $\Psi\colon H \times Y \to Y$ is defined on $(G \times H, X \times Y)$ by

$$(\Phi \times \Psi)((g,h),(x,y)) := (\Phi^g(x), \Psi^h(y)).$$

In the special case $G = H$ this gives rise to the *diagonal action* or *Cartesian product* of Φ and Ψ defined by $(g,(x,y)) \mapsto (\Phi^g(x), \Psi^g(y))$.

This clearly extends to finitely many factors and often to countable products, if an appropriate product structure is defined (such as product topology or measure).

f. Restrictions and inducing. An almost trivial construction is the *restriction* to an invariant subset: If $A \subset X$ and $\Phi(G,A) = A$ then A is said to be *invariant* under Φ. In this case there is a natural action of G on A. This is of interest when the subset A is well-behaved with respect to the structure on X, such as being a measurable set of positive measure in the measurable case, or being closed in the topological case. In the smooth case, one naturally encounters invariant sets that are compact, but not necessarily submanifolds (fractal sets, including strange attractors). This is one of the principal reasons for the widespread interest in these notions. See Section 5.1c and Section 5.2i below.

For cyclic systems a set A is *forward invariant* if $\Phi^t(A) \subset A$ for all $t \geq 0$. Such sets replace invariant ones for nonreversible systems.

At times, it is desirable to employ a procedure like this for sets that are not invariant. This is fraught with various difficulties, which can be resolved only in certain contexts and in ways that depend upon the setting, such as the first-return (or induced) map for a measure-preserving transformation or a Poincaré section map on a transversal near a periodic orbit. Therefore, these are addressed at the appropriate time (Section 3.4c, Section 5.2h).

A complementary restriction is in the group: One may restrict an action to a subgroup. In the case of transitive actions, for example, this often leads to dynamically interesting situations, such as in homogeneous dynamics (see Section 2.1b and, for more detail, [**S-KSS**]).

g. Irreducibility and decomposition into irreducible components. If the phase space or an essential part of it can be split into a well-behaved union (finite, countable, or uncountable) of invariant subsets, the action on these subsets may be studied separately. If no such decomposition is possible, then the dynamical system is said to be *irreducible*. An *a priori* stronger notion of irreducibility requires that there are no proper invariant subsets compatible with the structure, such as closed in the topological context or of positive but not full measure in the measurable context. Such dynamical systems are natural building blocks for more general ones. If a decomposition into irreducible components exists, there is also information on how the pieces or components are put together, but to a large extent the dynamics of an action is reduced to that of the components.

This approach generally fails in topological dynamics, except for special cases such as actions by isometries (Section 4.3d) or distal actions (Section 2.4a). But it works in ergodic theory (Section 3.4f). Other instances when it is applicable are the *spectral decomposition* for subshifts of finite type (Section 2.6d) and compact locally maximal hyperbolic sets (Theorem 6.7.1, [**S-H**], [**KH**, Section 18.3]). In smooth and symplectic dynamics the decomposition problem is closely related to the classical question of finding first integrals and complete integrability (Section 5.3h, Section 7.1e).

h. Factors and extensions. An action $\Psi\colon G \times Y \to Y$ is said to be a *factor* of $\Phi\colon G \times X \to X$, and Φ an *extension* of Ψ if there exists a surjective morphism $h\colon X \to Y$ such that $h(\Phi^g(x)) = \Psi^g(h(x))$ for all $(g, x) \in G \times X$. Ψ is said to be an *orbit factor* of Φ if there exists a surjective morphism $h\colon X \to Y$ mapping orbits onto orbits. This generalizes the isomorphism notion from Section 1.3a.

Similarly to invariant sets, factors, even natural ones, may lack the full structure of the original system. This happens, for example, with some topological factors of smooth systems.

i. Inverse limits. Iteration of extensions to a sequence of morphisms $\cdots \to X_3 \to X_2 \to X_1$ gives rise to the construction of *inverse limit*. The specifics of these constructions differ according to the structure considered on X. An important application of a version of the inverse limit construction is to produce an invertible system from a noninvertible one, which is then called the *natural extension* of the original system (Section 2.2h, Section 3.4j).

j. Suspension. Let G be a topological group, H a closed subgroup and $\Phi\colon H \times X \to X$ a left action on a space X that preserves some structure. It lifts to an action

$$\tilde{\Phi}\colon H \times (X \times G) \to (X \times G), \quad \tilde{\Phi}^h(x, g) = (\Phi^h(x), hg).$$

The action $R_{g_0}(x, g) = (x, gg_0)$ of G on $X \times G$ by right translations in G commutes with $\tilde{\Phi}$ and hence projects to the space $\tilde{\Phi}\backslash X \times G$ of $\tilde{\Phi}$-orbits. Since H is a closed subgroup, $\tilde{\Phi}\backslash X \times G$ usually inherits the structure from $X \times G$. The factor action of G on $\tilde{\Phi}\backslash X \times G$ is called the *suspension* Φ_G of Φ.

Naturally interesting cases appear when H is "sufficiently large" in G, *i.e.,* when the asymptotic behavior in H essentially captures that in G. The classical case is $G = \mathbb{R}$, $H = \mathbb{Z}$, in which case one speaks of the *suspension flow*. More generally, one may consider $G = \mathbb{R}^n$ and $H = \mathbb{Z}^n$. An even more general case is that of a *lattice* H in G (Section 3.3c).

k. Cocycles. A central role in many aspects of dynamical systems is played by *cocycles*.

A *1-cocycle* with values in a topological group H over an action $\Phi\colon G\times X\to X$ is defined to be a map $\alpha\colon G\times X\to H$, continuous in G, such that

$$\alpha(g_1g_2,x)=\alpha(g_2,\Phi^{g_1}(x))\alpha(g_1,x).$$

Two cocycles α,β are said to be *cohomologous* if there is a map $C\colon X\to H$, called a *transfer function*, such that

$$\alpha(g,x)=C(\Phi^g(x))\beta(g,x)C(x)^{-1}.$$

A cocycle is said to be a *coboundary* if it is cohomologous to the identity in H.

The notion of regularity of a cocycle as a function on the phase space depends on the structure of the phase space (measurable, topological, smooth). Sometimes it turns out to be natural to consider cohomology of cocycles in a sense weaker than the ambient structure, *i.e.*, the transfer function may only need to be of some lower regularity than the cocycles themselves.

Note that a cocycle independent of x is given by a homomorphism $G\to H$. If H is abelian then one can define a product of cocycles, coboundaries form a subgroup of the abelian group of all cocycles, and hence the set of cohomology classes has a group structure. Formally this is the first cohomology group of G acting on X with coefficients in H. In dynamics the regularity of the cocycles and transfer functions plays a central role and in the presence of nontrivial asymptotic behavior the calculation of the cohomology groups only rarely reduces to formal algebraic manipulations. Higher cohomology groups can be defined following the general prescription of homological algebra [**Bn**].

If H is nonabelian the set of cohomology classes does not possess any group structure.

Depending on the structure of the space on which the dynamics is defined, there are cocycles naturally associated with the dynamics, such as the Radon–Nikodym cocycle for transformations with quasi-invariant measures (the Jacobian cocycle in the case of smooth dynamics, Section 5.2k).

l. Skew products and cocycles. An important particular kind of extension is given by the *skew product* construction, which generalizes the product construction: Consider an extension Φ of $\Psi\colon G\times Y\to Y$ to $X=Y\times Z$ with $h=\pi_1$ the projection to the first coordinate. Then

$$\Phi^g((y,z))=(\Psi^g(y),\alpha(g,y)z),$$

and α must be a 1-cocycle over Ψ whose values are morphisms of Z:

$$\begin{aligned}(1.1)\quad &(\Psi^{g_1g_2}(y),\alpha(g_1g_2,y)z)=\Phi^{g_1g_2}(y,z)=\Phi^{g_2}(\Phi^{g_1}(y,z))\\ &=(\Psi^{g_2}(\Psi^{g_1}(y)),\alpha(g_2,\Psi^{g_1}(y))\alpha(g_1,y)z)=(\Psi^{g_1g_2}(y),\alpha(g_2,\Psi^{g_1}(y))\alpha(g_1,y)z).\end{aligned}$$

Diagonal actions (Cartesian products of Ψ with actions of G on Z) correspond to cocycles α independent of y.

This construction is quite useful when a group H acts on Z by morphisms. Then any 1-cocycle with values in H gives rise to a skew product. Examples are compact groups Z with H the left translations, or affine or projective spaces Z with H the linear group.

If one considers skew products Φ_1, Φ_2 over Ψ defined by cohomologous cocycles α_1 and α_2, then there is a bijection $c(y,z) = (y, C(y)z)$ between these extensions:

$$\Phi_1^g(c(y,z)) = \Phi_1^g(y, C(y)z) = (\Psi^g(y), \alpha_1(g,y)C(y)z) \\ = (\Psi^g(y), C(\Psi^g(y))\alpha_2(g,y)z) = c(\Phi_2^g(y,z)). \tag{1.2}$$

If the transfer function C respects the structure then this bijection is an isomorphism.

Skew products also provide the natural setting for "random dynamics" [**S-F**].

m. Orbit equivalence and cocycles. Cocycles also appear in connection with orbit equivalence, in particular time changes. In this case $H(\Phi^g(x)) = \Psi(\alpha(g, H(x)), H(x))$ and α is a 1-cocycle over Φ with values in G' such that $\alpha(\cdot, x)$ is bijective:

$$\begin{aligned}\Psi(\alpha(g_1g_2, H(x)), H(x)) &= H(\Phi^{g_1g_2}(x)) = H(\Phi^{g_2}(\Phi^{g_1}(x))) \\ &= H(\Phi^{g_2}(H^{-1}(\Psi(\alpha(g_1, H(x)), H(x))))) \\ &= \Psi(\alpha(g_2, H(\Phi^{g_1}(x))), \Psi(\alpha(g_1, H(x)), H(x))) \\ &= \Psi(\alpha(g_2, H(\Phi^{g_1}(x)))\alpha(g_1, H(x)), H(x)).\end{aligned}$$

If $G = G'$ and α is cohomologous to the identity map $G \to G'$ then, similarly to the previous situation, Ψ and Φ are isomorphic via the bijection $x \mapsto \Phi^{C(x)}(x)$, where C is the transfer function.

n. Induced action and Mackey range. Cocycles also provide a natural generalization of the suspension construction. If $\alpha\colon H \times X \to G$ is a cocycle, then $\tilde{\Phi}_\alpha^h(x,y) := (\Phi^h(x), \alpha(h,x)g)$ is an action of H, according to Section 1.3l. The action Φ_α commutes with the right action $R_{(\cdot)}$, which hence projects to the space of $\tilde{\Phi}_\alpha$-orbits.

Notice that we do not assume that H is a closed subgroup. The case of the cocycle $\alpha(h,x) = h \in H \subset G$ over Φ independent of x, *i.e.*, the inclusion $H \hookrightarrow G$, gives the standard suspension.

Sometimes the orbit space inherits a nice structure from X. In this case the resulting action can be viewed as a direct generalization of the suspension and it is usually called *the action induced by cocycle* α, or the twisted product [**S-FK**]. However, the orbit space may not always have the right structure, and then one is forced to consider a proper "hull" of the orbit space, the resulting right G-action on which is called a *Mackey range* Φ_α of Φ. The specifics appear in due course. If two cocycles are cohomologous in the proper category, then the corresponding induced actions or Mackey ranges are isomorphic in that category.

o. Special flow, integral map, induced map. The case $G = \mathbb{R}$, $H = \mathbb{Z}$, $\varphi := \alpha(1,\cdot) > 0$ gives a flow with a natural *fundamental domain*

$$\{(x,t) \mid 0 \le t \le \varphi(x)\} \subset X \times \mathbb{R},$$

which is called the *flow under a function* or *special flow* over the map f generating the $\mathbb{Z}$-action. We denote this flow by f_φ. The flow can be described as going along "vertical" lines $x = \mathrm{const}$ with unit speed and jumping from $(x, \varphi(x))$ to $(f(x), 0)$.

Another special case is $G = H = \mathbb{Z}$ with $\varphi > 0$, called the *integral map*.

On the other hand, if α takes values in $\{0,1\}$ this construction can be identified with the *first-return map* to $\varphi^{-1}(\{1\})$, also called the *induced map*, which may not be defined everywhere (Section 3.4c).

4. Asymptotic behavior and averaging

a. Dissipative and conservative behavior. The core issue in dynamics is to understand, how a sufficiently "large", *i.e.,* noncompact group acts on a "small" space, such as a finite measure space, a compact topological space or a compact subset of a differentiable manifold. At its center lies the general concept of *recurrence*, *i.e.,* the phenomenon that some points come back to certain parts of the phase space again and again as time goes to infinity. Section 2.3 and Section 3.4c present the most basic manifestations of this phenomenon. However, some points never return. This happens quite naturally if the phase space itself is not "sufficiently small", *i.e.,* only locally compact but not compact, or of σ-finite but not finite measure, in which case no recurrence is guaranteed for any orbit. But this may also happen in a compact or finite measure space, although in the latter case the presence of a positive-measure set of nonrecurrent points implies that the measure is not invariant. This type of behavior is called *dissipative*, because in mechanical systems it appears when energy dissipates in some way, *e.g.,* via friction. The opposite type of behavior, when orbits return again and again to where they came from, is called *conservative*, since it appears in mechanics when the total energy of the system is preserved and a hypersurface of fixed energy is compact (hence of finite volume) in the phase space.

The dichotomy between dissipative and conservative behavior is central to dynamics. In the former case, the emphasis is on the limit behavior of orbits, which is often (but far from always) simple (steady state, limit cycle, regular escape to infinity). Of course, it may also be complicated (strange attractors), in which case the study of dissipative orbits splits into two parts: existence of limit regimes, which are themselves rather complicated conservative motions, and the study of the conservative dynamics on this limit set.

Except for the trivial cases of fixed points and periodic orbits, conservative behavior is not simple. Hence various branches of dynamics develop an appropriate set of concepts and invariants to describe this behavior. A central role in this circle of ideas is played by averaging.

This survey, as well as the others in this volume, concentrates primarily on the conservative case.

b. Averaging. If one considers the roots of dynamical systems in mechanics, where the state of a system evolves in time (a point in phase space moves along an orbit), then an experimental observation of some observable quantity associated with the dynamical system corresponds to the evaluation of a function (on the phase space) at a point of the orbit. Repeated measurements correspond to multiple samplings of the function. In numerous systems one has come to expect that averages of such measurements settle down. Specifically, in the case of a map f and a function φ one wants to study the *Birkhoff averages*

$$\frac{1}{n}\sum_{i=0}^{n-1}\varphi(f^i(x))$$

and their convergence. These are also called *ergodic*, *time*, or *Cesaro* averages. Similarly one defines the Birkhoff average for a continuous-time system. The convergence of such averages plays a central role in ergodic theory and its applications to other branches of dynamics.

This idea admits a natural degree of generalization. Suppose G is a discrete semigroup. A sequence $(F_n)_{n\in\mathbb{N}} \subset G$ of finite sets is said to be *left-Følner* if $\operatorname{card}(L_g(F_n)\triangle F_n)/\operatorname{card}(F_n) \to$

0 as $n \to \infty$ for every $g \in G$. Here $L_g \colon G \to G,\ \gamma \mapsto g\gamma$ is the left translation. (Right-Følner sets are defined analogously.) Given a G-action Φ such a sequence induces a notion of averaging of a function φ by setting

$$\mathcal{F}_n(\varphi) := \frac{1}{\operatorname{card}(F_n)} \sum_{g \in F_n} \varphi \circ \Phi^g. \tag{1.1}$$

The Følner condition gives $\lim_{n\to\infty} \mathcal{F}_n(\varphi) - \mathcal{F}_n(\varphi \circ \Phi^g) \to 0$ for bounded φ and any $g \in G$. The question is whether $\mathcal{F}_n(\varphi)$ converges.

For continuous groups one can do the same: A left-Følner is a sequence $(F_n)_{n\in\mathbb{N}}$ of sets of nonzero finite Haar measure ν such that $\nu(L_g(F_n) \triangle F_n)/\nu(F_n) \to 0$ as $n \to \infty$ for every $g \in G$. Then let $\mathcal{F}_n(\varphi) := \int_{F_n} \varphi \circ \Phi^g \, d\nu(g)/\nu(F_n)$.

c. Amenability. Two obvious questions arise from this argument. First of all, which groups possess Følner sequences? We call such groups *amenable*. Partial answers can be given ad hoc: $\mathbb{Z}$ and $\mathbb{R}$ do (consider sequences of ever longer intervals) and this property is preserved under taking products, so $\mathbb{Z}^n$ and $\mathbb{R}^n$ also have Følner sequences. More generally, finitely generated discrete groups of subexponential growth (of the number of group elements expressible in words of generators of a given length) have Følner sequences, *e.g.*, balls in the word length metric. Shifts of balls lie inside larger balls and too many large symmetric differences would imply exponential growth of the cardinality of balls with the radius. Abelian groups are amenable whether they are finitely generated or not. Furthermore, amenability is inherited by extensions (semidirect products) if both the base and the fiber are amenable. This gives amenability for all solvable groups. Notice that many such groups that are finitely generated in fact have exponential growth. Thus, Følner sets need not resemble balls in a word metric. Among connected Lie groups, compact extensions of solvable groups are the most general amenable groups. Typical examples of nonamenable discrete groups are free groups with more than one generator and groups containing them, such as $SL(n, \mathbb{Z})$, $n \geq 2$ [**Gl, Z**].

The second question is how Følner sequences may look like in a given group. The Følner sequences we just proposed for $\mathbb{Z}$ were quite simple, but already in this context some rather complicated sets would also satisfy the definition: Any sequence of sets that are unions of sufficiently long intervals, plus possibly some "sparse" further appendages, would qualify, because the symmetric difference is dominated by neighborhoods of the ends of the intervals. This is the reason why in general Følner sets are not the best device for studying more subtle issues in ergodic theory, such as pointwise convergence. Nevertheless, every Følner sequence contains a subsequence for which the Birkhoff ergodic theorem holds [**Li**].

We presently give one characterization of amenability and we present another when we discuss existence of invariant measures for group actions (Theorem 4.2.2).

d. Characterizations of amenability. Amenability can be characterized in other ways that are illuminating and useful in various situations. One of these is the *Kakutani–Markov fixed point property* [**Gl, Z**]:

THEOREM 1.4.1. *A group is amenable if and only if every affine action has a fixed point.*

Here, an *affine action* is an action on a weak*-compact convex subset X of the unit ball in the dual E^* of a separable Banach space that arises from X being an invariant set in E^* under the adjoint representation of some continuous isometric representation on E.

PROOF OF "ONLY IF". We show (in the discrete case) that existence of a left-Følner gives the desired fixed point. Let G be a group, $\{F_n\}$ a left-Følner, E a separable Banach space, $I\colon G \to \mathrm{Iso}(E)$ a representation, $\Phi\colon G \times E^* \to E^*$, $(g,\varphi) \mapsto \varphi \circ I^g$ the adjoint representation, and $X \subset E^*$ a weak*-compact Φ-invariant convex subset of the unit ball. If $\varphi \in X$ then $\mathcal{F}_n(\varphi) \in X$ by convexity; weak*-compactness implies that there is a sequence $n_k \to \infty$ such that $\mathcal{F}_{n_k}(\varphi) \to \varphi_0 \in X$. By the Følner condition $\Phi^g(\varphi_0) = \varphi_0$ for all $g \in G$. □

This immediately implies an alternative description in terms of existence of invariant measures for actions on compact spaces that we exhibit in Theorem 4.2.2. It is this characterization that is most transparently responsible for the fact that amenable group actions are the most general setting in which many aspects of dynamics in the "standard" sense can be pursued.

Here we give another characterization that refers only to the group.

THEOREM 1.4.2. *A group G is amenable if and only if it has an invariant mean,* i.e., *a positive linear functional of norm 1 on the space $C_b(X)$ of bounded continuous functions $G \to \mathbb{R}$ that is invariant under left translations.*

SKETCH OF PROOF OF "ONLY IF" IN THE DISCRETE CASE. Set $l = 0$ on the closure V of $\mathrm{span}\{\varphi - \varphi \circ L_g \mid \varphi \in C_b(X)\}$. Applying (1.1) over a Følner shows that no $\psi \in V$ has positive infimum and hence $1 \notin V$. Thus set $l(1) = 1$ and extend by the Hahn–Banach Theorem. □

For the converse, see [**Gl**].

CHAPTER 2

Topological dynamics

1. Setting and examples

a. Topological dynamical systems. Topological dynamics considers groups of homeomorphisms and semigroups of continuous transformations of topological spaces. We suppose X is a topological space, G a topological semigroup and $\Phi\colon G \times X \to X$ continuous such that $\Phi^{g_1 g_2} = \Phi^{g_2} \circ \Phi^{g_1}$ for all $g_1, g_2 \in G$.

Topological dynamics provides many basic concepts and paradigms of asymptotic behavior that are central for dealing with more refined settings such as smooth, symplectic, and homogeneous dynamics. Some of these concepts also serve as models for more quantitative counterparts in ergodic theory.

The prevalence of topological notions in the study of various classes of smooth systems is quite remarkable. For example, the concept of structural stability is quite substantial in differentiable dynamics: There are many differentiable dynamical systems whose C^1-perturbations are topologically conjugate (Section 6.7h), whereas the analogous smooth stability is vacuous in the classical setting of cyclic systems, although relevant for actions of larger groups beginning with higher-rank abelian ones [**S-FK**].

Furthermore, much of the description of the orbit structure of smooth systems is made in topological terms: Periodic orbits, recurrence, topological entropy, structural stability, attractors, *etc.* (We associate periodic points with the topological category by way of contrast to the measurable one, where individual points may not be meaningful.)

Several general observation concerning the general setting of topological dynamics are in order.

1. *Standing assumptions.* We henceforth make the standing assumptions that X is a complete metric space with countable base and G is a locally compact noncompact second countable topological (semi-) group. We call such an action a *topological dynamical system*.

Usually the metric on X is less important than the uniform structure it entails. The latter is needed to define notions of relative asymptotic behavior of orbits. The leading case is that of a compact Hausdorff space X with countable base, which is hence metrizable. Therefore we usually intend "compact" to mean compact Hausdorff with countable base.

2. *The compactness principle.* The essential reason for the compactness assumption is that packing "large" orbits into a compact space provides for some nontrivial asymptotic accumulation and hence recurrent behavior. Accordingly, results about existence of various kinds of recurrence require compactness of the phase space (Theorem 2.2.1, Proposition 2.3.1, the Kryloff–Bogoliouboff Theorem 4.2.2 *etc.*), whereas those pertaining to the

description of various relationships between diverse properties often hold in greater generality (*e.g.,* Proposition 2.3.5, Proposition 2.3.7, Lemma 2.3.9). We sometimes refer to the former observation as the *compactness principle*.

3. *Cyclic dynamical systems.* Special attention is given to the case of cyclic dynamical systems, *i.e.,* actions of $\mathbb{Z}$, $\mathbb{N}_0$, or $\mathbb{R}$. There is no universal rule that determines the generality in which any notion makes sense or any given result holds, but a useful guiding principle is that in the cyclic case there is only one way of going to infinity (except the distinction of $\pm\infty$), in that leaving compact sets is pertinent to all asymptotic behavior, whereas in larger groups the notion of asymptotic behavior involves many ambiguities, such as a choice of "directions" (determined by a generator, subgroup, factor, or an element of a group boundary) or of a growing family of compact sets exhausting the (semi-) group.

Topological dynamics plays an essential role in the surveys [**S-FM, S-B, S-KSS**].

We presently introduce a few classes of standard examples of topological dynamical systems that play an important role throughout this survey and elsewhere in these volumes.

b. Homogeneous dynamics. (See also [**S-KSS, S-FK**].) For a locally compact second countable metrizable group H with a right-invariant metric and a closed subgroup K the factor $M := H/K$ has a metric. Given another topological group G and a continuous homomorphism $\rho\colon G \to H$ there is a natural action $\Phi\colon G \times M \to M$ defined by left translations $\Phi(g, x) = \rho(g)x$. In particular, G may be a subgroup of H. The case of compact M, *i.e.,* that of cocompact subgroups $K \subset H$, is of particular interest. Compact H give a special case. The G-action on M is isometric if there is a left-invariant metric on M. This happens when H is abelian or, more generally, possesses a bi-invariant metric, or if K is compact. However, in many interesting situations this is not the case. Basic examples of this kind are in Section 4.3f and Section 6.5e. For further discussion, see Section 3.3c.

EXAMPLE 2.1.1. Consider the n-torus $\mathbb{T}^n = \mathbb{R}^n/\mathbb{Z}^n$ and for $\gamma \in \mathbb{R}^n$ (as generator of an embedding of $\mathbb{Z}$) the *translation* $T_\gamma\colon \mathbb{T}^n \to \mathbb{T}^n$, $x \mapsto x + \gamma \pmod 1$, which defines a $\mathbb{Z}$-action. In particular, for $n = 1$ we obtain a *rotation of the circle* $R_\alpha\colon x \mapsto x + \alpha \pmod 1$, arguably the most basic nontrivial example of a dynamical system.

Similarly, a one-parameter subgroup of the torus generates a flow of translations, the *linear flow*, which is the basic building block in integrable behavior of Hamiltonian dynamics.

c. Group automorphisms and endomorphisms. (See also [**S-LS, S**].) Another important class of examples consists of actions defined by discrete groups of automorphisms or semigroups of endomorphisms of a group H. This gives interesting dynamics already in some simple cases, *e.g.,* the *linear expanding maps* $E_m\colon x \mapsto mx \pmod 1$ ($m \in \mathbb{Z}$, $|m| \geq 2$) on the circle or automorphisms of the torus (defined by the action of an integer matrix with determinant ± 1 on $\mathbb{R}^n/\mathbb{Z}^n$). This subject is discussed further in Section 3.7g and Section 6.5a.

d. Shifts and symbolic systems. (See also Section 2.6, [**S-LS**].) For a discrete (semi-) group Γ and a compact (Hausdorff second countable) K let $H = K^\Gamma$ be the space of all maps $\eta\colon \Gamma \to K$ with the product topology. Then Γ acts on H via $(\gamma, \eta) \mapsto \eta \circ L_\gamma$. This action is called the *shift* or *Bernoulli action*.

The standard cases are those of finite K and $\Gamma = \mathbb{Z}$ or $\mathbb{N}_0$. They give rise to the *N-shift* σ_N on $\Omega_N = \{0, \dots, N-1\}^{\mathbb{Z}}$, where $N = \operatorname{card} K$ and, in case of $\Gamma = \mathbb{N}_0$ to the *one-sided N-shift* σ_N^R on $\Omega_N^R = \{0, \dots, N-1\}^{\mathbb{N}_0}$.

A *symbolic dynamical system* is the restriction of a (one-sided) N-shift to a closed invariant subset. Although this definition looks innocuous, even for $N = 2$ it produces a rich class of dynamical systems, which is not tractable in full generality (see Theorem 4.3.10).

When K has a topological group structure then so does K^Γ, and a shift acts by automorphisms or endomorphisms, so shifts also provide examples of actions by automorphisms or endomorphisms of compact groups.

For a comprehensive treatment of symbolic dynamical systems see [**LM**].

2. Basic concepts and constructions

This section revisits the basic notions and constructions from Section 1.3 in the topological setting.

a. Topological conjugacy and orbit equivalence. Let $\Phi\colon G \times X \to X$, $\Psi\colon G \times Y \to Y$ be topological dynamical systems. Φ and Ψ are said to be *topologically conjugate* if there exists a homeomorphism $h\colon X \to Y$ such that

$$h(\Phi(g, x)) = \Psi(g, h(x)) \text{ for all } g \in G, x \in X.$$

Φ and Ψ are said to be (topologically) *orbit equivalent* if there exists a homeomorphism $h\colon X \to Y$ sending orbits of Φ onto orbits of Ψ.

For actions of continuous groups, orbit equivalence is the more natural isomorphism notion. It classically appears in the qualitative theory of ordinary differential equations and reflects the aspects of the orbit structure transverse to orbits rather than the parametrization of orbits. This equivalence relation is more robust because, *e.g.*, in the case of flows, topological conjugacy preserves the periods of periodic points, whereas orbit equivalence does not. Accordingly, the concept of structural stability for flows involves topological orbit equivalence (Section 5.2f and [**S-H**]).

As was mentioned in Section 1.3c classification of general topological dynamical systems up to topological conjugacy or topological orbit equivalence is not feasible. The primary function of these notions in the framework of topological dynamics is to provide a background for describing various properties related to asymptotic behavior.

b. Invariant sets, inducing. The restriction of a topological dynamical system $\Phi\colon G \times X \to X$ to a closed invariant set A is again a topological dynamical system, which is sometimes denoted Φ_A.

The closure $\overline{\mathcal{O}(x)}$ of the orbit of a point $x \in X$ is a closed invariant set. If X is compact then the orbit itself is closed if and only if it is compact in the sense of Section 1.2d, *i.e.*, if $G/G(x)$ is compact.

As mentioned in Section 1.3f, one may try to "restrict" a map f to a noninvariant set A. This results in the *induced* or *first-return map* $x \mapsto f^{\min\{n \in \mathbb{N} \,|\, f^n(x) \in A\}}(x)$ defined on a possibly empty subset of A. There are problems with this construction other than that it may be defined for no point: Even if A is closed and the induced map is defined on a nonempty subset, it often fails to be continuous. In some cases this construction is nevertheless useful and we return to it in the setting of smooth dynamics (Section 5.2h).

c. Topological transitivity and minimality. A semigroup action $\Phi\colon G\times X\to X$ is said to be *topologically transitive* if there is an $x\in X$ such that for every $y\in X$ there is a sequence $g_k\to\infty$ such that $\Phi^{g_k}(x)\to y$. In particular, the orbit of x is dense. Topological transitivity is one of two natural notions of irreducibility in topological dynamics.

Nonempty closed invariant sets are partially ordered by inclusion. Any minimal element of this partial ordering is called a *minimal set*. Equivalently, $A\subset X$ is minimal if $\overline{\mathcal{O}(x)}=A$ for all $x\in A$. If X is minimal then Φ is said to be a *minimal dynamical system*. Minimality is the second and stronger irreducibility notion in topological dynamics.

If X is compact then any intersection of an ordered chain of closed invariant sets is nonempty, so by Zorn's Lemma there is a minimal element in the partial order. This implies

THEOREM 2.2.1. *Every topological dynamical system Φ of a compact metric space X has an invariant minimal subset.*

PROOF WITHOUT ZORN'S LEMMA. The collection $\mathcal{C}$ of closed invariant sets is compact with respect to the Hausdorff metric d_H on the spaces C_X of all closed subsets of X defined as $d_H(A,B)=\max\{\max_{x\in A}d(x,B),\max_{x\in B}d(x,A)\}$. Let $m(B)=\max\{d_H(A,B)\mid B\supset A\in\mathcal{C}\}$ for $B\in\mathcal{C}$ and take $M\in\mathcal{C}$ such that $m(M)=\min m=:m_0$. Then M is minimal, for otherwise $m_0>0$ and there exists a closed invariant $M_1\subset M$ such that $d_H(M_1,M)=m_0$. By assumption $m(M_1)\geq m_0$ so there is $M_2\subset M_1$ such that $d_H(M_2,M_1)\geq m_0$ and hence $d_H(M_2,M)\geq m_0$—inductively find M_i such that $d_H(M_i,M_j)\geq m_0$, contradicting compactness of C_X with respect to the Hausdorff metric. □

This result is the first instance of the "compactness principle" at work.

For noninvertible systems one can use forward invariant sets (Section 1.3f) to make the same definition and prove existence of minimal sets.

d. Examples of transitivity and minimality.

EXAMPLE 2.2.2. The translation T_γ on the torus (Section 2.1b) is topologically transitive if and only if the cyclic subgroup $\mathbb{Z}\gamma$ is dense or, equivalently, if the coordinates of the vector $\gamma=(\gamma_1,\dots,\gamma_n)$ and 1 are rationally independent.

This can be checked by the following general criterion:

PROPOSITION 2.2.3. *A translation L_g on a compact abelian group is topologically transitive iff only the trivial character is 1 at g.*

PROOF. If L_g is topologically transitive then it is minimal (all orbits are isometric). Thus, if χ is a character such that $\chi(g)=1$ then $\chi=1$ on $G=\overline{\{g^n\}}_{n\in\mathbb{Z}}$ by continuity. On the other hand, if $H:=\overline{\{g^n\}}_{n\in\mathbb{Z}}$ is a proper subgroup then a nontrivial character on G/H lifts to a nontrivial character χ on G with $\chi(g)=1$. □

PROPOSITION 2.2.4. *If Φ is an isometric group action then every orbit closure is a minimal set.*

PROOF. If $y,z\in\overline{\mathcal{O}(x)}$ then there exist $(g_n)_{n\in\mathbb{N}}$, $(\gamma_n)_{n\in\mathbb{N}}$ such that $d(\Phi^{g_n}(x),y)\to 0$ and $d(\Phi^{\gamma_n}(x),z)\to 0$. Therefore

$$d(\Phi^{g_n^{-1}\gamma_n}(y),z)\leq d(\Phi^{g_n^{-1}\gamma_n}(y),\Phi^{\gamma_n}(x))+d(\Phi^{\gamma_n}(x),z)=d(\Phi^{g_n}(x),y)+d(\Phi^{\gamma_n}(x),z)\to 0$$

and the orbit of y is also dense in $\overline{\mathcal{O}(x)}$. □

In particular, left group translations have this property because one can consider a left invariant metric.

EXAMPLE 2.2.5. Any translation T_γ with $(1, \gamma_1, \dots, \gamma_n)$ rationally independent is minimal.

Transitivity and minimality are distinct.

EXAMPLE 2.2.6. The shift σ_2 (Section 2.1d) is topologically transitive (concatenating all possible finite 0-1-sequences gives a point with dense orbit) but also has nondense orbits, such as fixed points (constant sequences). Furthermore periodic points are dense, so there are many nondense orbits.

e. Isolated sets and attractors. (See also [**S-FM**].) An invariant set A of an invertible dynamical system is said to be *isolated* or *locally maximal* if there exists an open neighborhood $U \supset A$, called an *isolating neighborhood*, such that $\mathcal{O}(x) \subset U \implies x \in A$. Equivalently, there is a neighborhood V of A such that any closed invariant set $B \subset V$ satisfies $B \subset A$.

For cyclic dynamical systems there is a particular class of invariant sets that occupies a central place in the study of dynamical systems, notably dissipative ones: A compact set $A \subset X$ is said to be an *attractor* for Φ if there is a neighborhood V of A and a T with $\Phi^T(V) \subset V$ and $A = \bigcap_{t>0} \Phi^t(V)$. In this case, the complete preimage $\{\Phi^{-t}(V) \mid t > 0\}$ is called the *basin of attraction* of A. Attractors are isolated in the invertible case.

A compact set $A \subset X$ is said to be a *repeller* if it has a neighborhood U such that for every $x \in U \setminus A$ there is a $T > 0$ with $\Phi^T(x) \notin U$. Repellers are also isolated invariant sets.

Note that products of isolated invariant sets are themselves isolated invariant sets under the product action, and that products of attractors are again attractors.

EXAMPLE 2.2.7. The origin is an isolated invariant set for a linear map $L\colon \mathbb{R}^n \to \mathbb{R}^n$ if and only if L is *hyperbolic*, *i.e.,* has no eigenvalues on the unit circle in $\mathbb{C}$ (this follows, *e.g.,* from the Jordan normal form).

It is an attractor if and only if all eigenvalues are inside the unit circle.

f. Factors and almost isomorphism. Let $\Phi\colon G \times X \to X$, $\Psi\colon G \times Y \to Y$ be topological dynamical systems. Ψ is said to be a (topological) *factor* of Φ if there exists a surjective continuous map $h\colon X \to Y$ such that $h(\Phi(g, x)) = \Psi(g, h(x))$ for all $g \in G, x \in X$. Accordingly, the action Φ is an *extension* of Ψ.

Ψ is said to be a (topological) *orbit factor* of Φ if there exists a surjective continuous map $h\colon X \to Y$ mapping orbits onto orbits.

In some important situations the factor map is injective on a large set, such as open dense or dense G_δ [**AdM**]. Although the spaces may be far from homeomorphic, the factor map is almost a conjugacy in these cases. We call a factor map an *almost-isomorphism* or *almost-conjugacy* if it is injective on a dense G_δ.

A simple natural example is related to the binary expansion of real numbers.

EXAMPLE 2.2.8. Let $\sigma_2^R\colon \{0,1\}^{\mathbb{N}_0} \to \{0,1\}^{\mathbb{N}_0}$ be the one-sided 2-shift (Section 2.1d). The map $E_2\colon x \mapsto 2x \pmod 1$ on $Y = S^1 = \mathbb{R}/\mathbb{Z}$ is a factor by binary expansion $h(\omega_0\omega_1\dots) = 0.\omega_0\omega_1\dots \pmod 1$. The factor map is injective away from binary rationals, hence on a dense G_δ. Although the Cantor set $\{0,1\}^{\mathbb{N}_0}$ and the circle are topologically

distinct, the main dynamical properties of both systems are similar, as the existence of this almost-conjugacy suggests.

A similar, but geometrically much more intersting example of the same kind is the coding of a hyperbolic automorphism of the 2-torus via a topological Markov chain, described, for example in [**KH**, Section 2.5]. Section 6.7g explains that coding is generally possible in hyperbolic dynamics. (See also [**S-C**].)

g. Inverse limits. Suppose X_i ($i \in \mathbb{N}$) are compact metrizable spaces with continuous surjections $h_i \colon X_{i+1} \to X_i$ and consider the compact metric space

$$\mathcal{X} := \{(x_i)_{i\in\mathbb{N}} \mid h_i(x_{i+1}) = x_i\}, \quad \operatorname{dist}(x,y) := \sum_i 2^{-i} \operatorname{dist}_{X_i}(x_i, y_i).$$

If the h_i are factor maps for group actions on the X_i then the action is naturally defined on $\mathcal{X}$, which is then called an *inverse limit*.

EXAMPLE 2.2.9. Let $X_i = \mathbb{Z}/2^i\mathbb{Z}$ with h_i the natural projection. These are factors for the $\mathbb{Z}$-action generated by adding 1. Then the inverse limit is the additive group $\mathbb{Z}_2$ of dyadic integers, which is the dual group to the discrete group of all binary rationals mod 1 and is homeomorphic to a Cantor set. The resulting dynamical system which is generated by the map $x \to x + 1$ on $\mathbb{Z}_2$ is an example of the class of sytems called the *adding machines* or *odometers*.

h. Natural extension. The *natural extension* of a continuous surjective map f of a compact metric space X is obtained by taking $X_i = X$ and $h_i = f$. The inverse limit $\hat{f}$ on $\hat{X} := \mathcal{X}$ is given by $\{x_1, x_2, \dots\} \mapsto \{f(x_1), f(x_2), \dots\} = \{f(x_1), x_1, x_2, \dots\}$.

EXAMPLE 2.2.10. Starting from the one-sided shift this gives the two-sided shift.

EXAMPLE 2.2.11. From $x \mapsto 2x$ on S^1 one gets an automorphism of a solenoid, namely the dual group to the discrete group $\mathbb{Z}[1/2]$ of all binary rationals. A smooth realization is given by the Smale attractor (Section 5.2i, Section 6.5c).

i. Isometric extensions. A topological dynamical system $\Phi \colon G \times X \to X$ is an *isometric extension* of $\Psi \colon G \times Y \to Y$ with respect to a metric d in X if Φ is an extension of Ψ with factor map h and in addition $d(\Phi(g, x_1), \Phi(g, x_2)) = d(x_1, x_2)$ for any $g \in G$ and any $x_1, x_2 \in X$ with $h(x_1) = h(x_2)$. Isometric extensions are building blocks in the classification of distal dynamical systems (Section 2.4c).

A particular case of an isometric extension appears when X is a locally trivial fiber bundle over Y with compact structure group H, acting transitively on the fibers $X_y = h^{-1}(y)$, $y \in Y$ in such a way that the metric in X is H-invariant and the extension Φ commutes with the action of H in the fibers. Then every fiber is naturally identified with a homogeneous space of the group H. A particularly simple situation of this type appears when X is the principal bundle, *i.e.,* the fibers can be identified with the group H itself. An extension from this class is called a *group extension*.

EXAMPLE 2.2.12. Let $\phi \colon S^1 \to S^1$ be a continuous map. Then $F \colon \mathbb{T}^2 \to T^2$, $F(x, y) = (x + \alpha, y + \phi(x))$ is an S^1-extension of the rotation R_α. For $\phi = \text{const}$ this gives a translation on the torus. For $\phi = E_m$, $m \in \mathbb{Z} \setminus \{0\}$, this is an affine map, which will appear on numerous occasions later on (Section 4.3e, Section 4.3i, Corollary 7.5.4)

j. Suspensions. In the topological category suspension produces a space that is a locally trivial fiber bundle over $H\backslash G$ with fiber X. It is compact if and only if X is compact and H is cocompact in G. Topological transitivity and minimality of the H-action are inherited by the suspension action.

k. Cocycles and skew products. Clearly cocycles (Section 1.3k) and their cohomology have to be considered in the topological category now. A new facet that arises in regard to skew products is that there are locally trivial bundles that are topologically nontrivial (such as tangent bundles when the derivative extension is being considered). In this case skew products are not generated by cocycles. In fact, group extensions of actions to nontrivial principal bundles and, more generally, to locally trivial bundles with an H action, provide a natural generalization of the notion of cocycle in the topological setting [**S-FK**].

l. Induced action. In the construction of an induced action, conditions on the cocycle are needed in order to produce a Hausdorff space of orbits for Φ_α. A convenient condition is a co-Lipschitz or bounded contraction property:

$$\exists C > 0\ \forall x \quad \operatorname{dist}_H(h_1, h_2) \le C \operatorname{dist}_G(\alpha(h_1, x), \alpha(h_2, x)),$$

where $\operatorname{dist}_{(\cdot)}$ are distances induced from left-invariant metrics.

A flow under a continuous function over a homeomorphism of a compact space is topologically orbit equivalent to the corresponding suspension flow.

m. Principal classes of asymptotic properties and invariants. The study of dynamical systems relies on a collection of notions describing various aspects of asymptotic behavior of individual orbits, pairs of orbits relative to each other, or larger collections of orbits. In topological dynamics one can separate several principal categories of such notions:

(1) Types of recurrence,
(2) behavior of orbits relative to each other,
(3) growth of the number of orbits of various kinds and the complexity of various families of orbits, and
(4) asymptotic distribution of orbits in a statistical sense.

The first three classes are of a purely topological nature and are discussed in this chapter. The last class is related to ergodic theory and invariant measures for topological dynamical systems. The corresponding notions are accordingly discussed in Chapter 4 after the introduction to ergodic theory.

Many of these notions give rise to invariants, which accordingly can be divided into the corresponding categories and are discussed in turn.

3. Recurrence

a. Limit points. If $\Phi\colon G \times X \to X$ is a semigroup action then $y \in X$ is said to be a *limit point* for $x \in X$ if there is a sequence $g_k \to \infty$ such that $\Phi^{g_k}(x) \to y$. The *limit set* of $x \in X$ is then the set of limit points of x. (See [**KH**, Section 3.3].)

PROPOSITION 2.3.1. *If X is compact then every limit set is nonempty.*

Thus every point sooner or later comes to any given neighborhood of its limit set and stays there. Note that the definition of topological transitivity amounts to requiring that the whole phase space is a limit set.

For cyclic dynamical systems the ordering provides for a distinction between $+\infty$ and $-\infty$ and accordingly one can define two notions:

For a cyclic dynamical system a point $y \in X$ is called an *ω-limit point* for $x \in X$ if there is a sequence $t_k \to +\infty$ such that $\Phi^{t_k}(x) \to y$. If Φ is an $\mathbb{R}$ or $\mathbb{Z}$ action then y is an *α-limit point* for x if it is an ω-limit point for x under reversal of time. The closed invariant sets

$$\omega(x) = \bigcap_{T\geq 0} \overline{\bigcup_{t\geq T} \Phi^t(x)}, \quad \alpha(x) = \bigcap_{T\leq 0} \overline{\bigcup_{t\leq T} \Phi^t(x)}.$$

of all ω-limit points and α-limit points for x are called its *ω-limit* and *α-limit set*.

b. Recurrence. For cyclic dynamical systems we say that $x \in X$ is *positively recurrent* if $x \in \omega(x)$. If Φ is a $\mathbb{Z}$- or $\mathbb{R}$-action then x is said to be *negatively recurrent* if $x \in \alpha(x)$, it is *recurrent* if it is both positively and negatively recurrent. Denote the closures of the sets of all positively recurrent, negatively recurrent, and recurrent points by $R^+(\Phi)$, $R^-(\Phi)$, and $R(\Phi)$.

Positive recurrence does not necessarily imply negative recurrence and the sets of all positively recurrent, negatively recurrent, and recurrent points need not be closed.

Periodic points represent the simplest recurrence. However, not every dynamical system has periodic orbits, even if the phase space is compact. The presence of nonperiodic recurrent points is often referred to as *nontrivial recurrence*, especially in the literature on ordinary differential equations. It is the first indication of complicated asymptotic behavior. In certain low-dimensional situations such as homeomorphisms of the circle and flows on surfaces it is possible to give a comprehensive description of the nontrivial recurrence that can appear [**KH**, Chapters 11, 14].

EXAMPLE 2.3.2. A left translation ($\mathbb{Z}$-action) by an element $h \in H$ on a group H (see Section 2.1b) has no recurrent points if the subgroup $(h^n)_{n\in\mathbb{Z}}$ is closed in H. Otherwise all points are recurrent.

Since every point of a minimal set for a cyclic system is obviously recurrent, Theorem 2.2.1 implies

COROLLARY 2.3.3. *If X is compact and Φ is cyclic then $R(\Phi) \neq \varnothing$.*

c. Minimality and uniform recurrence. Every point of a minimal set is recurrent. Indeed, minimality can be characterized by recurrence that is uniform in a very general sense:

If G is a locally compact topological (semi-) group then $S \subset G$ is said to be *syndetic* if there exists a compact $K \subset G$ such that $SK^{-1} = G$. If $\Phi\colon G \times X \to X$ is an action then $x \in X$ is said to be *uniformly recurrent* if for each neighborhood V of x the set $\{g \in G \mid \Phi^g(x) \in V\}$ is syndetic. In the cases of $\mathbb{Z}$, $\mathbb{N}_0$, $\mathbb{R}$, or $\mathbb{R}_0^+$ this means that there is an upper bound for the length of complementary intervals.

PROPOSITION 2.3.4. *Every point of a compact minimal set of a topological dynamical system is uniformly recurrent. Conversely, if X is locally compact and a point $x \in X$ is uniformly recurrent then the closure of its orbit is a compact minimal set* [**F1**, Section 1.4].

d. Nonwandering points, regional recurrence and the center. (See also [**Bi**].) So far we were concerned with recurrence properties directly associated with individual orbits. There are others related to the behavior of entire sets. The simplest such property is the following:

A point $x \in X$ is said to be *nonwandering* with respect to Φ if for any open set $U \ni x$ and $T > 0$ there is a $t > T$ such that $\Phi^t(U) \cap U \neq \varnothing$. The set of all nonwandering points of Φ is denoted by $NW(\Phi)$ or $\Omega(\Phi)$. Φ is said to be *regionally recurrent* if $\Omega(\Phi) = X$. For reversible Φ, a nonwandering point x, and open $V \ni x$ there are also arbitrarily large negative T such that $\Phi^T(V) \cap V \neq \varnothing$.

PROPOSITION 2.3.5. *$\Omega(\Phi)$ is closed and invariant and contains all ω- and α-limit points for all points.*

PROOF. We only consider maps. If $\Omega(f) \ni x_n \to x \in U$ open then $x_n \in U$ for large enough n, so $f^N(U) \cap U \neq \varnothing$ for arbitrarily large N and thus $x \in \Omega(f)$. If $x \in \Omega(f)$, $f(x) \in U$ and $V = f^{-1}(U)$ then $V \cap f^N(V) \neq \varnothing$ for some $N > 0$ and $\varnothing \neq f(V \cap f^N(V)) = U \cap f^N(U)$. If $x = \lim_{n_k \to \infty} f^{n_k}(y) \in U$ and n_k is increasing then $f^{n_k}(y), f^{n_{k+1}}(y) \in U$ for large k, so $U \cap f^{n_{k+1}-n_k}(U) \neq \varnothing$. The argument for α-limit points is similar. □

COROLLARY 2.3.6. *If X is compact then $\Omega(\Phi) \neq \varnothing$.*

PROPOSITION 2.3.7. *If Φ is regionally recurrent then $R(\Phi) = X$.*

Let $\Omega_1(\Phi) = \Omega(\Phi)$ and $\Omega_{n+1}(\Phi) = \Omega(\Phi_{\restriction \Omega_n(\Phi)})$. This yields a nested sequence with intersection $\Omega_\omega(\Phi)$ and then the construction can be started again, so (if there is a countable base) by transfinite induction up to a countable ordinal we obtain the *center* of the dynamical system. In virtually all interesting examples this construction stabilizes quickly, at most after one or two steps, so the center is either $\Omega(\Phi)$ or $\Omega_2(\Phi)$. It is not difficult, however, to construct examples where this is not so. Since recurrent points are defined intrinsically using only their own orbits, Proposition 2.3.5 can be applied inductively to see that $R(\Phi)$ is contained in the center. Since the construction of the center stabilizes, there are no wandering points in the center, and by Proposition 2.3.7 we find

PROPOSITION 2.3.8. *$R(\Phi)$ is the center of Φ.*

Thus $\Omega(\Phi)$ is the hub of recurrence behavior: It contains all α- and ω-limit points and recurrent points, including all periodic points.

Denote by $M(\Phi)$ the closure of the union of all invariant minimal sets for Φ. Then

$$\operatorname{Per}(\Phi) \subset M(\Phi) \subset R(\Phi) \subset R^+(\Phi) \cup R^-(\Phi) \subset \Omega(\Phi). \tag{2.1}$$

Each of these inclusions may be proper and by Theorem 2.2.1 all sets in (2.1), except possibly $\operatorname{Per}(\Phi)$, are nonempty for compact metric X.

e. Topological transitivity and topological mixing. One can define topological transitivity in terms of asymptotic behavior of sets.

LEMMA 2.3.9. *A dynamical system $\Phi\colon G \times X \to X$ on a complete separable metric space X is topologically transitive if and only if for any two nonempty open sets $U, V \subset X$ there exists $g \in G$ such that $\Phi^g(U) \cap V \neq \varnothing$.*

PROOF. For maps necessity goes as follows: If $\mathcal{O}_f(x)$ is dense then it intersects U and V, so $f^n(x) \in U$, $f^m(x) \in V$, where, say $m \geq n$. Consequently $f^{m-n}(U) \cap V \neq \varnothing$. This clearly works without invertibility and generalizes to groups (but not semigroups).

If the intersection condition holds let $U_1, U_2, \dots$ be a countable base of open subsets of X with $\overline{U}_1$ compact. There exists $g_1 \in G$ such that $\Phi^{g_1}(U_1) \cap U_2 \neq \varnothing$. If $V_1 \neq \varnothing$ is open and $\overline{V}_1 \subset U_1 \cap (\Phi^{g_1})^{-1}(U_2)$ then there exists $g_2 \in G$ such that $\Phi^{g_2}(V_1) \cap U_3 \neq \varnothing$. Again, take an open set V_2 such that $\overline{V}_2 \subset V_1 \cap \Phi^{-g_2}(U_3)$. Inductively, construct a nested sequence of open sets V_n such that $\overline{V}_{n+1} \subset V_n \cap \Phi^{g_{n+1}^{-1}}(U_{n+2})$. Then $V = \bigcap_{n=1}^{\infty} \overline{V}_n = \bigcap_{n=1}^{\infty} V_n \neq \varnothing$ because the $\overline{V}_n$ are compact. If $x \in V$ then $\Phi^{g_{n-1}}(x) \in U_n$ for every $n \in G$. □

COROLLARY 2.3.10. *A continuous open dynamical system (*i.e., *one that maps open sets to open sets) of a complete separable metric space is topologically transitive if and only if there are no two disjoint open nonempty invariant sets.*

COROLLARY 2.3.11. *If Φ is topologically transitive then there is no nonconstant invariant continuous function $\varphi\colon X \to \mathbb{R}$.*

Another aspect of asymptotic behavior is related to regularity of set recurrence with respect to time. Topological transitivity implies that iterates of any open set from time to time intersect any other open set. Here is a stronger property:

DEFINITION 2.3.12. A topological dynamical system Φ is said to be *topologically mixing* if for any two nonempty open $U, V \subset X$ there exists a compact set $K \subset G$ such that $\Phi^g(U) \cap V \neq \varnothing$ for every $g \in G \setminus K$.

By Lemma 2.3.9, every topologically mixing dynamical system is topologically transitive.

Notice that minimality, which is also a stronger property than topological transitivity, concerns the regularity of returns for individual orbits (Proposition 2.3.4). To demonstrate the difference between minimality and mixing, note that topological transitivity and minimality are equivalent for actions by isometries (Proposition 2.2.4), but on the other hand, mixing is impossible:

PROPOSITION 2.3.13. *Isometric actions are not topologically mixing if* $\operatorname{card} X > 1$.

PROOF. For $\operatorname{card} X = 2$ this is trivial. For $\operatorname{card} X > 2$ and isometric $\Phi^g\colon X \to X$ take $\{x_1, x_2, x_3\} \subset X$ such that $0 < \delta := \min_{i \neq j} d(x_i, x_j)/10$ and let $U_i = B(x_i, \delta)$ for $i \in \{1, 2, 3\}$. The diameter of $\Phi^g(U_1)$ is at most 2δ whereas $d(p, q) > \delta$ for $p \in U_2$, $q \in U_3$, so for $g \in G$ either $\Phi^g(U_1) \cap U_2 = \varnothing$ or $\Phi^g(U_1) \cap U_3 = \varnothing$. □

f. Homological and homotopical recurrence, asymptotic cycles. For flows, there is a way to quantify the character of recurrence by considering homotopical or homological properties of long orbit segments. Given a compact connected manifold M and $p \in M$ fix a family $\Gamma = \{\gamma_x \mid x \in M\}$ of arcs γ_x of bounded length connecting p and x. Then for a flow $\varphi^t\colon M \to M$ fix T and consider for each $x \in M$ the closed loop $l(x, T)$ consisting of the arc γ_x, the orbit segment $\{\varphi^t x\}_{t=0}^{T}$, and the reverse of the arc $\gamma_{\varphi^T(x)}$. Those loops represent elements of the fundamental group $\pi_1(M, p)$. Via the Hurewicz identification of the first homology group $H_1(M)$ with $\pi_1(M, p)/[\pi_1(M, p), \pi_1(M, p)]$, they also give homology classes $c(x, T)$. Any limit point of $\{c(x, T)/T\}_T$ is called an *asymptotic cycle*

for x. Existence and the value of the limit of a sequence $\{c(x,T_n)\}/T_n$ are independent of the choices of p and Γ. Ergodic theory implies that $c(x,T)/T$ converges for many $x \in M$, *i.e.,* the asymptotic cycle is uniquely defined. One may also consider the asymptotic behavior of the $l(x,T) \in \pi_1(M)$ directly, using such structures as boundaries of groups (e.g. the Aranson–Grines homotopy rotation class for flows on surfaces, Section 2.7b).

For discrete time systems, these constructions can be applied to the suspension flow in order to obtain asymptotic cycles in the homology group of the suspension manifold or limits on the boundary of the fundamental group.

We sketch a proof of convergence in the smooth setting [**KH**, Section 14.7b]. Suppose M is a differentiable manifold and φ^t preserves a finite measure μ. A homology class is fixed by its action on a basis of 1-forms, so pick a 1-form ω. Taking the arcs in Γ to have bounded length and denoting by X the vector field generating the flow φ^t we find that $\lim_{T\to\infty} \int_{l(x,T)} \omega = \lim_{T\to\infty} \int_0^T \omega(X(\varphi^t(x)))\,dt = \int \omega(X)\,d\mu$ for μ-a.e. $x \in M$. Here the first equality reflects boundedness of the contribution of the two arcs of Γ to the integral, and the second is a consequence of the Birkhoff Ergodic Theorem (Theorem 3.5.2). If there is only one invariant probability measure, convergence is uniform in x (Section 4.3a).

g. Rotation number. There is one classical case where the above construction always produces asymptotic cycles independently of the point x. This is that of circle homeomorphisms and flows on $\mathbb{T}^2$ without fixed points and periodic orbits. [**KH**]. Taking $p = 0 \in \mathbb{T}^2$ and constructing Γ from paths pieced together from horizontal intervals and orbit segments, one sees that the asymptotic cycle is the asymptotic speed ("rotation vector") of an orbit for the lift of the toral flow to $\mathbb{R}^2$. By periodicity and the fact that orbits cannot cross, this is well-defined and independent of all choices. For this the assumption that there are no periodic orbits is needed. (This also follows from the last remark above because there is only one invariant probability measure.) Starting from a circle homeomorphism without periodic orbits, one obtains this result via suspension, but because the suspension has unit speed one can consider simply the slope rather than the direction. This quantity can, in fact, be calculated directly and in this case there are no restrictions on periodic orbits.

For a homeomorphism $f\colon S^1 \to S^1 = \mathbb{R}/\mathbb{Z}$ and a lift $F\colon \mathbb{R} \to \mathbb{R}$ (*i.e.,* $f \circ \pi = \pi \circ F$ for the standard projection $\pi\colon \mathbb{R} \to S^1$) it is not hard to show (using essential subadditivity of $a_n := F^n(x) - x$) that $\rho(f) := \pi(\lim_{|n|\to\infty}(F^n(x) - x)/n)$ is well-defined independently of x. This is the *rotation number* of f. Note that it is defined via the Birkhoff average of the "displacement" function $F - \mathrm{Id}$. This admits numerous generalizations. Independence of x can be shown to imply that the rotation number is rational if and only if there is a periodic point. It is evidently a conjugacy invariant. For sufficiently smooth circle diffeomorphisms without periodic points it is a complete invariant (Theorem 5.1.1).

In general, possible orbits in the cases of rational or irrational rotation number, respectively, are described via the Poincaré classification [**S-JS**], [**KH**, Chapter 11]. If there are any periodic orbits (rational rotation number) then they have the same period and all orbits are ordered exactly as the orbits of the corresponding rotation. Any nonperiodic points are positively and negatively asymptotic to periodic orbits (to the same if there is only one). These facts can be seen by using that $f^n(p) = p$ implies that f^n can be identified with an orientation-preserving homeomorphism of $[0,1]$. For irrational rotation number all orbits are ordered as for the corresponding rotation. All orbits have the same ω-limit set Λ (Section 2.3a), which is a perfect set, *i.e.,* either a Cantor set or the circle. The corresponding

rotation is a factor of the restriction to Λ, and topologically equivalent if $\Lambda = S^1$. Orbits not in Λ are positively and negatively asymptotic to Λ. Circle homeomorphisms with no periodic point, *i.e.,* with irrational rotation number, are uniquely ergodic and measure-theoretically a rotation (Section 3.4a).

4. Relative behavior of orbits

a. Proximality and distality. (See also [**F1, F2**].) Let $\Phi\colon G \times X \to X$ be a (semi-) group action. Then $(x, y) \in X \times X$ is said to be *proximal* if there exists a sequence $(g_n)_{n\in\mathbb{N}}$ in G such that $d(\Phi^{g_n}(x), \Phi^{g_n}(y)) \to 0$, *distal* otherwise. A point x is said to be *distal* if (x, y) is proximal only for $y = x$.

Φ is said to be proximal if all pairs of points are proximal, distal if all points are distal.

If X is compact then proximality is equivalent to the orbit under $\Phi\times\Phi$ of (x, y) having a limit point on the diagonal, and (x, y) is distal if and only if its orbit closure is disjoint from the diagonal, *i.e.,* if $d(\Phi^g(x), \Phi^g(y))$ is bounded away from 0.

Obvious examples of distal dynamical systems are isometries and, more generally, equicontinuous dynamical systems (which have an invariant metric, so these classes are topologically the same).

In general, proximal and distal behavior is interspersed in the same system.

EXAMPLE 2.4.1. In the two-shift σ_2 for any $\omega \in \Omega_2$ the set of ω' such that (ω, ω') is proximal is dense. Take, for example, those ω' with $\omega'_i = \omega_i$ for large i to get $d(\sigma_2^n\omega, \sigma_2^n\omega') \to 0$.

On the other hand, if $\omega \neq \omega'$ are periodic then (ω, ω') is distal because the orbit of (ω, ω') in the product is compact and disjoint from the diagonal.

b. Examples of proximal actions. Natural examples of cyclic proximal systems have very simple recurrent behavior.

EXAMPLE 2.4.2. Identify the circle S^1 with the projective line $\mathbb{R}\cup\{\infty\}$. and consider the map $x \to x + 1$. It is proximal.

For proximal $\mathbb{Z}$-actions this example is fairly representative:

PROPOSITION 2.4.3. *A minimal set for a proximal $\mathbb{Z}$-action is a fixed point.*

PROOF. If x is in the minimal set of a transformation f then there is a sequence $n_k \to \infty$ such that $d(f^{n_k}(x), f^{n_k}(f(x))) \to 0$. By compactness there is an accumulation point z of $(f^{n_k}(x))_{k\in\mathbb{N}}$, which must then be fixed. Thus the minimal set contains, and hence is, a fixed point. □

On the other hand, for actions of some large groups there are natural proximal actions with complicated recurrence structure.

EXAMPLE 2.4.4. With the identification described in the previous example the group $SL(2, \mathbb{R})$ acts on the circle by projective (fractional–linear) transformations: For $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{R})$ we define $\Phi^A(x) = \dfrac{ax + b}{cx + d}$. This action is transitive and proximal. Its restriction to the group $SL(2, \mathbb{Z})$ gives an example of a minimal proximal action of a discrete (countable) group. (See Example 3.3.2)

c. Classification of distal systems. (See also [**El, F2**].) Distal systems can be viewed as a natural generalization of isometric actions. Compare the following property with Proposition 2.2.4:

PROPOSITION 2.4.5. *A distal point is uniformly recurrent and (by Proposition 2.3.4) any orbit closure in a distal system is a minimal set.*

COROLLARY 2.4.6. *A distal system uniquely decomposes into invariant minimal sets.*

PROPOSITION 2.4.7. *An isometric extension of a distal system is itself distal.*

Thus Example 2.2.12 provides a collection of distal systems. If $\deg \phi \neq 0$ such a system is obviously not equicontinuous. See Section 4.3e, Section 8.3a and Example 8.3.2 for further examples of distal systems.

Beginning from a minimal isometry one can take any isometric extension, decompose it into minimal components each of which is a minimal isometric extension of the original system and continue this process using induction, transfinite if necessary, but up to a countable ordinal. The result will always be a minimal distal system on a metrizable compact space. It is quite remarkable that every distal system can be obtained by such a process.

THEOREM 2.4.8. [**F2**] *Any distal minimal system is topologically conjugate to a system obtained from the dynamical system on a point by a (possibly transfinite) succession of isometric extensions.*

The crucial part of the proof which allows to begin the induction as well as carry out the inductive step is the following.

PROPOSITION 2.4.9. *Every minimal distal system has an equicontinuous factor.*

d. Expansiveness. (See also [**KH**, Section 3.2g].) An action $\Phi\colon G \times X \to X$ of a discrete (semi-) group G is said to be *expansive* if there exists a number $\delta > 0$, called an *expansivity constant*, such that if $d(\Phi^g(x), \Phi^g(y)) < \delta$ for all $g \in G$ then $x = y$. The maximal such number is called *the* expansivity constant of the action Φ. Equivalently, the action is expansive if the diagonal in $X \times X$ is an isolated invariant set for the diagonal action.

EXAMPLE 2.4.10. The shift σ_N is expansive because for $x \neq y$ there exists an $n \in \mathbb{Z}$ such that $\sigma_N^n(x)$ and $\sigma_N^n(y)$ have distinct zero coordinates and are hence more than a certain fixed distance away from each other.

The restriction of an expansive system to a closed invariant subset is clearly expansive. Thus, in particular, any symbolic system (see Section 2.6) is expansive. This shows that there is no hope for a classification or comprehensive description of expansive systems along the lines of that for distal systems. However, under some conditions on the phase space, expansivity becomes a strong property, which makes a good structural description possible. For example, expansive homeomorphisms of compact surfaces have been classified [**L**].

Products of expansive actions are again expansive.

PROPOSITION 2.4.11. *If f is an expansive continuous map and $h \circ f = f \circ h$, $d(h(x), x) < \delta$ for all $x \in X$ then $h = \mathrm{Id}$.*

PROPOSITION 2.4.12. *Let f be an expansive map, $n \in \mathbb{N}$. Then the set of periodic points of period up to n consists of isolated points.*

Defining expansivity for continuous groups requires allowing for a relative drift of orbits in the time direction in order to capture actual divergence of orbits. For flows one handles this as follows:

DEFINITION 2.4.13. A continuous flow φ^t is said to be *expansive* with expansivity constant $\delta > 0$ if for any two points x, y we have the following implication: If there is a continuous function $s\colon \mathbb{R} \to \mathbb{R}$ with $s(0) = 0$, $d(\varphi^t(x), \varphi^{s(t)}(x)) < \delta$ and $d(\varphi^t(x), \varphi^{s(t)}(y)) < \delta$ for all $t \in \mathbb{R}$ then $y \in \mathcal{O}(x)$.

For further uses of expansiveness see Section 2.5f6, Section 4.4e and Section 6.7c.

5. Orbit growth properties

Growth properties relate to the numbers of orbits or families of orbits of various kinds as time goes to infinity. Thus, these properties are usually defined unambiguously for cyclic systems, where there is essentially only one way of going to infinity. For more general systems, growth invariants may be of "global type", usually associated with a choice of a growing family of compact sets exhausting G, or of a "partial type", measuring growth in various directions. The difficulties and uncertainty of the choices involved are but one reason to restrict attention to the case of cyclic systems.

Nevertheless, there will be occasions to mention such notions in connection with higher-rank abelian groups, where these complications are manageable.

a. Periodic orbits. Periodic orbits (see Section 1.2d) represent the most distinctive special class of orbits. Finding periodic orbits and studying their asymptotic growth and spatial distribution is one of the central aims in dynamics. It is also closely related to various questions in number theory, algebraic geometry and statistical mechanics. Accordingly, various aspects of this subject are broadly represented in these volumes. ζ-functions (see below) and related topics are the main subjects of [**S-P**], and also appear briefly in [**S-C**] as well as in [**S-FM**], among whose main topics is to find and classify periodic points. Finding periodic orbits was the original goal of the variational approach to dynamics and remains central to that area, see [**S-BK**], [**S-R**] and [**HZ**]. There are also connections with number theory [**S-KSS**].

One can count periodic points or periodic orbits. In the discrete-time case it is frequently convenient to count periodic points with (not necessarily minimal) period n, whereas for continuous-time systems one has no choice but to count periodic orbits. These numbers are useful and finite when periodic orbits are isolated. While this is generically the case, there are classes of systems, such as systems with symmetries, where periodic orbits naturally appear in infinite families. It is natural to count connected components of such sets of orbits because usually calculations of the number of periodic points result in the number of connected components by solving systems of equations and obtaining joint level sets of certain functions. In the case of isolated periodic orbits this gives the same number.

DEFINITION 2.5.1. Let Φ be a discrete-time dynamical system. Then we denote by $P_n(\Phi)$ the number of connected components of the set of periodic points of Φ with (not necessarily minimal) period n. If Φ is a continuous-time dynamical system we denote by $\mathcal{P}_t(\Phi)$ the number of connected components of the set of periodic orbits of period up to t.

While there are situations where this formulation is necessary, we repeat that generically periodic orbits are isolated.

This construction becomes problematic when there are connected sets of periodic orbits with varying periods. This often happens in Hamiltonian systems and in that case more elaborate counting methods should be used.

The most natural measure of asymptotic growth of the number of periodic points is the exponential growth rate $p(\Phi)$ of $P_n(\Phi)$ and $\mathcal{P}_t(\Phi)$:

$$p(\Phi) = \begin{cases} \overline{\lim}_{n\to\infty} \log(P_n(\Phi))/n & \text{for discrete time} \\ \overline{\lim}_{t\to\infty} \log(\mathcal{P}_t(\Phi))/t & \text{for continuous time,} \end{cases} \tag{2.1}$$

where we set $\log 0 = 0$.

b. The ζ-function for discrete time systems. If $p(f) < \infty$, *i.e.*, if the growth rate of periodic points is at most exponential, one can incorporate all the information about the numbers of periodic points into the *zeta-function*

$$\zeta_f(z) = \exp \sum_{n=1}^{\infty} \frac{P_n(f)}{n} z^n, \tag{2.2}$$

where $z \in \mathbb{C}$ [**S-P, S-FM, S-C**]. This series converges for $|z| < \exp(-p(f))$ and always has a singularity at $\exp(-p(f))$. For some nice classes of systems this is an isolated simple pole and the only singularity on the circle $|z| = \exp(-p(f))$. In many cases the function ζ_f admits an analytic continuation, often to a meromorphic function in the whole complex plane, whose poles, zeroes, and residues provide topological invariants for f, thus encoding the countably many integer invariants given by numbers of periodic points by finitely many complex numbers. Although these are determined by the numbers of periodic points, they often provide nontrivial insights into the orbit structure.

EXAMPLE 2.5.2. For the linear expanding map $E_m\colon S^1 \to S^1$ (Section 2.1b) we have $P_n(E_m) = |m^n - 1|$ and hence

$$\zeta_{E_m}(z) = \frac{m - |m|z}{m - m|m|z}.$$

c. Index and algebraic ζ-function. One can motivate the introduction of the ζ-function and see why there is hope to organize the periodic data into a nice function by considering the *algebraic* ζ-function. It uses the notion of *index* $\operatorname{ind}_f(x)$ of an isolated fixed point of a continuous map $f\colon U \to M$ on a manifold [**KH**, Section 8.4]. The index of a fixed point may be interpreted as a multiplicity of the fixed point with a sign, *e.g.*, a point of zero index can be removed by a C^0 perturbation of f.

The sum of the indices of all fixed points can be found from the global behavior of the map via the Lefschetz Fixed Point Formula—it is given by the Lefschetz number $L(f)$, which can be calculated as follows. For $i = 1, \dots, \beta_k$ (the kth Betti number) let λ_{ki} be the eigenvalues of f_{k*} (on the kth homology), then [**S-FM**, Section 4]

$$L(f) = \sum_{k=0}^{\dim M} \sum_{i=1}^{\beta_k} \lambda_{ki}.$$

The Lefschetz Fixed Point Formula may be applied to iterates of f so long as their fixed points are still isolated. This yields the sum $P_n^A(f)$ of the indices of all periodic points of any given period in terms of a finite set of data, namely the eigenvalues λ_{ki}:

$$P_n^A(f) = \sum_{x \in \mathrm{Per}_n(f)} \mathrm{ind}_{f^k} x = \sum_{k=0}^{\dim M} \sum_{i=1}^{\beta_k} \lambda_{ki}^n.$$

If one defines the algebraic ζ-function of f by

$$\zeta_f^A(z) := \exp \sum_{n \in \mathbb{N}} \frac{P_n^A(f)}{n} z^n$$

then it is easy to check that this is always rational, indeed

$$\zeta_f^A(z) = \prod_{k=0}^{\dim M} (1 - \lambda_{ki} z)^{(-1)^{k+1}}.$$

REMARK. There are two important ways in which the dynamical ζ-function ζ_f and the algebraic ζ-function ζ_f^A differ: The indices of some periodic points may be large and the contributions of points with indices of different signs to ζ_f^A partially cancel each other. Nevertheless, there is a number of cases where ζ_f can be calculated along somewhat similar lines. This usually happens when one can guarantee that all (or all but finitely many) periodic orbits have index ± 1 and when the signs can be systematically calculated. A good example is [**S-FM**, Theorem 4.11], which follows [**Fr**] and proves rationality of the zeta-function under remarkably general conditions.

A simple example was given in the previous subsection: For $m > 1$ the indices of all periodic points of E_m are -1, whereas for $m < -1$ the indices are $(-1)^n$ for all n-periodic points. A similar calculation can be made for toral automorphisms with no roots of unity as eigenvalues (which guarantees that periodic points are isolated).

d. The ζ-function for flows. In the continuous-time case assume that periodic orbits come in families of constant period and let $l(\gamma)$ denote the smallest positive period of the orbits in such a family γ. Then we can set

$$\zeta_\Phi(z) = \prod_\gamma (1 - \exp(-z l(\gamma)))^{-1}, \tag{2.3}$$

where the product is taken over all families of nonfixed periodic points. This converges for $\Re(z) > p(\Phi)$ and has singularities on the critical line $\Re(z) = p(\Phi)$, one of which is always at $p(\Phi)$. As in the discrete-time case this is often the only singularity on that line and a simple pole and it is of particular interest. Again, often a meromorphic extension to $\mathbb{C}$ provides interesting insights. For a development of these facts see [**S-P**].

Using the power series form for the discrete-time case is a matter of convenience and the transformation $z \mapsto e^{-z}$ changes the discrete-time counterpart of (2.3) to (2.2).

REMARK. Unlike the discrete time case where the ζ-function is often quite simple, *e.g.*, rational, in the continuous time case ζ-functions do not usually belong to an easily characterized class of functions and in particular do not come from any finite-parameter families. The reason is that the periods are now real numbers and vary with perturbations.

Clearly the numbers $P_n(\Phi)$ and $\mathcal{P}_t(\Phi)$, and hence the ζ-function, are topological conjugacy invariants. The relative simplicity of ζ-functions for some discrete time systems is related to structural stability, which provides for few conjugacy classes, whereas in continuous time at best one has only orbit equivalence available, and hence still a large set of conjugacy classes. Nevertheless we have:

PROPOSITION 2.5.3. *Whether $p(\Phi)$ is zero, infinite, or neither, is an invariant of orbit equivalence.*

e. Entropy. The most important numerical invariant related to the orbit growth is topological entropy. It represents the exponential growth rate of the number of orbit segments distinguishable with arbitrarily fine but finite precision and thus describes in a crude but suggestive way the total exponential complexity of the orbit structure with a single number. We define it first for dynamical systems on compact metric spaces, though the definition is independent of the metric chosen. We then exhibit definitions that relax the assumptions on metrizability and compactness.

1. *Entropy by separated sets.* The details missing here can be found in [**KH**, Section 3.1]. Let (X, d) be a compact metric space, Φ a dynamical system. Define

$$d_t^\Phi(x, y) = \max_{0 \le \tau < t} d(\Phi^\tau(x), \Phi^\tau(y)), \tag{2.4}$$

measuring the distance between the orbit segments $\mathcal{O}^t(x) = \{\Phi^\tau(x) \mid 0 \le \tau < t\}$ and $\mathcal{O}^t(y)$. Let $N_d(\Phi, \epsilon, t)$ be the maximal number of points in X with pairwise d_t^Φ-distances at least ϵ. We call such a set of points (t, ϵ)-*separated*. Such points generate the maximal number of orbit segments of length t that are distinguishable with precision ϵ.

DEFINITION 2.5.4. We define the topological entropy by

$$h(\Phi) := h_{\text{top}}(\Phi) := \lim_{\epsilon \to 0} \overline{\lim_{t \to \infty}} \frac{1}{t} \log N_d(\Phi, \epsilon, t) = \lim_{\epsilon \to 0} \underline{\lim_{t \to \infty}} \frac{1}{t} \log N_d(\Phi, \epsilon, t).$$

It is not hard to show that these two expressions coincide and are independent of the metric. This also becomes apparent below when we give the original definition of topological entropy. Topological entropy is nonnegative and the primary distinction of levels of complexity of a dynamical system is between zero and positive entropy (Section 5.1g).

2. *Entropy by spanning sets.* Another way of measuring the exponential complexity of the orbit structure is to count the minimal number of orbit segments needed to approximate any orbit segment of a certain length to a given accuracy. This also gives topological entropy. A set $E \subset X$ is said to be (t, ϵ)-*spanning* if it is ϵ-dense for d_t^Φ. Let $S_d(\Phi, \epsilon, t)$ be the minimal cardinality of an (t, ϵ)-spanning set, or equivalently the cardinality of a *minimal* (t, ϵ)-*spanning set* or the minimal number of initial conditions whose behavior up to time t approximates the behavior of *any* initial condition up to ϵ. Then

$$h_{\text{top}}(\Phi) = \lim_{\epsilon \to 0} \overline{\lim_{t \to \infty}} \frac{1}{t} \log S_d(\Phi, \epsilon, t).$$

That one gets topological entropy both ways follows from the fact that a maximal (t, ϵ)-separated set is a (t, ϵ)-spanning set because otherwise it would be possible to increase the set by adding any point not covered, while on the other hand no ϵ-ball can contain two points 2ϵ apart, *i.e.,*

$$N_d(\Phi, \epsilon, t) \ge S_d(\Phi, \epsilon, t) \ge N_d(\Phi, 2\epsilon, t).$$

In the definition of h_{top} the $\underline{\lim}_t$ and $\overline{\lim}_t$ may disagree for positive ϵ in either of these cases. There is a third quantity, the minimal cardinality $D_d(\Phi, \epsilon, t)$ of a cover by sets whose d_t^Φ-diameter is less than ϵ, for which the limit exists by submultiplicativity: $D_d(\Phi, \epsilon, t+s) \leq D_d(\Phi, \epsilon, t) \cdot D_d(\Phi, \epsilon, s)$.

We will see soon (Section 2.5f3) that $h_{\text{top}}(\Phi^T) = |T| h_{\text{top}}(\Phi^1)$, so it suffices to develop entropy theory for discrete-time dynamical systems.

3. *Entropy as dimension.* One can interpret the preceding definition of entropy in a way that is reminiscent of the definition of the box dimension of a set in a metric space. Passing to an analog of Hausdorff dimension then leads to a more general definition of entropy [**PPi**].

Define the ϵ-size of $E \subset X$ as

$$\exp(-\sup\{t \geq 0 \mid \operatorname{diam}(\Phi^\tau(E)) < \epsilon \text{ for } 0 \leq \tau < t\}),$$

where $e^{-\infty} := 0$. Then $D_d(\Phi, \epsilon, t)$ is the minimal cardinality of a cover of X by sets of ϵ-size less than e^{-t} and one can re-express the definition of entropy through D_d by considering the asymptotics of the sum $\sum_i \delta^s$ as $\delta \to 0$ for any minimal cover of X by sets E_i of ϵ-size less than $\delta = e^{-t}$ and a parameter s. This diverges for $s > s_\epsilon$ and converges for $s < s_\epsilon$. Then $h_{\text{top}}(\Phi) = \lim_{\epsilon \to 0} s_\epsilon$. This is a calculation as it appears in the definition of box dimension. Note that by considering covers of a given subset we obtain a definition of entropy of a set. Passing to a Hausdorff-dimension analog and allowing infinite covers leads to a definition of the entropy of a not necessarily compact set and that of entropy of a dynamical system on some noncompact spaces. To that end assume X is precompact (*i.e.,* has finite covers of arbitrarily small diameter) and consider a set $Y \subset X$. Denote by $S(\epsilon, \delta)$ the infimum of $\sum_i \delta_i^s$ over countable covers of Y by open sets of ϵ-size $\delta_i < \delta$ and let $h(\Phi, Y, \epsilon) := \inf\{s \mid \lim_{\delta \to 0} S(\epsilon, \delta) = 0\}$ and

$$h_{\text{top}}(\Phi, Y) := \lim_{\epsilon \to 0} h(\Phi, Y, \epsilon).$$

We write $h_{\text{top}}(\Phi) := h_{\text{top}}(\Phi, X)$. If X is compact then this coincides with our earlier definition.

4. *Entropy via covers.* The original definition of topological entropy for discrete-time dynamical systems as given by Adler, Konheim, and McAndrew [**AdKM**] uses covers as follows. Let $\mathcal{A}$ be an open cover of a compact space X and $N(\mathcal{A})$ the minimal cardinality of a subcover. If $\mathcal{A}$ and $\mathcal{B}$ are covers, let $\mathcal{A} \vee \mathcal{B} := \{A \cap B \mid A \in \mathcal{A}, B \in \mathcal{B}\}$ and $\Phi^{-1}(\mathcal{A}) := \{\Phi^{-1}(A) \mid A \in \mathcal{A}\}$, where Φ^{-1} denotes the preimage under the map Φ^1. Then

$$h_{\text{top}}(\Phi) = \sup_{\mathcal{A}} \lim_{n \to \infty} \frac{1}{n} \log N(\mathcal{A} \vee \Phi^{-1}(\mathcal{A}) \vee \cdots \vee \Phi^{1-n}(\mathcal{A})).$$

By construction, this definition is topologically invariant, in particular independent of the metric. It is also clear from it that the entropy does not increase when one passes to a topological factor.

Considering the cover $\mathcal{A}$ by all sets of diameter less than ϵ leads to the previous definition using D_d, once one shows that taking the limit as $\epsilon \to 0$ amounts to the same as the $\sup$ over $\mathcal{A}$.

5. *Entropy for noncompact spaces.* In a similar vein one can pass to a definition, due to Bowen [**B2**], which does not require the space to be precompact by defining an analog to the ϵ-size above that requires no metric. To that end fix a finite open cover $\mathcal{A}$ of a set

$Y \subset X$ and define the $\mathcal{A}$-size of a set E as

$$\exp(-\sup\{t \geq 0 \mid \Phi^\tau(E) \prec \mathcal{A} \text{ for } 0 \leq \tau < t\}),$$

where $E \prec \mathcal{A}$ means $E \subset A$ for some $A \in \mathcal{A}$. As before this leads to a definition of the entropy $h_{\text{top}}(\Phi, Y)$ of Φ on a set Y. One should be careful to note that the definition of $h_{\text{top}}(\Phi, Y)$ is extrinsic to Y in the following sense: It may happen that two dynamical systems contain subsets Y_1 and Y_2 on which they are conjugate, but the respective subset-entropies disagree.

f. Basic properties of entropy. [**B2**], [**KH**, Proposition 3.1.6, Proposition 3.1.7, Corollary 3.2.13]

(1) $h_{\text{top}}(\Phi, \Lambda) \leq h_{\text{top}}(\Phi)$.
(2) If $\Lambda_i \subset X$ are invariant then $h_{\text{top}}(\Phi_{\restriction \bigcup \Lambda_i}) = \sup_i h_{\text{top}}(\Phi_{\restriction \Lambda_i})$.
(3) $h_{\text{top}}(\Phi^T, Y) = |T| h_{\text{top}}(\Phi, Y)$.
(4) If g is a factor of f, then $h_{\text{top}}(g) \leq h_{\text{top}}(f)$.
(5) $h_{\text{top}}(\Phi_1 \times \Phi_2, Y_1 \times Y_2) = h_{\text{top}}(\Phi_1, Y_1) + h_{\text{top}}(\Phi_2, Y_2)$.
(6) If f is expansive then $P_n(f) \leq N(f, \epsilon, n)$ and hence $p(f) \leq h_{\text{top}}(f)$. If δ is an expansivity constant then the $\lim_{\epsilon \to 0}$ in the definition of entropy is attained for any $\epsilon < \delta$.

The reason for 6. is that under an expansive map ϵ-separated sets are δ-separated after a few iterates, and that periodic orbits are always δ-separated.

g. Finiteness of entropy. Entropy can be related to the local expansion rate of a dynamical system in various ways. One of the simplest of these is based on the observation that any "direction" contributes no more to the entropy than the expansion rate in that direction, which is, of course, bounded by the maximal expansion rate, *i.e.*, the Lipschitz constant

$$\text{Lip}(f) := \sup_{x \neq y} d(f(x), f(y))/d(x, y). \tag{2.5}$$

PROPOSITION 2.5.5. [**KH**, Theorem 3.2.9] *Let f be a map of a compact metric space X with* box dimension $D(X)$. *Then $h_{top}(f) \leq D(X) \max(0, \log(\text{Lip}(f)))$.*

Thus, in particular, smooth maps of compact manifolds have finite entropy [**Ks**].

h. Growth of separated and spanning sets. In systems with zero topological entropy, particularly those with parabolic behavior (Section 8.2a), the most straightforward way to measure the complexity of the orbit structure is to look at the subexponential asymptotic growth of the quantities $N_d(\Phi, \epsilon, t)$ and $S_d(\Phi, \epsilon, t)$ with t that were used in the definition of topological entropy in Section 2.5e. Various scales of growth can be used, and we briefly describe a convenient general scheme for producing a numerical invariant. We treat continuous and discrete time in a homogeneous fashion.

A function $a\colon (0, \infty) \times (0, \infty) \to (0, \infty)$ is said to be a *scale function* if $a(\cdot, t)$ is increasing for all t and $\lim_{t \to \infty} a(s, t) = \infty$ for all s. For the case of parabolic systems the power scale $a(s, t) = t^s$ is the most suitable. Define the *upper a-entropy* as

$$\lim_{\epsilon \to 0} \sup\{s \mid \overline{\lim_{t \to \infty}} N_d(\Phi, \epsilon, t)/a(s, t) > 0\}.$$

The *lower a-entropy* is defined analogously, with $\underline{\lim}$ instead of $\overline{\lim}$. If both agree then this defines the *a-entropy* $\mathrm{ent}_a(\Phi)$ of Φ. For $a(s,t) = t^s$ the a-entropy is called the *power entropy* and denoted by $\mathrm{ent}_p(\Phi)$. Evidently the power entropy of an isometry is zero, because $N_d(\Phi, \epsilon, t)$ is bounded for given ϵ.

i. Slow entropy and the Hamming metric. A more robust approach, which works for both the topological and measure-theoretic situation (Section 3.7l), is based on replacing the supremum metric d_t^Φ from (2.4) by an integral metric

$$\eth_t^\Phi(x,y) = \frac{1}{t}\int d(\Phi^s(x), \Phi^s(y))\, ds, \tag{2.6}$$

where, as usual, the integral stands for summation in the discrete time case. The construction then proceeds exactly as above. The results of this modified definition of topological a-entropy with the metric $\eth_t^\Phi$ are denoted by $\overline{\mathrm{ent}}_a^{\mathrm{top}}(\Phi)$, $\underline{\mathrm{ent}}_a^{\mathrm{top}}(\Phi)$ and $\mathrm{ent}_a^{\mathrm{top}}(\Phi)$, according to whether we use upper or lower limits, or these coincide.

j. Weighted zeta-functions. Fix a bounded function $\varphi\colon X \to \mathbb{C}$ on the phase space X of a dynamical system. In order to use this function to assign weight to periodic orbits as we count them, we assume that periodic orbits are isolated. The weight of a periodic orbit is then given by the integral of φ over the periodic orbit. In the discrete-time case this is just a sum of the values of the function along the orbit, but for a monolithic treatment we use integrals throughout.

If we replace the exponent $-zl(\gamma)$ in (2.3) by $\int_0^{l(\gamma)}(\varphi(\Phi^t(x)) - z)\, dt$ we obtain the weighted zeta-function

$$\zeta_{\Phi,\varphi}(z) = \prod_\gamma \Big(1 - \exp\big(\int_0^{l(\gamma)} (\varphi(\Phi^t(x)) - z)\, dt\big)\Big)^{-1}, \tag{2.7}$$

The analog to our earlier discrete-time version is

$$\zeta_{\Phi,\varphi}(z) = \exp \sum_{n=1}^\infty \frac{z^n}{n} \sum_{x\in \mathrm{Fix}(\Phi^n)} \exp \sum_{k=0}^{n-1} \varphi(\Phi^k(x)),$$

which explains more clearly why this is appropriately described as taking weighted sums. Note that for $\varphi = 0$ we recover the original zeta-function. See [**S-P, PP**].

k. Pressure. Similarly, such a function φ can be added to the data used in the definition of entropy by counting orbits with weights. This leads to the definition of pressure. Specifically we assign the weight $\exp\int_0^T \varphi(\Phi^t(x))\, dt$ to an orbit segment $\mathcal{O}^t(x)$.

DEFINITION 2.5.6. The topological pressure of φ is

$$P_\varphi(f) := \lim_{\epsilon\to 0} \overline{\lim_{t\to\infty}} \frac{1}{t} \log N_d(\Phi, \varphi, \epsilon, t),$$

where

$$N_d(\Phi, \varphi, \epsilon, t) := \sup\Big\{ \sum_{x\in E} \exp\big(\int_0^t \varphi(\Phi^t(x))\, dt\big) \mid E \subset X \text{ is } (t,\epsilon)\text{-separated}\Big\}$$

and in the discrete time case integrals are replaced by sums.

Note that we can similarly modify the alternative definitions of entropy and that we recover entropy as the special case $\varphi = 0$.

The definition of pressure is often used with potential functions φ that are naturally related to the dynamics in some way. The principal importance of pressure appears in connection with the study of a special class of invariant measures for topological (in particular, smooth and symbolic) dynamical systems, *equilibrium states*, see Section 4.4g, Section 6.7c, [**S-C**], [**KH**, Chapter 20].

l. Higher rank abelian actions. The notions of entropy and pressure can be allow a strightforward extension to the case of $\mathbb{Z}^k$ and $\mathbb{R}^k$ actions. This is important for applications that involve the thermodynamical formalism on lattice models as well as the study of actions by automorphisms of compact abelian groups [**S-LS**]. The basic point is that we can define metrics

$$d_t^{\Phi}(x,y) = \max_{0\le\tau_i<t} d(\Phi^{(\tau_1,\dots,\tau_k)}(x), \Phi^{(\tau_1,\dots,\tau_k)}(y)),$$

also for these actions. The notions of separated and spanning sets become immediately natural. Since the cubes in $\mathbb{Z}^k$ and $\mathbb{R}^k$ used here tile the respective group, the arguments from the cyclic case go through to prove the existence of the expressions defining entropy and pressure [**Ru3, Mi**].

m. Complexity of families of orbits. In addition to considering the growth of discrete families of orbits one can measure the growth of continuous families of orbits. To that end one may consider their topological complexity. This idea leads to several algebraic counterparts of entropy. The first invariant of this kind is related to the growth of homotopical complexity of iterates for a closed loop.

1. *Fundamental group entropy.* To define the entropy of an endomorphism $F\colon \pi \to \pi$ of a finitely generated group π let $\Gamma = \{\gamma_1, \dots, \gamma_s\}$ be a system of generators and for $\gamma \in \pi$ set

$$L(\gamma,\Gamma) = \min\{\sum_{j=1}^{ks} |i_j| \mid \gamma = \gamma_1^{i_1}\gamma_2^{i_2}\cdots\gamma_s^{i_s}\gamma_1^{i_{s+1}}\cdots\gamma_s^{i_{2s}}\cdots\gamma_s^{i_{ks}}\},$$

$L_n(F,\Gamma) = \max_{1\le i\le s} L(F^n\gamma_i,\Gamma)$ and $h_A(F) := \lim_{n\to\infty} \log L_n(F,\Gamma)/n$. This is independent of Γ and is called the *algebraic entropy* of F. Clearly it is invariant under conjugacy of group endomorphisms.

Now consider a continuous map f of a compact connected manifold M and let $p \in M$. Fix a continuous path α connecting p with its image $f(p)$, *i.e.*, a map $\alpha\colon [0,1] \to M$ such that $\alpha(0) = p$, $\alpha(1) = f(p)$. Then define an endomorphism $f_*^{\alpha}\colon \pi_1(M,p) \to \pi_1(M,p)$, $[\gamma] \mapsto [\alpha f(\gamma)\alpha^{-1}]$, which is represented by the path α followed by the loop $f\circ\gamma$ and then by α taken in the opposite direction. Define the *fundamental-group entropy* of f as $h_*(f) := h_A(f_*^{\alpha})$. This is independent of the choice of α and p and clearly a topological invariant [**KH**, Section 3.1]. It turns out [**S-FM**], [**KH**, Section 8.1] that

$$h_*(f) \le h_{\text{top}}(f).$$

2. *Homological entropy.* Other useful topological growth invariants come from considering the linear maps f_{*i} induced by f on the homology groups $H_i(M, \mathbb{R})$. The spectral radii $r(f_{*i})$ are topological invariants of f. It follows immwdiately from the Hurewicz identification $H_1(M) \sim \pi_1(M,p)/[\pi_1(M,p), \pi_1(M,p)]$ that

$$\log r(f_{*1}) \le h_*(f).$$

See[**S-FM**] for other results in this direction.

3. *Homotopical entropy.* For continuous-time dynamical systems the invariants defined above are vacuous since every element of the flow is homotopic to the identity and hence induces trivial maps of the fundamental group and homology groups. There are, however, different ways to measure the growth of topological complexity. For example, on a compact connected manifold X one can fix a point $p \in X$ and a family of arcs $\Gamma = \{\gamma_x \mid x \in X\}$ of bounded length connecting p with various points of X. Then for a flow $\Phi = \varphi^t \colon X \to X$ one fixes T and considers for each $x \in X$ the closed loop $l(x,T)$ consisting of the arc γ_x, the orbit segment $\{\varphi^t x\}_{t=1}^T$, and the reverse of the arc $\gamma_{f_T x}$. Those loops represent different elements of the fundamental group $\pi_1(X,p)$. Their number $\Pi(\Phi, p, \Gamma, T)$ grows at most exponentially and the exponential growth rate $h_{\text{hom}}(\Phi) := \overline{\lim}_{T\to\infty} \log \Pi(\Phi, p, \Gamma, T)/T$ is independent of p and Γ and is called the *homotopical entropy* of Φ. It is obviously invariant under flow equivalence and similarly to before we have

$$h_{\text{hom}}(\Phi) \le h_{\text{top}}(\Phi).$$

In Section 2.3f, similar ideas were used from the point of view of recurrence.

6. Symbolic dynamical systems

We now look more carefully at the structure of the n-shift introduced in Section 2.1d. See [**LM**], [**KH**, Section 1.9] for more detailed accounts.

a. Metrics and functions of exponential type. For $N \ge 2$ consider the Cantor set $\Omega_N = \{0, 1, \dots, N-1\}^{\mathbb{Z}}$ of two-sided sequences of N symbols and the one-sided space $\Omega_N^R = \{0, 1, \dots, N-1\}^{\mathbb{N}_0}$ with the product topology. Since the set of "states" $\{0, 1, \dots, N-1\}$ can be identified with the cyclic group $\mathbb{Z}/N\mathbb{Z}$, the spaces Ω_N^R also possess the structure of a compact abelian topological group. For $n_1 < n_2 < \cdots < n_k$ and $\alpha_1, \dots, \alpha_k \in \{0, 1, \dots, N-1\}$ we call

$$C_{\alpha_1,\dots,\alpha_k}^{n_1,\dots,n_k} = \{\omega \in \Omega_N \mid \omega_{n_i} = \alpha_i \text{ for } i = 1, \dots, k\} \tag{2.1}$$

a *cylinder* and k the *rank* of that cylinder. Cylinders in Ω_N^R are defined similarly. Cylinders form a base for the product topology. Every cylinder is also closed because the complement of a cylinder is a finite union of cylinders. The most general open set is a countable union of cylinders. The topology is given by any metric

$$d_\lambda(\omega, \omega') = \lambda^{\max\{n \in \mathbb{N}_0 \,\mid\, \omega_k = \omega'_k \text{ for } |k| \le n\}}$$

with $\lambda \in (0,1)$. Then any symmetric cylinder $C_{\alpha_{-n},\dots,\alpha_n}^{-n,\dots,n}$ of rank $2n+1$ is a λ^n-ball.

The different metrics d_λ define the same topology on Ω_N (although they are not equivalent as metrics) and also determine a Hölder structure. This means that the notion of Hölder-continuous function with respect to the metric d_λ does not depend on λ. The class of Hölder-continuous functions plays an important role in applications to differentiable dynamics and can be described as follows. Let φ be a continuous complex-valued

function defined on Ω_N or on a closed subset and write $\omega = (\dots, \omega_{-1}, \omega_0.\omega_1, \dots)$ and $\omega' = (\dots, \omega'_{-1}, \omega'_0.\omega'_1, \dots)$. Then for $n \in \mathbb{N}$ let

$$V_n(\varphi) := \max\{|\varphi(\omega) - \varphi(\omega')| \mid \omega_k = \omega'_k \text{ for } |k| \le n\}.$$

Since Ω_N is compact, φ is uniformly continuous and $V_n(\varphi) \to 0$ as $n \to \infty$. We say that φ has *exponential type* if $V_n(\varphi) \le ce^{-an}$ for some $a, c > 0$.

PROPOSITION 2.6.1. *φ has exponential type if and only if it is Hölder continuous with respect to some (and hence any) metric d_λ.*

All this translates to Ω_N^R and has obvious analogs for $\mathbb{Z}^k$ and $\mathbb{Z}_+^k$ as index sets. This more general setting is motivated by lattice models in statistical mechanics.

b. Shifts. Topological entropy and periodic orbit growth coincide for shifts. It is easy to calculate

$$h_{\text{top}}(\sigma_N) = p(\sigma_N) = \log N.$$

Note that these maps are expansive and

$$\zeta_{\sigma_N}(z) = \zeta_{\sigma_N^R}(z) = \exp \sum_{n=1}^{\infty} \frac{N^n}{n} z^n = \frac{1}{(1 - Nz)}.$$

Orbit closures are easy to characterize: If $\omega \in \Omega_N$ then

$$\overline{\mathcal{O}(\omega)} = \{\omega' \in \Omega_N \mid \forall m \in \mathbb{N}\, \exists k \in \mathbb{Z} \colon \omega'_i = \omega_{k+i} \text{ for } |i| \le m\}.$$

However, they may be rather complicated.

EXAMPLE 2.6.2. The one-sided shift on two symbols arises naturally from coding in simple examples: It is topologically conjugate to the restriction of the tripling map $E_3 \colon x \mapsto 3x \pmod 1$ to the ternary Cantor set (in $[0,1]$ embedded into $S^1 = \mathbb{R}/\mathbb{Z}$) as well as to the restriction of $f_a \colon \mathbb{R} \to \mathbb{R}$, $x \mapsto ax(1-x)$ for $a > 4$ to the invariant set $\Lambda := \bigcap_{n \in \mathbb{N}} f^{-n}([0,1])$.

These are simple instances of the fact that shifts are standard models for some closed invariant sets in smooth dynamical systems. This is one of the central themes in hyperbolic dynamics, see Section 6.7g, [**S-C**, Chapter 8], [**KH**, Section 18.7].

Recall that the restriction of the shifts σ_N or σ_N^R to any closed invariant subset Λ of Ω_N or Ω_N^R, respectively, is called a *symbolic dynamical system*. Properties of symbolic dynamical systems vary widely. They are a rich source of examples and counterexamples for topological dynamics and ergodic theory.

Any symbolic dynamical system can be characterized by the existence of a collection S of "forbidden" blocks, *i.e.*, of finite sequences $\alpha = (\alpha_0, \dots, \alpha_{n_\alpha - 1})$, such that

$$\Lambda = \{\omega \in \Omega^N \mid (\omega_k, \dots, \omega_{k+n_\alpha}) \ne \alpha \text{ for all } k \in \mathbb{Z}, \alpha \in S\}.$$

c. Topological Markov chains and subshifts of finite type. It is natural to try to look at those symbolic systems for which the collection S of forbidden blocks is simple, in particular those with finite S. We begin with the situation where S contains only blocks of length two.

Let $A = (a_{ij})_{i,j=0}^{N-1}$ be a 0-1 matrix, *i.e.*, with entries $a_{ij} \in \{0, 1\}$ and

$$\Omega_A := \{\omega \in \Omega_N \mid a_{\omega_n \omega_{n+1}} = 1 \text{ for } n \in \mathbb{Z}\}. \tag{2.2}$$

In other words, the matrix A determines all admissible transitions between the symbols $0, 1, \ldots, N-1$. The set Ω_A is obviously shift invariant.

The restriction $\sigma_N{\restriction}_{\Omega_A} =: \sigma_A$ is called the *topological Markov chain* determined by the matrix A. Let $\mathcal{A}\colon \{1, \ldots, N\}^{n+1} \to \{0, 1\}$ and $\Omega_{\mathcal{A}} := \{\omega \in \Omega_N \mid \mathcal{A}(\omega_m, \ldots, \omega_{m+n}) = 1 \text{ for } m \in \mathbb{Z}\}$. Then the restriction $\sigma_{\mathcal{A}}$ of σ_N to $\Omega_{\mathcal{A}}$ is called an *n-step topological Markov chain* or a *subshift of finite type*. The latter terminology derives from the fact that these shifts can be described by giving a finite list of forbidden words (of length up to $n+1$), *i.e.*, a subshift of finite type can be described as the set of sequences containing none of a finite list of excluded words. Some authors, however, intend "subshift of finite type" to be synonymous with "topological Markov chain".

Topological Markov chains constitute a special (although important) class of symbolic dynamical systems. From the point of view of their intrinsic dynamics n-step topological Markov chains are the same as topological Markov chains, since they can be described as topological Markov chains over the alphabet $\{1, \ldots, N\}^n$ by taking the matrix A given by $A_{(i_1,\ldots,i_n),(j_1,\ldots,j_n)} = 1$ if $j_k = i_{k+1}$ for $k = 1, \ldots, n-1$ and $\mathcal{A}(i_1, \ldots, i_n, j_n) = 1$.

Subshifts of finite type and hence topological Markov chains are of interest *e.g.*, because of the following.

PROPOSITION 2.6.3. *A closed shift-invariant set $\Lambda \subset \Omega_N$ is locally maximal (isolated) if and only if $\sigma_N{\restriction}_\Lambda$ is a subshift of finite type.*

This expresses the fact that checking blocks of finite length corresponds to fixing a point up to a finite error.

d. Properties of topological Markov chains. There is a useful geometric representation for topological Markov chains. Connect i with j by an arrow if $a_{ij} = 1$. This way we obtain a directed graph G_A with N vertices. We call a finite or infinite sequence of vertices of G_A an *admissible path* or *admissible sequence* if any two consecutive vertices in the sequence are connected by an oriented arrow. A point of Ω_A corresponds to a doubly infinite path in G_A with marked origin, and the topological Markov chain σ_A corresponds to moving the origin to the next vertex. Here is a simple example:

$$\begin{array}{ccc} 0 & \longrightarrow & 1 \\ \uparrow\downarrow & \searrow & \downarrow \\ 3 & \longrightarrow & 2 \end{array} \quad \text{corresponds to} \quad \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}.$$

This topological Markov chain consists of a single period-2 orbit $\overline{03}$.

PROPOSITION 2.6.4. *$h_{top}(\sigma_A) = \log r(A)$, the spectral radius of A, and $P_n(\sigma_A) = \operatorname{tr}(A^n)$, in particular the zeta-function is rational.*

REMARK. This resembles the algebraic ζ-function discussed earlier. One could say that the appearance of only one matrix in the formula (compared to one per homology dimension) reflects the fact that the sequence space is zero-dimensional.

Assume from now on that A is a 0-1 $N \times N$ matrix which has at least one 1 in each row and each column. If $i \in \{0, \ldots, N-1\}$ then $\Omega_{A,i} := \{\omega \in \Omega_A \mid \omega_0 = i\} \neq \varnothing$. If there is an element $\omega \in \Omega_A$ that contains the symbol i at least twice then we call i *essential*. Otherwise i is said to be *transient*. This is equivalent to the existence of a periodic point

$\omega' \in \Omega_A$ such that $\omega'_0 = i$. Any ω-limit point (see Section 2.3a) of any element of Ω_A contains only essential symbols. We call two essential symbols i and j *equivalent* if there exist $\omega, \omega' \in \Omega_A$, $k_1 < k_2$, $l_1 < l_2$ such that $\omega_{k_1} = \omega'_{l_2} = i, \quad \omega_{k_2} = \omega'_{l_1} = j$, *i.e.*, they occur in the same cycle. Thus the set of all essential symbols splits into disjoint equivalence classes. Now σ_A has a dense positive semiorbit if and only if all symbols are essential and equivalent. This gives

PROPOSITION 2.6.5. *If σ_A has a dense positive semiorbit then there exists $m \in \mathbb{N}$ and a decomposition of Ω_A into closed disjoint subsets $\Lambda_1, \dots, \Lambda_m = \Lambda_0$ such that $\sigma_A \Lambda_i = \Lambda_{i+1}$ for $i = 0, 1, \dots, m-1$ and the restriction of $(\sigma_A)^m$ to each Λ_i is topologically mixing. Moreover, the decomposition of Ω_A into Λ_i's corresponds to a decomposition of the set $\{1, \dots, N\}$ into m equal groups such that every element $\omega \in \Omega_A$ has only symbols from one group in positions equal modulo m. The nonwandering set is a disjoint union corresponding to different classes of essential symbols.*

This is an instance of the so-called *spectral decomposition*, where the phase space decomposes into transitive sets, each of which is a union of cyclically permuted sets on which the appropriate iterate is topologically mixing.

SKETCH OF PROOF. Let m be the greatest common divisor of lengths of cycles (sequences beginning and ending at the same symbol) and identify two symbols if they are connected by a path whose length is a multiple of m. Let Λ_i be the equivalence classes. For mixing assume without loss of generality that $m = 1$. □

The preceding proposition shows that the primary case of interest is the mixing one.

PROPOSITION 2.6.6. *A is a* transitive matrix, i.e., *there is a power of A all of whose entries are positive, if and only if σ_A is topologically mixing. In this case A has a single eigenvalue $\lambda_{\max} = r(A)$ of maximal absolute value and $|P_n(\sigma_A) - \lambda^n_{\max}| \le C\lambda^n$ for some $C > 0$ and $\lambda < \lambda_{\max}$.*

REMARK. As we have just seen, topological Markov chains and subshifts of finite type represent a class of symbolic systems that allow a comprehensive structural description. However, the classification of those systems up to topological conjugacy is a difficult algebraic and combinatorial problem and engendered considerable activity in the last 25 years, which produced highly nontrivial invariants and finally led to counterexamples to the leading classification conjecture (see [**LM**, Chapter 7]).

e. Some subshifts of infinite type. While subshifts of finite type are the simplest symbolic systems in terms of enumeration of forbidden blocks, there are numerous other symbolic systems which can be algorithmically described and some of them produce rather simple dynamics. We refer the reader to [**LM**] for a survey of such constructions and present here only some characteristic examples.

EXAMPLE 2.6.7. The Prouhet–Thue–Morse sequence [**LM**] is a uniformly recurrent point for σ_2^R, which produces minimal non-Markov systems. It is constructed as follows. Beginning with the word $W_0 = 0$ recursively define words $W_{n+1} = W_n \overline{W}_n$, where $\overline{W}_n$ is obtained from W_n by replacing 0 by 1 and vice versa. For $n \ge k$ the word W_n begins with the word W_k, so the limit is a well-defined one-sided sequence W. Indeed,

$$W = 0\,1\,10\,1001\,10010110\,1001011001101001\dots$$

Equivalently, one can start from the word 0 and repeatedly apply the substitution scheme $0 \to 01, 1 \to 10$. Again, the limit is W. Indeed, this is an example of an interesting class of non-Markov symbolic systems called substitution shifts [**LM**].

As W is made up of pairs 01 and 10, any two 0's are separated by at most two entries. Therefore any two words 01 (obtained from 0 by the substitution scheme) are separated by at most four entries, any two words 0110 by at most eight, and so on. These initial words include all allowed words, so W is uniformly recurrent for σ_2^R.

To obtain a minimal non-Markov shift extend W to a two-sided sequence $x = \alpha W \in \Omega_2$ and let $\Lambda = \omega(x)$ be the ω-limit set of x under σ_2. Then $\sigma_{2\restriction_\Lambda}$ is minimal by Proposition 2.3.4. Λ can equivalently be described as the collection of sequences all of whose subwords appear in W.

EXAMPLE 2.6.8. Toeplitz shifts are defined as orbit closures of Toeplitz sequences. A sequence $x \in \Omega_N$ is said to be a Toeplitz sequence if $\mathbb{Z}$ decomposes into arithmetic progressions on which x is constant. Toeplitz shifts are always minimal [**LM**, p. 460], but may have positive entropy. (There are also smooth minimal dynamical systems with positive entropy [**Hm4**].)

EXAMPLE 2.6.9. For $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ consider the coding of the circle rotation R_α by $[0, 1/2)$, $[1/2, 1)$. Although the words 01 and 10 are allowed, this symbolic system S is not the full shift because there are no periodic sequences. It is not hard to check minimality and that S is not a subshift of finite type.

Since $h_{\text{top}}(S) = 0$, we can try to study a subexponential rate of orbit growth in this case. For symbolic dynamical systems one can define the power entropy (Section 2.5h) via the number of nonempty cylinders of rank n instead of $N_d(S, \epsilon, n)$. In the present example there are $2n$ nonempty cylinders of length n, so $\operatorname{ent}_p(S) = 1$.

On the other hand, rotations are isometries, so $\operatorname{ent}_p(R_\alpha) = 0$.

EXAMPLE 2.6.10. Let

$$B_k = \{\omega \in \Omega_2 \mid m, n \in \mathbb{Z}, m > n \Rightarrow \Big|\sum_{i=n}^{m} (-1)^{\omega_i}\Big| \le k\}.$$

It is easy to see that B_k is a closed σ_2-invariant subset of Ω_2. Denote $S_k = \sigma_{2\restriction_{B_k}}$. Partial sums of ω give (in a nonunique way) sequences that vary within $[0, k]$ up to translation. Let A_k be the $(k+1) \times (k+1)$ 0-1 matrix with $a_{ij} = 1$ if $|i - j| = 1$ and $a_{ij} = 0$ otherwise. The corresponding topological Markov chain Ω_{A_k} represents the *discrete random walk* on the interval $[0, k]$ because for each element $\omega \in \Omega_{A_k}$ successive entries differ by exactly 1. Notice that Ω_{A_k} is topologically transitive but not topologically mixing, because it interchanges even and odd symbols. Coding decreasing transitions by zeroes and increasing transitions by ones defines a map $H_k \colon \Omega_{A_k} \to \Omega_2$. The image of H_k is B_k and

$$H_k \circ \sigma_{A_k} = S_k \circ H_k. \tag{2.3}$$

The map H_k is in fact an almost isomorphism (Section 2.2f). The set of nonuniqueness consists of the walks that do not cover the whole interval $[0, k]$. For example for $k = 2$ this set contains exactly two period two orbits $\overline{01}$ and $\overline{12}$, which are both mapped to $\overline{01}$. Similarly, for $k > 2$ there are two copies of $\Omega_{A_{k-1}}$ inside Ω_{A_k} that differ by translation and are identified by the map H_k. Thus S_k^2 is not transitive but B_k is the union of two transitive invariant sets that are the closures of their interiors and whose intersection is

an invariant nowhere dense set. And this is incompatible with spectral decomposition (Proposition 2.6.5).

Thus, we have shown that S_k is not a subshift of finite type although it is an almost isomorphic factor of one. Symbolic systems that are factors of subshifts of finite type are called *sofic*.

f. Complexity of symbolic systems. A natural notion closely related to slow topological entropy (Section 2.5h) is the *complexity function* $p_\Lambda(n)$ for a symbolic dynamical system, which gives the number of standard cylinders of rank n which intersect the set Λ. This notion has been extensively studied both with an eye to describing possible complexity functions and characterizing systems with low complexity. The area is surveyed in [**Fe2**]. An interesting open question is the characterization of symbolic systems with complexity function growing at most linearly.

g. The Furstenberg Reduction Principle and multiple recurrence. Any $\omega \in \Omega_N$ defines a partition of $\mathbb{Z}$ into subsets $S_i = \{m \in \mathbb{Z} \mid \omega_n = i\}$ $(0 < i < N)$. Accordingly, $\sigma^m\omega \in C_i^0$ if and only if $m \in S_i$. This simple observation lies at the root of an important class of applications of topological dynamics to combinatorics (and vice versa). Rather loosely we can say that any statement asserting that one of the sets from any finite partition of $\mathbb{Z}$ possesses a certain large scale structure is equivalent to a certain statement about recurrence behavior in topological dynamical systems. Furstenberg made systematic use of this observation (as well as its quantitative counterpart, Section 4.2f [**Pt**]), so we refer to both as the Furstenberg Reduction Principle [**F1, S-B, F4**].

Let us illustrate this by an important example. As the statement about partitions of $\mathbb{Z}$ take the van der Waerden Theorem [**PY**]: If $\mathbb{Z}$ is partitioned (or covered) by finitely many subsets then at least one of these contains arbitrarily long arithmetic progressions.

Let $f\colon X \to X$ be a continuous map of a compact metric space. A point $x \in X$ is said to be *multiply nonwandering* if for any open neighborhood U and k there exists l such that $\bigcap_{j=0}^{k-1} f^{-jl}(U) \neq \varnothing$.

THEOREM 2.6.11. *The van der Waerden Theorem is equivalent to the existence of a multiply nonwandering point for every map.*

PROOF. To any partition $\mathbb{Z} = S_0 \cup \cdots \cup S_{N-1}$ corresponds an $\omega \in \Omega_N$ by setting $\omega_n = i$ if $n \in S_i$. Then S_i contains an arithmetic progression $(m, m+l, \ldots, m+(k-1)l)$ if and only if $\sigma^{jl}(\sigma^m(\omega)) \in C_i^0$ for $0 < j < k$.

Assuming that the restriction of the shift σ to the invariant compact set $\overline{\mathcal{O}(\omega)}$ has a multiply nonwandering point ω^0 we obtain

$$\bigcap_{j=0}^{k-1} \sigma^{-jl}(C_i^0 \cap \overline{\mathcal{O}(\omega)}) \neq \varnothing.$$

But since cylinders are closed this implies that also

$$\bigcap_{j=0}^{k-1} \sigma^{-jl}(C_i^0 \cap \mathcal{O}(\omega)) \neq \varnothing.$$

and hence $\sigma^m\omega \in \bigcap_{j=0}^{k-1} \sigma^{-jl}(C_i^0)$ for some $m \in \mathbb{Z}$, implying the statement of the van der Waerden Theorem: the set $S_{\omega_0^0}$ contains an arbitrary long arithmetic progression.

Conversely, assume the van der Waerden Theorem holds and consider an arbitrary continuous map f of a compact metric space. Without loss of generality we may assume that f is invertible by considering the natural extension of f because multiply recurrent points of the natural extension project to multiply recurrent points of the map. Now consider a nested sequence of partitions $(\mathcal{P}_n)_{n\in\mathbb{N}}$ into sets of diameter less than 2^{-n}. Using the coding associated with the partition $\mathcal{P}_n$ and applying the van der Waerden Theorem one finds for each n an element $C_n \in \mathcal{P}_n$ such that for any $k \in \mathbb{N}$ there exists an $l \in \mathbb{N}$ with $\bigcap_{j=0}^{k-1} f^{-jl}(C_n) \neq \varnothing$. By compactness one can find a sequence of points $x_{n_m} \in C_{n_m}$ converging to x. For any open neighborhood U of the point x one has $C_{n_m} \subset U$ for sufficiently large m. This proves that x is multiply nonwandering. □

h. Topological Markov chains and subshifts of finite type for other groups. Symbolic systems are naturally defined over groups other than $\mathbb{Z}$, and subshifts of finite type are defined by "localized" interactions, *i.e.,* by prohibiting the appearance of some finite patterns. The case $\mathbb{Z}^k$ is important and relates to statistical mechanics (lattice models) [**Ru2**] and automorphisms of abelian groups [**S**].

This subject has many more intricacies than the cyclic situation and is surveyed in [**S-LS**].

7. Low-dimensional topological dynamical systems

In most of this chapter topological dynamical systems were considered in a generic way without any particular regard to the topology of the phase space save for compactness or an even weaker assumption. When one begins to look into topological dynamics in a more thorough fashion specific properties of the phase space come to the fore. Symbolic systems represent the most important class of dynamical systems acting on the totally disconnected and hence zero-dimensional phase space. Since topological dynamics grew out of attempts to distill some natural general properties of more classical systems, dynamical systems acting on manifolds and similar spaces are of particular interest. Algebraic topology plays an essential role in this area of dynamics, see [**S-FM**].

An area where the impact of topology is particularly well understood is dynamics on *low-dimensional connected* phase spaces. The primary cases are discrete-time (invertible and, more interestingly, noninvertible) systems in dimension one, especially on the interval, the circle, and the line, homeomorphisms of two-dimensional manifolds, flows on two-dimensional manifolds and, to a limited extent, flows on three-dimensional manifolds.

a. One-dimensional dynamics. For a detailed overview of topological dynamics in one dimension see [**S-JS**, Chapter 1]. The comprehensive monograph [**MS**] on one-dimensional dynamics pays considerable attention to the topological aspects of the subject. [**ALM**] covers one-dimensional topological dynamics in even greater depth.

The key topological property of connected one-dimensional spaces is the fact that a small connected neighborhood of a point becomes disconnected when the point itself is removed. A closely connected fact is the *intermediate value theorem* for functions of one real variable. This theorem allows to use simple combinatorial data to construct invariant sets for an interval map closely associated with topological Markov chains determined by these data. Here is a simple illustrative example.

Let I be an interval and $f\colon I \to I$ be continuous map. We say that $J \subset I$ *covers* $K \subset I$ (under f) if $K \subset f(J)$ and we denote this situation by $J \to K$.

Consider a collection $\mathcal{C} = \{I_1, \dots, I_n\}$ of closed subintervals of I with pairwise disjoint interiors. The relation "$\to$" then yields the edges of a directed graph, the *Markov graph*, associated to $\mathcal{C}$, whose vertices are the intervals in $\mathcal{C}$. Let A be the 0-1 matrix determined by the graph and σ_A^R the one-sided topological Markov shift defined by A. We will say that A is *associated* to the collection $\mathcal{C}$.

THEOREM 2.7.1 ([**KH**, Theorem 15.1.5]). *Let $\mathcal{C} = \{I_1, \dots, I_n\}$ be a collection of pairwise disjoint closed subintervals of I, $J := \bigcup I_i$, and A the matrix associated to $\mathcal{C}$. Then there exists a closed f-invariant subset $S \subset J$ such that σ_A^R is a factor of $f_{\restriction S}$ via a map $h\colon S \to \Omega_A^R$. There are at most countably many points with more than one preimage under h and the preimages of these points are intervals.*

Markov covers of this type are the basic tool for studying the topological dynamics of noninvertible maps in one dimension (for invertible ones the rotation number from Section 2.3g plays a similar role). In particular, they play a central role in the proof of the following definitive result by Misiurewicz and Szlenk which connects two fundamental growth invariants, the growth of periodic orbits and the topological entropy.

THEOREM 2.7.2 ([**KH**, Corollary 15.2.2], [**ALM**, Theorem 4.3.14]). *$h_{top}(f) \le p(f)$ for any continuous map f of the interval or the circle.*

b. Flows and homeomorphisms on surfaces. The Jordan curve theorem plays a crucial role in two-dimensional topological dynamics. In particular, it is responsible for the simplicity of flows on the plane and sphere (Poincaré–Bendixon theory [**KH**, Section 14.1]) as well as for the limited complexity of flows on other surfaces (Section 8.4, [**KH**, Chapter 14]). In these situations the principal tool is the construction of closed transversals and return maps to these. The Jordan curve theorem is used to show that certain arrangements of transversals must cut the surface into pieces, thus producing limitations on recurrent behavior. The main invariants for flows on surfaces of genus greater than one are of a homological nature (Section 8.4d) and can be viewed as generalizations of asymptotic cycles and rotation number for flows on the torus (Section 2.3f, Section 2.3g). An alternative (in fact, historically earlier) approach was suggested by Aranson and Grines in the form of the homotopy rotation class (see [**NZ**] for a comprehensive modern account). In that approach the Jordan curve theorem is used to associate with a flow on a compact surface M of genus $g \ge 2$ a subset of the circle that is identified with the boundary of the hyperbolic plane, the universal cover of M. For a broad class of flows the homotopy rotation class is a complete invariant of topological orbit equivalence, albeit not an easily computable one.

While the orbit structure of surface homeomorphisms may be much more complicated than that of flows it is remarkable that the homotopy type of such a homeomorphism, absolute or modulo an invariant set consisting of several periodic orbits, provides substantial infornation about the dynamics (see [**S-FM**]). This connection is the basis of *Nielsen (or Nielsen–Thurston) Theory*. (See [**KH**, Section 8.7] for a basic introduction, [**Ji**] for a comprehensive account and also [**FLP**].)

CHAPTER 3

Ergodic theory

1. Introduction

a. Invariant measures and asymptotic distribution. In the most general terms, the subject of ergodic theory is the study of groups and semigroups of nonsingular transformations of measure spaces. Most attention is given to the case where the measure in question is both finite and invariant. This assumption is somewhat akin to the compactness assumption in topological dynamics. Its implications are more powerful and include a quantitative analog of regional recurrence (Section 2.3d).

To be more specific for the purpose of providing motivation, we consider only the case of cyclic systems with discrete time. For a Borel measure μ on a topological space X define the *support* of μ as the set of all points for which any neighborhood has positive measure. As the complement of the maximal open null set, it is a closed set. The Poincaré Recurrence Theorem (Section 3.4c), which is often hailed as the first true result of ergodic theory, implies that the restriction of a continuous map to the support of a finite invariant measure is regionally recurrent.

However the main reason ergodic theory has such a powerful presence in other major areas of dynamics, lies in *ergodic theorems*, which assert that for almost every initial condition the distribution of iterates among various parts of the phase space satisfies a certain asymptotic law. This explains why interest in invariant measures arises naturally in the study of smooth or topological dynamical systems (which are the primary applications).

Let X be a set, $f\colon X \to X$, and for $x \in X$ and $U \subset X$ let $F_U(f,x,n) = \sum_{k=0}^{n-1} \chi_U(f^k(x)) = \operatorname{card}\{k \in [0,n-1] \mid f^k(x) \in U\}$, *i.e.,* the number of visits to the set U under the first n iterates of x. The limit

$$F_U(f,x) := \lim_{n\to\infty} \frac{F_U(f,x,n)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \chi_U(f^k(x)),$$

if it exists, gives the *asymptotic density of the distribution* of the iterates between the set U and its complement $X \smallsetminus U$. It is called the *time average* or *Birkhoff average* of χ_U. As a corollary of the Birkhoff Ergodic Theorem (Theorem 3.5.2), we conclude that for any measurable set U the time average exists for almost every initial condition $x \in X$. Furthermore, it is positive for almost every $x \in U$.

As mentioned in Section 1.4b, one can consider time averages of functions other than characteristic functions. Bounded functions, or integrable ones, are the most natural candidates. Thus, for a given $x \in X$ and a bounded measurable function φ the time average

is defined as

$$I_x(\varphi) := \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi(f^k(x)),$$

and again the Birkhoff Ergodic Theorem implies existence of the limit for almost every $x \in X$.

See [**S-B**] for a survey and [**Kg**],[**Tp**] for extensive treatments of the subject.

b. Quantitative recurrence properties. Beginning from the parallel between regional recurrence and existence of a finite invariant measure one arrives at a panoply of quantitative or "statistical" counterparts of other recurrence properties in topological dynamics.

The premier among those is *ergodicity*, which is modeled on topological transitivity but vastly exceeds its topological cousin in importance. The reasons are twofold. First, in a certain sense the study of arbitrary measure-preserving systems may be reduced to ergodic ones by *ergodic decomposition* (Section 3.4f), while there is no decomposition into topologically transitive components for a general topological dynamical system. Second, another corollary of the Birkhoff Ergodic Theorem gives the Birkhoff average for an ergodic system: it is simply the space average of the function, *i.e.,* the measure of the set in the case of a characteristic function.

A natural counterpart of topological mixing is *mixing* (see Section 3.6h). Even the terminology suggests that it is a more basic and fundamental property than the topological one. Mixing means roughly that any set of initial conditions of positive measure becomes almost uniformly distributed throughout the phase space in the long run.

There are various refinements of the mixing property. In our context the most important one being the *K-property* (see Section 3.6k and Section 3.7j), which can be formulated as uniform mixing with respect to the distant past but is also intimately connected with *entropy*, the natural statistical counterpart (and precursor) of topological entropy, an important theme in the previous chapter.

c. The classification problem versus applications. Ergodic theory is a more extensive and better developed discipline in its own right than topological dynamics. One of the reasons is that the measure spaces which appear as phase spaces in the ergodic theory are standard (Section 3.2b). Ergodic theory may be considered from the general structural point of view as the branch of dynamics that deals with the structure of groups of measure-preserving (and more generally, nonsingular) transformations of a measure space up to metric isomorphism (see Section 3.4a). From this point of view its main goal is to describe invariants and models that classify maximally broad classes of systems with respect to metric isomorphism and certain weaker natural equivalence relations such as orbit equivalence (Section 3.4a) or Kakutani equivalence (Section 3.4e, Section 3.4p) as completely as possible. However, when ergodic theory is viewed in the broad context of dynamics the usefulness of the structural approach has to be qualified by two reasons.

First, and most importantly, measurable coordinate changes form too broad a class from the topological or differentiable point of view. In particular they do not preserve any "local" properties. Thus, the existence of a metric isomorphism with a certain "model" often provides only limited information about asymptotic properties important in the topological or differentiable context.

Second, while within ergodic theory an enormous variety of invariants and isomorphism types exist, those that appear in (or are at least typical for) interesting more "classical" situations are more limited. These two factors account for a perceptible difference between "abstract" and "applied" ergodic theory. The former is centered around the structural approach. It is presented in the surveys on the isomorphism problem and spectral and combinatorial constructions [**S-T, S-KT**]. The latter mostly concerns itself with several principal invariants of metric isomorphism, such as ergodicity, mixing, entropy, K-property and a variety of *noninvariant* properties.

Applications of ergodic theory to smooth dynamics and other areas appear in the surveys [**S-C, S-BKP, S-K, S-LS, S-KSS**]. Ergodic theory also plays a central role in [**S-B**].

In this chapter we describe some basics of ergodic theory having mostly "applied" aspects in mind. We briefly outline the structural approach but concentrate on notions and examples that are of particular relevance for other branches of dynamics.

d. Dynamical systems and random processes. There are two principal ways to introduce a surrogate of local properties in the general context of measurable transformations. One is to select in the space of bounded measurable functions an invariant subspace, separable but not closed in the L^∞ topology, which plays the role of regular (*e.g.,* differentiable or Hölder) functions. The other way is to fix a *partition* of the phase space into a finite or countable number of measurable sets. Then two points are "close" if they belong to the same element of the partition. This suggests to define nearby orbit segments, at least for discrete time systems, as those that visit the same elements of the partition during a prescribed time interval. One can view such a partition as a measuring device and say that only points in different elements of the partition may by identified as distinct. This is connected with *symbolic representation*: The partition defines an alphabet with respect to which every orbit gives a well-defined sequence. When this correspondence is injective, one obtains a natural conjugacy to a subshift, otherwise the subshift is a factor.

Both approaches are related to the representation of measure-preserving transformations as *stationary random processes*. These are the primary objects of probability theory, which provides the insights for selecting properties important for ergodic theory. Obviously the same or isomorphic measure-preserving transformations may be represented as stationary random processes in many different ways. The probability theory of such processes concerns itself with properties or numerical quantities that may happen to be invariant under metric isomorphism, such as mixing, entropy, or the law of large numbers, which is another name for ergodicity, or that are not invariant, such as various kinds of strong regularity, central limit theorem, exponential decay of correlations (Section 3.6l), large deviations estimates, the law of iterated logarithm *etc.* [**S-F**].

e. Entropy. The entropy of a finite state random process (or, equivalently, entropy of a measure-preserving transformation with respect to a given finite partition) can be defined as the average amount of information obtained on one step given complete knowledge of the past, *i.e.,* the sequence of partition elements to which preimages of a given point belong (Section 3.7, [**Rk1, Pa1**], [**KH**, Section 4.3]). In the case of an ergodic transformation, entropy can be characterized in a way parallel to its topological counterpart as the exponential growth rate for the number of *statistically significant* distinguishable orbit segments [**K4**].

2. Measure spaces, maps, and Lebesgue spaces

Natural measure spaces are standard and represent only a few classes up to an intrinsically defined equivalence. There is only one model of real significance for most of ergodic theory, namely the nonatomic Lebesgue space with finite measure, essentially an interval (Section 3.2b). Naturally, understanding the structure of the phase space helps develop dynamical notions and results. This is the reason an exposition of ergodic theory is usually preceded by a discussion of the structure of measure spaces.

This contrasts with the case of topological dynamics with its great variety of different phase spaces, where only basic concepts and principles of general topology (compactenss, completeness *etc.*) are relevant.

a. Measure spaces and maps. A measure space $(X, \mathcal{A}, \mu)$ consists of a set X, a σ-algebra $\mathcal{A}$ of subsets of X and a monotone σ-additive function $\mu\colon \mathcal{A} \to [0, \infty]$ with respect to which $\mathcal{A}$ is complete and σ-finite [**Rk1, H2**]. μ is said to be a probability measure and (X, μ) a probability space if $\mu(X) = 1$. Let $(X, \mathcal{A}, \mu)$, $(Y, \mathcal{B}, \nu)$ be measure spaces.

A map $T\colon X \to Y$ is said to be *measurable* if $T^{-1}(\mathcal{B}) \subset \mathcal{A}$. In this case we set $(T^*\nu)(T^{-1}(A)) := \nu(A)$ and $(T_*\mu)(A) := \mu(T^{-1}(A))$ for $A \in \mathcal{B}$. $T^*\nu$ is thus only defined on $T^{-1}(\mathcal{B})$.

T is said to be *nonsingular* if furthermore $T_*\mu \ll \nu$. In this case we define the *Radon–Nikodym derivative* $\rho_T\colon Y \to [0, \infty]$ by $T_*\mu = \rho_T\nu$.

A measurable map T is said to be *surjective* if $T(\mathcal{A}) = \mathcal{B}$, *measure-preserving* if $T_*\mu = \nu$, an *equivalence* if T is surjective with $T^{-1}(\mathcal{B}) = \mathcal{A}$ (up to discarding a single null set), and an *isomorphism* if it is a measure-preserving equivalence.

Two measures on the same space are said to be *equivalent* if the identity is an equivalence, *i.e.*, if they have the same null sets.

If T is measure-preserving then (Y, ν) is said to be a factor of (X, μ) and sometimes T is said to be a factor. To any factor $T\colon (X, \mathcal{A}, \mu) \to (Y, \mathcal{B}, \nu)$ one can associate the σ-algebra $T^{-1}(\mathcal{B}) \subset \mathcal{A}$.

(μ, ν) is said to be *quasi-equivariant* if $\nu(T(E)) = 0 \iff \mu(E) = 0$. If $\mu = \nu$ this single measure is then said to be *quasi-invariant*.

Throughout, we use the "mod-0-convention" that equalities between sets, transformations, *etc.*, are to be understood modulo null sets, *i.e.*, two sets agree if their symmetric difference is a null set, *etc.* In particular, one cannot assign definite meaning to a specific point or null set. Occasionally it is nevertheless necessary to emphasize an instance where a single null set can be discarded, rather than possibly uncountably many. Implicit in this convention and assumed therefore throughout, is completeness of the measure considered, *i.e.*, that subsets of null sets are always measurable.

There is a natural distance in the space of equivalence classes mod 0 measurable sets: $d(A, B) = \mu(A \triangle B)$, *i.e.*, the L^1 distance between corresponding characteristic functions

b. Lebesgue spaces. (See also [**Rk1**].) In ergodic theory the phase space X is a measure space with a finite or σ-finite measure μ, often a probability measure. Several results, including some important ones, can be proved in such a bare setting. However, in order to develop a comprehensive theory it is necessary to work in a context where the relation between σ-algebra and points of the space can be clarified sufficiently. The proper setting for this is that of a Lebesgue space. This assumption is nothing like as

restrictive as any of those we make about topological spaces. Indeed any measure space worth considering, as it were, is a Lebesgue space:

A measure space $(X, \mathcal{A}, \mu)$ is said to be a *Lebesgue space* if for some $a \in [0, \infty]$ it is isomorphic to the union of $[0, a]$ with Lebesgue measure and at most countably many points of positive measure.

There is also an intrinsic characterization of Lebesgue spaces that does not involve the isomorphism to the standard space. It includes *separability* (which for complete measures is equivalent to separability of the corresponding L^1) and a certain *completeness* property (which is different from completeness of the measure mentioned above). Similarly to the case of metric spaces there is a uniquely defined *completion* of a separable measure space which is a Lebesgue space [**Rk1**].

We describe the structural theory of these in case of finite measure, which is all that is needed in most of our applications. Infinite Lebesgue spaces primarily arise in connection with quasi-invariant measures, in which case one may consider an equivalent finite measure instead. Notice however difficulties which arise in the the description of measurable partitions (end of Section 3.2e).

An illustration of the generality of the notion of Lebesgue space is that any σ-finite Borel measure on a Borel subset of a separable metric space gives a Lebesgue space by completion.

c. Lebesgue points. Although individual points are not easy to get a handle on in the pure measure-theoretic context, Lebesgue spaces offer a way of finding a set of points of full measure that are individually "meaningful". These are Lebesgue points. Atoms are always Lebesgue points and in the standard model $[0, a]$ with Lebesgue measure one can define the Lebesgue points of a function $f \in L^1([0, a], \lambda)$ as those x for which

$$f(x) = \frac{1}{2\epsilon} \lim_{\epsilon \to 0} \int_{x-\epsilon}^{x+\epsilon} f \, d\lambda.$$

or $f(x) = (d/dx) \int_0^x f \, d\lambda$. This is a set of full measure. Thus if a countable collection of functions on a Lebesgue space is fixed, almost every point is a Lebesgue point for all of these functions. See the proof of Theorem 3.6.3 below for an application of this method.

d. Measurable partitions. For a partition ξ of a Lebesgue space $(X, \mathcal{A}, \mu)$ *with finite measure* let $\mathcal{A}(\xi) := \{\bigcup_\alpha C_\alpha \in \mathcal{A} \mid C_\alpha \in \xi\}$ be the σ-algebra of unions of elements of ξ. A *measurable partition* of a Lebesgue space $(X, \mathcal{A}, \mu)$ is a partition $\xi \subset \mathcal{A}$ for which $\mathcal{A}(\xi)$ contains a countable subset that separates any two elements of ξ (up to discarding a *single* null set).

Among the useful properties of Lebesgue spaces is that the relation between factors and sub-σ-algebras goes both ways. Following [**Rk1**] we have:

THEOREM 3.2.1. *For every measurable partition ξ of a Lebesgue space $(X, \mathcal{A}, \mu)$ the triple $(\xi, \mathcal{P}(\xi), \nu)$, where $\mathcal{P}(\xi) := \{A \subset \xi \mid \bigcup_{C \in A} C \in \mathcal{A}\}$ and $\nu(A) := \mu(\bigcup_{C \in A} C)$, is a Lebesgue space.*

Furthermore every sub-σ-algebra arises in the above fashion:

If $(X, \mathcal{A}, \mu)$ is a Lebesgue space and $\mathcal{B} \subset \mathcal{A}$ a sub-σ-algebra there exists a unique (mod 0) measurable partition ξ such that $\mathcal{B} = \mathcal{A}(\xi)$.

This implies that for any partition ξ the σ-algebra $\mathcal{A}(\xi)$ is of the form $\mathcal{A}(\eta)$ for a unique measurable partition η, which is called the *measurable hull* of ξ.

There is an obvious partial-ordering relation between partitions: We say that η is a *refinement* of ξ and that ξ is *subordinate* to η, and we write $\xi \leq \eta$, if $\mathcal{A}(\xi) \subset \mathcal{A}(\eta)$, *i.e.*, for all $D \in \eta$ there exists a $C \in \xi$ such that $D \subset C$.

For measurable partitions ξ, η we define the *joint partition*

$$\xi \vee \eta := \{C \cap D \mid C \in \xi,\ D \in \eta\}.$$

For any family $\{\xi_\alpha\}_{\alpha \in A}$ of partitions define $\bigvee_{\alpha \in A} \xi_\alpha$ to be the measurable partition defined by the smallest σ-algebra containing $\bigcup_{\alpha \in A} \xi_\alpha$.

A complementary definition is that of $\bigwedge_{\alpha \in A} \xi_\alpha$ via $\mathcal{A}(\bigwedge_{\alpha \in A} \xi_\alpha) = \bigcap_{\alpha \in A} \mathcal{A}(\xi_\alpha)$.

We say that ξ and η are *independent* if $\mu(C \cap D) = \mu(C) \cdot \mu(D)$ for all $C \in \mathcal{A}(\xi)$, $D \in \mathcal{A}(\eta)$.

To each measurable partition corresponds the space $L^2(\mathcal{A}(\xi), \mu)$. Note that $L^2(\mathcal{A}(\xi), \mu)$ is *multiplicative* in the sense that the algebra $L^2(\mathcal{A}(\xi), \mu) \cap L^\infty(\mathcal{A}(\xi), \mu)$ (with multiplication defined pointwise) is dense in $L^2(\mathcal{A}(\xi), \mu)$. Conversely, every linear subspace generated by bounded functions and closed under multiplication of its bounded elements is of the form $L^2(\mathcal{A}(\xi), \mu)$ for a unique measurable partition ξ.

e. Conditional measures. The central property of Lebesgue spaces is that for every measurable partition ξ there exists a system of *conditional measures*:

THEOREM 3.2.2. *Let $(X, \mathcal{A}, \mu)$ be a Lebesgue space and ξ a measurable partition. Then on each $C \in \xi$ there is a probability measure μ_C defined on a σ-algebra $\mathcal{A}_C$ such that for every $A \in \mathcal{A}$ we have:*

(1) *$A \cap C \in \mathcal{A}_C$ for almost every $C \in \xi$,*
(2) *$\mu(A) = \int_\xi \mu_C(A \cap C)\, d\nu$, where ν is as in Theorem 3.2.1.*

In Section 4.2d we indicate a proof using *continuous realization* (Section 4.2e, [**F1**, Proposition 5.3, Theorem 5.8]). This approach is equivalent to one using sub-σ-algebras and conditional expectation.

Any two systems of conditional measures coincide outside a null set. The projection operator on $L^1(\mathcal{A}, \mu)$ to ξ-measurable functions is obtained by integration with respect to the conditional measures. On $L^2(\mathcal{A}, \mu)$, this is an orthogonal projection onto the subspace $L^2(\mathcal{A}(\xi), \mu)$ of functions constant on elements of ξ. The measures μ_C are naturally defined on X via $\mu_C(C \cap (\cdot))$.

This and the preceding subsection contain the essence of the theory of Lebesgue spaces and provide sufficient measure theoretic background for the study of ergodic theory.

One can describe measurable partitions of a Lebesgue space with finite measure up to equivalence of the Lebesgue space. The leading case is that where there are no elements of positive measure (atoms) and almost every conditional measure is nonatomic. In this case the partition is isomorphic to the standard partition of the unit square into vertical fibers. The atomic parts of the conditional measures can also be organized measurably. This in particular gives the following observation:

PROPOSITION 3.2.3. *Any measurable partition of a Lebesgue space such that almost all elements (C, μ_C) are isomorphic to a fixed Lebesgue space (Y, ν), is isomorphic to the direct product $(X(\xi), \mu) \times (Y, \nu)$. Thus, every skew product (space with a factor) is a product.*

For infinite measure Lebesgue spaces the theory of measurable partitions presents certain difficulties. While the notion of a measurable partition can be defined by using an equivalent finite measure the factor measure and conditional measures are not always defined. On the other hand, there are important cases where the measure on the factor space is defined and is either infinite or finite: projection of Lebesgue measure from the real line $\mathbb{R}$ to the circle $\mathbb{R}/\mathbb{Z}$ is an example where the factor measure is finite and the conditional measures are infinite. A more general example of this situation appears in Section 3.3c below (lattices in locally compact groups).

f. The Radon–Nikodym cocycle. As an application and an illustration of the utility of the restriction to Lebesgue spaces, we now construct a Radon–Nikodym cocycle on the preimage. Suppose $T\colon (X,\mathcal{A},\mu) \to (Y,\mathcal{B},\nu)$ with (μ,ν) quasi-equivariant probability measures (Section 3.2a) so that the Radon–Nikodym derivative $\rho_T\colon Y \to [0,\infty]$ is defined by $T_*\mu = \rho_T\nu$. Consider $T^{-1}(\mathcal{B}) \subset \mathcal{A}$ and the corresponding measurable partition ξ such that $T^{-1}(\mathcal{B}) = \mathcal{A}(\xi)$. Then almost every conditional measure μ_C is atomic (because otherwise there would be a null set with nonnull image) and hence defines a function on X accociating to almost every point $x \in X$ the conditional measure $\mu_{C(x)}(x)$ of x on the element of $C(x) \in \xi$ that contains x. We then obtain

$$\nu(T(A)) = \int_A \frac{1}{\rho_T(T(x))\mu_{C(x)}(x)} d\mu,$$

so $J^T := 1/(\rho_T \circ T \cdot \mu_C)$ is called *the Radon–Nikodym Jacobian* for T. In Section 5.2m we use this to find a criterion for existence of an invariant measure in an invariant measure class.

g. Relative products. Here is another application of conditional measures. Let $\varphi_1\colon (X_1,\mu_1) \to (Y,\nu)$ and $\varphi_2\colon (X_2,\mu_2) \to (Y,\nu)$ be measure-preserving maps of Lebesgue spaces (*i.e.*, (X_1,μ_1) and (X_2,μ_2) have a common factor). Then the *relative product of* (X_1,μ_1) *and* (X_2,μ_2) *over* (X,ν) is the space $X = X_1 \underset{\nu}{\overset{Y}{\times}} X_2 := \{(x_1,x_2) \in X_1 \times X_2 \mid \varphi_1(x_1) = \varphi_2(x_2)\}$ with the measure $\mu = \mu_1 \underset{\nu}{\times} \mu_2$ defined by the conditions $\pi_*\mu = \nu$, where $\pi(x_1,x_2) = \varphi_1(x_1)$, and the conditional measure on $\pi^{-1}(\{y\})$ is the product of the conditional measures for μ_1 on $\varphi_1^{-1}(\{y\})$ and μ_2 on $\varphi_2^{-1}(\{y\})$.

In particular, the case $\operatorname{card} Y = 1$ gives the standard product measure. Furthermore, since every measurable partition (or sub-σ-algebra) ξ defines a factor $\pi\colon (X,\mu) \to (X(\xi),\mu)$, the case $\varphi_1 = \varphi_2 = \pi$ defines the relative product of a Lebesgue space with itself over a sub-σ-algebra.

3. Setting and examples

a. Measurable actions. A measurable action of a second countable locally compact topological group or countable semigroup G is a measurable map $\Phi\colon G \times X \to X$ with $\Phi^{g_1 g_2} = \Phi^{g_2} \circ \Phi^{g_1}$ for all $g_1, g_2 \in G$.

We consider only actions of groups and discrete semigroups. Actions of continuous semigroups are omitted because they are not sufficiently prominent in ergodic theory and present some technical difficulties. In both cases there is a natural measurable structure on $G \times X$ (obvious in the discrete case and given by left Haar measure in case of a group).

Ergodic theory may study two possible structures on the phase space, namely either a measure itself or its equivalence class (with respect to mutual absolute continuity), *i.e.*, the collection of all sets of measure zero. Accordingly, ergodic theory concerns groups or semigroups of measurable transformations of X that either preserve μ (μ is invariant) or transform it into an equivalent measure.

Ergodic theory provides the appropriate paradigms and tools for studying the asymptotic distribution and statistical behavior of orbits for continuous and smooth dynamical systems and as such impinges upon virtually all dynamics. The central point is existence of Borel invariant measures for topological dynamical systems on compact spaces (Section 4.2 and Section 4.4) and of an invariant measure class for smooth systems.

b. Quasi-invariant measures. The most general natural settings are semigroup actions by nonsingular surjective maps and group actions by equivalences. In the latter case we get the Radon–Nikodym Jacobian $J_{g,\mu}$ by $\Phi^{g*}\mu = J_{g,\mu} \cdot \mu$. This is a multiplicative cocycle over Φ whose values are positive reals (see Section 1.3k). There exists an equivalent invariant measure if the Jacobian is a coboundary. The density is the transfer function, because it gives the multiple of μ for which the Jacobian is 1.

For a semigroup action by transformations with quasi-invariant measure μ the density θ of an absolutely continuous invariant measure $\nu = \theta\mu$ is a fixed point of the *Perron–Frobenius operator* or *transfer operator* $\mathcal{T}(\cdot)$ defined by

$$\mathcal{T}(\theta)(x) := \sum_{\Phi^g(y)=x} \frac{\theta(y)}{J_{g,\mu}(y)}$$

(see Section 3.2f, Section 5.2m).

Now we look at the examples introduced in Section 2.1 from the point of view of ergodic theory and introduce some more examples relevant to the subject.

c. Homogeneous dynamics. (See also [**S-KSS**].) Left Haar measure on a group H is naturally invariant under any group of left translations. This measure is finite if H is compact. In particular, translations on compact abelian groups (Section 2.1b, especially Example 2.1.1) give nice examples of transformations with finite invariant measures. These play an important role in parts of the later discussion (Theorem 3.6.3, Section 4.3c, Section 4.3d, Section 7.1).

If H is not compact, then Haar measure is infinite. It produces a measure on the factor $M = H/K$ (where K is a closed subgroup), which is quasi-invariant under left translations. However, it does not necessarily produce a finite or σ-finite measure invariant under all or any left translations on M, even if K is cocompact (*i.e.*, M is compact).

Several cases are worth special attention. Assume that there is a two-sided Haar measure on H. Any discrete subgroup $K < H$ such that $M = H/K$ has a finite invariant measure, is said to be a *lattice*. If K is a discrete cocompact subgroup of H then it is a lattice: Simply take any *fundamental domain* $\mathcal{F}$ for K with compact closure, restrict Haar measure to it, and project to the factor.

EXAMPLE 3.3.1 (A lattice that is not cocompact). Let $H = SL(n, \mathbb{R})$ be the group of all real $n \times n$ matrices with determinant 1. Then $K = SL(n, \mathbb{Z})$, the subgroup with integer matrix entries, is a lattice that is not cocompact: There is an unbounded fundamental domain for K that has finite Haar measure.

EXAMPLE 3.3.2 (An example without invariant measure on a compact factor). This is another aspect of Example 2.4.4. Consider $H = SL(2,\mathbb{R})$ and K the upper triangular subgroup. The factor can be identified with a projective line (*i.e.,* the circle) with action by projective transformations. There is no invariant measure in the class obtained from projecting Haar measure, which is in this case the Lebesgue class. Moreover, there is no σ-finite invariant measure at all. Even for *countable* subgroups such as $SL(2,\mathbb{Z})$ there are no finite invariant measures (Section 4.2b).

d. Group automorphisms. For any automorphism of a compact group, normalized Haar measure is invariant by uniqueness. In fact, this holds for endomorphisms as well. For a noncompact group, the Haar measure is multiplied by a constant and is hence quasi-invariant. Examples of this are toral automorphisms and endomorphisms, which preserve Lebesgue measure.

e. Bernoulli shifts. For any Borel probability measure μ on a compact set K the *product measure* μ_Γ on K^Γ is invariant under the shift. Such measures are said to be *Bernoulli measures* and the measure preserving transformation given by the shift is then called *Bernoulli shift*. In particular, if K is a compact group and μ the Haar measure then μ_Γ is the Haar measure on K^Γ, providing an example of the previous situation.

If $K = \{0,\dots,N-1\}$, then any probability measure on K, hence any Bernoulli measure on K^Γ, is defined by a probability distribution $p = (p_0,\dots,p_{N-1}) \in \mathbb{R}^N$ (*i.e.,* $\mu_i \geq 0$, $\sum \mu_i = 1$). The product Bernoulli measure is denoted by μ_p.

f. Markov measure. (See also [**KH**, Section 4.4c].) A more general class of invariant measures for the N-shift and topological Markov chains are *Markov measures*. Let $\Pi := (\pi_{ij})_{i,j=0,\dots,N-1}$ be an $N \times N$ matrix with nonnegative entries such that $\sum_{i=0}^{N-1} \pi_{ij} = 1$ for $j = 0,\dots,N-1$. Such matrices are said to be *stochastic*. Similarly to the case of 0-1 matrices, we say that a stochastic matrix Π is *transitive* if for some m all entries of Π^m are positive. Every stochastic matrix Π has an invariant vector p with nonnegative coordinates. If Π is transitive, such a vector is unique (up to rescaling), 1 is a simple eigenvalue, and all other eigenvalues of Π have absolute values less than 1 [**KH**, Theorem 1.9.11].

Given a stochastic matrix Π and an invariant probability vector p we define the Markov measure $\mu_{\Pi,p}$ on Ω_N by

$$\mu_{\Pi,p}(C^m_\alpha) = \Big(\prod_{i=-m}^{m-1} \pi_{\alpha_i\alpha_{i+1}}\Big) p_{\alpha_m}. \tag{3.1}$$

Let us emphasize that π_{ij} represents the proportion of the measure of the cylinder C^0_j (whose measure is p_j) that is transported to C^0_i. (Compare with $\sum_j \pi_{ij} p_j = p_i$.) This makes stochasticity of the matrix an obvious necessary condition for invariance of the measure. Calculation shows that $\Pi p = p$ guarantees σ_N-invariance of $\mu_{\Pi,p}$.

Suppose now A is a 0-1 matrix and suppose that a stochastic matrix Π is such that $\pi_{ij} = 0$ if $a_{ij} = 0$. Then $\operatorname{supp}\mu_{\Pi,p} \subset \Omega_A$ and hence $\mu_{\Pi,p}$ can be viewed as an invariant measure for the topological Markov chain σ_A. If Π is a transitive matrix we denote the measure $\mu_{\Pi,p}$ simply by μ_Π since the vector p is unique in this case.

4. Basic concepts and constructions

Now we revisit the basic notions and constructions introduced in Section 1.3 and add some more that are specific to the present setting.

a. Isomorphism and orbit equivalence. The isomorphism notions from Section 1.3a and Section 1.3b can be made specific to the present context in two ways, depending on whether one concentrates on a measure or a measure class.

Let $\Phi\colon G \times (X,\mu) \to (X,\mu)$ and $\Psi\colon G \times (Y,\nu) \to (Y,\nu)$ be measure-preserving actions. Φ and Ψ are said to be *metrically isomorphic* if there exists an isomorphism $R\colon (X,\mu) \to (Y,\nu)$ with $\Psi \circ R = R \circ \Phi$.

Φ and Ψ are said to be *orbit equivalent* if there exists an equivalence $R\colon (X,\mu) \to (Y,\nu)$ that maps orbits onto orbits.

If Φ and Ψ are actions with quasi-invariant measures then they are said to be *metrically isomorphic* if there is an equivalence $R\colon (X,\mu) \to (Y,\nu)$ that is injective $(\mathrm{mod}\ 0)$ and $\Psi \circ R = R \circ \Phi$.

b. Joinings. (See also [**S-T**].) An important general construction in ergodic theory is that of a joining, which, in particular, unifies those of isomorphism and products:

Let $\Phi\colon G \times (X,\mu) \to (X,\mu)$ and $\Psi\colon G \times (Y,\nu) \to (Y,\nu)$ be measure-preserving actions. A *joining* of these actions is the action $\Phi \times \Psi\colon X \times Y \to X \times Y$ together with an invariant measure η that projects properly: If π_X, π_Y are the projections from $X \times Y$ then $\pi_{X*}\eta = \mu$ and $\pi_{Y*}\eta = \nu$.

The product measure always gives a joining of two actions and there are cases in which this is the only one (Section 3.6f6, [**S-T**]). On the other hand, the diagonal measure μ_Δ always is a joining of an action on (X,μ) with itself, and, more generally, for two isomorphic actions a joining other than the product is given by $(\mathrm{Id} \times R)_*\mu$, where R is the isomorphism. In particular, any transformation commuting with an action produces a corresponding *self-joining* in this way.

c. Poincaré Recurrence and induced maps. In the context of transformations preserving a finite measure, one can define restrictions of a map to sets of positive measure even if these are not invariant. This is due to the following basic fact.

THEOREM 3.4.1. *(Poincaré Recurrence Theorem). Almost every point from a set A of positive measure returns to A infinitely often,* i.e., $\mu(\{x \in A \mid \{T^n(x)\}_{n\geq N} \subset X \smallsetminus A\}) = 0$ *for any* $N \in \mathbb{N}$.

PROOF. To see this, note that replacing T by T^N we may assume $N = 1$. The set

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

is measurable, $T^{-n}(\tilde{A}) \cap \tilde{A} = \varnothing$ for every n and hence $T^{-n}(\tilde{A}) \cap T^{-m}(\tilde{A}) = \varnothing$ for all $m, n \in \mathbb{N}$, so

$$\infty > \mu(X) \geq \mu\left(\bigcup_{n=0}^{\infty} T^{-n}(\tilde{A})\right) = \sum_{n=0}^{\infty} \mu(T^{-n}(\tilde{A})) = \sum_{n=0}^{\infty} \mu(\tilde{A})$$

and $\mu(\tilde{A}) = 0$, as needed. □

Thus, if T preserves a probability measure and $A \subset X$ has positive measure then for almost every $x \in A$ we can define a *first return time* $n_A(x) := \min\{n \in \mathbb{N} \mid T^n(x) \in A\}$ and hence an *induced map* $T_A : (A, \mu_A) \to (A, \mu_A)$ by $x \mapsto T^{n_A(x)}(x)$.

d. Ergodicity. The irreducibility notion in the present setting that corresponds to both topological transitivity and minimality is that of ergodicity: An action is said to be *ergodic* if every measurable invariant set is null or conull.

For this definition, and more so for what follows, the following result is useful:

THEOREM 3.4.2. **[Vs]** *For an action of a locally compact group on a Lebesgue space a measurable set is invariant mod 0 under the entire action if and only if it is invariant mod 0 under the action of each element.*

The point is that no potentially dangerous uncountable union of null sets is needed. In particular, for countable groups and semigroups this result holds because countable unions of null sets are null sets.

e. Kakutani (monotone) equivalence. (See also [**S-T, K1, ORW**].) For ergodic transformations there is another equivalence relation which is modelled on topological orbit equivalence for flows.

Ergodic measure preserving transformations $T\colon (X, \mu) \to (X, \mu)$ and $S\colon (Y, \nu) \to (Y, \nu)$ are *Kakutani equivalent* or *monotone equivalent* if there are measurable sets of positive measure $A \subset X$ and $B \subset Y$ such that the induced maps T_A and S_B are metrically isomorphic. See Section 3.4p for another definition.

It is quite remarkable that Kakutani equivalence is both much weaker than metric isomorphism and much stronger than orbit equivalence. In particular, there is a unique natural simplest class of Kakutani equivalent maps on a nonatomic Lebesue space.

f. Ergodic decomposition. Given a measurable action Φ of G on a Lebesgue space the *invariant σ-algebra* $\mathbb{I}(\Phi)$ is the collection of Φ-invariant measurable sets. By Theorem 3.2.1 and Theorem 3.2.2 this gives rise to a measurable partition $\eta(\Phi)$ with a system of conditional measures, together called the *ergodic decomposition* of (Φ, μ). The partition is invariant by construction and the uniqueness assertion in Theorem 3.2.1. The conditional measures are invariant by uniqueness. The terminology is justified by the following main result (Section 4.2d, Section 4.2e):

THEOREM 3.4.3. *For a measurable action Φ of G on a Lebesgue space the conditional measures in the ergodic decomposition are ergodic.*

g. Factors. Let $\Phi\colon G \times (X, \mu) \to (X, \mu)$ and $\Psi\colon G \times (Y, \nu) \to (Y, \nu)$ be measure-preserving actions. Ψ is said to be a (metric) *factor* of Φ if there exists a factor $R\colon (X, \mu) \to (Y, \nu)$ such that $\Psi \circ R = R \circ \Phi$.

Every measurable partition defines a factor via the partition $\xi^T := \bigvee_{n\in\mathbb{Z}} T^n\xi$. If the partition is finite or countable then the factor is represented as a stationary random process with a finite or countable set of states correspondingly.

Every factor defines a self-joining of an action with itself via the construction of the relative product from Section 3.2g. Let $\Psi\colon G \times (Y, \nu) \to (Y, \nu)$ be a factor of $\Phi\colon G \times (X, \mu) \to (X, \mu)$ and consider the action on $X \overset{Y}{\times} X \hookrightarrow X \times X$ induced by the diagonal action on $X \times X$. The measure $\mu \overset{\nu}{\times} \mu$ is invariant under this action and projects to μ under

the factors, hence is a self-joining. More generally, if two actions of the same group have a common factor (or, rather, isomorphic factors), the relative product construction similarly provides for a joining of the two actions.

There are some canonically defined factors in measure-preserving systems, of which we have already encountered the ergodic decomposition, which can be described as the maximal factor over the identity. Other examples are the Pinsker algebra (Section 3.7j) and the maximal algebra with discrete spectrum for ergodic systems (Section 3.6d).

h. Generators. A family Ξ of partitions is said to be *sufficient* or *exhaustive* if for any $\epsilon > 0$ any measurable set can be approximated in the symmetric difference metric (Section 3.2a) to within ϵ by measurable unions from a partition $\xi_l := \bigvee_{i=-l}^{l} T^i \xi$ for some $\xi \in \Xi$, $l \in \mathbb{N}$. A partition ξ is said to be a *generator* if $\{\xi\}$ is sufficient.

Generators are needed for effective symbolic representation (Section 3.1d) and the following nontrivial results [**Kr1, Pa1**] limit the complexity of the symbolic systems required:

THEOREM 3.4.4. *For ergodic systems there is always a countable generator. If the entropy (Section 3.1e, Section 3.7c) is finite then there is a finite generator, and there is an optimal bound for its cardinality expressible in terms of entropy. In particular, zero-entropy systems have a 2-element generator.*

i. Inverse limits. Suppose $h_i \colon X_{i+1} \to X_i$ ($i \in \mathbb{N}$) are factors of probability spaces $(X_i, \mathcal{A}_i, \mu_i)$ and consider the space

$$\mathcal{X} := \{(x_i)_{i\in\mathbb{N}} \mid h_i(x_{i+1}) = x_i\}$$

with the measure induced by pulling back the measures μ_i to the σ-algebra $\mathcal{A}$ generated by the pullbacks of the σ-algebras $\mathcal{A}_i$. If the h_i define factors of actions on the X_i then the action on $\mathcal{X}$ is naturally defined and is called the *inverse limit*.

EXAMPLE 3.4.5. The case $X_i = \mathbb{Z}/2^i\mathbb{Z}$ discussed in the topological setting fits into the present category by using the uniform measures on the X_i, which gives rise to Haar measure on $\mathbb{Z}_2$, an invariant measure for the binary adding machine (Example 2.2.9). It is easy to see that this is the only invariant Borel probability measure because the projections, which determine it uniquely, are invariant measures for finite cyclic permutations. This is an example of *unique ergodicity* (Section 4.3a).

j. Natural extensions. The *natural extension* of a measure-preserving transformation T is obtained by applying the inverse limit construction with $X_i = X$, $h_i = T$. The map is defined as in the topological context (Section 2.2h). The resulting map is an invertible measure-preserving transformation that has the original map as a factor.

Corresponding to the topological situation we now find by way of example that from any Bernoulli measure for the one-sided shift one obtains a corresponding Bernoulli measure for the two-sided shift, and from Lebesgue measure for $E_2 \colon x \mapsto 2x$ on S^1 one gets Haar measure on the solenoid.

k. Cocycles. A particularly important role of cocycles in ergodic theory stems from the fact that in the measurable setting every skew product is a direct product. In the case of a general measurable structure this is due to Proposition 3.2.3, but the same holds even if the fibers have a more special structure, such as that of a smooth manifold, or a homogeneous space.

l. Isometric extensions. Thus, isometric extensions may be identified with cocycles with values in the isometry group of the fiber. In the case of a compact fiber there is a natural invariant product measure for isometric extensions and the ergodic decomposition as well as some other basic properties of the extension can be well understood from the properties of the base and the structure of the cocycle. In particular, the structure of the ergodic components in the fiber direction is fairly regular. For example, in the case of a compact group extension of an ergodic finite measure-preserving action the ergodic decomposition and the partition into fibers are independent, and in a typical fiber the ergodic decomposition induces a homogeneous partition by cosets of a closed subgroup.

The situation is quite different for isometic extensions with noncompact fibers, even as simple as $\mathbb{R}$-extensions of rotations on the circle, which are sometimes called *cylindrical cascades*. The ergodic decomposition for such an extension may have nothing to do with the homogeneous structure in the fibers.

m. Suspensions. In the measurable situation the suspension space is isomorphic to the product $X \times (H\backslash G)$ and in the case of finite-measure-preserving actions on X the suspension preserves the product of the invariant measure with the right-invariant measure on $H\backslash G$ (if it exists). Then according to Section 3.3c, if H is a lattice in G, in particular $G = \mathbb{R}^n$, $H = \mathbb{Z}^n$ or $G = SL(n, \mathbb{R})$, $H = SL(n, \mathbb{Z})$, such a measure always exists and is finite.

Ergodicity of the H-action is equivalent to ergodicity of the suspension.

n. Mackey range. (See also [**Z**].) For the general Mackey range construction in the measurable category one considers the measurable hull of the partition into orbits. In general there is no invariant measure. For the flow under a function over an invertible measure-preserving transformation, however, the restriction of the product measure to the fundamental domain provides an invariant measure for the flow, which is finite or infinite according to whether $\alpha(1, \cdot)$ is integrable or not.

o. Sections, special representations for flows and cocycle representations for $\mathbb{R}^k$ actions. Given a measurable (in particular measure-preserving) action Φ of a continuous group G and a lattice $\Gamma \subset G$, it is natural to ask whether Φ is metrically isomorphic to an action induced by a cocycle from an action of Γ. In order to achieve this, one needs to construct a proper "section" of the action Φ and then introduce a measurable Γ-structure on the intersections of the section with the orbits of Φ. If one wants the resulting cocycle to be nice, one should also make sure that this Γ-structure is somehow related to the G-structure on each orbit. The most basic and classical result of this kind is the Ambrose–Kakutani theorem, which we present in a stronger form.

THEOREM 3.4.6 ([**AmK**]). *Every measure-preserving flow is isomorphic to a special flow over a measure-preserving transformation. Furthermore, for an aperiodic flow one can choose this special representation in such a way that the roof function is arbitrarily close to a given constant in the uniform topology.*

Thus, the order of points on the orbits of the flow agrees with the order of the points on the orbits of the section map, and the time distortion is almost constant with a given precision.

A natural generalization to higher-rank groups is the special representation theorem for flows [**K2**].

THEOREM 3.4.7. *Given $\epsilon > 0$, any essentially free measure-preserving action of $\mathbb{R}^k$ on a space X is isomorphic to an action of $\mathbb{Z}^k$ induced by a cocycle $\alpha\colon \mathbb{Z}^k \times X \to \mathbb{R}^k$ such that*

$$(1-\epsilon)\|n_1 - n_2\| \le \|\alpha(n_1, x) - \alpha(n_2, x)\| \le (1+\epsilon)\|n_1 - n_2\|$$

for all $n_1, n_2 \in \mathbb{Z}^k$, $x \in X$.

For an ergodic flow *any* section may serve as the "base" for the representation as a special flow. This is not the case for higher-rank groups, as was conjectured by Furstenberg and proved by Burago and Kleiner [**BK**] and independently by McMullen [**MM**].

On the other hand, for actions of some nonamenable groups, such as noncompact simple Lie groups or groups of real rank greater than one, a representation as an induced action is not always possible, even with a measurable cocycle. This follows from the Zimmer Cocycle Superrigidity Theorem [**S-FK, Z**].

p. Kakutani equivalence for flows. The Ambrose–Kakutani Theorem 3.4.6 makes the following question natural: When can two flows be represented as special flows over the same measure-preserving transformation? A related question is: What do different sections of the same flow have in common? For ergodic flows these questions are related to Kakutani equivalence, which first appeared in Section 3.4e.

DEFINITION 3.4.8. Two ergodic flows ϕ and ψ are said to be *Kakutani equivalent* if they are metrically isomorphic to special flows over the same measure-preserving transformation.

The basic properties of Kakutani equivalence can be summarized as follows:

THEOREM 3.4.9. [**K1**] *Two flows are Kakutani equivalent if and only if they are isomorphic to special flows over Kakutani equivalent transformations. In particular, if a flow is isomorphic to special flows over different transformations these transformations are Kakutani equivalent.*

Two flows ϕ and ψ are Kakutani equivalent if and only if there exists a measurable equivalence h between the phase spaces that maps orbits of ϕ onto orbits of ψ and preserves order on almost every orbit. Furthermore, if such an equivalence exists it can be chosen differentiable along almost every orbit.

q. Induced action on L^p. Any measure-preserving action Φ of a group G on a measure space (X, μ) generates an isometric representation of G on $L^p(\mu)$ by $U_g\colon \varphi \mapsto \varphi \circ \Phi^{g^{-1}}$ with the property that $U_g(\varphi \cdot \psi) = U_g(\varphi) \cdot U_g(\psi)$ (multiplicativity). For an action of $\mathbb{Z}$ generated by $T\colon X \to X$ the notation U_T for the operator U_1 is commonly used.

The case $p = 2$ is interesting because of the well developed theory of unitary group representations in Hilbert space. If two actions are isomorphic then the corresponding unitary representations on 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 sometimes called *spectrally isomorphic*

Metric isomorphism between actions is equivalent to unitary isomorphism between the corresponding unitary representations which in addition preserves the multiplicative

structure in L^2. Thus in general one does not expect spectrally isomorphic actions to be metrically isomorphic. As we go through the list of principal invariants of metric isomorphism we will indicate whether they are spectral or not.

For actions on probability spaces, ergodicity is a spectral invariant: It is equivalent to one-dimensionality of the space of invariant functions (because only constant functions are invariant).

A general approach to spectral invariants is to consider the decomposition of the representation in $L^2(X,\mu)$ into irreducible ones [**S-FK**, Theorem 4.1] and to look at the spectral measure on the space of irreducible representations and multiplicities. For abelian groups all irreducible representations are one-dimensional and are identified with the characters in the group. Thus, a complete set of spectral invariants consists of a measure class on the group of characters, called the *maximal spectral type* and a multiplicity function [**S-KT**].

For actions of semigroups the associated operators are isometric but not unitary.

Even for actions with a quasi-invariant measure μ one can define unitary operators by $U_g\varphi := \sqrt{[\Phi^g_*\mu/\mu]}\varphi\circ\Phi^{g^{-1}}$, but this construction is less useful because isomorphic actions may produce different representations.

See [**S-KT**] and [**S-FK**] for a more detailed discussion of spectral invariants.

r. Rokhlin Lemma. No properties save for positive measure sets of periodic orbits can be expressed in terms of fixed length orbit segments. The most striking manifestation of this observation, which is also one of the most useful devices in ergodic theory, is the celebrated Rokhlin Lemma [**S-T, H1**][**Pt**, Lemma 4.7]:

THEOREM 3.4.10. *Let T be an aperiodic measure-preserving transformation of (X,μ),* i.e., *the set of periodic points has zero measure. Then for $n\in\mathbb{N}$, $\epsilon>0$ there exists a set A such that $T^i(A)\cap T^j(A)=\varnothing$ for $i\neq j$, $0\le i,j<n$ and $\mu(\bigcup_{k=0}^{n-1}T^k(A))>1-\epsilon$.*

An important negative consequence of the Rokhlin Lemma is that properties of any aperiodic transformation can be changed in an arbitrary way by altering it on a set of arbitrarily small positive measure.

As a corollary of the Rokhlin Lemma one has the following general spectral property of measure preserving transformations.

THEOREM 3.4.11. *For an aperiodic measure-preserving transformation T the spectrum of the operator U_T on L^2 is the entire unit circle.*

SKETCH OF PROOF. For $n\in\mathbb{N}$ and $k\in\{0,\dots,n-1\}$ pick $\epsilon>0$ and construct A as in Theorem 3.4.10. Set $f=\exp(-2\pi ikm/n)$ on $T^m(A)$ for $m\in\{0,\dots,n-1\}$. Then $\|U_Tf-\exp(2\pi ik/n)f\|<4\epsilon\|f\|$. Hence every rational point on the unit circle belongs to the spectrum of U_T. □

The Rokhlin Lemma is related to amenability. While direct generalizations are possible only to groups with tiling properties of a Følner sequence, such as $\mathbb{R}^n$ or $\mathbb{Z}^n$, a weaker version serves as the basis for ergodic theory of amenable group actions [**S-T, OW**].

5. Ergodic theorems

We restrict ourselves to cyclic systems. See [**S-B**] for in-depth discussion of the subject, including more general types of averaging and ergodic theorems for noncyclic systems.

For a measure-preserving transformation T of a probability space $(X, \mathcal{B}, \mu)$ denote by $\mathbb{I} := \{A \in \mathcal{A} \mid T^{-1}(A) = A\}$ the invariant σ-algebra (Section 3.4f) and by $P_{\mathbb{I}}$ the projection to the space $\mathbb{I}$ of invariant functions. (For any sub σ-algebra $\mathcal{B}$ there is a projection operator from $L^p(\mathcal{A}, \mu)$ to $L^p(\mathcal{B}, \mu)$.)

a. The von Neumann mean ergodic theorem. In the introduction we already discussed pointwise convergence of asymptotic distribution. Sometimes convergence of certain averages (in the mean) is more natural and immediate. It is also easier to prove than pointwise convergence, so we begin there.

THEOREM 3.5.1. *Consider an invertible measure-preserving transformation T of a measure space (X, μ). Then*

$$B_n(f) := \frac{1}{n} \sum_{k=0}^{n-1} U_T^k f \xrightarrow{L^p} P_{\mathbb{I}} f$$

for all $f \in L^p$, $1 \le p < \infty$.

PROOF. The case $p = 2$ is particularly nice: Since $\|B_n(f)\| \le \|f\|$ it suffices to check the claim for a dense set of $f \in L^2$. Note that $B_n(f) = f$ for $f \in \mathbb{I}$. Next, if $f \in \{U_T g - g \mid g \in L^2\}$ then $\|B_n(f)\| = \|\frac{1}{n}(U_T^n g - g)\| \le 2\|g\|/n \to 0$ and consequently the same holds for $f \in L_0 := \ker P_{\mathbb{I}}$, the linear hull of $\{U_T g - g \mid g \in L^2\}$. Finally, for $f \in L_0^{\perp}$ we have $0 = \langle f, g\rangle = \langle f, U_T g\rangle = \langle U_T^{-1} f, g\rangle$ for all $g \in L_0$. □

This is a Hilbert space argument, *i.e.*, it uses only the unitary nature of the operators, in fact, most of the argument is independent of unitarity and uses only that U_T does not expand norms, which helps prove the corresponding result in the noninvertible case.

Since this is an averaging argument (Section 1.4b) it works for amenable groups. The averaging procedure of (1.1) gives $F_n(f) \to P_{\mathbb{I}}(f)$ in L^2 by the same argument: Show that if $f = g \circ \Phi^g - g$ then $F_n(f) \to 0$, and if $\langle f, g\rangle = 0$ for all $g \in L_0$ then $f \circ \Phi^g = f$.

b. The Birkhoff pointwise Ergodic Theorem.

THEOREM 3.5.2. *Let $T\colon (X, \mu) \to (X, \mu)$ be a measure-preserving transformation of a probability space, $\varphi \in L^1(X, \mu)$. Then*

$$\varphi_T := \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi \circ T^k = \varphi_{\mathbb{I}} := P_{\mathbb{I}}(\varphi) \tag{3.1}$$

for μ-a.e. $x \in X$.

PROOF. Let $f := \varphi - \varphi_{\mathbb{I}} - \epsilon \in L^1(\mu)$ and $F_n := \max_{k\le n} \sum_{i=0}^{k-1} f \circ T^i$. Then

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k \le \overline{\lim_{n\to\infty}} \frac{F_n}{n} \le 0 \text{ off } A := \{x \mid F_n(x) \to \infty\} \in \mathbb{I}, \tag{3.2}$$

but $F_{n+1} - F_n \circ T = f - \min(0, F_n \circ T) \downarrow f$ on A, so (Dominated Convergence) $0 \le \int_A (F_{n+1} - F_n)\, d\mu = \int_A (F_{n+1} - F_n \circ T)\, d\mu \to \int_A f\, d\mu = \int_A f_{\mathbb{I}}\, d\mu_{\restriction_{\mathbb{I}}}$ and $\mu(A) = 0$ since $f_{\mathbb{I}} = -\epsilon < 0$. Thus $\overline{\lim}_{n\to\infty} \sum_{k=0}^{n-1} (\varphi \circ T^k)/n - \varphi_{\mathbb{I}} - \epsilon \le 0$ μ-a.e. by (3.2). Replacing φ by $-\varphi$ gives $\underline{\lim}_{n\to\infty} \sum_{k=0}^{n-1} \varphi \circ T^k/n \ge \varphi_{\mathbb{I}} - \epsilon$ μ-a.e. □

If T is invertible, the Birkhoff Ergodic Theorem applies to T^{-1} and implies a.e. convergence of negative time averages $\sum_{k=0}^{n-1} \varphi(T^{-k}(x))/n$ and hence also for the two-sided time average $\sum_{k=1-n}^{n-1} \varphi(T^k(x))/(2n-1)$.

There are generalizations of the Birkhoff Ergodic Theorem in several directions: There are results about actions of different groups on one hand and about various types of subsequences over which to average on the other. Furthermore it is possible to make some statements in situations that do not admit a finite invariant measure [**Ho, CO**].

c. Typical points and recurrence. The points from the set of full measure in the Birkhoff Ergodic Theorem 3.5.2 are said to behave *typically*. This is to some extent analogous to recurrence in topological dynamics: Consider a measurable set U and its characteristic function χ_U. If x is a point for which $(\chi_U)_T$ exists then this function is positive a.e. on U because otherwise the average of $(\chi_U)_T$ and χ_U over $(\chi_U)_T^{-1}(\{0\})$ disagree. This is a strengthening of the conclusion of the Poincaré Recurrence Theorem.

d. Orbit equivalence. We are now in a position to address the question of orbit equivalence. The ergodic decomposition is an invariant of orbit equivalence, and hence we may restrict attention to ergodic actions. For transformations preserving an ergodic finite measure the orbit equivalence problem is solved completely by a result of Dye:

THEOREM 3.5.3. *Any two ergodic measure-preserving transformations of a Lebesgue probability space are orbit equivalent* [**Dy**].

An orbit equivalence can be constructed by inductive application of the Rokhlin Lemma, Theorem 3.4.10

It turns out that amenability is the proper framework for treating this problem:

THEOREM 3.5.4. *A discrete group G is amenable if and only if every ergodic action of G preserving a finite measure is orbit equivalent to an ergodic $\mathbb{Z}$-action* [**CFW**].

To see this connection note that every group action on a space X generates an equivalence relation (whose classes are the orbits). Ergodicity of an action renders the measurable hull of this partition trivial. The fact that the action is measure-preserving can be seen from an intrinsic property (type II_1) of the equivalence relation [**FM**]. Finally, there is a certain property of countable equivalence relations, called hyperfiniteness, which can be paraphrased by saying that the relation is well-approximated by finite equivalence relations (*i.e.,* with finite classes). By the Rokhlin Lemma this property holds for the equivalence relations generated by measure-preserving $\mathbb{Z}$-actions. If the group is amenable then the resulting equivalence relation is still hyperfinite due to a partial extension of the Rokhlin Lemma [**OW**]. For actions of amenable groups with infinite invariant measure the situation is parallel and deals with hyperfinite equivalence relations of type II_∞.

Interesting aspects of the equivalence question arise for actions not preserving a measure or actions of groups that are not amenable. In the former case one deals with hyperfinite equivalence relations that do not have an invariant measure (type III). They allow a complete and nontrivial classification due to Krieger and Connes [**Kr1, Cn**]. The invariants are the so-called ratio set, which distinguishes types III_1, III_α $(0 < \alpha < 1)$ and III_0, and in the latter case the isomorphism type of a measure-preserving flow.

For measure-preserving actions of nonamenable groups, the equivalence relations are still of type II_1 but no longer hyperfinite. Many nonequivalent classes are known, and a

complete classification does not seem to be feasible. Finally, the classification question for non-hyperfinite type III relations is wide open. See the survey [**S-FK**] for a treatment of some questions related to orbit equivalence of nonamenable groups, and also [**Z, Fm**].

6. Quantitative recurrence and principal spectral properties

a. Ergodicity. Ergodicity is a property previously encountered in regard to irreducibility, but due to ergodic theorems it also provides for quantitative recurrence. Here are several equivalent characterizations of ergodicity for transformations with finite invariant measure.

(1) $TA = A \implies \mu(A)\mu(X \smallsetminus A) = 0$
(2) If $f\colon X \to \mathbb{R}$ is measurable (or $f \in L^p$) and $f \circ T = f$ then $f \equiv \text{const}$
(3) $\sum_{k=0}^{n-1} f \circ T^k/n \to \int f$ a.e. for all $f \in L^1$
(4) $\limsup_{n\to\infty} \sum_{k=0}^{n-1} f \circ T^k/n > 0$ for all $f \geq 0$, $f \not\equiv 0$.
(5) $\sum_{k=0}^{n-1} f \circ T^k/n \to \int f$ a.e. for all f from a family of functions whose linear hull is dense in L^1.
(6) $\sum_{k=0}^{n-1} \langle f \circ T^k, g\rangle/n \to \langle f, 1\rangle\langle 1, g\rangle$ for all $f, g \in L^2$ (or a family whose linear hull is dense in L^2)
(7) The only invariant measure absolutely continuous with respect to μ is a multiple of μ.

Furthermore, it suffices to check any of the conditions 3–6 for a dense family of functions or sets.

That these characterizations are equivalent is obvious in some cases and otherwise a consequence of the Birkhoff Ergodic Theorem, which is the device that establishes the connection between the qualitative description in terms of irreducibility and the quantitative recurrence descriptions.

b. Speed of convergence in ergodic theorems. For a measure-preserving transformation T and a given L^1 function f it is natural to ask how fast the Birkhoff averages $B_n(f)$ converge to $\int f d\mu$. It turns out that no *uniform* estimate is possible for *any* T, even for bounded fuctions; convergence can be arbitrary slow. On the other hand for any function of the form $U_T f - f + \text{const}$ convergence is as fast as it can be and such functions are dense. Thus no property reflecting the speed of convergence in the ergodic theorem is invariant under metric isomorphism.

On the other hand, if an extra structure is present one may ask for a speed of convergence for a a certain family of fucntions related to this structure, such as differentiable, Hölder or of exponential type in symbolic dynamics (Section 2.6a).

c. Correlation coefficients and spectral measures. Let $T\colon (X, \mu) \to (X, \mu)$ be a measure-preserving transformation. For a function $f \in L^2$, the scalar products

$$\langle f \circ T^n, f\rangle = \langle U_T^{-n} f, f\rangle$$

are called the *correlation coefficients* of the function f. The correlation coefficients are the Fourier coefficients of a certain measure λ_f on the unit circle which is called *the spectral measure* of f:

$$\langle U_T^{-n} f, f\rangle = \int_{S^1} x^n d\lambda_f.$$

All spectral measures are subordinate to a measure of the maximal spectral type and in fact belong to that type for a dense set of f.

Since $\langle U_T^{-n} f, f\rangle = \overline{\langle U_T^{-n} \bar{f}, \bar{f}\rangle} = \int_{S^1} x^{-n} d\lambda_f = \int_{S^1} x^n d(I_*\lambda_f)$ where $I(x) = \bar{x}$ we obtain

PROPOSITION 3.6.1. *The maximal spectral type of the operator induced by an ergodic measure-preserving transformation is symmetric with respect to complex conjugation* i.e., *the reflection of the unit circle in the real axis.*

For a flow the correlation coefficients are defined similarly, the spectral measure of a function is defined on the real line, and the maximal spectral type is symmetric with respect to the origin.

d. Eigenfunctions. As we pointed out, ergodicity is a spectral invariant: It is equivalent to 1 being a simple 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, eigenfunctions and eigenvalues for an ergodic transformation form a group. Eigenfunctions determine a canonical factor of T, the maximal factor with discrete spectrum [**W1, CFS**], which is usually called the *Kronecker factor*.

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. Hence we obtain the following general spectral property of measure-preserving transformations.

THEOREM 3.6.2. *The maximal spectral type of the operator U_T induced by an ergodic measure-preserving transformation T is invariant under multiplication by any eigenvalue.*

THEOREM 3.6.3 (Discrete Spectrum Theorem [**CFS**]). *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 $\Gamma < S^1$ of eigenvalues: A transformation 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.*

SKETCH OF PROOF. Let $T\colon (X,\mu) \to (X,\mu)$ 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 $V\colon L^2(X,\mu) \to L^2(\Gamma^*,\lambda)$, which is multiplicative on the eigenfunctions by (3.1). Their finite linear combinations are dense in $L^2(X,\mu)$, so V is generated by a measure-preserving invertible transformation $H\colon (X,\mu) \to (\Gamma^*,\lambda)$. 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 $HTH^{-1} = L_{s_0}$. □

PROPOSITION 3.6.4. *A measure-preserving transformation T (not necessarily ergodic) has pure point spectrum if and only if the closure G of the sequence $(U_{T^n})_{n\in\mathbb{Z}}$ is compact in the strong operator topology. In this case, G is abelian and T is metrically isomorphic to the translation on G induced by multiplication by U_T, with Haar measure.*

SKETCH OF PROOF. The "if" part follows from Theorem 3.6.3. The "only if" direction is a fact about unitary operators and can be easily deduced from the spectral theorem (see *e.g.,* [**S-KT, Pa2**]). □

Dynamical systems with pure point spectrum include translations and linear flow on tori. These are basic building blocks for completely integrable Hamiltonian systems and principal models for elliptic behavior in dynamics.

e. Rigidity and good periodic approximation. Rigidity is a recurrence property, expressed in terms of unitary operators and hence spectral and weaker than the one given by pure point spectrum.

A transformation is said to be *rigid* if the identity is an accumulation point of $(U_T^n)_{n\in\mathbb{N}}$ in the strong operator topology.

Recall that a sequence of measurable partitions $\{\xi_n\}$ is said to be *exhaustive* if for any measurable set A and any $\epsilon > 0$ one can find $N \in \mathbb{N}$ such that for any $n \geq N$ unions from the partition ξ_n approximate A within ϵ in the symmetric difference metric (Section 3.4h).

DEFINITION 3.6.5. A measure-preserving transformation $T\colon (X,\mu) \to (X,\mu)$ is said to allow a *good periodic approximation* if there exists a exhaustive sequence of finite partitions of the form $\xi_n = \{C_{1,n}, \dots, C_{q_n,n}, d_n\}$ with $\mu(C_{1,n}) = \cdots = \mu(C_{q_n,n})$ such that

$$\sum_{i=1}^{q_n} \mu(T(C_{i,n}) \triangle C_{i+1,n}) = o(q_n^{-1}),$$

where we set $C_{q_n+1,n} = C_{1,n}$.

One may specify the *speed* of approximation by replacing $o(q_n^{-1})$ by a specific function [**CFS, K6, KS**].

Good periodic approximation implies rigidity but it is in fact stronger. For example it also implies simple spectrum, which rigidity does not. It is not known whether it is a spectral property and quite likely it is not.

f. Weak mixing. A measure-preserving action is said to be *weakly mixing* if the orbit under the unitary operators U_g ($g \in G$) for a function f orthogonal to the constant functions never has compact closure. In particular, there are no eigenfunctions other than constants, which implies ergodicity.

Equivalently, weak mixing means that the unitary representation induced by the action does not have any finite-dimensional representation as a direct summand. If the group is abelian then all irreducible representations are one-dimensional and this is equivalent to the

absence of eigenfunctions or, equivalently, for the maximal spectral type being a nonatomic measure.

For a measure-preserving transformation T the following are equivalent [**H1, S-KT**]:

(1) T is weakly mixing.
(2) For any two measurable sets A, B there exists a sequence $n_k \to \infty$ such that $\mu(T^{-n_k}(A) \cap B) \to \mu(A) \cdot \mu(B)$ as $k \to \infty$.
(3) $T \times T$ is ergodic.
(4) $T \times S$ is ergodic for any ergodic S.
(5) $\sum_{k=0}^{n-1} |\langle f \circ T^k, g\rangle - \langle f, 1\rangle\langle 1, g\rangle|/n \to 0$ for all $f, g \in L^2$.
(6) Any joining of T with a measure-preserving transformation S with pure point spectrum is a product.

As in Section 3.6a it suffices to check the convergence conditions in 2 and 5 for a dense collection of functions or sets.

Let us note the following nontrivial fact:

THEOREM 3.6.6 ([**F1**, Theorem 4.11]). *A weakly mixing measure-preserving transformation T is* multiply weakly mixing, i.e., *for any finite collection $A_0, \dots, A_m$ of measurable sets there is a sequence $n_k \to \infty$ such that $\mu(A_0 \cap T^{-n_k} A_1 \cap T^{-2n_k} A_2 \cap \dots \cap T^{-mn_k} A_m) \to \mu(A_0)\mu(A_1) \dots \mu(A_m)$.*

g. Mild mixing. To define the next stronger mixing notion we say that a function $f \in L^2$ is *rigid* if it is recurrent under the action of U_T on L^2. This holds for eigenfunctions and their linear combinations, but there exist weakly mixing transformations for which some (or all) functions are rigid. Rigidity is a spectral property: $U_T^{n_k} f \to f \iff \langle U_T^{k_k} f, f\rangle \to \|f\|^2$. In fact, the spectral type [**S-KT**] of a rigid function is singular [**KS, CFS**]. A transformation is rigid (Section 3.6e) if and only if all functions are rigid. Hence rigid transformations have singular maximal spectral type.

We say that a measure-preserving transformation T is *mildly mixing* if it has no nonconstant rigid functions.

THEOREM 3.6.7. *T is mildly mixing if and only if $T \times S$ is ergodic whenever S is an ergodic transformation preserving a finite or infinite measure* [**FW**].

Further characterizations of mild mixing are in [**S-B**].

h. Mixing. A measure-preserving transformation $T\colon (X, \mu) \to (X, \mu)$ is said to be *mixing* if for any two measurable sets A, B

$$\mu(T^{-n}(A) \cap B) \to \mu(A) \cdot \mu(B) \quad \text{as} \quad n \to \infty.$$

Again, it suffices to verify this for a dense collection of sets.

In terms of functions mixing means that for any function with zero average the correlation coefficients tend to zero as $n \to \pm\infty$. Thus mixing is a property of the maximal spectral type. In particular, any factor of a mixing map is mixing.

The following characterization allows to deduce mixing from a seemingly weaker property [**KH**, Proposition 20.3.6].

PROPOSITION 3.6.8. *Suppose T is a measure-preserving transformation such that*

$$\limsup_{n\to\infty} \frac{\mu(T^{-n}(A) \cap B)}{\mu(A) \cdot \mu(B)} < c.$$

for some constant c and for any two sets A, B of positive measure from a dense collection of sets. Then T is mixing.

Note that a map with a mixing invariant Borel probability measure is topologically mixing on the support.

i. Multiple mixing. Analogously to multiple weak mixing one can define a notion of multiple mixing in two ways, either via proportional returns as above, or via independent returns. It is not known whether mixing implies either of the resulting notions. This is one of the oldest unsolved problems in ergodic theory. While there is no compelling structural reason to believe that the answers are positive (other than the validity of the corresponding fact for weak mixing), in most natural classes, such as homogeneous and affine systems, mixing is equivalent to mixing of all orders [**S-KSS**]. On the other hand, there is a number of deep results excluding what may look like natural candidates for counterexamples. For rank one maps mixing implies mixing of all orders [**S-KT, Kl**]. Kalikow's result was extended to a wider class of finite rank transformations by Ryzhikov [**Ry**]. Host proved that a mixing transformation with singular spectrum is mixing of all orders [**Hos, Na**] and Thouvenot showed that some of the mixing flows on higher genus surface are also mixing of all orders (unpublished). Mozes achieved the same for actions of Lie groups with finite center whose adjoint representation is a proper map [**Mz**].

j. Absolutely continuous spectrum. Another property stronger than mixing is to have the maximal spectral type [**S-KT**] absolutely continuous. This is stronger because the Fourier coefficients of any absolutely continuous measure on the circle converge to zero as $n \to \pm\infty$ (Riemann–Lebesgue Lemma). A quite widespread behavior of that type is *countable Lebesgue spectrum*: the maximum spectral type is Lebesgue and the multiplicity is ∞ (Section 3.4q). In more pedestrian terms, countable Lebesgue spectrum means existence of an orthonormal basis $\{e_{m,n}\}_{m\in\mathbb{N},n\in\mathbb{Z}}$ in $L^2_0(X,\mu)$ such that $U_T e_{m,n} = e_{m,n+1}$.

There are some rather special (although not unnatural) examples, whose spectrum has a Lebesgue component of finite multiplicity (at least two, see [**MN**]). It is quite remarkable that it is unknown whether one can have simple Lebesgue spectrum or even a spectrum that consists of a simple Lebesgue and singular part. Together with the multiple mixing proplem the simple Lebesgue spectrum problem is one of the oldest open questions in ergodic theory.

k. The K-property. (See also [**Rk2**].) An automorphism T is said to have the *K-property* (after Kolmogorov) or simply is a *a K-automorphism* , if

$$\lim_{n\to\infty}\lim_{N\to\infty}\sup\{|\mu(A\cap C)-\mu(A)\mu(C)| \mid C\in\mathcal{A}(\bigvee_{i=n}^{N} T^i\xi)\}=0$$

for any finite partition ξ and every measurable set A.

The most effective criterion for the K-property is existence of a σ-algebra of measurable sets $\mathcal{A}$ such that $\mathcal{A}\subset T\mathcal{A}$, $\bigcup_{n=0}^{\infty} T^n\mathcal{A}$ is dense in the σ-algebra $\mathcal{B}$ of all measurable sets, and $\bigcap_{n=0}^{\infty} T^{-n}\mathcal{A} = \mathcal{N}$, the trivial subalgebra of null sets and their complements. Thus any shift with a Bernoulli (Section 3.3e) or transitive Markov (Section 3.3f) measure is a K-automorphism.

But the K-property can also be characterized in terms of entropy (Section 3.7j). Any factor of a K-automorphism is again a K-automorphism.

THEOREM 3.6.9. *For any K-automorphism T the operator U_T has countable Lebesgue spectrum in the space $L^2_0(X,\mu)$.*

SKETCH OF PROOF. Let L be the orthogonal complement to $L^2(\mathcal{A},\mu)$ in $U_T L^2(\mathcal{A},\mu)$. The space L is infinite-dimensional, orthogonal to all its images and the sum of these images generates $L^2_0(X,\mu)$. □

As will be seen later (Section 3.7j, Section 8.3b) not every automorphism with countable Lebesgue spectrum is K. Thus the K-property is not a spectral one.

l. Decay of correlations. The speed of mixing can be measured by the decay of correlations. Lower bounds on these quantities (*i.e.*, limitations on how fast correlations may decay for particular functions) give spectral invariants. However, there are always L^1 functions with arbitrarily slow decay of correlations, which means that this notion is useful only when one restricts to subfamilies of L^1 (that exclude functions with pathologically slow decay of correlations). The choice of such families is determined by context and typically consists, for example, of Hölder continuous or smooth functions in the presence of a differentiable structure. See[**S-C**] and [**S-P**] for the principal results in this direction in the case of symbolic systems and systems with hyperbolic behavior.

Evidently the resulting decay rates are not measure-theoretic invariants, because the respective subfamilies may not be equivariant under measure-theoretic equivalence. This situation is similar to the question of the speed of convergence of Birkhoff averages discussed in Section 3.6b. In fact, the two types of questions are often related.

For example, the maximal spectral type is Lebesgue if and only if the correlations vanish for a dense family of functions. For Bernoulli measures (Section 3.3e) this happens for all functions which depend only on finitely many coordinates.

7. Entropy

Now we introduce the entropy of a measure-preserving transformation, often called the *Kolmogorov* (or sometimes *Kolmogorov–Sinai*) *entropy* [**Pa1, Rk2**], [**KH**, Section 4.3]. The comments about the central role of topological entropy as a growth invariant for topological dynamical systems apply with even greater force to the role of entropy of a transformation with respect to an invariant measure. This justifies a comparatively detailed treatment of entropy in a general survey.

a. Entropy and conditional entropy of partitions. We need to start with some elementary preparations. The *entropy* of a finite or countable measurable partition ξ is given by

$$H(\xi) := H_\mu(\xi) := -\sum_{C\in\xi} \mu(C)\log\mu(C) \geq 0,$$

where $0\log 0 := 0$. For countable ξ the entropy may be infinite. In most cases we suppress the dependence of entropy on the measure.

Let $\xi(x)$ be the element of ξ that contains x and

$$I_\xi\colon X\to\mathbb{R},\ I_\xi(x) := -\log\mu(\xi(x)), \tag{3.1}$$

the *information function* of ξ. Then

$$H_\mu(\xi) = \int_X I_\xi \, d\mu. \tag{3.2}$$

This illuminates and makes natural the following notion of conditional entropy of a partition with respect to another partition, which plays the central role in the entropy theory for measure-preserving transformations.

DEFINITION 3.7.1. Let ξ, η be measurable partitions of (X, μ). The *conditional entropy* of ξ with respect to η is $H(\xi \mid \eta) := -\sum_{D\in\eta} \mu(D) \sum_{C\in\xi} \mu(C \mid D) \log \mu(C \mid D)$, where $\mu(A \mid B) := \mu(A \cap B)/\mu(B)$.

REMARK. If $\nu = \{X\}$ then $H(\xi) = H(\xi \mid \nu)$. If ξ_D is the partition of D into the intersections $D \cap C$, $C \in \xi$, then $H(\xi \mid \eta) = \sum_{D\in\eta} \mu(D) H_{\mu_D}(\xi_D)$. Similarly to (3.2) one gets

$$H(\xi \mid \eta) = \int_X I_{\xi,\eta} \, d\mu,$$

where $I_{\xi,\eta}$ is the *conditional information function* defined by

$$I_{\xi,\eta}(x) = -\log \mu(C_\xi(x) \mid C_\eta(x)).$$

Formula (3.2) allows us to define conditional entropy even in some cases when ξ is a continuous partition.

Misiurewicz [**Mi2**] developed a concept of conditional topological entropy which however is not nearly as useful as the measurable version.

b. Basic properties of entropy and conditional entropy of a partition. (See also [**KH**, Section 4.3].) Let $(X, \mathcal{B}, \mu)$ be a probability space and let ξ, η, ζ be finite or countable measurable partitions of X. Then:

(1) $0 < -\log(\sup_{C\in\xi} \mu(C)) \le H(\xi) \le \log \operatorname{card} \xi$. If ξ is finite then $H(\xi) = \log \operatorname{card} \xi$ if and only if all elements of ξ have equal measure.
(2) $0 \le H(\xi \mid \eta) \le H(\xi)$, $H(\xi \mid \eta) = H(\xi)$ if and only if ξ and η are independent. $H(\xi \mid \eta) = 0$ if and only if $\xi \le \eta$. If $\zeta \ge \eta$ then $H(\xi \mid \zeta) \le H(\xi \mid \eta)$.
(3) $H(\xi \vee \eta \mid \zeta) = H(\xi \mid \zeta) + H(\eta \mid \xi \vee \zeta)$. In particular, for $\zeta = \{X\}$ we obtain $H(\xi \vee \eta) = H(\xi) + H(\eta \mid \xi)$.
(4) $H(\xi \vee \eta \mid \zeta) \le H(\xi \mid \zeta) + H(\eta \mid \zeta)$. In particular $H(\xi \vee \eta) \le H(\xi) + H(\eta)$.
(5) $H(\xi \mid \eta) + H(\eta \mid \zeta) \ge H(\xi \mid \zeta)$.
(6) If λ is another measure on X, ξ a measurable partition for both μ and λ and $p \in [0, 1]$ then

$$p\, H_\mu(\xi) + (1-p) H_\lambda(\xi) \le H_{p\mu+(1-p)\lambda}(\xi).$$

(7) On the set of (all equivalence classes mod 0 of) measurable partitions with finite entropy

$$d_R(\xi, \eta) := H(\xi \mid \eta) + H(\eta \mid \xi) \tag{3.3}$$

defines a metric, called the *Rokhlin metric*.

c. Entropy of a transformation relative to a partition. For a measurable partition ξ and a measure-preserving (not necessarily invertible) transformation T we define the *joint partition* by

$$\xi_{-n}^T := \bigvee_{i=1}^{n} T^{1-i}(\xi).$$

From now on, unless stated otherwise, we assume that ξ is a finite or countable measurable partition with finite entropy. Since $H(\xi_{-n-m}^T) \le H(\xi_{-n}^T) + H(\xi_{-m}^T)$ by Section 3.7b4, $\lim_{n\to\infty} H(\xi_{-n}^T)/n$ exists.

DEFINITION 3.7.2. $h(T,\xi) := h_\mu(T,\xi) := \lim_{n\to\infty} H(\xi_{-n}^T)/n$ is called the *metric entropy* of the transformation T *relative* to the partition ξ.

The definition of the entropy of T (Section 3.7f) is immediate from here, but some properties of the entropy relative to a partition are worth exploring now.

The following proposition gives an alternative proof of existence of the limit $h(T,\xi)$ as well as another expression for it.

PROPOSITION 3.7.3. $H(\xi \mid T^{-1}(\xi_{-n}^T)) \downarrow h(T,\xi)$.

SKETCH OF PROOF. Using Section 3.7b4. we obtain

$$\begin{aligned} H(\xi_{-n}^T) &= H(T^{-1}(\xi_{-n+1}^T)) + H(\xi \mid T^{-1}(\xi_{-n+1}^T)) \\ &= H(\xi_{-n+1}^T) + H(\xi \mid T^{-1}(\xi_{-n+1}^T)) = H(\xi_0^T) + \sum_{k=0}^{n-1} H(\xi \mid T^{-1}(\xi_{-k}^T)). \end{aligned}$$

Since $T^{-1}(\xi_{-k}^T)$ is refined as k increases, $b_n := H(\xi \mid T^{-1}(\xi_{-n}^T))$ is nonincreasing by Section 3.7b2. Thus $h_\mu(T,\xi) = \lim_{n\to\infty} \sum_{k=0}^{n-1} b_k/n = \lim_{n\to\infty} b_n$. □

In fact, we have

PROPOSITION 3.7.4. $h(T,\xi) = H(\xi \mid \xi_{-\infty}^T)$.

d. Properties of entropy with respect to a partition. (See also [**KH**, Section 4.3].)

(1) $0 \le \overline{\lim}_{n\to\infty}(-1/n)\log(\sup_{c\in\xi_{-n}^T} \mu(C)) \le h(T,\xi) \le H(\xi)$.
(2) $h(T,\xi\vee\eta) \le h(T,\xi) + h(T,\eta)$.
(3) $h(T,\eta) \le h(T,\xi) + H(\eta \mid \xi)$. In particular, if $\xi \le \eta$ then $h(T,\xi) \le h(T,\eta)$.
(4) $|h(T,\xi) - h(T,\eta)| \le H(\xi \mid \eta) + H(\eta \mid \xi)$ (the *Rokhlin inequality*).
(5) $h(T,T^{-1}(\xi)) = h(T,\xi)$ and if T is invertible then $h(T,\xi) = h(T,T(\xi))$.
(6) $h(T,\xi) = h(T,\bigvee_{i=0}^k T^{-i}(\xi))$ for $k \in \mathbb{N}$ and if T is invertible then $h(T,\xi) = h(T,\bigvee_{i=-k}^k T^i(\xi))$ for $k \in \mathbb{N}$.
(7) If ν is another measure and $p \in [0,1]$ then

$$p h_\mu(T,\xi) + (1-p) h_\nu(T,\xi) \le h_{p\mu+(1-p)\nu}(T,\xi).$$

REMARK. Property 4 means that $h(T,\cdot)$ is a Lipschitz function with Lipschitz constant 1 on the space of partitions with finite entropy provided with the Rokhlin metric (3.3).

e. The Shannon–McMillan–Breiman Theorem. The entropy of a transformation relative to a partition measures the exponential rate of the average size of the elements of iterated partitions. In other words, the average size of the elements shrinks exponentially with a rate given by entropy. The following result shows that for ergodic transformations deviations from this average size are rather rare, or, conversely, that on a set of arbitrarily large measure all partition elements have close to average size.

THEOREM 3.7.5. [**S-T**] *Let T be an ergodic measure-preserving transformation of X and ξ a partition. Denote by $\xi_{-n}^T(x)$ the element of ξ_{-n}^T containing x. Then $-\frac{1}{n}\log\mu(\xi_{-n}^T(x)) \to h_\mu(T,\xi)$ almost everywhere.*

Section 2.5i and Section 3.7l suggest modifying the definition of entropy by introducing averaging into the process. Using the Shannon–McMillan–Breiman Theorem one can show that for regular entropy this makes no difference. (This is proved in [**S-T**].)

There is a corresponding statement for nonergodic transformations as well. In this case the limit exists almost everywhere and coincides with the entropy with respect to the corresponding conditional mesure on the ergodic component.

Existence of the limit in Shannon–McMillan–Breiman theorem is a fairly straightforward consequence of the Martingale Theorem for stationary random processes in probability theory [**Bi**]. Identification of the limit with entropy in the ergodic case is immediate by comparing averages.

f. Entropy of a measure-preserving transformation. The *entropy* of T with respect to μ (or the entropy of μ) is

$$h(T) := h_\mu(T) := \sup\{h_\mu(T,\xi) \mid \xi \text{ is a measurable partition with } H(\xi) < \infty\}.$$

Entropy is invariant under metric isomorphism. This definition is more constructive than it seems. In many cases $h_\mu(T) = h_\mu(T,\xi)$ for an appropriately chosen ξ (Corollary 3.7.10).

Recalling the definition of the partition entropy through the information function (3.1)–(3.2) we can interpret the entropy $h_\mu(T,\xi)$ as the average amount of information provided by the knowledge of the "present state" in addition to the knowledge of an arbitrarily long past. Thus, a system with zero entropy can be viewed as strongly deterministic in the sense that an approximate knowledge of the entire past (*i.e.*, the past itinerary with respect to a finite partition) precisely determines the future itinerary. Obviously, it is sufficient to know only the arbitrarily distant past. The K-property which can be characterized using entropy describes the situation where the more and more distant past carries less and less information about the present and future and hence the arbitrarily distant past carries no information at all.

g. Examples. (See also [**KH**, Section 4.4].)

EXAMPLE 3.7.6. For the N-shift σ_N with the Bernoulli measure μ_p (Section 3.3e) one has

$$h_{\mu_p}(\sigma_N) = -\sum_{i=0}^{N-1} p_i \log p_i.$$

EXAMPLE 3.7.7. For the Markov measure $\mu_{\Pi,p}$ (Section 3.3f) one has

$$h_{\mu_{\Pi,p}}(\sigma_N) = -\sum_{i,j=0}^{N-1} p_j \pi_{ij} \log \pi_{ij}.$$

EXAMPLE 3.7.8. Let A be an invertible integer $m \times m$ matrix, $\lambda_1, \ldots, \lambda_m$ the eigenvalues of A counted with multiplicities, F_A the corresponding endomorphism of the torus $\mathbb{T}^m$, μ Lebesgue measure on $\mathbb{T}^m$. Then

$$h_\mu(F_A) = \sum_{i:|\lambda_i|>1} \log |\lambda_i|.$$

In particular, for the linear expanding map $E_m : x \mapsto mx \pmod 1$ ($m \in \mathbb{Z}$, $|m| \geq 2$) of the circle

$$h_\mu(E_m) = \log |m|.$$

h. Calculation of entropy. We present criteria for calculating the entropy of a measure-preserving transformation.

THEOREM 3.7.9. $h_\mu(T) = \sup_{\xi \in \Xi} h_\mu(T, \xi)$ *for any sufficient family* Ξ *of partitions (Section 3.4h).*

SKETCH OF PROOF. Let η be an arbitrary measurable partition of X with $H_\mu(\eta) < \infty$. Fix $\epsilon > 0$, $\xi \in \Xi$ and $k \in \mathbb{N}$ such that $d_R(\eta, \zeta) = H(\eta \mid \zeta) + H(\zeta \mid \eta) < \epsilon$ for some partition $\zeta \leq \bigvee_{i=0}^k T^{-i}(\xi)$ if T is noninvertible and $\zeta \leq \bigvee_{i=-k}^k T^i(\xi)$ if T is invertible. By Section 3.7d4,3,6 we obtain in the noninvertible case $h_\mu(T, \eta) \leq h_\mu(T, \zeta) + \epsilon \leq h_\mu(T, \bigvee_{i=0}^k T^{-i}(\xi)) + \epsilon = h_\mu(T, \xi) + \epsilon$ (and similarly if T is invertible). Since ϵ is arbitrary, the statement follows. □

The following corollary is the best-known and simplest criterion for calculating entropy.

COROLLARY 3.7.10. *If* ξ *is a generator (Section 3.4h) for* T *then* $h_\mu(T) = h_\mu(T, \xi)$.

At this point it is useful to stress the difference between the invertible and the noninvertible case. Let us call a partition ξ a *one-sided generator* for an invertible measure-preserving transformation T if partitions subordinate to partitions of the form $\bigvee_{i=0}^k T^{-i}(\xi)$ ($k \in \mathbb{N}$) are dense in the metric d_R.

PROPOSITION 3.7.11. *If an invertible measure-preserving transformation possesses a one-sided generator then* $h_\mu(T) = 0$.

PROOF. A one-sided generator ξ is obviously a generator for T so by Corollary 3.7.10 it suffices to check $h_\mu(T, \xi) = 0$. By Section 3.7d5 this is equivalent to $h_\mu(T, T\xi) = 0$. Suppose ξ is a one-sided generator and $\epsilon > 0$. Then take $k \in \mathbb{N}$ and $\zeta \leq \bigvee_{i=0}^k T^{-i}(\xi)$ such that $d(T(\xi), \zeta) < \epsilon$ and hence $H(T(\xi) \mid \bigvee_{i=0}^k T^{-i}(\xi)) \leq H(T(\xi) \mid \zeta) < \epsilon$. Thus since the sequence $a_n := H(T(\xi) \mid \bigvee_{i=0}^n T^{-i}(\xi))$ is nonincreasing, $h_\mu(T, T(\xi)) < \epsilon$ by Proposition 3.7.3. Since ϵ is arbitrary, we have $h(T, T(\xi)) = 0$. □

i. Properties of entropy. (See also [**KH**, Section 4.3].)

(1) If $S\colon (Y,\nu)\to(Y,\nu)$ is a factor (see Section 3.4a) of $T\colon (X,\mu)\to(X,\mu)$ then $h_\nu(S)\le h_\mu(T)$.
(2) If A is invariant for T and $\mu(A)>0$ then $h_\mu(T)=\mu(A)h_{\mu_A}(T)+\mu(X\smallsetminus A)h_{\mu_{X\smallsetminus A}}(T)$.
(3) If μ, ν are invariant probability measures for T and $p\in[0,1]$ then $h_{p\mu+(1-p)\nu}(T)=ph_\mu(T)+(1-p)h_\nu(T)$.
(4) $h_\mu(T^k)=kh_\mu(T)$ for any $k\in\mathbb{N}$. If T is invertible then $h_\mu(T^{-1})=h_\mu(T)$ and hence $h_\mu(T^k)=|k|h_\mu(T)$ for any $k\in\mathbb{Z}$.
(5) $h_{\mu\times\lambda}(T\times S)=h_\mu(T)+h_\lambda(S)$.
(6) If Φ^t is a flow then $h(\Phi^t)=|t|h(\Phi^1)$. This motivates defining the entropy of a flow as that of its time one map.
(7) Suppose T is an ergodic measure-preserving transformation of (X,μ). For a measurable set $A\subset X$ denote by T_A the induced map. Then $h_{\mu_A}(T_A)=h_\mu(T)/\mu(A)$.
(8) If $0\le\varphi\in L^1(X,\mu)$ and T^t_φ is the special flow over T under the function φ then $h(T^t_\varphi)=h(T)/\int\varphi$ (the Abramov formula).
(9) Consider an ergodic skew product $S\colon (X\times Y,\mu\times\nu)\to(X\times Y,\mu\times\nu)$, $S(x,y)=(T(x),S_x(y))$. Pick a finite partition η of Y and consider $\xi^x_n:=\eta\vee S_x\eta\vee S_{T(x)}S_x\eta\vee\cdots\vee S_{T^{n-1}(x)}\dots S_x\eta$. Then

$$h_x(S,\eta):=\lim\frac{H_\nu(\eta^x_n)}{n}$$

exists for almost every $x\in X$ and is independent of x. It is called the *relative entropy* of η and the supremum $h^*_\nu(S)$ over η is called the *relative entropy*.
(10) $h_{\mu\times\nu}(S)=h_\mu(T)+h^*_\nu(S)$
(11) If $(\eta(T),\mathcal{P}(\eta(T)),\nu_\eta)$ (Theorem 3.2.1) is the ergodic decomposition (Section 3.4f) of (T,μ) then $h_\mu(T)=\int_{\eta(T)}h_{\mu_C}(T)\,d\nu_\eta(C)$, where μ_C is as in Theorem 3.2.2.

As corollaries of 7. and 8. correspondingly one sees that Kakutani equivalence (Section 3.4e, Section 3.4p) for both maps and flows preserves the property of entropy to be zero, a finite positive number, or infinity.

j. Pinsker algebra, K-property and entropy. [**S-T**] Note that the join of two partitions with respect to which a measure preserving transformation T has zero entropy again is a zero entropy partition. Therefore, one can define the *Pinsker algebra* $\pi(T)$ to be the maximal zero entropy partition for T.

THEOREM 3.7.12. *A measure-preserving transformation T has the K-property if and only if the Pinsker algebra is trivial.*

Another way to express this is that T is a K-automorphism if and only if it has *completely positive entropy*, *i.e.,* that it has positive entropy with respect to any nontrivial partition (*i.e.,* with more than one element of positive measure).

While the K-property is interesting for individual systems and looks fairly strong, it turns out that in regard to classification it imposes surprisingly few restrictions. Indeed, any ergodic positive-entropy transformation has many subsets on which the induced map

has the K-property [**OS**]. Furthermore, every ergodic positive-entropy flow can be time-changed to a K-flow [**OS**]. Thus any Kakutani equivalence class of automorphisms or flows with positive (finite or infinite) entropy contains a K-system.

Entropy is not a spectral invariant. For example, all Bernoulli shifts (Section 3.3e) have countable Lebesgue spectrum in the orthogonal complement to constants but may have an arbitrary positive value of entropy. Furthermore, there are also zero entropy systems with countable Lebesgue spectrum. Nevertheless there is an important connection between entropy and spectral properties. It is proved similarly to Theorem 3.6.9.

THEOREM 3.7.13. *For a measure-preserving transformation with nonzero (*i.e., *positive or infinite) entropy there is a countable Lebesgue component in the spectrum,* i.e., *the maximal spectral type (Section 3.4q) dominates Lebesgue measure and the multiplicity function is infinite at the Lebesgue part.*

Thus there is a sufficient spectral criterion for the vanishing of entropy that can, in fact, be expressed in terms of the maximal spectral type (Section 3.4q).

COROLLARY 3.7.14. *If the maximal spectral type of the operator U_T is singular (pure point, continuous or mixed), then $h_\mu(T) = 0$.*

k. Noninvertible maps. (See also [**Pa2**].) Any measure-preserving transformation with zero entropy is invertible. This follows from Proposition 3.7.4

A noninvertible measure-preserving transformation that has positive entropy with respect to any nontrivial partition is said to be an *exact endomorphism*. A measure-preserving transformation is exact if and only if its natural extension (Section 3.4j) is a K-automorphism. Conversely for a K-automorphism T and any partition ξ the increasing partition $\xi^T_{-\infty}$ deternines a noninvertible factor which is an exact endomorphism.

In general a measure-preserving transformation T has unique *maximal invariant sub-algebra* $\mathcal{B}^I_T \supset \pi(T)$ such that the corresponding factor is invertible.

l. Slow metric entropy. With Example 2.6.9 in mind, where the power entropy differs for two very closely related systems, it is not surprising that definitions of a measure-theoretic counterpart of the a-entropy along the lines of Section 3.7 encounter serious difficulties for measure-preserving transformations T with zero entropy. The problem is that the sublinear asymptotic of the entropy of the joint partition $H(\xi^T_{-n})$ in Section 3.7c is very sensitive to the choice of partition and, in fact, for an aperiodic transformation *any* sublinear growth is exceeded by some ξ. This can be shown using the Rokhlin Lemma (Theorem 3.4.10). An approach to overcome this difficulty using an "economical" choice of partition was developed by Blume [**Blu**].

However, these problems disappear, when one develops a definition in analogy to Section 2.5i, using an averaged rather than maximum metric on orbit segments. For simplicity, we present this measure-theoretic construction only for the case of discrete time. Details are given in [**KT**]. See also [**Fe1**], where the corresponding concept is called *measure-theoretic complexity*. Let $T\colon (X,\mu) \to (X,\mu)$ be a measure-preserving transformation and $\xi = \{C_0, \dots, C_{N-1}\}$ a measurable partition. Define the "coding map" $\phi_{T,\xi}\colon X \to \Omega_N$ by $T^n(x) \in C_{(\phi_{T,\xi}(x))_n}$. Let $\phi^n_{T,\xi}$ be the projection of $\phi_{T,\xi}$ to the coordinates $0, \dots, n-1$ and $\lambda_n = (\phi^n_{T,\xi})_*\mu$.

The measure-theoretic counterpart of the metric $\eth_n^\Phi$ defined in (2.6) is the *Hamming metric*

$$d_n^H(\omega,\omega') = \frac{1}{n}\sum_{i=0}^{n-1}(1-\delta_{\omega_i\omega_i'})$$

on $\Omega_{N,n} := \{0,\dots,N-1\}^{\{0,\dots,n-1\}}$. Here we used the Kronecker delta. Denote by $S_\xi^H(T,\epsilon,n,\delta)$ the minimal number of d_n^H-ϵ-balls in $\Omega_{N,n}$ whose union has λ_n-measure at least $1-\delta$. This is a measure-theoretic counterpart to the numbers $S_d(T,\epsilon,n)$ used to define topological entropy in Section 2.5e. From here the construction works as expected. For a scale function $a(s,n)$ define the upper a-entropy of T with respect to ξ by

$$\overline{\mathrm{ent}}_a^\mu(T,\xi) := \lim_{\delta\to 0}\overline{\lim_{\epsilon\to 0}}\{s \mid \overline{\lim_{n\to\infty}} S_\xi^H(T,\epsilon,n,\delta)/a(s,n) > 0\}$$

and the upper a-entropy of T by $\overline{\mathrm{ent}}_a^\mu(T) := \sup_\xi \mathrm{ent}_a^\mu(T,\xi)$. Unlike the sublinear asymptotic of the entropy of a joint partition, the a-entropy can be calculated using any convenient partition. With the notations from Section 3.4h we then have [**KT**]

PROPOSITION 3.7.15. $\overline{\mathrm{ent}}_a^\mu(T) = \sup_m \overline{\mathrm{ent}}_a^\mu(T,\xi_m)$.

COROLLARY 3.7.16. *If ξ is a generator then* $\overline{\mathrm{ent}}_a^\mu(T) = \overline{\mathrm{ent}}_a^\mu(T,\xi)$.

These results work equally well for the "lower" counterparts, which use $\underline{\lim}_{n\to\infty}$ in the definition.

As before, the power entropy is defined as the a-entropy for $a(s,n) = n^s$. Power entropy is not inherited by ergodic components, for example in the map $S^1\times[0,1] \to S^1\times[0,1]$, $(x,y)\mapsto(x+y,y)$.

m. Entropy for amenable groups. Entropy theory has been developed for general discrete amenable groups [**S-T**]. This theory serves as the basis for an isomorphism theory for Bernoulli actions of such groups. Note, however, that the notion of a "past" is not available for general group actions. Therefore the characterization of entropy with respect to a partition given by Proposition 3.7.4 as well as the definition of the K-property from Section 3.6k do not have counterparts for arbitrary amenable groups. However for some groups in which an algebraic past (usually nonunique) exists, such as $\mathbb{Z}^k$, the remaining aspects of entropy theory also can be carried out [**PiS**]. On the other hand, the characterization of K-property via completely positive entropy (Theorem 3.7.12) can be used to extend the notion to greater generality.

n. Entropy for continuous groups. An effective way to define entropy for actions of amenable locally compact groups is to follow the definition based on the Hamming metric in the space of codes. Naturally the sums in the averages have to be replaced by integrals.

o. Entropy function. For actions of discrete of continuous groups "larger" than $\mathbb{Z}$ or $\mathbb{R}$ the entropy of an action whose individual elements have finite entropy is usually equal to zero. For those actions that appear naturally in differentiable and homogeneous dynamics, where individual elements have finite entropy, a natural growth invariant is the *entropy function* h, which is defined on the acting group G and associates to $g\in G$ the entropy of the corresponding element of the action. The entropy function is positive homogeneous, $h(g^n) = |n|h(g)$ (Section 3.7i, 4), but little else is known about it in full generality.

However for some important classes of actions (but not always) it is also *subadditive* for commuting maps: if $g_1 g_2 = g_2 g_1$ then $h(g_1 g_2) \leq h(g_1) + h(g_2)$. [**S-FK**].

CHAPTER 4

Invariant measures in topological dynamics

1. Introduction

a. Existence of invariant measures. In many situations both topology and invariant measures are present. Sometimes there is only one invariant measure (*e.g.,* minimal rotations, see Section 4.3b), in other examples there are many (shifts, Markov chains, toral automorphisms). Here we consider the question of existence of invariant measures for a broad class of topological dynamical systems. Since we are looking for finite invariant measures the analogous topological property of compactness is essential. Indeed, even such simple and natural transformations as translations on $\mathbb{R}$ have no finite invariant measure. Considering this question for group actions leads once again to amenability as the essential property—as was mentioned in Theorem 3.5.1, where amenability appears as the natural setting for the Mean Ergodic Theorem. On the other hand, in Section 3.3c we saw that there may be no invariant measure for actions of a nonamenable group.

b. Topological versus measure-theoretic properties. While in general measure-theoretic properties of invariant measures and topological behavior of orbits in their support may not correspond too closely (see, *e.g.,* Section 4.3i below), there is remarkable agreement in the case of entropy. Topological entropy and pressure are the supremum over invariant measures of their measurable counterparts, and in the expansive case the supremum is attained, *i.e.,* there exists an invariant measure of maximal entropy/pressure (Section 4.4d). For systems with the additional property of *specification* this measure is unique and has strong mixing properties. This latter case covers both transitive topological Markov chains and dynamical systems with hyperbolic behavior [**S-C**], [**KH**, Chapter 20].

c. Smooth measures. Smooth dynamical systems are a natural setting for some of these considerations because they come with a natural invariant measure class (absolutely continuous measures, including volume), and furthermore, some member of this class may be invariant. This is automatically the case for the important classes of Hamiltonian, contact and Lagrangian systems, which provided much of the motivation for the development of ergodic theory.

2. Existence of invariant measures

a. The Kryloff–Bogoliouboff Theorem.

THEOREM 4.2.1 ([**KB**]). *Any continuous map f of a metrizable compact space to itself has an invariant Borel probability measure.*

SKETCH OF PROOF. Beginning from the point mass δ_x construct the sequence of averages

$$\delta_{x,n} = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{f^n x}$$

and, using compactenss of the collection $\mathfrak{M}$ of Borel probability measures in the weak* topology, find an accumulation point of this sequence, which is f-invariant □

This argument can be generalized direcly to the actions of amenable groups using the Følner property (Section 1.4b). We give an alternative argument for the general amenable group case that emphasizes the Riesz Representation Theorem and works with positive linear functionals and the Kakutani–Markov fixed point property (Section 1.4d), which implies that the natural affine action on positive linear functionals has a fixed point. Furthermore, amenability is necessary.

THEOREM 4.2.2. *A locally compact group is amenable if and only if any continuous action on a compact metrizable space has an invariant Borel probability measure.*

SKETCH OF PROOF. We prove one direction. For any action Φ^g on X the affine action on positive linear functionals with norm 1 defined by $(\Phi_g^* F)(\varphi) := F(\varphi \circ \Phi^g)$ has a fixed point by Theorem 1.4.1. By the Riesz Representation Theorem this fixed point is given by a Borel probability measure, which is invariant by construction. □

b. Nonamenability. Example 2.4.4 (and Example 3.3.2) illustrates the necessity of amenability: Take $H = SL(2,\mathbb{R})$ and K the upper triangular subgroup. The factor can be identified with a projective line (*i.e.,* the circle) with action by projective transformations.

EXAMPLE 4.2.3. There is no finite invariant measure for $SL(2,\mathbb{Z})$-actions by projective transformations. For, any parabolic element, such as $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, has a unique invariant measure (Section 4.3b), the atom at its fixed point, and the fixed points are not all the same, such as for $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}$.

c. Ergodicity. For a compact topological space X the collection $\mathfrak{M}$ of Borel probability measures is weak*-compact and convex and therefore the same holds for the closed subset $\mathfrak{M}(\Phi)$ of measures invariant under a group action Φ.

THEOREM 4.2.4. *A Φ-invariant probability measure is ergodic if and only if it is an extreme point of $\mathfrak{M}(\Phi)$.*

PROOF. If $\mu \in \mathfrak{M}(\Phi)$ is not ergodic, $A \subset X$ Φ-invariant, $0 < \mu(A) < 1$ then $\mu = \mu(A)\mu_A + (1-\mu(A))\mu_{X\smallsetminus A}$, where $\mu_A(Y) := \mu(Y \cap A)/\mu(A)$. If $\mu = \lambda\mu_1 + (1-\lambda)\mu_2$, $0 < \lambda < 1$, $\mu_1 \neq \mu_2$, then the density of $\mu_1 \ll \mu$ is a nonconstant Φ-invariant $L^1(\mu)$-function. □

d. Ergodic decomposition. The advantage of this point of view is that one can bring to bear results from functional analysis in order to gain insight into invariant measures. For example, existence of ergodic measures follows immediately from existence of extreme points. Moreover, one directly obtains the ergodic decomposition of invariant measures by using the following result:

THEOREM 4.2.5 (Choquet). *If C is a compact metrizable convex set C in a locally convex topological vector space and $x \in C$ then there is a probability measure μ supported on $\operatorname{ex} C$ such that $x = \int_{\operatorname{ex} C} z \, d\mu(z)$.*

THEOREM 4.2.6. *Every invariant Borel probability measure for an action Φ on a metrizable compact space X can be decomposed into an integral of ergodic invariant Borel probability measures in the following sense: There is a partition (modulo null sets) of X into invariant subsets X_α, $\alpha \in A$ with A a Lebesgue space, and each X_α carrying an f-invariant ergodic measure μ_α such that $\int \varphi \, d\mu = \iint \varphi \, d\mu_\alpha \, d\alpha$ for any function φ.*

SKETCH OF PROOF. The ergodic decomposition theorem is reduced to the Choquet theorem as follows. As before let $\mathfrak{M}(\Phi)$ be the closed subset of $\mathfrak{M}$ consisting of Φ-invariant probability measures. It is itself compact and convex. By Theorem 4.2.4 the ergodic Φ-invariant measures are the *extreme* points of $\mathfrak{M}(\Phi)$, i.e. they are those points which cannot be written as a nontrivial convex combination of two Φ-invariant probability measures. We denote by $\mathfrak{M}_E(\Phi)$ the space of extreme points (ergodic measures) in $\mathfrak{M}(\Phi)$. Then by Theorem 4.2.5 for any Φ-invariant probability measure μ on X there exists a probability measure ν on $\mathfrak{M}_E$ such that

$$\mu = \int_{\mathfrak{M}_E} \alpha \, d\nu(\alpha).$$

For the details, see [**Ph**]. □

While in the purely measure-theoretic context points and null sets are somewhat elusive, and hence sets are difficult to pin down precisely, the present setting allows an explicit description of the ergodic decomposition due to Oxtoby [**Ox**] (which can to some extent be traced to [**KB**]): For each ergodic measure consider the G_δ set of typical points with respect to all continuous functions, *e.g.*, points for which the Birkhoff averages for each continuous function converge to the integral of this function with respect to the measure in question (Theorem 3.5.2). This is a null set for all other ergodic measures and these sets are evidently pairwise disjoint. They are called *ergodic sets*. This essential uniqueness of the ergodic decomposition shows that $\mathfrak{M}(\Phi)$ is essentially a simplex.

REMARK. Notice that the above arguments relied on the Birkhoff Ergodic Theorem in an essential way and hence are not applicable to general nonsingular measurable actions. The proof of the ergodic decomposition theorem in that case uses that ergodic decomposition is an invariant of orbit equivalence and uses an appropriate element from the full group [**S-FK**].

e. Continuous representation. While the preceding arguments expressly invoked the topological structure of the space, there is a general scheme by which one can obtain analogous results for measure-preserving actions. It is the device of *continuous representation*.

THEOREM 4.2.7. [**F1**, Theorem 5.15]. *If $(X, \mathcal{A}, \mu)$ is a separable measure space and T a measure-preserving transformation then there is a compact topological space Y with a Borel measure λ and a λ-preserving homeomorphism $f : Y \to Y$ such that $(Y, \mathcal{B}, \lambda, f)$ is equivalent to $(X, \mathcal{A}, \mu, T)$*

Here a measure space $(X, \mathcal{A}, \mu)$ is said to be separable if $\tilde{\mathcal{A}} := \mathcal{A}/\sim$ is generated by a countable set, where $\sim$ denotes equivalence mod 0.

SKETCH OF PROOF. To construct the continuous representation let $\mathcal{A}_0 = \{A_m\}_{m\in\mathbb{N}} \subset \mathcal{A}$ be a countable T-invariant algebra of distinct sets with $\tilde{\mathcal{A}}_0$ dense in $\tilde{\mathcal{A}}$, $Y := \Omega_2^R$, $\pi_m \colon Y \to \{0,1\}$ the mth coordinate projection, $A'_m := \pi_m^{-1}(\{1\})$. For $N \in \mathbb{N}$ construct a Borel measure λ_N on Y such that $\lambda_N(\bigcap_{j=1}^k A'_{i_j}) = \mu(\bigcap_{j=1}^k A_{i_j})$ whenever $i_1 < i_2 < \cdots < i_k \le N$ and let λ be an accumulation point of $(\lambda_N)_{N\in\mathbb{N}}$. Define $f \colon Y \to Y$ by $(f(\omega))_k = \omega_{l(k)}$, where $l(k)$ is defined by $A_{l(k)} = T^{-1}(A_k)$. Then f is continuous and $f^{-1}(A'_k) = A'_{l(k)}$, so f preserves λ. It is easy to see that $(Y, \mathcal{B}, \lambda, f)$ is equivalent to $(X, \mathcal{A}, \mu, T)$. □

For actions of more general groups the corresponding continuous representation result has been established by Varadarajan [**Va**].

These continuous representation results together with Theorem 4.2.6 yield teh Ergodic Decomposition Theorem 3.4.3 for general measure-preserving actions.

f. The Furstenberg Correspondence Principle; the Szemerédi Theorem. (See also [**S-B, Pt**].) In Section 2.6g we discussed a connection between recurrence behavior in topological dynamics and the appearance of a large scale structure in at least one of the sets which appear from a finite partition of $\mathbb{Z}$. A counterpart of this principle connects the large scale structure of *positive upper density* subsets of $\mathbb{Z}$ or $\mathbb{N}$ with quantitative recurrence properties of measure-preserving transformations. For a subset S of natural numbers the *upper density* is defined as

$$\overline{\lim_{n\to\infty}} \operatorname{card}([1,\dots,n] \cap S)/n.$$

The general construction connecting such statements is as follows. Any $S \subset \mathbb{N}$ defines a one-sided sequence $\omega \in \Omega_2^R$ of zeros and ones where $\omega_i = 0$ if $i \in S$ and $\omega_i = 1$ if $i \notin S$. Using the averaging of the point mass δ_ω along the subsequence realizing the upper density $d > 0$ of S one constructs an invariant mesure μ for the one-sided shift σ_2^R such that $\mu(C_0^0) = d$.

This construction connects the pattern of returns of the cylinder C_0^0 with the pattern of appearances of natural numbers in the set S.

The argument in the other direction uses coding similarly to the proof of Theorem 2.6.11.

We illustrate the operation of this version of the Furstenberg correspondence principle by describing its original and most famous application [**F1**].

The Szemerédi Theorem asserts that any set of natural numbers of positive upper density contains an arbitrary long arithmetic progression. It is stronger than the van der Waerden Theorem (Section 2.6g) because the upper density is subadditive and hence one of the sets in a finite partition has to have positive upper density.

The corresponding property of a measure-preserving transformation $T \colon (X,\mu) \to (X,\mu)$ is *multiple Poincaré recurrence*: For every set A of positive measure and every $k \in \mathbb{N}$ there exists an $n \in \mathbb{N}$ such that

$$\mu(\bigcap_{i=0}^{k-1} T^{-in} A) > 0.$$

Notice that then there are infinitely many such n.

THEOREM 4.2.8. *The Szemerédi theorem is equivalent to the multiple Poincaré recurrence propery for any measure-preserving transformation.*

PROOF. Assuming multiple Poincaré recurrence and having a subset S of natural numbers of positive upper density construct an invariant measure μ for the shift σ_2^R as above. By construction μ is the weak limit of the sequence of measures

$$\mu_m = \frac{1}{n} \sum_{i=0}^{n_m-1} \delta_{(\sigma_2^R)^i\omega}$$

for some sequence $n_m \to \infty$ and $\mu(C_0^0) > 0$. By multiple recurrence for any natural number k there exists an l such that

$$\mu(\bigcap_{j=0}^{k-1} \sigma^{-jl}(C_0^0)) > 0,$$

hence by closeness of the cylinders and weak convergence for a sufficiently large m one also has

$$\mu_m(\bigcap_{j=0}^{k-1} \sigma^{-jl}(C_0^0)) > 0.$$

But this implies that $\delta_{(\sigma_2^R)^i\omega}(\bigcap_{j=0}^{k-1} \sigma^{-jl}(C_0^0)) > 0$, for some $i \in \{0, \dots, n_m - 1\}$, *i.e.,* the set S contains the arithmetic progression $(i, i+l, \dots, i+(k-1)l)$.

Conversely, assume multiple Poincaré recurrence does not hold. This means that for some $k \in \mathbb{N}$ there exists a measure-preserving transformation T and a set A of positive measure such that

$$\mu(\bigcap_{i=0}^{k-1} T^{-in}A) = 0 \text{ for all } n \in \mathbb{N}.$$

Using ergodic decomposition one can then find an ergodic transformation with the same property. Let

$$B = \bigcup_{m=0}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{i=0}^{k-1} T^{-m-in}A$$

Since $\mu(B) = 0$ there exists a typical point x for A outside of B. For such a point we define the set of natural numbers S by $i \in S$ if $T^i x \in A$. Since x is typical the set S actually has positive density equal to $\mu(A)$ but does not contain any arithmetic progression of length k since $x \notin B$, contradicting the assertion of the Szemerédi theorem. □

3. Unique ergodicity

a. Definition and uniform convergence. A topological dynamical system Φ is said to be *uniquely ergodic* if $\operatorname{card} \mathfrak{M}(\Phi) = 1$, that is, there is a unique invariant Borel probability measure. By Theorem 4.2.4 this measure is ergodic.

This is evidently a topological property and it can be described in topological terms:

THEOREM 4.3.1. *For a uniquely ergodic action and any continuous function the Birkhoff averages in* (1.1) *(or* (3.1)*) converge uniformly. The converse holds for topologically transitive actions.*

PROOF. Nonuniform convergence for some continuous function φ gives sequences $x_k \to x, y_k \to y \subset X$ for which the averages $\mathcal{F}_{n_k}$ lie on either side of some interval (a, b). By the diagonal process there is a subsequence $(n_{k_l})_{l\in\mathbb{N}}$ such that $J(\varphi) := \lim_{l\to\infty} \mathcal{F}_{n_{k_l}}(\varphi)$

exists at both x and y and gives different values. By the Riesz Representation Theorem we thus have two distinct invariant measures. □

Transitivity is needed for the converse because it fails for the map $f\colon S^1 \times [0,1] \to S^1 \times [0,1]$, $(x,t) \mapsto (x+\alpha, t)$ for $\alpha \notin \mathbb{Q}$, that is, the product of a uniquely ergodic map (Proposition 4.3.3) with the identity.

Since by the Kryloff–Bogoliouboff Theorem 4.2.2 every minimal set is the support of an invariant measure we observe

THEOREM 4.3.2. *A uniquely ergodic action has only one minimal set; in particular a topologically transitive uniquely ergodic action is minimal.*

Note the analogy to Proposition 2.3.4.

b. Unique ergodicity with trivial recurrence. The simplest examples of uniquely ergodic systems are those for which the invariant measure is atomic, *i.e.*, is concentrated on a single point (or periodic orbit in continuous time). A specific instance is the map $x \mapsto x+1$ of the projective plane (Example 2.4.2) or the diffeomorphism $f\colon S^1 \to S^1$ induced by the map $x \mapsto x + \frac{1}{10}\sin^2 \pi x \pmod 1$ for which the fixed point 0 is not an attractor but $\alpha(f) = \omega(f) = \{0\}$, or a parabolic projective map on S^1 as in Example 2.4.4, Example 2.4.4 and Section 4.2b. The only invariant probability measure is the atom at the fixed point.

Next, we give some examples with nontrivial recurrence.

c. Minimal translations of compact abelian groups. These are the simplest nontrivial class of topologically transitive examples.

PROPOSITION 4.3.3. *Let G be a compact abelian group, $g \in G$. Then the following properties of the translation L_g are equivalent:*

(1) *if $\chi \in G^*$ and $\chi(g) = 1$ then $\chi = e$, the identity,*
(2) *topological transitivity,*
(3) *minimality,*
(4) *ergodicity of Haar measure,*
(5) *unique ergodicity.*

PROOF. Equivalence of the first three properties is lies entirely within the topological realm, see Proposition 2.2.3 and Proposition 2.2.4. The first property is equivalent to ergodicity of Haar measure: The characters are the eigenfunctions with their values at g the eigenvalues, so the assumption implies that 1 is a simple eigenvalue, which is equivalent to ergodicity (invariant functions are constant). Ergodicity of Haar measure implies unique ergodicity: If $h \in G$ is typical for Haar measure (in the sense of the Birkhoff Ergodic Theorem) then so is kh for any $k \in G$ by invariance of Haar measure under L_k. □

d. Isometries. More generally, consider an isometry $f\colon X \to X$ of a compact metric space. As we pointed out in Proposition 2.2.4, every orbit closure is a minimal set. This implies the following "minimal decomposition" into orbit closures:

COROLLARY 4.3.4. *Let Φ be an action by isometries on a compact metric space X. Then X is uniquely partitioned by closed invariant sets X_α on each of which Φ is topologically transitive.*

PROPOSITION 4.3.5. *Let $f\colon X \to X$ be a minimal isometry. Then f is topologically conjugate to a translation of a compact abelian group.*

This is a topological counterpart of the Discrete Spectrum Theorem 3.6.3.

COROLLARY 4.3.6. *A minimal isometry is uniquely ergodic.*

COROLLARY 4.3.7. *Any ergodic invariant measure for an isometry on a compact metric space has pure point spectrum.*

PROOF. Use Proposition 4.3.5 and the Discrete Spectrum Theorem 3.6.3. □

PROOF OF PROPOSITION 4.3.5. Consider the space $C(X,X) = \{T\colon X \to X \mid T \text{ continuous}\}$ with the uniform topology. $(f^n)_{n\in\mathbb{Z}}$ is equicontinuous, hence the closure is a compact abelian group H. Minimality implies that for $x, y \in X$ there exists $(n_k)_{k\in\mathbb{Z}}$ such that $f^{n_k}(x) \to y$, so by passing to a subsequence we obtain $g \in H$ such that $g(x) = y$, *i.e.,* H acts transitively. Then $\varphi\colon H/H_x \to X$, $gH_x \to g(x)$, where H_x is the stationary subgroup, is an isomorphism and by definition $f \circ \varphi = \varphi \circ L_f$. □

This argument immediately translates to isometric actions of other discrete or continuous abelian groups. For the discussion of the nonabelian case see the survey [**S-FK**].

Because of its appearance in transitive components of isometries unique ergodicity is an important paradigm for *elliptic dynamics*, which will be discussed in Chapter 7. But it also appears quite often in systems with *parabolic* behavior (Chapter 8). Without going into details we exhibit several characteristic examples.

e. Unipotent affine maps of the torus. For $\alpha \notin \mathbb{Q}$ the map $A_{n,\alpha}\colon \mathbb{T}^{n+1} \to \mathbb{T}^{n+1}$, $(x_1, \dots, x_{n+1}) \mapsto (x_1 + \alpha, x_2 + x_1, \dots, x_{n+1} + x_n) \pmod 1$ is uniquely ergodic (with Lebesgue measure invariant). This is closely related to uniform distribution of the fractional parts $(\{p(n)\})_{n\in\mathbb{Z}}$ of a polynomial, see [**KH**, Exercises 2.4.3–7]. The original proof is due to Weyl and uses estimates of trigonometric sums. A purely qualitative argument using ergodic theory and generalizing the above argument for translations is due to Furstenberg.

These maps have zero entropy, but one can apply the subexponential orbit growth concepts from Section 2.5h, Section 2.5i and Section 3.7l. The power scale is appropriate. One easily checks that $\mathrm{ent}_p(A_{n,\alpha}) = n$. Section 8.3a presents more detail as well as a more general class of examples of this type.

f. Horocycle flows on surfaces of negative curvature. In the setting of Section 2.1b and Section 3.3c consider $H = SL(2,\mathbb{R})$, $\Gamma \subset G$ a cocompact lattice, $h^t := \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$. The homogeneous flow Φ^{h^t} on H/Γ given by left translations is uniquely ergodic [**S-KSS, F3**]. The name *horocycle flow* will be explained in Section 6.5e.

Here, too, one easily finds the power entropy (Section 2.5h): $\mathrm{ent}_p(h^t) = 2$. Section 8.3b describes more general examples of this nature in some detail.

g. Interval exchanges. (See also [**S-MT**].) Interval exchanges are maps defined by rigidly permuting finitely many subintervals of $[0,1]$: Consider a permutation π of $\{1,\dots,n\}$, a vector $v = (v_1, \dots, v_n)$ in the interior of the unit simplex, *i.e.,* such that $v_i > 0$ for $i = 1,\dots,n$ and $\sum_{i=1}^n v_i = 1$, and $\epsilon = (\epsilon_1, \dots, \epsilon_n) \in \{-1,1\}^n$. Let

$u_i = \sum_{j=1}^{i} v_j$ and $\Delta_i = (u_{i-1}, u_i)$ for $i = 1, \dots, n$. The *interval exchange transformation* $I_{v,\pi,\epsilon}\colon [0,1] \to [0,1]$ is the map that is an isometry on every interval Δ_i, rearranges those intervals according to the permutation π, and preserves or reverses orientation on Δ_i according to the sign of ϵ_i ($i = 1, \dots, n$). If $\epsilon_i = 1$ for all i we write $I_{v,\pi}$ for $I_{v,\pi,\epsilon}$ and call it an *oriented interval exchange transformation*.

Although interval exchanges are not continuous, they properly belong here because they are closely connected to smooth dynamics. They arise as sections of smooth area-preserving flows on surfaces of genus greater than 1, which are a standard example of systems with parabolic behavior (see Section 8.4 and [**KH**, Section 14.6]).

In the present context the central features of interest are:

(1) An exchange of m intervals has at most m nonatomic ergodic invariant measures [**KH**, Theorem 14.5.14].
(2) Under a natural irreducibility condition on the permutation, (Lebesgue-) almost all interval exchanges are uniquely ergodic [**V2, Ms1**].

Any aperiodic interval exchange transformation I has $\mathrm{ent}_p(I) = 1$.

EXAMPLE 4.3.8. A circle rotation by $\alpha \in (0,1)$ naturally corresponds to an exchange of the intervals $[0, 1-\alpha)$ and $[1-\alpha, 1)$.

EXAMPLE 4.3.9 ([**KS**]). The map induced by the circle rotation by $\alpha \in (0,1)$ an any interval naturally corresponds to an oriented exchange of three intervals with the permutation $\pi(k) = 4 - k$, $k = 1, 2, 3$. Conversely, any interval exchange of this kind naturally corresponds to a map induced by a circle rotation on an interval. Since any other oriented exchange of three intervals reduces to an exchange of two intervals or preserves one of the end intervals we deduce that for oriented exchanges of three intervals topological transitivity implies unique ergodicity.

h. Uniquely ergodic realization. In the preceding examples unique ergodicity is related to rather special behavior with respect to the invariant measure. In particular, topological and metric entropy are zero. This is, however, not the case in general, even for symbolic systems.

THEOREM 4.3.10. [**J, Kr3, BF**] *For any ergodic measure-preserving transformation there is a metrically isomorphic uniquely ergodic homeomorphism of a Cantor set.*

For any ergodic measure-preserving transformation with finite entropy there is a metrically isomorphic uniquely ergodic symbolic system.

Thus unique ergodicity does not impose any restrictions of ergodic peoperties with respect to unique invariant measure.

i. Minimal systems with many invariant measures. Unique ergodicity (coupled with topological transitivity) implies minimality and may be viewed as a stronger quantitative counterpart of the latter. While in algebraic systems like those described above (Section 2.1b, Section 2.2c, Section 3.3c) minimality implies unique ergodicity (with Haar measure being the only invariant one), this is not the case even for such natural classes of systems as isometric extensions of rotations (see [**KH**, Corollary 12.6.4] and Section 7.5a) and interval exchange transformations ([**KH**, Corollary 14.5.18] and Section 8.4d). These are instances of a more general pathology, which often appears in connection with an abnormally fast periodic approximation or "Liouvillian" behavior. See Section 7.5 for a more extensive discussion.

4. Metric and topological entropy

a. Averaging versus maximizing. Topological entropy was discovered after measure-theoretic entropy. Measure-theoretic entropy gives a quantitative measure of the complexity of a dynamical system as seen via an invariant measure. Topological entropy was found by extracting from the same concept an invariant of the topological dynamics only. Though there are some analogies in the definitions, the absence of a natural measure of the size of sets in topological dynamics leads to some differences between the two notions. Notably the measure-theoretic entropy of the union of two invariant sets is the average of the entropies of the invariant sets, weighted by their measures (Section 3.7i), whereas for topological entropy the entropy of a union is the maximum of the entropies of the two components (Section 2.5f). In other words, topological entropy measures the maximal dynamical complexity versus an average complexity reflected by measure-theoretic entropy. Therefore, one expects measure-theoretic entropy to be no greater than topological entropy. This is indeed the case:

PROPOSITION 4.4.1. *Let $f\colon X \to X$ be a homeomorphism of a compact metric space X. Then $h_\mu(f) \le h_{top}(f)$ for all $\mu \in \mathfrak{M}(f)$.*

SKETCH OF PROOF. [**Mi**] If $\xi = \{C_1, \dots, C_k\}$ is a measurable partition of X, μ a Borel measure, then $\mu(C_i) = \sup\{\mu(B) \mid B \subset C_i \text{ closed}\}$, so there are compact sets $B_i \subset C_i$ such that $H(\xi \mid \beta) < 1$ for $\beta = \{B_0, B_1, \dots, B_k\}$ with $B_0 = X \smallsetminus \bigcup_{i=1}^k B_i$. By Section 3.7d3 $h_\mu(f, \xi) \le h_\mu(f, \beta) + H_\mu(\xi \mid \beta) \le h_\mu(f, \beta) + 1$.

$\mathcal{B} := \{B_0 \cup B_1, \dots, B_0 \cup B_k\}$ is an open cover of X. By Section 3.7b1 $H_\mu(\beta^f_{-n}) \le \log \operatorname{card} \beta^f_{-n} \le \log(2^n \operatorname{card} \mathcal{B}^f_{-n})$. The Lebesgue number δ_0 of $\mathcal{B}$ is also that of $\mathcal{B}^f_{-n}$ with respect to d^f_n. Since $\mathcal{B}^f_{-n}$ is a minimal cover, every $C \in \mathcal{B}^f_{-n}$ contains a point x_C not in any other element of $\mathcal{B}^f_{-n}$. The x_C form a δ_0-separated set. Consequently $h_\mu(f, \beta) \le h_{\text{top}}(f) + \log 2$ and $h_\mu(f, \xi) \le h_\mu(f, \beta) + 1 \le h_{\text{top}}(f) + \log 2 + 1$. Therefore $h_\mu(f) = h_\mu(f^n)/n \le (h_{\text{top}}(f^n) + \log 2 + 1)/n = h_{\text{top}}(f) + (\log 2 + 1)/n$ for all $n \in \mathbb{N}$ by Section 3.7i4 and Section 2.5f3, and hence $h_\mu(f) \le h_{\text{top}}(f)$. □

b. Slow entropy. Example 2.6.9 shows that contrary to the situation of exponential growth, subexponential orbit growth cannot always be derived from properties of invariant measures. However, using the modified topological a-entropy $\operatorname{ent}^{\text{top}}_a(\Phi)$ from Section 2.5i eliminates the disparity between the topological and ergodic behavior in Example 2.6.9. In this case the modified topological power entropy and the metric power entropy are both zero.

Indeed, Proposition 4.4.1 extends to a-entropy.

PROPOSITION 4.4.2. *Let f be a continuous map of a compact metric space X, μ an f-invariant Borel probability measure. Then $\overline{\operatorname{ent}}^\mu_a(T) \le \overline{\operatorname{ent}}^{top}_a(T)$ for any scale function a and likewise for the "lower" counterparts $\underline{\operatorname{ent}}$.*

The proof uses an inequality between the cardinalities of spanning sets in the topological situation and in the Hamming metric. We give an inequality that is slightly stronger than needed. Let $S_d(f, \epsilon, n, \delta)$ be the minimal number of $\eth^f_n$-ϵ-balls whose union has measure at least $1 - \delta$. For a finite partition $\xi = \{C_0, \dots, C_{N-1}\}$ such that $\mu(\partial\xi) = 0$ and $\delta > 0$ take $\alpha > 0$ such that $\mu(U_\alpha(\partial\xi)) < \delta^2$, where $U_\alpha(A)$ denotes the open α-neighborhood of A.

PROPOSITION 4.4.3. **[KT]** $S^H_\xi(f, \delta+\sqrt{\alpha}, n, \beta+\delta) \le S_d(f, \alpha/2, n, \beta)$ *for any* $\beta > 0$.

It is quite probable that a counterpart of the Variational Principle below holds for the modified topological a-entropy.

c. Measures of high complexity. The Kryloff–Bogoliouboff Theorem gives invariant measures for standard dynamical systems. In the absence of unique ergodicity it is, by analogy, natural to look for distinguished ones among the invariant measures. As we saw, ergodic measures are distinguished as extreme points of the compact convex set of invariant Borel probability measure, but in a particular context some of these may be more interesting than others. A natural idea is to consider maxima of continuous functionals on measures, such as integrals of action functions or forms, which are attained by compactness (and, as extreme points, ergodic). This issue comes up in Lagrangian dynamics, see the survey [**S-BK**] and [**Mt**].

At this stage, entropy suggests itself as a functional on measures for which one should implement this scheme. More generally one may consider the "pressure" functional $\mu \mapsto \int \varphi \, d\mu + h_\mu(f)$ for a given continuous function φ. Unfortunately, entropy does not always depend continuously on the measure, and therefore neither boundedness nor existence of a maximum can a priori be expected. We will see, however, that boundedness is not exceptional and that, while there is not always a maximum, the supremum can be described explicitly.

The underlying reason is that measures assigning most weight to regions of high complexity should have measure-theoretic entropy (or pressure) close to the topological entropy (or pressure). This is indeed true, *i.e.*, the topological entropy is the supremum of the measure-theoretic entropies, and likewise for pressure (Section 2.5k). This is the content of the Variational Principle. It is an easy observation that existence of the maximum follows from expansivity, and there is a further criterion for uniqueness of the maximizing measure.

d. The Variational Principle.

THEOREM 4.4.4. *Let* $f\colon X \to X$ *be a homeomorphism of a compact metric space* X *and* $\varphi \in C(X)$. *Then* $P(\varphi) = \sup\{h_\mu(\varphi) + \int \varphi \, d\mu \mid \mu \in \mathfrak{M}(f)\}$.

We show this for the simpler case of entropy, *i.e.*, $\varphi = 0$, via a proof due to Misiurewicz [**Mi**]. Including the function φ in the arguments adds enough bulk to the argument to recommend this shortcut, while the ideas are the same. The argument is complete except that we use the properties of entropy from Section 3.7b as well as Proposition 4.4.1.

PROOF. By Proposition 4.4.1 we need $\overline{\lim}_{n\to\infty} N_X(f, \epsilon, n)/n \le \sup_\mu h_\mu(f)$ or:

Let $E_n \subset X$ be an (n, ϵ)-separated set, $\nu_n := \sum_{x\in E_n} \delta_x / \operatorname{card}(E_n)$, where δ_x is the probability measure supported on $\{x\}$, and $\mu_n := \sum_{i=0}^{n-1} f^i_* \nu_n / n$, where $f_*\mu(A) := \mu(f^{-1}(A))$. Then there is an f-invariant weak*-accumulation point μ of $(\mu_n)_{n\in\mathbb{N}}$ with $\overline{\lim}_{n\to\infty}(1/n)\log\operatorname{card}(E_n) \le h_\mu(f)$.

For a sequence $(n_k)_{n\in\mathbb{N}}$ with $\lim_{k\to\infty}\log\operatorname{card}(E_{n_k}) = \overline{\lim}_{n\to\infty}\log\operatorname{card}(E_n)$ take any accumulation point μ of $(\mu_{n_k})_{k\in\mathbb{N}}$ (by weak*-compactness of $\mathfrak{M}$); $\mu \in \mathfrak{M}(f)$ since $f_*\mu_n - \mu_n = (f^n_*\nu_n - \nu_n)/n$ and ν_n are probability measures. It is not hard to see that there is a finite partition ξ with elements of diameter less than ϵ and $\partial\xi := \bigcup_{C\in\xi} \partial C$ a μ-null set (this implies $\lim_{k\to\infty} H_{\mu_{n_k}}(\xi^f_{-q}) = H_\mu(\xi^f_{-q})$). Now $\log\operatorname{card}(E_n) = H_{\nu_n}(\xi^f_{-n})$

since each $C \in \xi^f_{-n}$ contains at most one $x \in E_n$, so there are $\operatorname{card}(E_n)$ elements of ξ^f_{-n} with ν_n-measure $1/\operatorname{card}(E_n)$. For $0 \le k < q < n$ let $a(k) := \lfloor (n-k)/q \rfloor$ (integer part) and $S = \{0, 1, \dots, k, k + a(k)q + 1, \dots, n-1\}$. Then $\operatorname{card}(S) \le 2q$ because $k + a(k)q \ge n - q$. Since $\{0, 1, \dots, n-1\} = \{k + rq + i \mid 0 \le r < a(k), 0 < i \le q\} \cup S$ we have

$$\xi^f_{-n} = \Big(\bigvee_{r=0}^{a(k)-1} f^{-(rq+k)}(\xi^f_{-q}) \Big) \vee \Big(\bigvee_{i \in S} f^{-i}(\xi) \Big)$$

and

$$\begin{aligned} \log \operatorname{card}(E_n) = H_{\nu_n}(\xi^f_{-n}) &\le \sum_{r=0}^{a(k)-1} H_{\nu_n}(f^{-(rq+k)}(\xi^f_{-q})) + \sum_{i \in S} H_{\nu_n}(f^{-i}(\xi)) \\ &\le \sum_{r=0}^{a(k)-1} H_{f_*^{rq+k}\nu_n}(\xi^f_{-q}) + 2q \log \operatorname{card}(\xi) \end{aligned}$$

by Section 3.7b1,4. Thus by Section 3.7b6

$$\begin{aligned} q \log \operatorname{card}(E_n) = \sum_{k=0}^{q-1} H_{\nu_n}(\xi^f_{-n}) &\le \sum_{k=0}^{q-1} \Big(\sum_{r=0}^{a(k)-1} H_{f_*^{rq+k}\nu_n}(\xi^f_{-q}) + 2q \log \operatorname{card}(\xi) \Big) \\ &\le n H_{\mu_n}(\xi^f_{-q}) + 2q^2 \log \operatorname{card}(\xi) \end{aligned}$$

and $\overline{\lim}_{n\to\infty} (1/n) \log \operatorname{card}(E_n) \le \lim_{k\to\infty} H_{\mu_{n_k}}(\xi^f_{-q})/q = H_\mu(\xi^f_{-q})/q$. Therefore $\overline{\lim}_{n\to\infty} (1/n) \log \operatorname{card}(E_n) \le h_\mu(f, \xi) \le h_\mu(f)$. □

REMARK. For a proof for pressure see [**KH**, Section 20.2]. Furthermore, this proof of the variational principle immediately extends to the actions of $\mathbb{Z}^k_+$ [**Mi**].

e. Existence of a maximizing measure. It is clear that a measure of maximal entropy and measures of maximal pressure exist whenever metric entropy is an upper semicontinuous function of the measure. Here are two situations where this is the case.

If in the preceding construction of a measure of large entropy we use maximal (n, ϵ)-separated sets E_n we obtain a measure μ such that $\overline{\lim}_{n\to\infty} N_X(f, \epsilon, n)/n \le h_\mu(f)$. If f is expansive and ϵ an expansivity constant then the left hand side is $h_{\text{top}}(f)$ (Section 2.5f6). Therefore expansive homeomorphisms of a compact metric space have a measure of maximal entropy. This extends to the case of pressure.

A trivial case of existence of a measure of maximal pressure is that of $h_{\text{top}}(f) = 0$. Any measure has zero entropy and for $\varphi \in C(X)$ the functional $h_\mu(f) + \int \varphi \, d\mu = \int \varphi \, d\mu$ attains its supremum by continuity and weak*-compactness.

f. Specification. Uniqueness of maximizing measures needs an assumption from topological dynamics. The corresponding notion was not introduced earlier because it is mostly used when studying invariant measures.

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

called the *spacing* of S. We say that S *parameterizes* the collection $\{P_I \mid I \in \tau\}$ of orbit segments of f.

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

Thus, a specification is a parameterized union of orbit segments $P_{\restriction_{I_i}}$ of f.

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

An example of a specification in the full shift is given by fixing the entries of a sequence on finitely many index intervals, such as: $\omega_{-3} = 0, \omega_0 = 1, \omega_1 = 0, (\omega_{17}\omega_{18}\omega_{19}\omega_{20}\omega_{21}) = (10110)$ (three intervals, spacing 2). Full shifts have the specification property; the required spacing is related to the rank of a cylinder of size ϵ. Transitive subshifts of finite type also have this property, but the spacing may need to be increased a little to allow for intermediate states to make a transition that is disallowed in a single step. There are other classes of symbolic systems with specification, such as *sofic* systems [**LM**]. The linear expanding maps E_m of S^1 and hyperbolic automorphisms of the torus (Section 6.5a) provide further examples of maps with this property. See Section 6.7c for the principal application of specification.

g. Uniqueness of maximal measures. Going beyond existence of maximizing measures by expansivity requires further hypotheses. The main assumption is that the map f have the specification property. The secondary assumption is that, when one considers pressure, the continuous function in question be in

$$C^f := \big\{\varphi \in C(X) \mid \exists K, \epsilon > 0 \text{ such that } d_n^f(x, y) \leq \epsilon \Rightarrow |S_n\varphi(x) - S_n\varphi(y)| \leq K\big\}.$$

This means that the statistical sums $S_n\varphi(x) := \sum_{i=0}^{n-1} \varphi(f^i(x))$, which are used to define pressure (Proposition 2.5.6), change with the orbit segment in a way that can be controlled entirely in terms of closeness of the orbit segments and independently of the length of the segments considered. Since $0 \in C^f$, entropy is a special case. This class is sufficiently large and naturally defined in applications (Hölder functions for hyperbolic sets [**KH**, Proposition 20.2.6]). This subject is presented in detail in [**S-C**].

Now we can present Bowen's result about uniqueness of maximizing measures:

THEOREM 4.4.5. *Let (X, d) be a compact metric space, $f\colon X \to X$ an expansive homeomorphism with the specification property and $\varphi \in C^f(X)$. Then there is exactly one $\mu_\varphi = \mu \in \mathfrak{M}(f)$ with $h_\mu(f, \varphi) + \int \varphi\, d\mu = P(f, \varphi)$. It is mixing and, when counted with φ-weights, periodic orbits are equidistributed.*

A always, the special case $\varphi = 0$ is interesting and gives a unique measure of maximal entropy, called the Bowen measure. One should note, by the way, that as a byproduct of the proof one obtains positive topological entropy for f [**KH**, Theorem 18.5.5], unless $\operatorname{card} X \leq 1$.

SKETCH OF PROOF FOR $\varphi = 0$. Some results about periodic points in expansive systems with specification are needed. In Section 2.5f6 we found that $P_n(f) \leq N(f, \epsilon, n)$. Adding specification gives $c_1 e^{nh_{\text{top}}(f)} \leq P_n(f) \leq c_2 e^{nh_{\text{top}}(f)}$ [**KH**, Theorem 18.5.5] by showing that the growth of periodic points is multiplicative. In essence, specification is

used to glue together orbits in $\mathrm{Fix}(f^{n_i})$, $(i = 1, \dots, l)$ to an orbit in $\mathrm{Fix}(f^{lM_\epsilon + \sum n_i})$. Up to a constant this gives $P_{\sum n_i}(f) = \prod P_{n_i}(f)$.

Now let μ_n be the f-invariant measure obtained by giving equal weights to the points $x \in \mathrm{Fix}(f^n)$. The preceding preliminaries give the following important fact. If $B = B_f(y, \epsilon, n)$ is the ϵ-ball around y for the metric d_n^f in (2.4) and $r \geq n$ then

$$\mu_r(\bar{B}) = \frac{\operatorname{card} \mathrm{Fix}(f^r) \cap \bar{B}}{P_r(f)} \geq \text{const} \, \frac{e^{(r-n)h_{\text{top}}(f)}}{e^{rh_{\text{top}}(f)}} = e^{-nh_{\text{top}}(f)}/C_\epsilon, \tag{4.1}$$

where the inequality essentially uses that by specification there are enough r-periodic orbits in B to ϵ-shadow any $(r-n)$-periodic orbit. The same statement evidently holds for any limit point μ of $(\mu_n)_{n \in \mathbb{N}}$. One can show that any such μ is ergodic (mixing, even), and uniqueness follows by showing that if $h_\nu(f) = h_{\text{top}}(f)$ for any invariant Borel probability measure ν (such measures exist by expansivity) then $\nu = \mu$. To prove this claim one need only verify (by ergodicity) that $\nu \perp \mu$ implies $h_\nu(f) < h_{\text{top}}(f)$. This argument uses convexity of $\phi(x) = x \log x \geq -1/e$ via

$$-\sum_{i=1}^{m} \phi(a_i) \leq \sum_{i=1}^{m} a_i \log m + \frac{1}{e}. \tag{4.2}$$

Fix $A = f(A) \subset X$ such that $\mu(A) = 0$, $\nu(A) = 1$ and for $n \in \mathbb{N}$ take a partition $B_n = \{\beta_i\}_{i=1}^{k_n}$ such that $B_f(x_i, \epsilon, n) \subset \beta_i \subset B_f(x_i, 2\epsilon, n)$ for some x_i. Expansivity implies that there is a finite union C_n of elements of B_n such that $(\mu + \nu)(C_n \triangle A) \to 0$ as $n \to \infty$. B_n is a generating partition for f^n, so, using (4.2)

$$\begin{aligned} nh_\nu(f) \leq H_\nu(B_n) &= -\sum_{i=1}^{k_n} \phi(\nu(\beta_i)) = -\sum_{\beta_i \subset C_n} \phi(\nu(\beta_i)) - \sum_{\beta_i \not\subset C_n} \phi(\nu(\beta_i)) \\ &\leq \nu(C_n) \log \operatorname{card}\{i \mid \beta_i \subset C_n\} + \nu(C_n) \log \operatorname{card}\{i \mid \beta_i \not\subset C_n\} + \frac{2}{e} \\ &\leq \nu(C_n) \log(C_\epsilon \mu(C_n)) + \nu(X \smallsetminus C_n) \log(C_\epsilon \mu(X \smallsetminus C_n)) + nh_{\text{top}}(f) + \frac{2}{e} \end{aligned}$$

by (4.1). Since $\nu(C_n) \to 1$ and $\mu(C_n) \to 0$ we get $n(h_\nu(f) - h_{\text{top}}(f)) \to -\infty$ as $n \to \infty$. □

For a detailed presentation see [**KH**, Section 20.3], which follows the original proof by Bowen [**B3**].

REMARK. Even for some systems with the specification property there are some functions naturally related to dynamics, such as the logarithm of the unstable Jacobian (Section 6.7d), which do not belong to the class C^f defined above and in fact the corresponding maximal measure may not be unique (see [**K3**]).

CHAPTER 5

Smooth, Hamiltonian and Lagrangian dynamics

1. Differentiable dynamics

a. Differentiable dynamical systems. Differentiable dynamics deals with groups of diffeomorphisms and semigroups of smooth transformations of finite-dimensional differentiable (smooth) manifolds. Usually the "time", *i.e.,* the acting group or semigroup G, is also assumed to possess a differentiable structure. This is trivially satisfied for discrete groups, so any countable group or semigroup qualifies (countability follows from discreteness by our standing assumption of second countability of the topology on G). Noninvertible systems with continuous time are rarely considered in the setting of finite-dimensional differentiable dynamics, so we exclude this possibility (as we did in our treatment of ergodic theory). Accordingly, in the invertible case G is assumed to be a Lie group and the standing assumption of differentiability in the time direction means that the action is generated by infinitesimal generators, *i.e.,* by a homomorphism of the Lie algebra $\mathrm{Lie}(G)$ of G into the Lie algebra $\Gamma(TM)$ of vector fields on the phase space M. In the case of cyclic dynamical systems we thus deal with a single vector field generating a smooth flow, *i.e.,* an $\mathbb{R}$-action, via solving an ordinary differential equation on M.

The preceding description must be qualified by noticing that, besides the natural global situation $\Phi\colon G \times M \to M$ of a G-action on the whole phase space, an important role in differentiable dynamics is played by semilocal and local situations as described in Section 5.1c and Section 5.1d. Furthermore, there are situations of "dynamics without time", where one looks at asymptotic behavior of noncompact leaves of a smooth foliation of a compact space. This was first mentioned in Section 1.2a and is developed further in Section 5.1e.

A differentiable dynamical system is also a topological one and thus concepts and results from topological dynamics are applicable. The study of topological properties, in the form of invariants as well as conjugacies, is one of the central themes in differentiable dynamics. In fact, differentiable dynamical systems are more amenable to topological classification than general topological dynamical systems. The theory of structural stability is an outstanding example (see Section 6.2e, Section 6.7i and [**S-H**]).

For invertible and, more generally, nonsingular differentiable dynamical systems there is also a natural invariant *measure class*, namely the Lebesgue, or smooth, class represented by any measure that is given by a smooth positive density in any local coordinate system. Existence of an invariant *measure* within this class is in general a highly nontrivial problem, although for many classes of systems the answer is trivially negative. Criteria for existence of such a measure are given by functional or differential equations, which are derived in Section 5.2m. For a nontrivial application see Section 6.7e. On the other hand, there are important specific classes of smooth dynamical systems that possess a smooth invariant

measure connected to an invariant geometric structure. Some homogeneous and affine dynamical systems, which appeared in Section 2.1b, Section 2.1c and Section 3.3c, are of this kind. Other classes include Hamiltonian, Lagrangian and contact systems described correspondingly in Section 5.3, Section 5.4 and Section 5.5. Hamiltonian dynamics is the main subject of [**S-LL, S-R, HZ**], and [**S-BK**] deals with both the Hamiltonian and Lagrangian case.

Finally, there is an important class of differentiable dynamical systems that usually do not possess a smooth invariant measure. It is *holomorphic systems* acting on a complex manifold, which are briefly discussed in Section 5.7.

b. Linearization. For every point p of the phase space we have a linear map $Df_p\colon T_pM \to T_pM$ between tangent spaces, which is invertible if p is a regular point. In this case the map may be approximated by Df_p in a small neighborhood of p. Under iteration we obtain the derivative $Df_p^n = Df_{f^{n-1}(p)} \cdots Df_{f(p)}Df_p\colon T_pM \to T_{f^n(p)}M$. In the simplest case of a periodic point $p = f^k(p)$ the behavior of the maps Df_p^n is largely determined by the single linear map $Df_p^k\colon T_pM \to T_pM$. The eigenvalues and Jordan block structure are invariant.

For a nonperiodic point one deals with the product of a growing number of different linear maps and for an individual orbit this product may not exhibit any regularity of behavior in terms of n. Nevertheless, for "most" points the asymptotic growth is quite definite, when measured on an exponential scale [**S-BKP**, Oseledets Multiplicative Ergodic Theorem], [**Os, Rg, W2**]. Of course, the quality of approximation of f^n by Df_p^n on a given neighborhood generally deteriorates as n increases. One of the central issues in differentiable dynamics is to what extent conclusions about the asymptotic behavior of orbits of nonlinear systems can be made based on the asymptotic behavior of the derivative.

c. Semilocal analysis. Unlike topological dynamics, where the restriction of a dynamical system to a closed invariant set belongs to the same category, natural invariant sets for smooth systems (such as attractors and other isolated sets) are often not submanifolds. See Section 5.2i and [**R**], [**KH**, Chapter 17] for typical examples. Still, the smooth origin of these sets is often reflected in many features, such as approximate self-similarity.

These observations motivate the following framework of semilocal analysis, which we first formulate for cyclic discrete time dynamical systems.

Let M be a differentiable manifold, $U \subset M$ open, $\Lambda \subset U$ closed (usually compact), and $f\colon U \to M$ a differentiable map such that $f(\Lambda) = \Lambda$ (but in the noninvertible case not necessarily $f^{-1}(\Lambda) = \Lambda$). In the invertible case f is assumed to be an embedding. Semilocal analysis studies orbits in Λ itself as well as orbit segments contained in a small open neighborhood $V \subset U$ of Λ (which may differ from U).

A simple illustration comes from the linear map $f(x, y) = (2x, y/2)$ of the plane with $\Lambda = \{0\}$ and V an r-ball around $(0, 0)$. While there is not much to be said about the behavior of Λ itself, there are the sets $A = \{(0, y) \mid 0 < |y| < r\}$ of points whose positive iterates stay in V and $B = \{(x, 0) \mid 0 < |x| < r\}$ of points whose negative semiorbit lies in V. Any other point outside Λ leaves V in finite time, both in the positive and negative direction, *i.e.*, has only a finite orbit segment in V.

A particular setting for semilocal analysis that is suitable for a large class of problems of great interest is provided by the Conley theory of isolating blocks [**S-FM**].

For semilocal analysis in the case of groups or semigroups G more general than $\mathbb{Z}$ or $\mathbb{N}$ we assume that the action of some generating set in G that includes a neighborhood of the identity is defined in a common open neighborhood U of Λ. Then for each element of G, we can study a neighborhood $V \subset U$ of Λ, which in general depends on the element.

d. Local analysis. For a cyclic dynamical system the simplest case of local analysis is the special case of the above semilocal situation where Λ consists of a single periodic orbit. This is a classical problem, some aspects of which are described later (Section 5.2c and Section 5.2h). The interest lies not with the orbit itself but with nearby orbits and orbit segments, the primary issue being the description of the sets of points whose entire orbit or whose positive or negative semiorbit are contained in a small neighborhood of the periodic orbit. This extends to noncyclic dynamical systems in the same way as above.

However, there are compelling reasons to extend the domain of local analysis to *nonperiodic orbits*. We restrict this discussion to cyclic systems to avoid technicalities and to look at the situation that is dominant in applications of local analysis anyway. Thus we have a reference orbit $(\Phi^t(p))_{t\in G}$ and a *tube* $\mathcal{T}$ around it. In the cases $G = \mathbb{Z}$ and $G = \mathbb{N}$ the tube can be described as a sequence of neighborhoods of the points $\Phi^n(p)$ of uniform "size" with coordinate systems inherited from appropriate coordinate charts. It is important to keep in mind that these coordinate neighborhoods can, and in a compact space must, overlap. In the case of a flow on an m-dimensional manifold M a tube is a smooth map $\mathcal{F}\colon \mathbb{R} \times D^{m-1} \to M$, where D^{m-1} is the unit ball in $\mathbb{R}^{m-1}$, such that $\mathcal{F}(t, 0) = \Phi^t(p)$ and $\mathcal{F}\restriction_{\{t_0\}\times D^{m-1}}$ is an embedding transverse to the flow. If M is compact or Riemannian we assume that $\mathcal{T}$ is of "uniform thickness", *i.e.,* that with respect to the Riemannian metric (any Riemannian metric in the compact case) the derivatives of $\mathcal{F}$ and their inverses are uniformly bounded.

The point of local analysis is to trace orbit segments in the given tube or a smaller one, *i.e.,* to study orbits of points for time intervals during which they stay close to the reference orbit. In this generality this is not particularly specific to the paradigms of dynamics. For example, for the case of flows it is essentially the framework for the local study of nonautonomous ordinary differential equations. What makes it specific is the juxtaposition of its results to compactness of the phase space, which forces recurrence. This can make it possible to use such nonstationary local analysis as a tool for the study of global or semilocal properties. Various aspects of local analysis is discussed later in Section 7.3e, Section 5.2c and Theorem 6.3.1 and also appear in [**S-H, S-BKP**].

e. Foliations and holonomy. A situation somewhat reminiscent of semilocal and local analysis appears in the study of asymptotic properties of foliations of compact manifolds by (in general) noncompact submanifolds.

One approach to studying such foliations is to consider a finite system of transversals $\Gamma = \{\Gamma_i\}_{i=1}^N$ such that for sufficiently large R any R-ball on any leaf intersects at least one of the transversals. Such a system is said to be relatively compact. If $x \in \Gamma_i$ and $y \in \Gamma_j$ are connected by a curve γ in one leaf then we can define a *local holonomy* map $H_{x,y,\gamma}\colon U \to V$ between some neighborhoods $U \subset \Gamma_i$ and $V \subset \Gamma_j$ of x and y, respectively, by mapping $x' \in U$ to a point $y' \in V$ connected to x' by a path γ' close to γ and lying in a leaf. Clearly this local map depends only on the homotopy class of γ, so in the case of contractible leaves it is independent of γ. If one identifies holonomy maps that coincide on the intersection of their domains, one obtains the *holonomy semigroup* with

respect to the natural composition structure. Although different elements of the semigroup have different domains, many ideas from dynamics can be used for studying foliations via this semigroup.

This approach has been particularly successful with respect to *codimension one* foliations, where local transversals are one-dimensional and the holonomy semigroup displays many features of one-dimensional dynamical systems.

On the other hand, the case of foliations with *one-dimensional leaves* is quite similar to the study of flows up to orbit equivalence. If the foliation is *orientable*, one can find a vector field that is tangent to it and hence generates a flow, whose orbits are leaves. All such flows are orbit equivalent. In the nonorientable case one can construct a double cover with a vector field whose orbits project to the leaves of the foliation. A more general situation appears for one-dimensional foliations with singularities. An important special case is discussed in Section 8.4b.

f. Derivative extension and other bundle extensions. A differentiable manifold comes with a variety of natural structures in the form of fibered bundles, which provide invariant (coordinate free) global expressions of various differential operations. The functorial nature of these objects provides for canonical extensions of morphisms (differentiable maps) to these bundles.

The following bundle extensions often appear in dynamics:

(1) the tangent and cotangent bundles of a manifold M, which are denoted by TM and T^*M, with their projectivizations SM (the sphere bundle) and S^*M, and
(2) *tensor bundles*, which are tensor products of various copies of TM and T^*M, some of their subbundles, especially those of contravariant symmetric 2-tensors and contravariant skew-symmetric tensors, and
(3) the *frame bundle*.
All these objects are of first order, *i.e.*, they depend only on the C^1 structure on the manifold.
(4) the jet bundles $J^k(M)$, which are defined for a C^k structure.

All of these objects generate natural extensions of differentiable dynamical systems acting on M.

Of primary importance among these extensions is the *derivative* or *differential* Df of the map $f\colon U \to M$ in the general semilocal setting of Section 5.1c (which, of course, includes the case $U = M$). It is a linear extension of f to T_UM. Other bundle extensions appear naturally in connection with the existence of various invariant structures which cna be identified with sections of certain bundles. Examples are a smooth volume element (mth exterior power of the cotangent bundle, where $m = \dim M$), a Riemannian or pseudo-Riemannian metric (symmetric 2-tensors), or a symplectic form (skew-symmetric 2-forms).

g. Elliptic, parabolic, hyperbolic and partially hyperbolic behavior. We utilize three paradigms to focus the discussion of dynamical systems on classes of systems for which particular phenomena appear and for which special techniques apply. We call these *elliptic*, *parabolic* and *hyperbolic*. Adjoined to the latter is a mixed *partially hyperbolic* case. They are named in analogy with the behavior of linear maps. However, unlike in the linear situation, these three classes do not give an exhaustive description, nor are the

distinctions always unambiguous. The boundary between elliptic and parabolic dynamics cannot be precisely defined in full generality.

The overview of these situations constitutes the content of the three final chapters of this survey. The central point of this analysis is that each type of behavior of the linearized system produces corresponding effects for the nonlinear system, including topological, measure-theoretic and differentiable properties.

1. *Hyperbolic systems.* Linear hyperbolic maps are those with no eigenvalues of absolute value 1. Note that this is an open condition.

For a smooth hyperbolic map f the derivative Df^n grows exponentially in some directions and shrinks exponentially in others. There are no "slow" (subexponential) directions. In a compact phase space, hyperbolicity produces complex patterns of recurrence as well as an abundance of invariant measures and many features of exponential complexity, so long as at least some nontrivial recurrence is present. Positive entropy is among the consequences.

2. *Partially hyperbolic systems.* Linear partially hyperbolic maps are those with a mixture of eigenvalues off and on the unit circle.

Nonlinear partially hyperbolic systems are those with some fast growing/shrinking exponential directions as well as "slow" directions. While linear partially hyperbolic maps are simply those that do not fall into any of the three basic categories, nonlinear partially hyperbolic systems are studied not so much with universality in mind, but with a focus on those dynamical features that are dominated by the presence of exponential behavior. For example, positive metric entropy (and hence topological entropy by Section 4.4d) forces at least partial hyperbolicity, and it is natural to approach such systems with techniques developed for hyperbolic dynamical systems.

3. *Elliptic systems.* Linear elliptic maps are those with all eigenvalues of absolute value 1 and no Jordan block of size two or more.

For nonlinear maps ellipticity can be described by having in mind a certain similarity to being locally isometric. In terms of the derivative this would be the case when $\|Df^n\|$ does not grow with n or exhibits irregular oscillatory behavior of slowly growing magnitude without persistent growth.

4. *Parabolic systems.* Linear parabolic maps are those with all eigenvalues of absolute value 1 but some Jordan blocks of size at least 2.

For nonlinear parabolic maps one has to allow subexponential (usually polynomial) growth of Df^n with n. While the distinction from the elliptic case may not be entirely unambiguous in these terms, the core of the parabolic paradigm is the local "shear" pattern of the orbit structure, as exhibited in the linear case.

h. Prototype examples. Useful nontrivial examples of the three main classes of behavior appear immediately and in a natural way when one takes an affine example and forces recurrence by compactification. In other words, one projects an affine example to a torus. Specific examples are accordingly:

(1) Hyperbolic examples arise from projecting an appropriate linear map to a toral automorphism (Section 6.5a).
(2) Partially hyperbolic examples arise the same way.
(3) Translations of $\mathbb{R}^n$ projected to $\mathbb{T}^n$ give elliptic examples in the form of the translations T_γ (Example 2.1.1, Section 7.1d),

(4) Unipotent affine maps project to parabolic examples, such as the maps $A_{n,\alpha}$ (Section 4.3e) or $A_{L,v}$ (Section 8.3a).

Thus, the basic models of all types of behavior appear in the setting of affine maps of the torus. In all cases under natural assumtions the maps are topologically transitive and ergodic with respect to Lebesgue measure.

Further examples where the local structure of the phase space is not "flat", but the behavior of the derivative is uniform throughout the space appear in the setting of homogeneous dynamics introduced in Section 2.1b and Section 3.3c. See Section 4.3f and Section 6.5e for typical parabolic examples of that kind and Section 6.5e for typical hyperbolic ones.

i. Low-dimensional and conformal dymanics. Since properties of the derivative play the central role in differentiable dynamics it is natural to consider the situation where the derivative has a particularly simple structure. The simplest kind of a linear map between two Euclidean spaces is a *conformal map* which is simply a scalar multiple of an isometry. Correspondingly, a dynamical system acting on a manifold provided with Riemannian metric is called *conformal* if its derivative at every noncritical point is a conformal linear map. Since in dimension one any linear map is conformal this definition includes all differentiable systems on one-dimensional manifolds. Conformality is an extra ingredient which is added to the consequences of the intermediate value theorem to generate specific properties of differentiable dynamical systems in one dimension (*cf.* Section 2.7a). The Denjoy Theorem 5.1.1 is a classical example of such a property. See [**S-JS**] for a detailed overview of differentiable dynamics in one dimension and [**MS**] for an in-depth account.

Conformal systems in real dimension two become holomorphic maps on a one-dimensional complex manifold after introducing an appropriate complex structure (see Section 5.7).

For connected manifolds in dimension higher than two there are few conformal maps and the dynamics of such maps is rather simple. However, conformal actions of some groups such as actions of fundamental groups of compact hyperbolic manifolds on the sphere at infinity, possess interesting dynamical properties. Furthermore, in the semilocal setting there are many nontrivial examples of conformal maps (*e.g.*, expanding ones) in any dimension [**P**, Section 20].

j. Degree of differentiability of smooth dynamical systems. In virtually all situations that arise in differentiable dynamics it is safe and innocuous to assume that the phase space possesses a C^∞ structure, *i.e.*, that in local coordinates one can differentiate as many times as needed. The situation is quite different for the dynamical system itself. Similarly to the rather large gap between general topological dynamical systems, even when acting on a nice phase space such as a differentiable manifold, and differentiable dynamical systems, there are considerable distinctions within the realm of differentiable dynamics according to the degree of differentiability (either the differentiability of the dynamical system or the number of derivatives used to define the topology in a space of dynamical systems). In the subsequent chapters specific instances of these appear, so for now we only make brief general comments.

C^1 regularity is often sufficient for certain topological properties when the derivatives exhibit sufficiently robust behavior [**KH**, Chapter 18]. On the other hand, a considerable

amount of pathology may appear in C^1 systems, in particular in relation to ergodic behavior [**RY, Pu2**]. When one moves from properties of individual systems to those prevalent in various classes of systems, C^1 regularity and the C^1 topology are distinguished from higher regularity by the abundance of positive results. The Closing Lemma (Theorem 5.2.10 [**Pu1**]) and necessary conditions for C^1-structural stability [**S-H, M3**] are outstanding examples. One can express this by saying that there are many C^1 perturbation constructions available that allow to control dynamical properties.

$C^{1+\epsilon}$ for some $\epsilon > 0$ (differentiability with Hölder continuous derivatives) is a standard regularity assumption in aspects of hyperbolic dynamics (both uniform and nonuniform) dealing with behavior with respect to invariant measures and other properties. In fact, in the nonuniformly hyperbolic situation this assumption is needed throughout.

Conditions around C^2 (*e.g.*, bounded variation, absolute continuity, or a Zygmund property of the first derivative) often appear in low-dimensional dynamics out of the need to control the distortion caused by growing numbers of iterates. Many fundamental qualitative results depend on this kind of information [**S-JS**]. Here is a classical example.

THEOREM 5.1.1 (Denjoy Theorem). [**S-JS**], [**KH**, Theorem 12.1.1] *A C^1 circle diffeomorphism without periodic points whose derivative has bounded variation is topologically conjugate to a rotation.*

In this result, one can replace bounded variation by a Zygmund condition [**S-JS**], but not by a Hölder condition [**KH**, Section 12.2].

Varying finite numbers of derivatives (from 3 to a number growing with dimension) are often required in elliptic problems in order to offset "small denominator" effects for the type of return associated with "diophantine" behavior [**S-LL**], [**Ms2**].

Finally, a C^∞ condition appears in a number of remarkable results where polynomial approximation plays an essential role. The solution of the "Entropy Conjecture" [**S-FM, Sh**] is one outstanding example. The existence of a measure with maximal entropy for surface diffeomorphisms is another [**Yd, N1, N2**].

2. Basic concepts and constructions

a. Conjugacy. The natural functorial notion of conjugacy between smooth systems is smooth conjugacy. In semilocal form (which includes global and local ones as special cases) it is defined as follows: For $k \le l \le r$, M, N C^r manifolds, $U \subset M, V \subset N$ open, two C^l maps $f\colon U \to M$, $g\colon V \to N$ with invariant sets
$\Lambda_f \subset U, \Lambda_g \subset V$ are said to be (locally) C^k conjugate if there is an open neighborhood $O \subset g(V)$ of Λ_g and a C^k diffeomorphism $h\colon O \to U$ such that $\Lambda_f \subset h(O)$ and $f \circ h = h \circ g$. The global case is $\Lambda_f = M$, $\Lambda_g = N$.

It would seem particularly appropriate to concentrate on the case $k = l$ of conjugacy in the natural category, and there are important situations where this or the more general case $k \ge l \ge 1$ are useful notions. In general, however, the most tractable notion is topological conjugacy, *i.e.*, the case $k = 0$ (Section 5.2f). The reason is that there are too many invariants of smooth conjugacy, which causes smallness and irregularity of smooth conjugacy classes.

Analogously to smooth conjugacy one can define smooth orbit equivalence (Section 2.2a), smooth factors, and smooth orbit factors (Section 2.2f) by replacing continuity of the conjugating map with differentiability. The distinction between smooth conjugacy

versus smooth orbit equivalence of flows can be expressed by saying that smooth orbit equivalence conjugates the transverse behavior of orbits and is closely related to conjugacy of corresponding section maps. For actions of Lie groups other than $\mathbb{R}$ the study of orbit equivalence is closely related to foliation theory.

b. Equivalence of measures. Any differentiable manifold M carries a natural measure class. This class is represented in particular by any *smooth positive* measure, *i.e.,* a measure which is given by a positive differentiable density in any smooth coordinate system on the manifold. If M is orientable, such a measure is determined by a volume form. In general, a smooth positive measure on an n-dimensional manifold can be identified with an *odd n-form* [**KH**, Section 5.1a]. Suppose μ and ν are two smooth positive measures on the same manifold M (possibly with boundary). If M is not compact assume in addition that $\mu = \nu$ outside a compact set K.

THEOREM 5.2.1. *If $\mu(K) = \nu(K)$ then there exists a compactly supported diffeomorphism $f\colon M \to M$ such that $f_*\mu = \nu$*

In particular if M is compact then for any smooth positive measures μ and ν such that $\mu(M) = \nu(M)$ there exists a diffeomorphism f such that $f_*\mu = \nu$.

This is a more general form of Moser's theorem [**KH**, Theorem 5.1.27]. For the treatment of the case with boundary see [**AK**].

c. Local conjugacy and normal forms. (See also [**KH**, Section 6.6].) In the local setting, where maps are considered in a neighborhood of a single periodic orbit, smooth local conjugacies are more tractable than in the global case. The reason is that usually no nontrivial recurrence appears in this picture and that those C^k conjugacy invariants discussed below are in this case only attached to the single reference orbit.

Though this is a wide subject ([**Bl**] is a good survey), the basic plan for establishing the possibility of smooth conjugacy is easy to outline. Suppose two maps f and g have a common fixed point, which we may take to be $0 \in \mathbb{R}^m$. In order for f and g to be smoothly conjugate in a neighborhood, the differentials at 0 must be conjugate linear maps (by differentiating the conjugacy equation, see Section 5.2d) and may therefore be assumed equal after a linear coordinate change. Furthermore, there are other local invariants of C^k conjugacy associated with the kth jet at the reference orbit. These data are related to coefficients in the kth order Taylor polynomial at the point in question—which of these coefficients play a role is related to eigenvalue data.

The strategy now is to conjugate both f and g to their respective "normal forms" and then to see whether these are equal. The normal form is uniquely defined and collects all smooth local conjugacy invariants. To find the normal form of f at 0 write it as an as yet unknown power series N with the same linear part as f. Solve the conjugacy equation $h \circ f = N \circ h$ for the power series of N and the conjugacy h by taking $Dh\big|_0 = \mathrm{Id}$ and then setting coefficients in N to zero whenever possible while comparing coefficients on either side to inductively determine the coefficients of h.

The simplest normal form is $N = Df$, and usually many higher order coefficients in N can actually be eliminated. But it is possible that for some coefficients there are "accidental" cancellations of the h-coefficients in such a way that the corresponding N-coefficient is uniquely determined by the corresponding one in f. These cancellations arise when products of some eigenvalues coincide with an eigenvalue ("resonances") and they

give terms in the normal form that cannot be removed. Thus, one obtains power series for h and N.

Now consider the case when the map f is analytic. If one can show that both series converge, one has an analytic conjugacy of f to a normal form. This normal form is uniquely defined, so local analytic conjugacy between two maps is achieved by obtaining and comparing normal forms, whose coefficients define higher jet invariants of local analytic conjugacy.

For C^∞ maps one starts from the (possibly divergent) Taylor series and goes through the same formalism to obtain a normal form and formal conjugacy. One then needs to prove that this corresponds to genuine maps. For C^k conjugacies one can go through the same scheme but terminate at terms of order k. Again, one gets a (truncated) normal form containing the invariants associated with the kth jet.

This scheme can be carried out in the general smooth hyperbolic case. In the analytic category the issue of convergence of the normal form and the conjugacy, while tractable, is highly nontrivial. The leading paradigm here as well as in the elliptic case is that of "small denominators". These correspond to "almost cancellations" and usually result in relatively large coefficients in the conjugacy [**Br**].

d. Invariants. Various invariants of topological conjugacy discussed in Chapter 2 are useful in the context of smooth dynamics. It is a great help, for example, that local analysis (Section 5.1d) is available as a tool for the study of relative behavior of orbits including such properties as expansiveness (Section 2.4d).

We mention some invariants of C^k conjugacy for $k \geq 1$. This in particular will help explain why in the global and most semilocal settings this conjugacy notion is usually too narrow to be useful.

e. Periodic eigenvalue data. The simplest smooth conjugacy invariant is referred to as periodic data or the Lyapunov cocycle: If Φ^g and Ψ^g are actions on M that are C^1 conjugate via h, *i.e.*, $\Phi^g \circ h = h \circ \Psi^g$, and if $x \in M$ is such that $\Psi^g(x) = x$ for some $g \in G$, then differentiation gives $D\Phi^g(h(x))Dh(x) = Dh(x)D\Psi^g(x)$. Thus, $D\Phi^g(h(x))$ and $D\Psi^g(x)$ are conjugate as linear maps and in particular they have the same eigenvalues, *etc.* Therefore, under smooth conjugacy, the family of conjugacy classes of differentials over periodic orbits is invariant. As we noted in Section 5.2c, even when the eigenvalue data agree, there are higher jet invariants of C^k conjugacy associated with periodic points. There are several interesting situations where in the global setting coincidence of all periodic eigenvalue data gives (for *global* reasons) coincidence of all higher jet invariants and C^∞ conjugacy. An example is that of area-preserving C^∞ Anosov diffeomorphisms of $\mathbb{T}^2$ (Theorem 6.7.6, [**KH**, Theorem 20.4.3]).

An easy but important observation is that these data can be changed by arbitrarily small C^k-perturbations of a smooth system. Therefore there are no open smooth conjugacy classes containing maps with periodic points. For C^1 conjugacy this statement can be strengthened considerably due to the C^1-closing Lemma of Pugh [**Pu1**], which implies that systems with a periodic point are C^1-dense. Thus there are no C^1 open C^k-conjugacy classes at all (for $k \geq 1$). This issue is discussed in the next subsection.

Still, smooth conjugacy and even some classification up to smooth conjugacy, often on a subset of the phase space, appears in various natural situations. Such situations arise in elliptic dynamics, typically by way of absence of periodic points, Theorem 7.3.5, and in

hyperbolic dynamics, by explicitly or implicitly fixing periodic data (Section 6.7h, [**S-H**], [**KH**, Theorem 20.4.3]). In the elliptic situation, however, it is almost always the case that smooth classification excludes some values of invariants, such as the rotation number for circle maps, or is achieved only on part of the phase space, as is the case in KAM theory, where regions of instability are excluded.

Open C^k conjugacy classes for $k \geq 1$ do appear in noncyclic dynamical systems beginning from actions of $\mathbb{Z}^m$ and $\mathbb{R}^m$ for $m \geq 2$ and even more characteristically for actions of higher-rank semisimple Lie groups and lattices in such groups [**S-FK**].

f. Stability. A dynamical system is said to be *C^k-structurally stable* if it is topologically conjugate to all sufficiently small C^k-perturbations. The case $k = 1$ is of primary importance and is often referred to as just structural stability. For discrete time dynamical systems this is the only natural notion of stability. For continuous time systems, *e.g.*, flows there is a weaker and in fact more natural notion of stability when any perturbed system is topologically orbit equivalent to the unperturbed system, or, equivalently topologically conjugate to a continuous time change of it. It is this weaker notion which is called structural stability for flows. The periods of periodic orbits are typical examples of invariants of topological conjugacy which do not prevent topological orbit equivalence.

C^k-structural stability implies that all topological features of the transverse orbit structure of a dynamical system (described in terms of orbit conjugacy invariants) are impervious to C^k perturbation. A related notion that reflects the point of view of stability of topological properties is *topological stability*: A dynamical system is said to be topologically stable if it is a topological (orbit) factor of any sufficiently small C^0 perturbation. This means, in essence, that all topological features persist under perturbation, although additional complexity might appear.

In the spirit of semilocal analysis there is a related stability notion. A differentiable dynamical system is said to be *Ω-stable* or *NW-stable* if for any sufficiently small C^k perturbation the restrictions of both systems to their nonwandering set (Section 2.3d) are topologically orbit equivalent. The notion is natural because all nontrivial recurrence takes place on the nonwandering set. C^1-structural stability and Ω-stability have been characterized in terms of uniformly hyperbolic behavior of orbits (Section 6.7i, [**S-H, R, M3**]).

A related wide open question is about C^k structural stability for $k \geq 2$. Nothing is known about systems with this property and how much larger these classes are than those of C^1 structurally stable systems. It is unlikely that new phenomena appear, but for $r \geq 2$ a striking lack of constructions of C^r-perturbations with controlled dynamical properties and the failure of the closing lemma to hold in C^r for $r \geq 2$ [**Gt**] (see Section 5.2p) impedes progress in this direction.

g. Invariant manifolds and normal forms. The most natural kind of invariant subsets for differentiable dynamical systems are embedded invariant submanifolds. Even if the phase space is a compact manifold, such a submanifold may not be compact, *e.g.*, an orbit connecting two saddles of a vector field. However, compact and, more generally, complete invariant submanifolds are of primary interest. In the compact case invariant manifolds of the lowest dimesion are periodic orbits for maps and fixed points and periodic orbits for flows. The first natural question concerning an invariant submanifold is its stability under small perturbations of a dynamical system. A natural generalization of stability of a hyperbolic periodic orbit is stability of *transversely hyperbolic* compact invariant manifolds

[**HPS**]. Another case of stability is that of Diophantine invariant tori in Hamiltonian and some other types of systems (Theorem 7.3.6, Section 7.4c).

To describe the behavior near an invariant manifold N the normal forms approach of Section 5.2c can be extended. Its applicability depends both on the dynamics on the invariant manifold and on the properties of the linearized system in the transverse direction. One might expect that when the dynamics on N is sufficiently simple the normal form can be more readily available. One case where this approach is particularly successful is when N is an invariant torus with a Diophantine translation (Section 7.2c).

h. Sections. The concept of sections was introduced in Section 1.2e and that of "restricting" a dynamical system to a noninvariant set via the first-return map in Section 1.3f. In the context of topological dynamics we noted the danger of discontinuities in the first-return map (Section 2.2b), aside from the most basic problem that the induced map may not be defined anywhere. A special choice of subset makes this construction useful for local analysis in smooth flows.

For a nonwandering point p (periodic ones are of special interest) consider a small hypersurface H containing p and transverse to the vector field V generating the flow φ^t, a *transversal*. Then the first-return map ϕ_H to H is defined on an open set in a neighborhood of p. While the continuity problems mentioned earlier do not disappear, they do not affect the possibility of carrying out local analysis near a point of continuity of the first-return map. Moreover, ϕ_H is smooth at continuity points (by transversality and the implicit function theorem). Making convenient choices can be helpful: If the initial orbit is periodic then for a sufficiently small transversal the first return time of this orbit is the period and the orbit do not intersect the boundary of the transversal. Then the return map is a local diffeomorphism in a neighborhood of its fixed point p and is amenable to local analysis, which in turn gives information about the behavior of φ^t near the orbit od p. For example, determining asymptotic stability of a periodic orbit can be carried out this way [**R**, Theorem 8.4].

i. Inverse limits. The construction of inverse limits, introduced in Section 1.3i and described for topological dynamical systems in Section 2.2g, can be applied in the setting of smooth dynamics. Evidently this construction is purely topological in that the inverse limit space usually does not have any smooth structure. On the other hand, for the natural extension (Section 2.2h) some smoothness can often be salvaged by embedding the inverse limit space into a bona fide smooth system as an attractor [**Wl**].

As an example, one may consider the covering $E_2\colon S^1 \to S^1$, $x \mapsto 2x \pmod 1$. The natural extension is obtained as the limit of twofold, fourfold, eightfold... covers of S^1 and can be seen in various ways to have a Cantor structure. It is locally the product of a Cantor set with an interval. It can be represented as the attractor of the embedding $f\colon M \to M := S^1 \times D^2$, where D^2 is the unit disk in $\mathbb{R}^2$, given by $f(\varphi, x, y) = \Big(2\varphi, \frac{1}{10}x + \frac{1}{2}\cos\varphi, \frac{1}{10}y + \frac{1}{2}\sin\varphi\Big)$, where $\varphi \in S^1$ and $(x, y) \in D^2$, *i.e.*, $x^2 + y^2 \le 1$ [**KH**, Section 17.1]. This is known as the *Smale attractor*, *solenoid*, or a map *derived from expanding (DE)* [**Sm**] (Figure 6.4, p. 140).

j. Suspensions. The suspension construction described in Section 1.3j can be carried out for topological dynamical systems (Section 2.2j) and hence for smooth systems. It can

be performed in such a way as to retain the smooth structure. For example, this construction can be used to obtain a smooth action of a Lie group from a smooth action of a cocompact discrete subgroup (lattice).

k. Cocycles and extensions. Cocycles were introduced in Section 1.3k and can be studied in the topological context (Section 2.2k), hence also in the smooth one. In this setting various cocycles associated with an action appear quite naturally, usually easily expressed in terms of canonical extensions to various bundles. Probably the most immediate example is the differential itself. There are also several other constructs related to the differential that have a multiplicative nature. An important example is the *Jacobian* Jf with respect to a volume form Ω on the manifold defined by $Jf\Omega = f^*\Omega$ (pullback), which is a measure of the volume distortion by f as measured in terms of Ω. That this is a cocycle is immediate. It is useful for the study of invariant measures.

l. Isometric extensions. Isometric extensions of differentiable dynamical systems appear naturally in various problems coming from geometry, mechanics and number theory. Isometric extensions of hyperbolic systems constitute a basic class of examples of partially hyperbolic systems. If the orbit structure of the system in the base is sufficently robust, isometric extensions often possess a rather regular orbit structure, *e.g.*, their ergodic components are smooth subbundles over the base. On the other hand, among such extensions over systems with more fragile orbit structure one finds examples of nonstandard behavior (such as minimal but not uniquely ergodic).

m. Smooth invariant measures. As mentioned in Section 5.1a, the smooth measure class is invariant under smooth dynamical systems, but ascertaining the existence of an invariant absolutely continuous measure requires some effort. Analogously to the unit determinant criterion for volume preservation in $\mathbb{R}^n$, a volume form Ω on a manifold is f-invariant if the *Jacobian* Jf of f with respect to Ω is (identically) 1. Failing this, one looks for an invariant measure of the form $\rho\Omega$ for a nonnegative density function $\rho\colon M \to \mathbb{R}$. Invariance of $\rho\Omega$ means $\int_A \rho\Omega = \int_{f^{-1}(A)} \rho \circ f\Omega$ or $Jf\rho \circ f = \rho$.

In the noninvertible case one needs to consider all preimages of a point and thus requires $\rho(x) = \sum_{y\in f^{-1}(\{x\})} \dfrac{\rho(y)}{Jf(y)}$ for all $x \in M$.

A useful point of view is that existence of a smooth invariant measure for a nonsingular map is equivalent to existence of a fixed point of the *Perron–Frobenius operator*

$$(\mathcal{F}\rho)(x) = \sum_{y\in f^{-1}(\{x\})} \frac{\rho(y)}{Jf(y)}$$

on nonnegative measurable functions. This is a particular case of the general situation described in Section 3.3b

In the case of a diffeomorphism an obvious necessary criterion for existence of an invariant continuous positive density ρ is boundedness of $\{Jf^n(x) \mid n \in \mathbb{N}, x \in M\}$: $Jf^n(x) = \dfrac{\rho(x)}{\rho(f^n(x))} \le \dfrac{\max_{x\in M} \rho(x)}{\min_{x\in M} \rho(x)}$.

This criterion is actually more than sufficient for existence of an invariant measure in the Lebesgue class:

PROPOSITION 5.2.2. *Suppose $f\colon M \to M$ is an orientation-preserving diffeomorphism, Ω a volume form. If $\{Jf^n(x) \mid n \in \mathbb{Z}\}$ is bounded for almost every $x \in M$ then there is a Borel function $\omega\colon M \to \mathbb{R}_+$ such that $\omega \geq 1/Jf$ and $\omega\Omega$ is f-invariant. If $\{Jf^n(x) \mid n \in \mathbb{Z}, x \in M\}$ is bounded then ω is bounded.*

PROOF. The solution $\varphi := \sup_{n\in\mathbb{N}} \sum_{i=0}^{n} \Phi \circ f^{-i} \geq \Phi := -\log Jf$ of the cohomological equation $\varphi \circ f^{-1} - \varphi = -\Phi$, is a well-defined almost everywhere finite Borel function. Write $\omega := e^{\varphi} \geq 1/Jf$ to get

$$f^*\omega\Omega - \omega\Omega = e^{\varphi\circ f^{-1}}(Jf)^{-1}\Omega - e^{\varphi}\Omega = (e^{\varphi\circ f^{-1}}e^{\Phi} - e^{\varphi})\Omega = 0.$$

□

Note that the boundedness criterion implies in particular that $Jf^n = 1$ on $\mathrm{Fix}(f^n)$. In the case of hyperbolic (Anosov) dynamical systems, this control of periodic data is sufficient [**S-H**], [**KH**, Theorem 19.2.7]. On the other hand in that case the above periodic conditions $Jf^n = 1$ happen to be necessary for existence of even an absolutely continuous invariant measure. Since any finite collection among these conditions is independent one concludes that among Anosov systems (which are open in the space of diffeomorphisms) those with an absolutely continuous invariant measure form a closed submanifold of infinite codimension.

For noninvertible maps one can formulate analogous criteria, but the necessity of successively tracking the possibilities of multiple preimages renders the description and the arguments more involved. (See [**KH**, Section 5.1c,d] for the most basic example.)

n. Invariant distributions. Since smooth dynamical systems on a space X preserve the class of smooth functions, one can look for invariant distributions on $C^k(X)$, where $0 < k \leq \infty$ (k is then the *order* of the distribution). Other spaces of functions, such as Hölder continuous or Sobolev, may also be considered.

1. *Invariant distributions determined by measures.* Invariant measures are evidently a special case corresponding to $k = 0$, and there are also distributions that are obtained as limits of linear combinations of invariant measures in the corresponding distribution topologies, which are weaker than the weak*-topology for measures. We call such distributions *invariant distributions determined by measures*. Often, all invariant distributions are of this kind.

2. *Uniqueness of invariant distribution.* On the one end of the complexity scale are minimal translations and linear flows on the torus where Lebesgue measure is the only invariant distribution. This can be proven by looking at the Fourier coefficients of a distribution, *i.e.*, its values on the characters, and verifying that invariance under a minimal translation forces all but one of these to vanish.

3. *Hyperbolic systems.* On the other end of the complexity scale are various kinds of hyperbolic systems, where the totality of invariant measures is quite rich (Section 6.7c). Still, due to the Livschitz Theorem 6.7.2 for a locally maximal hyperbolic set, all invariant distributions on the spaces of Hölder and smooth functions are determined by atomic measures on periodic orbits.

4. *Distributions not determined by measures.* In the parabolic situation, however, there usually are invariant distributions that are not determined by measures.

EXAMPLE 5.2.3. For a circle diffeomorphism with a single parabolic fixed point at 0, such as $x \mapsto x + \frac{1}{10}\sin^2 \pi x \pmod 1$, the distribution $D\colon C^1(S^1) \to \mathbb{R}$, $f \mapsto f'(0)$ is not determined by a measure. (This is the same example as Example 2.4.2, which also appeared in Section 4.3b.)

More interesting manifestations of this phenomenon appear in Section 8.2g, Section 8.4f and Section 8.3b6.

5. *Distributions and cocycle stability.* If the space of functions is a Banach space then the common kernel of invariant distributions is the closure of the space of coboundaries: The latter is clearly in the kernel, but the Hahn–Banach Theorem gives equality.

Given a dynamical system and a class of functions (*e.g.*, Hölder, smooth, Sobolev, or analytic) on the phase space, a natural notion is *cocycle stability*. It can be rather vaguely described as follows: Every function from the given class, on which all invariant distributions vanish, is a coboundary with a transfer function of controllable, though possibly lower, regularity.

Stability for C^0 cocycles is impossible except when the space consists of finitely many points [**K7**]. On the other hand, smooth cocycle stability is shared by such diverse classes of systems as Diophantine translations on the torus (Proposition 7.3.2), locally maximal hyperbolic sets (Theorem 6.7.2), parabolic systems, such as many affine unipotent maps on the torus and flows on higher genus surfaces (Section 8.4f), and partially hyperbolic systems, such as ergodic automorphisms of the torus [**V1**].

o. Transversality and Kupka–Smale theorem. The transversality notion from differential topology has natural applications to smooth dynamics. The most basic one is to periodic points.

DEFINITION 5.2.4. [**KH**, Chapter 7] Let M be a smooth manifold and K, $N \subset M$ smooth submanifolds. K and N are said to be *transverse* at $x \in M$ if $x \notin K \cap N$ or $T_xK + T_xN = T_xM$. We write $K \pitchfork_x N$. In particular, if $\dim K + \dim N = \dim M$ and $x \in K \cap M$ the latter condition is equivalent to $T_xK \cap T_xN = \{0\}$.

We say that K and N are transverse (to each other), written $K \pitchfork N$, if $K \pitchfork_x N$ for all $x \in K \cap N$. Manifolds K and M with boundary are said to be transverse if ∂K, $K \smallsetminus \partial K$, ∂M, $M \smallsetminus \partial M$ are pairwise transverse.

Let $0 \le r \le \infty$ and M a C^r manifold. Two submanifolds K_1 and K_2 of M are said to be C^r-close if there exist a C^r manifold K_0 and C^r embeddings $f_i\colon K_0 \to K_i$ such that f_1 and f_2 are C^r-close.

A fixed point $p = f(p)$ of a smooth map $f\colon M \to M$ is said to be *transverse* if $\operatorname{graph} f \pitchfork_{(p,p)} \Delta$ in $M \times M$, where Δ is the diagonal, *i.e.*, if 1 is not an eigenvalue of Df_p.

If K and M are compact transverse manifolds (possibly with boundary) then any sufficiently small C^1-perturbations $\tilde K$ and $\tilde M$ are transverse.

Transversality of periodic points implies persistence of such periodic points under C^1 perturbations. Furthermore, transverse periodic points can be easily perturbed to hyperbolic ones.

The notion of transversality as such is of interest because of its innate persistence and prevalence:

THEOREM 5.2.5 (Transversality Theorem). [**KH**, Theorem A.3.20] *Let M be a C^∞ manifold of dimension m, and $N \subset M$ a submanifold of dimension n. Then among the k-dimensional submanifolds $K \subset M$, those transverse to N are C^∞-dense.*

This fact is the major ingredient for the genericity result below.

Kupka–Smale diffeomorphisms are diffeomorphisms with only hyperbolic periodic points and a transversality condition involving their stable and unstable manifolds. These manifolds are discussed in detail in Chapter 6. They are the sets of points positively and negatively asymptotic (respectively) to the periodic orbit, and they are injectively immersed disks tangent to the contracting and expanding subspaces of the linearization correspondingly.

DEFINITION 5.2.6. Suppose M is a C^k manifold, $f \in \mathrm{Diff}^k(M)$. f is said to be *Kupka–Smale to order n* (with respect to a given Riemannian metric) if all periodic points of f of period at most n are hyperbolic and the ball of radius n in the stable manifold of any $x \in \mathrm{Fix}\, f^n$ is transverse to the ball of radius n in the unstable manifold of any $y \in \mathrm{Fix}\, f^n$. f is called a *Kupka–Smale diffeomorphism* if it is Kupka–Smale to all orders.

THEOREM 5.2.7 (Kupka–Smale Theorem). *Let $0 < r \leq k \leq \infty$ and M a σ-compact C^k manifold. Then for any $n \in \mathbb{N}$, Kupka–Smale diffeomorphisms of order n are a C^r-dense C^1-open set in $\mathrm{Diff}^k(M)$ and hence Kupka–Smale diffeomorphisms are a C^r-dense C^1-G_δ set in $\mathrm{Diff}(M)$.*

This also works for flows, with the appropriate changes.

DEFINITION 5.2.8. A fixed point p of a local flow is said to be *transverse* if the differential at p of any time-t map for $t \neq 0$ does not have 1 as an eigenvalue. Equivalently, the linear part of the vector field at p does not have 0 as an eigenvalue.

A periodic point p of period $t > 0$ for a flow is said to be *transverse* if 1 is a simple eigenvalue of the differential at p of the time-t map of the flow. Equivalently, p is a transverse fixed point for the Poincaré map on a transversal to the flow near p.

A smooth flow is said to be a *Kupka–Smale flow to order t* if all fixed points and all periodic orbits of period less than t are hyperbolic and the t-balls in their stable and unstable manifolds are pairwise transverse. It is called a *Kupka–Smale flow* if it is a Kupka–Smale flow to order t for all $t > 0$.

THEOREM 5.2.9 (Kupka–Smale Theorem). *Let $0 < r \leq k \leq \infty$ and M a compact C^k manifold. Then for any $t > 0$, Kupka–Smale flows of order t are a C^r-dense C^1-open set and hence Kupka–Smale flows are a C^r-dense C^1-G_δ set in the space of C^r flows.*

p. Persistence of recurrence and closing lemma. For an individual dynamical system there are no connections between different kinds of recurrent behavior except for the general relations discussed in the context of topological dynamics (Section 2.2). However, one may ask whether a certain kind of recurrence present in a system persists under perturbations, or whether there are perturbations that produce a stronger type of recurrence. A particularly natural question is whether an orbit that almost returns to its initial position can be approximated by a closed orbit of either the system itself or of a perturbed system. Assertions that such things are possible are known as *closing lemmas*.

The closing of orbits for a given system is possible under certain conditions of a generally hyperbolic type. In the uniformly hyperbolic case there is the Anosov closing

lemma (Theorem 6.6.1). In the nonuniformly hyperbolic situation one can apply the closing lemma for regular orbits [**KH**, Theorem S.4.13] and the Ergodic Closing lemma by Mañé [**M2**]. All of these statements are sophisticated cousins of the fairly simple fact that a nonlinear map with hyperbolic linear part and an almost fixed point has in fact a fixed point nearby (Theorem 6.1.1). In all of these closing lemmas the existence statement is accompanied by an exponential estimate of closeness of the periodic orbit to the original orbit segment.

In general, one cannot expect to be able to close an orbit segment by a perturbation of the initial condition only, as irrational circle rotations show. Thus, a perturbation of the system may be needed. The existence of such perturbations, coupled with the persistence of the periodic points thus obtained (since every periodic point can be perturbed further to become transverse and hence persistent), are the base of various genericity results that go beyond Kupka–Smale type theorems. In order to prove a closing lemma one needs to develop a general construction of perturbations that allows to control the properties of long orbit segments. The crucial problem is that such a segment contains points that are close in the phase space but separated by long time intervals. The fewer points on the segment are affected by the perturbation, the better are the chances of controlling the outcome. Thus, an effective construction would include sharply defined shifts in certain places and not disturbing others. While such perturbations can be easily made small in C^0 topology their derivatives would tend to be large. As it turns out, a more sophisticated inductive approach leads to a construction of C^1-small perturbations with effectively controlled returns. The result is known as the Pugh Closing Lemma

THEOREM 5.2.10. [**Pu1**] *Let f be a C^r diffeomorphism of a compact differentiable manifold, $1 \leq r \leq \infty$, and $x \in NW(f)$. Then there exists a C^r diffeomorphism g arbitrarily close to f in the C^1 topology, for which x is a periodic point. Furthermore, given an open set $U \ni x$ one can choose g in such a way that $g = f$ outside U.*

Naturally the period of x with respect to the perturbed map g depends on (and in general grows with) the closeness of g to f and the smallness of the set U.

An important application of the Pugh closing lemma is the following genericity result.

THEOREM 5.2.11. [**Pu1**] *In the space of C^1 diffeomorphisms of a compact differentiable manifold there is a dense G_δ subset of Kupka-Smale diffeomorphisms for which periodic points are dense in the set of nonwandering points.*

As was mentioned above, in this form the closing lemma is not true for C^2 small perturbations. In Gutierrez's examples [**Gt**] there is a special point x_0 for a diffeomorphism f of $\mathbb{T}^2$ such that a C^2-small perturbation cannot have periodic points near a nonwandering point $x \neq x_0$ if it agrees with f on a neighborhood of x_0 or even if $f - g$ vanishes to sufficiently high order at x_0. It is not known however, whether a closing lemma in higher regularity holds without the localization condition, *i.e.*, whether for any $r > 1$ there are any C^r-small perturbations of f that make a nonwandering point periodic.

The Mañé Ergodic Closing Lemma is a further development that was obtained in the course of Mañé's efforts to prove the Stability Conjecture, which characterizes structural stability by hyperbolic behavior. It is of independent interest:

THEOREM 5.2.12. *Let f be a C^1 diffeomorphism of a compact manifold. For every f-invariant Borel probability measure almost every $x \in M$ can be closed by approximation,*

i.e., *for every C^1 neighborhood U of f and all $\epsilon > 0$ there are an $m \in \mathbb{N}$, a $g \in U$ and a $y \in \mathrm{Fix}(g^m)$ such that $\max_{0\le i\le m} d(f^i(x), g^i(y)) \le \epsilon$ and $g = f$ outside an ϵ-neighborhood of $\mathcal{O}_f(x)$.*

Combining both of these closing results, Hayashi proved his Connecting Lemma, which properly belongs to the hyperbolic setting [**S-H**]. It is the central ingredient for proving the stability conjecture for flows.

3. Hamiltonian dynamics

Hamiltonian systems are classical objects in the theory of dynamical systems. They naturally preserve volume. A detailed presentation of Hamiltonian and Lagrangian mechanics is given in [**AM**]; [**HZ**] presents new aspects in Hamiltonian dynamics closely connected with the modern symplectic geometry.

a. Linear symplectic geometry. A 2-tensor $\alpha\colon E \times E \to \mathbb{R}$ on a Euclidean space E is said to be *nondegenerate* if $\alpha^\flat\colon E \to E^*$, $v \mapsto \alpha(v, \cdot)$ is an isomorphism. It is said to be *antisymmetric* or *skew-symmetric* if $\alpha(v, w) = -\alpha(w, v)$. An antisymmetric 2-form is nondegenerate if and only if E is even-dimensional and the nth exterior power α^n is not zero. A nondegenerate antisymmetric 2-form is called a *symplectic* form and a linear space with a symplectic form, a *symplectic* linear space. The volume form α^n is determined by a symplectic form. If (E, α), (F, β) are symplectic linear spaces then a map $T\colon E \to F$ is said to be *symplectic* if $T^*\beta = \alpha$.

Symplectic maps preserve volume and orientation and hence is invertible with Jacobian 1, so the set of symplectic maps $(E, \alpha) \to (E, \alpha)$ is a group, called the *symplectic group* of (E, α). A subspace V of a symplectic linear space (E, α) is said to be *isotropic* if $\alpha_{\restriction V} = 0$. An isotropic subspace has dimension at most $n = \dim E/2$, in which case it is said to be *Lagrangian*.

If a scalar product $\langle \cdot, \cdot \rangle$ on E is fixed then α has a matrix representation A with $\alpha(\cdot, \cdot) = \langle \cdot, A\cdot \rangle$. If α is a symplectic form then $\dim E = 2n$ for some $n \in \mathbb{N}$ and there is a basis $e_1, \dots, e_{2n}$ of E such that $\alpha(e_i, e_{n+i}) = 1$ if $i = 1, \dots, n$ and $\alpha(e_i, e_j) = 0$ if $|i - j| \ne n$. Hence, if one fixes a scalar product with respect to which $e_1, \dots, e_{2n}$ is an orthonormal basis, then $A = J := \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$ with respect to this basis, where I is the $n \times n$ identity matrix. Thus this "adapted" basis above gives a decomposition of E as a direct sum of two Lagrangian subspaces. In particular J defines the canonical symplectic form on $\mathbb{R}^{2n}$. Alternatively, identify $\mathbb{R}^{2n}$ with $\mathbb{C}^n$ and take the imaginary part of the standard Hermitian inner product.

If $T\colon (E, \alpha) \to (F, \beta)$ is a symplectic map and λ is an eigenvalue of T then so are $\bar\lambda$, $1/\lambda$, $1/\bar\lambda$. If T has the form $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ with respect to a basis for which $\alpha(v, w) = \langle v, Jw \rangle$ then A^tC and B^tD are symmetric and $A^tD - C^tB = I$.

b. Symplectic geometry. Let M be a smooth manifold. A differential 2-form ω is said to be *nondegenerate* if it is nondegenerate at every point. If ω is nondegenerate with $d\omega = 0$ then ω is said to be a *symplectic form* and (M, ω) a *symplectic manifold*. Then $[\![X, Y]\!] := \omega(X, Y)$ is called the *Lagrange bracket* of X and Y, and a subbundle of TM is said to be *isotropic* (*Lagrangian*) if at every point $p \in M$ it defines an isotropic

(Lagrangian) subspace of T_pM. A smooth submanifold is said to be *isotropic* (*Lagrangian*) if its tangent bundle is isotropic (Lagrangian).

DEFINITION 5.3.1. A diffeomorphism $f\colon (M,\omega) \to (N,\eta)$ between symplectic manifolds with $f^*\eta = \omega$ is a *symplectic diffeomorphism* or *symplectomorphism*. If $(M,\omega) = (N,\eta)$ it is also called a *canonical transformation*.

A differentiable embedding (*i.e.,* injective nonsingular map) $f\colon (M,\omega) \to (N,\eta)$ such that $f^*\eta = \omega$ is called a *symplectic embedding*.

If (M,ω) is a symplectic manifold then M is even-dimensional and ω^n is a volume form. In particular M is orientable.

Unlike in the case of a Riemannian metric, it is possible to find a local chart such that the symplectic form is in standard form at *every* point of the chart:

THEOREM 5.3.2 (Darboux, [**KH**, Theorem 5.5.9]). *Let (M,ω) be a symplectic manifold. For every point $x \in M$ there exists a neighborhood U of x and coordinates $\varphi\colon U \to \mathbb{R}^{2n}$, referred to as* Darboux *or* symplectic coordinates*, such that ω is in standard form $\sum_{i=1}^n dx_i \wedge dx_{i+n}$ with respect to the basis $\left\{\partial/\partial x_1, \dots, \partial/\partial x_{2n}\right\}$ at every point $y \in U$.*

Thus there are no *local* invariants for symplectic diffeomorphisms or symplectic embeddings.

c. Examples of symplectic manifolds. The simplest example is $\mathbb{R}^{2n}$ with the standard symplectic form $\sum_{i=1}^n dx_i \wedge dx_{i+n}$.

1. *Tori.* Since the standard symplectic form is invariant under translations it can be projected to any torus $\mathbb{R}^{2n}/\Gamma$, where Γ is a lattice. However, on the torus the Darboux Theorem is not true globally since there is an invariant, the *cohomology class* of $\omega \in H^2(\mathbb{T}^{2n},\mathbb{R})$. For example, for different $\alpha = (\alpha_1,\dots,\alpha_n) \in \mathbb{R}^n$ the forms $\omega_\alpha = \sum \alpha_i\, dx_i \wedge dx_{i+n}$ belong to different cohomology classes.

2. *Cotangent bundles.* For any differentiable manifold M the cotangent bundle T^*M possesses a canonically defined symplectic form ω, which is furthermore exact, *i.e.,* $\omega = d\theta$, where the *Poincaré–Cartan* form θ is canonically defined as well: An element of TT^*M can be viewed as a pair $(v,w) \in T_x \times T_x^*M$. Then $\omega(v,w) = w(v)$. In local coordinates, if $(q_1,\dots,q_n)$ are coordinates on M and $(p_1,\dots,p_n)$ the corresponding coordinates in T^*M with respect to the basis $(dq_1,\dots,dq_n)$ then $\theta = \sum_{i=1}^n p_i\, dq_i$ and $\omega = d\theta = \sum_{i=1}^n dp_i \wedge dq_i$. This class of symplectic manifolds plays a central role in Lagrangian dynamics and hence in classical mechanics.

3. *Kähler manifolds.* Another important class of symplectic manifolds is related to complex geometry. As was pointed out in Section 5.3a, the imaginary part of a Hermitian form is a symplectic form. Thus, if one considers a complex n-dimensional manifold M with Hermitian metric as a $2n$-dimensional real manifold N, then this defines a nondegenerate two-form on TM. This form is closed if and only if the Hermitian metric is *Kähler*. Thus any Kähler manifold is a symplectic manifold. The simplest compact examples are Riemann surfaces (one-dimensional complex manifolds) with any Hermitian metric, and complex projective space $\mathbb{C}P(n)$ with the symmetric metric.

d. Hamiltonian vector fields and flows. Let (M,ω) be a symplectic manifold, and $H\colon M \to \mathbb{R}$ a smooth function. Then the vector field $X_H = dH^\#$ defined by $\omega \lrcorner X_H =$

dH is called the *Hamiltonian vector field* associated with H or the *symplectic gradient* of H. The flow φ^t with $\dot\varphi^t = X_H$ is called the *Hamiltonian flow* of the *Hamiltonian* H.

Thus, one can identify the space of C^r Hamiltonian flows, which is a closed linear subspace of $\Gamma^r(TM)$, with the space $C^{r+1}(M,\mathbb{R})$.

Hamiltonian flows are actions by symplectomorphisms and hence preserve volume:

$$\frac{d}{dt}\varphi^{t*}\omega = \varphi^{t*}(\pounds_{X_H}\omega) = \varphi^{t*}(d(\omega\lrcorner X_H) + (d\omega\lrcorner X_H))$$
$$= \varphi^{t*}(d(\omega\lrcorner X_H)) = \varphi^{t*}(ddH) = 0.$$

On the other hand, there are symplectic flows that are not Hamiltonian. A flow generated by a vector field v preserves the symplectic form ω if and only if

$$0 = \pounds_v\omega = d\omega\lrcorner v + d(\omega\lrcorner v) = d(\omega\lrcorner v),$$

hence, if and only if $\omega\lrcorner v$ is closed, whereas, if v generates a Hamiltonian flow then $\omega\lrcorner v$ is exact. Thus, if $H^1(M,\mathbb{R}) = 0$ then any symplectic flow is Hamiltonian, while otherwise there are symplectic non-Hamiltonian flows. A simple example is given by a linear flow on the two-dimensional torus with the standard volume 2-form $dx\wedge dy$. It preserves area and is hence symplectic. Its velocity vector field is constant $\neq 0$, so the Hamiltonian would have to have constant nonzero gradient, which is false at its maximum. Equivalently, the form $\omega\lrcorner v$ for such a flow has constant coefficients and is closed but not exact.

An important observation is that H is constant along orbits of the Hamiltonian vector field X_H (preservation of energy):

$$\pounds_{X_H}H = dH\lrcorner X_H = \omega\lrcorner X_H\lrcorner X_H = 0.$$

See also Section 5.3f.

Notice that if two Hamiltonian fuctions have a common level surface then the corresponding Hamiltonian vector fields *on that surface* are collinear and hence the Hamiltonian flows on that surface are obtained by a time change from each other.

Hamiltonian flows arise in classical mechanics, where such systems are described by the usual Hamiltonian equations

$$\dot q_i = \frac{\partial H}{\partial p_i}, \qquad \dot p_i = -\frac{\partial H}{\partial q_i}. \tag{5.1}$$

This is indeed a local representation of the Hamiltonian differential equation because the vector field $X_H := (\dfrac{\partial H}{\partial p_i}, -\dfrac{\partial H}{\partial q_i})$ satisfies $\omega\lrcorner X_H = dH$ in Darboux (symplectic) coordinates.

e. Symplectic invariants. For closed compact $2n$-dimensional symplectic manifolds the cohomology class of the symplectic form is the most obvious invariant. In particular, this class determine the cohomology class of the nth exterior power of the symplectic form,*i.e.,* the total volume $v(M)$.

The volume, however, is an invariant for arbitrary symplectic manifolds, compact or not, with or without boundary. In the noncompact case the volume may be infinite. Furthermore the volume is *monotone, i.e.,* if $f\colon (M,\omega)\to(N,\eta)$ is a symplectic embedding then $v(N)\le v(M)$.

By Theorem 5.2.1 the volume and Euler characteristic form a complete set of invariants of 2-dimensional compact symplectic manifolds with respect to a symplectic diffeomorphism. For $n \geq 2$ there are further nontrivial monotone symplectic invariants of $2n$-dimensional symplectic manifolds called *symplectic capacities* [**HZ**]. Let $B(r)$ be the open ball or radius r in $\mathbb{R}^{2n}$ with the standard symplectic form $\omega_0 = \sum_{i=1}^n x_i \wedge x_{i+n}$ and $Z(r)$ the cylinder $\{(x, \dots, x_{2n}) \in \mathbb{R}^{2n} \mid x_1^2 + x_{n+1}^2 < 1\}$ with the same form.

DEFINITION 5.3.3. A function c defined on the class of symplectic manifolds of dimension $2n$, possibly with boundary, with values in $\mathbb{R}_+ \cup \infty$ is called a *symplectic capacity* if it is monotone, $c(M, \alpha\omega) = |\alpha| c(M, \omega)$ for any $\alpha \in \mathbb{R} \smallsetminus \{0\}$ and

$$c(B(1), \omega_0) = \pi = c(Z(1), \omega_0).$$

An example of symplectic capacity is *the Gromov width* of (M, ω), which is defined as the supremum of the values of πr^2 for such r that the ball $B(r)$ can be symplectically embedded into (M, ω) [**HZ**].

The existence of capacities lies at the root of global rigidity properties of symplectic structures and maps. Capacities also play the central role in the number of powerful variational results about the existence of periodic orbits for Hamiltonian systems [**S-HZ**].

f. Poisson brackets. On a symplectic manifold (M, ω) the *Poisson bracket* of $f, g \colon M \to \mathbb{R}$ is defined by $\{f, g\} := [\![X_f, X_g]\!] = df(X_g)$. In Darboux coordinates $\{f, g\} = \sum_{i=1}^n \Big(\dfrac{\partial f}{\partial q_i} \dfrac{\partial g}{\partial p_i} - \dfrac{\partial f}{\partial p_i} \dfrac{\partial g}{\partial q_i} \Big)$. f, g are said to be *in involution* if $\{f, g\} = 0$.

Important pertinent facts are [**AM, KH**]

(1) $(C^\infty(M), \{\cdot, \cdot\})$ is a Lie algebra, *i.e.*, $\{\{f, g\}, h\} + \{\{g, h\}, f\} + \{\{h, f\}, g\} = 0$ (Jacobi identity).
(2) f is an integral of H, *i.e.*, invariant under the Hamiltonian flow of H, if and only if $\{f, H\} = 0$; in particular H is constant on orbits.
(3) If f, g are integrals of H then so is $\{f, g\}$ (by the Jacobi identity).

The last property suggests a way of finding new integrals once several are known. This may help at times, but often one only obtains integrals that are functions of known ones.

g. The Noether Theorem. Using Poisson brackets it is easy to obtain the following result that symmetries produce integrals: If H is invariant under a one-parameter family of symplectic transformations generated by a Hamiltonian f, then f is an integral of H (because both statements are equivalent to $\{f, H\} = 0$).

This is easy to use when the phase space is a cotangent bundle and the symmetries come from diffeomorphisms in the base. Standard applications are that translation invariance of H implies constant velocity of the center of mass; rotation invariance gives constant angular momentum [**KH**, Section 5.5d].

h. Completely integrable systems, the Liouville–Arnold Theorem. Suppose (M, ω) is a $2n$-dimensional symplectic manifold, $H = f_1, f_2, \dots, f_n \in C^\infty(M)$, $\{f_i, f_j\} = 0$ $(i, j = 1, \dots, n)$, and $x \in M$ is such that the differentials Df_i are (pointwise) linearly independent on $M_z := \{x \in M \mid f_i(x) = z_i, i = 1, \dots, n\}$, *i.e.*, there are n independent integrals in involution. Then in a neighborhood of M_z one can find a symplectic change

of coordinates to *action-angle coordinates* $(y_1, \dots, y_n, \varphi_1, \dots, \varphi_n)$ such that H depends only on $(y_1, \dots, y_n)$ [**AM**, Section 5.2].

This implies also that whenever M_z is compact, the given coordinate neighborhood is foliated by invariant tori, on each of which the flow is linear.

The key idea of the proof is to use the first integrals in involution as Hamiltoninans whose vector fields generate an $\mathbb{R}^n$ action on their common level manifolds. This manifolds are Lagrangian. The action is transitive on every connected component. If any of these manifolds is compact then the stationary subgroup of a point is a lattice, hence the manifold is a torus. The φ coordinates come from parametrization of the level manifolds, and the y coordinates from a proper combination of the first integrals.

Such systems are said to be *completely integrable* because explicit formulas for their solutions can be found (the equations of motion can be integrated) by quadrature, *i.e.*, in terms of roots of inverses of antiderivatives [**An1**].

Since Hamiltonian systems preserve volume this result is also interesting from the point of view of ergodic theory: The invariant manifolds described here capture much of the ergodic decomposition of volume, because (in the generic case of nondegenerate frequency function) the flow is an irrational linear flow on almost every invariant torus, hence ergodic with respect to the (preserved) volume form on that torus. The few tori with nontransitive flows can be relatively easily analyzed. The point is that the ergodic decomposition, which in general consists of a partition into complicated sets, is presented here in a smooth fashion with maximally regular conditional measures. We return to this discussion in Chapter 7.

4. Lagrangian systems

a. The Euler–Lagrange equation. A *Lagrangian dynamical system* on a manifold M, the *configuration space*, has *phase space* TM and the system is determined by the *Lagrangian* $L\colon TM \to \mathbb{R}$ via the *Lagrange equation* or *Euler–Lagrange equation*

$$\frac{d}{dt}\frac{\partial L}{\partial v} = \frac{\partial L}{\partial x}. \tag{5.1}$$

The form is independent of the local coordinate chart: If $x = f(y)$ then $v = \dot{x} = Df\dot{y} = Dfw$ and

$$\frac{\partial L}{\partial y} = \frac{\partial L}{\partial x}\frac{\partial x}{\partial y} + \frac{\partial L}{\partial v}\frac{\partial v}{\partial y}, \quad \frac{\partial L}{\partial w} = \frac{\partial L}{\partial v}\frac{\partial v}{\partial w} + \frac{\partial L}{\partial x}\frac{\partial x}{\partial w} = \frac{\partial L}{\partial v}\frac{\partial x}{\partial y}$$

since $\dfrac{\partial x}{\partial w} = 0$. Along any curve we have $\dfrac{d}{dt}x = v$, so $\dfrac{d}{dt}\dfrac{\partial}{\partial y}x = \dfrac{\partial v}{\partial y}$ and both sides of

$$\frac{d}{dt}\frac{\partial L}{\partial w} - \frac{\partial L}{\partial y} = \Big(\frac{d}{dt}\frac{\partial L}{\partial v}\Big)\frac{\partial x}{\partial y} + \frac{\partial L}{\partial v}\Big(\frac{d}{dt}\frac{\partial}{\partial y}x\Big) - \frac{\partial L}{\partial x}\frac{\partial x}{\partial y} - \frac{\partial L}{\partial v}\frac{\partial v}{\partial y} = \Big(\frac{d}{dt}\frac{\partial L}{\partial v} - \frac{\partial L}{\partial x}\Big)\frac{\partial x}{\partial y}$$

vanish together since $\dfrac{\partial x}{\partial y} = Df$ is nonsingular.

In general, the dynamics is only determined for finite time. However, in the case of compact M and L of the form $L(x, v) = K(v) - V(x)$ with *kinetic energy* given by a positive definite quadratic form $K(v) = g_x(v, v)/2$, it is defined for all times and determines a *complete* flow on TM, *i.e.*, a flow defined for all t. This flow is essentially

the Hamiltonian flow for the *total energy* $H = g_x(v,v)/2 + V(x)$. In particular, H is invariant.

According to the Variational Principle of Hamilton (principle of least action), the solutions of (5.1) are exactly the critical points of the *action functional* $\int_a^b L(c(t), c'(t))\,dt$ with given endpoints (Section 5.6b).

Two Lagrangians L and $\tilde{L}$ generate the same flow if and only if $L = \tilde{L} + \alpha + \text{const}$ for some closed one-form $\alpha\colon TM \to \mathbb{R}$.

In classical mechanics Lagrangians of the form $L(x,v) = g_x(v,v)/2 - V(x)$ for some Riemannian metric g appear for the equations of motion of systems with holonomic constraints with potential forces. Here g represents the kinetic energy and V the potential energy of the system. The special case $V = 0$ gives the geodesic flow (Section 5.4c) for the metric g. However, for all sufficiently high energy levels the Lagrangian flow is essentially geodesic also when $V \neq 0$:

THEOREM 5.4.1. *If $c > \sup V$ then the solution curves with energy c for the Lagrangian $L(x,v) = g_x(v,v)/2 - V(x)$ are reparametrized geodesics for the metric $(c - V)g$.*

Put differently, motion at sufficiently high energies looks like free motion in a slightly distorted space.

b. The Legendre transform. Over a Riemannian manifold M there is a natural identification of the *cotangent bundle* T^*M with TM via a map $\mathcal{L}\colon TM \to T^*M$ induced by $v \mapsto \langle v, \cdot \rangle$. The coordinates of a vector $v \in T_xM$ are given by the coefficients with respect to the canonical basis $\partial/\partial x_i$ and the coordinates of a form $\omega \in T_x^*M$ are given by the coefficients with respect to the standard basis dx_i dual to $\partial/\partial x_i$. Then

$$\mathcal{L}(x,v) := \Big(x, \frac{\partial K}{\partial v}\Big), \tag{5.2}$$

where $K(v) = g_x(v,v)/2$. If the x_i are viewed as coordinates in the configuration space M, then the v_i are velocities and the variables $p_i = \partial K/\partial v_i$ are called *momenta*.

The map $\mathcal{L}$ transforms the Lagrange equation into the Hamiltonian equations (5.1) with H the total energy because the Lagrange equation is $\dot{p} = \partial L/\partial q$ and $H = \langle p, v\rangle - L$, hence

$$\frac{\partial H}{\partial p}dp + \frac{\partial H}{\partial q}dq = dH = d(p\dot{q} - L) = \dot{q}\,dp - \frac{\partial L}{\partial q}dq = \dot{q}\,dp - \dot{p}\,dq.$$

The transformation $\mathcal{L}$ can be defined for Lagrangians that are C^2 convex functions of v (notice that we could use a Lagrangian L rather than K in the definition of $\mathcal{L}$). Thus such a Lagrangian L determines a transformation $\mathcal{L}(x,v) = (x, \partial L/\partial v)$ called the *Legendre transform*. Notice, however, that in this case the Legendre transform is not linear in v any more.

Another way of describing the Legendre transform is to call $\mathcal{L}\colon TM \to T^*M$ the *fiber derivative* of the Lagrangian K and note that E and L are real-valued functions on TM and H is a real-valued function on T^*M. The orbits of the Lagrangian and Hamiltonian system project to the same curves in M [**AM**].

c. Geodesic flows. A particular Lagrangian system is free particle motion in the configuration space M given by the Lagrangian $L(x,v) = g_x(v,v)/2$. This Lagrangian system as well as its restriction to the unit tangent bundle SM is called the *geodesic flow* of (M,g). It preserves the total energy $g_x(v,v)/2$ and hence the length of tangent vectors. Its orbits project to geodesics in M. The geodesic flow on any compact or homogeneous (transitive isometry group) manifold is a complete flow.

Geodesic flows are Hamiltonian flows in a natural way via the Legendre transform because the cotangent bundle is naturally identified with the tangent bundle. Geodesic flows preserve the product of the volume form on the manifold and the Euclidean volume defined in the tangent space by the Riemannian metric as well as a volume form on every hypersurface $H = \text{const}$, which is a sphere bundle $\{\|v\| = \text{const}\}$. This volume is the product of the Riemannian volume and the canonical (angular) volume on the spheres, called the *Liouville measure*. If M is compact then the Liouville measure is finite and can hence be normalized.

Geodesic flows are among the favorite subjects of study in dynamical systems. In particular they provide excellent illustrations of

(1) completely integrable behavior (flat tori, ellipsoids, surfaces of revolution),
(2) uniform hyperbolicity (manifolds of negative sectional curvature, including symmetric spaces of noncompact type of rank one),
(3) partial hyperbolicity (symmetric spaces of noncompact type of rank greater than one)[**S-FK**], and
(4) nonuniform hyperbolicity (manifolds of nonpositive sectional curvature of geometric rank one)[**S-K**].

5. Contact systems

Contact structures are odd-dimensional counterparts of symplectic structures [**By**], [**KH**, Section 5.6].

a. Contact forms and contact structures. Let M be a $(2n-1)$-dimensional manifold. A differential 1-form θ is called a *contact form* if the $(2n-1)$-form $\theta\wedge(d\theta)^{n-1}$ is nondegenerate. A contact form determines the codimension one distribution $\mathcal{D}{:=}\text{Ker}\theta \subset TM$ which is *totally nonintegrable*: Any two nearby points can be connected by a curve tangent to the distribution. Differentiably this is expressed by the fact that iterated Lie brackets of vector fields in the distribution generate the entire tangent space. Such a distribution is called *a contact structure* on M and the pair $(M,\mathcal{D})$ is called a *contact manifold*. The same contact structure can be defined by different contact forms; any two such forms are obtained from each other by multiplication by a nonvanishing scalar function.

Locally a contact form, similarly to a symplectic form, can be brought into a standard form. In fact, the following result is a simple consequence of the Darboux Theorem 5.3.2 for symplectic forms.

Let $\theta_0 = x_1dy_1 + \cdots + x_ndy_n + dz$ be the canonical contact form on $\mathbb{R}^{2n+1}$.

THEOREM 5.5.1. *Let (M,θ) be a contact $(2n+1)$-manifold. Then for $x \in M$ there exists a neighborhood U of x with coordinates in which $\theta = \theta_0$.*

PROOF. For $x \in M$ pick a neighborhood V_0 of 0 in $\ker\theta_x$ and let $V = V_0 \times (-\epsilon,\epsilon)$, $U' = \exp V$, $U'_t = \exp(V_0 \times \{t\}) \subset M$. $d\theta$ restricted to U'_t is a symplectic form so

by the Darboux Theorem Theorem 5.3.2 each $y \in U_t'$ has a neighborhood $U_t \subset U_t'$ on which there are Darboux coordinates $x_1, \dots, x_n, y_1, \dots, y_n, z$, *i.e.*, $d\theta = \sum dx_i \wedge dy_i$. On $U := \bigcup_{-\epsilon<t<\epsilon} U_t$ we thus have $d(\theta - \sum dx_i \wedge dy_i) = 0$ whence $\theta = \sum dx_i \wedge dy_i + dz$ and $x_1, \dots, x_n, y_1, \dots, y_n, z$ are the desired coordinates. □

Unlike a symplectic manifold which admits a variety of Hamiltonian vector fields, a contact manifold comes furnished with a canonical vector field v which is defined by $v \lrcorner \theta = 1$ and $v \lrcorner d\theta = 0$. This is unique because the kernel of $d\theta^n$ is one-dimensional and disjoint from that of θ by the nondegeneracy assumption. Note that the Lie derivative $\pounds_v \theta$ vanishes since $v \lrcorner \theta = \text{const}$, so the flow of v, which is called the *characteristic flow* of the contact form, preserves θ and hence all structures defined in terms of θ, in particular the volume. Thus the characteristic flow provides a canonical example of a volume-preserving flow.

Thus, the proper counterpart of a symplectic manifold is a contact *structure* whereas a contact *manifold* corresponds to a symplectic manifold together with a Hamiltonian vector field.

PROPOSITION 5.5.2. *Suppose (M, θ) is a contact manifold. Then M can be embedded into a symplectic manifold (N, ω) in such a way that the restriction of the ambient symplectic form to M is $d\theta$.*

REMARK. A contact manifold embedded in this way is called a *submanifold of contact type*.

PROOF. If $N = M \times \mathbb{R}$ and $\omega_{x,t} = d(e^t \theta_x)$ then $\omega^n = e^{nt}(n dt \wedge \theta \wedge (d\theta)^{n-1})$ is a volume, so (N, ω) is a symplectic manifold and ω restricted to $M \times \{0\}$ is $d\theta$. □

The following characterization is useful in applications of variational methods to Hamiltonian mechanics.

PROPOSITION 5.5.3. *Let ω be the standard symplectic form on $\mathbb{R}^{2n}$ and $M = f^{-1}(c) \subset \mathbb{R}^{2n}$ a level set of a smooth function $f : \mathbb{R}^{2n} \to \mathbb{R}$ with c as a regular value. Then M is a submanifold of contact type if and only if on a neighborhood U of M there is a vector field ξ transverse to M for which $\pounds_\xi \omega = \omega$.*

b. Hamiltonian systems preserving a 1-form. From the point of view of classical mechanics the most important (or at least the most traditional) symplectic manifolds are $\mathbb{R}^{2n}$ with the standard symplectic structure and the cotangent bundle of a differentiable manifold M (the *configuration space* of a mechanical system) with the symplectic form ω described in Section 5.3c2. Notice that in both cases the symplectic manifold (phase space) itself is not compact, although in the second case the configuration space M may be compact; this is true in many important classical problems such as the motion of a rigid body. Of course $\mathbb{R}^{2n}$ can also be viewed as $T^*\mathbb{R}^n$, so the first case is a particular instance of the second.

There is an important situation when the invariant 1-forms can be described in a particularly natural way. Notice that in the case of both $\mathbb{R}^{2n}$ and T^*M the form ω is not only closed, but also *exact*. The 1-form θ defined by $\sum_{i=1}^n p_i dq_i$—globally in the first case, locally in the second—obviously satisfies $d\theta = \omega$. Notice that θ is defined on T^*M independently of the choice of local coordinates. Of course in general a Hamiltonian system

on T^*M does not preserve θ or any other 1-form whose exterior derivative is equal to ω. Let us see what conditions the invariance of θ imposes on the Hamiltonian. One has

$$\pounds_{X_H}\theta = d\theta \lrcorner X_H + d(\theta \lrcorner X_H) = dH + d(\theta \lrcorner X_H).$$

Thus a 1-form θ is invariant if $\theta \lrcorner X_H = -H$. Local calculation in Darboux coordinates gives $\theta \lrcorner X_H = -\sum p_i \dfrac{\partial H}{\partial p_i}$. Notice that the choice of Hamiltonian for a given vector field X_H is unique up to an additive constant. Thus we have proved the following fact:

PROPOSITION 5.5.4. *The Hamiltonian vector field X_H on T^*M preserves the* 1*-form θ if and only if the Hamiltonian can be chosen as positively homogeneous in p of degree one,* i.e., *$H(q, \lambda p) = \lambda H(q, p)$ for $\lambda > 0$.*

There is a broader class of Hamiltonians that preserves the form θ *along* the hypersurfaces $H = \text{const}$ In this case the invariance condition becomes

$$d(\theta \lrcorner X_H)(\xi) = 0 \text{ if } dH(\xi) = 0,$$

or in other words the function $\theta \lrcorner X_H$ is constant on every connected component of the hypersurface $H = \text{const}$ This is satisfied if

$$\theta \lrcorner X_H = \varphi(H),$$

i.e., using Darboux coordinates,

$$H(q, \lambda p) = \Phi(\lambda) H(q, p),$$

where $\Phi' = \varphi$. If $\varphi(\lambda) \neq 0$ then such Hamiltonians will be called *generalized homogeneous Hamiltonians (in p).* Away from the zero section every such Hamiltonian is a function of a homogeneous Hamiltonian of degree one, namely, $H_1(q, p) = \Phi^{-1}(H(q, p))$, where Φ^{-1} is the inverse of Φ. An immediate calculation shows that $X_{\rho(H)} = \rho' X_H$ for any C^1 function ρ, so the flow generated by a generalized homogeneous Hamiltonian is obtained by a time change from the flow generated by a Hamiltonian that preserves θ and the time change is constant on every surface $H = \text{const}$

c. Geodesic flows as contact systems. In particular, since the Hamiltonian of a geodesic flow is a quadratic function of p, it preserves the restriction of θ to any energy surface. The phase space of the geodesic flow can be identified with the sphere bundle (the bundle of positive rays) over the configuration space, which can be defined independently of the choice of Riemannian metric. The corresponding contact structure does not depend on the choice of metric either. However, the specific contact form whose characteriscic flow is identified with the geodesic flow does depend on the metric.

6. Variational methods in dynamics

a. Variational description of orbits. It turns out that interesting orbits in some dynamical systems often can be found as special critical points of functionals defined on appropriate auxiliary spaces of potential orbits. This idea goes back to the variational principles in classical mechanics (Maupertuis, d'Alembert, Lagrange, Hamilton) which still remain the foundation of most variational methods in dynamics. Variational principles describe *all orbit segments* of a Lagrangian or Hamiltonian system as critical points of a

functional. At that level a variational description does not have particular dynamical significance. This appears when one extends the variational description from orbit segments to special types of orbits such as periodic or heteroclinic ones [**S-R**], [**KH**, Chapter 9].

The study of critical points of functions in finite- or infinite-dimensional spaces has two aspects: The local one dealing with the structure and stability of isolated critical points, and the global one, sometimes called *Morse theory*, which deals with the relation between the global topological properties of the space and the structure of critical points of functions on that space.

The prototypical finite-dimensional local result is the Morse Lemma:

PROPOSITION 5.6.1. *Let p be a nondegenerate critical point of a C^r function, $r \geq 2$, on a smooth manifold M. Then there exist $0 \geq k \geq n$ and a local C^{r-2} coordinate system $(x_1, \ldots, x_n)$ with p as the origin such that in these coordinates f is given by*

$$f(x) = f(0) + \sum_{i=1}^{k} x_i^2 - \sum_{i=k+1}^{n} x_i^2$$

The number k is called the *Morse index* of the point p. There are genereralizations of the Morse Lemma and the notion of index for certain kinds of critical points of functions in infinite-dimensional spaces.

In the infinite-dimensional situation sometimes orbits can identified as critical points of finite Morse index. Various minimax or mountain pass arguments are used to find such points. For a long time the applicability of variational methods to global and semilocal problems in dynamics was restricted to situations of this type that appear in Lagrangian dynamics and to Hamiltonian systems where a separation between the coordinates and momenta can be made. Advances in the critical point theory allow to treat situations where critical points have infinite index. See [**S-R, S-HZ, HZ**] for a more detailed discussion.

b. The least action principle in Lagrangian mechanics. (See also [**S-BK**].) Consider a connected configuration space M and a Largangian L.

THEOREM 5.6.2 (Variational Principle of Hamilton/principle of least action). *The solution $(x(t), v(t))$ of* (5.1)*, where $v(t) = x'(t)$ for $a \leq t \leq b$, is a critical point of the* action functional

$$\int_a^b L(c(t), c'(t))\, dt$$

defined of the space of smooth curves $c\colon [a, b] \to M$ such that $c(a) = x(a)$ and $c(b) = x(b)$. Conversely, any critical point is a solution of (5.1)*.*

Under proper convexity assumptions (the strong Legendre condition), which in particular are satisied in the classical case $L(x, v) = g_x(v, v)/2 - V(x)$, the action functional always has a minimum, which is unique if the endpoints are close enough. Thus for a global solution $(x(t), v(t))$ of (5.1) with $-\infty < t < \infty$, any sufficiently small segment is the unique minimum of the corresponding action functional.

In order to apply the least action principle to finding periodic orbits one should consider the action functional on a space of periodic curves where the existence of sufficiently nondegenerate critical points is guaranteed. The basic example of the difficulty inherent in this approach appears when one considers all curves with a given period: The minima are simply constant solutions corresponding to the maxima of V.

On the other hand, if one considers periodic curves in a given nontrivial homotopy class then at least for compact configuration spaces there are always nontrivial periodic solutions. Thus, for example for a compact Rieamnnian manifold M there is always at least one closed geodesic (periodic orbit of the geodesic flow) in each nontrivial free homotopy class of curves on M. If the fundamental group of M is sufficiently complicated then this guarantees some growth of the number of closed geodesics measured by their length [**KH**, Sections 9.6, 9.7].

A more sophisticated minimax type argument is involved in showing that any compact Riemannian manifold has a closed geodesic.

The least action principle is also quite effective in finding special families of non-closed geodesics both for tori and for compact surfaces of higher genus. In the former case the geodesics form invariant laminations corresponding to the Aubry–Mather sets (Section 7.2b) for a section map for the geodesic flow. In the latter the geodesics globally approximate the geodesics for the conformally equivalent metric of constant negative curvature.

c. The action principle in Hamiltonian dynamics. (See also [**S-HZ, HZ**].) There are several reasons to consider more general Hamiltonian systems than those discussed in Section 5.3, namely those whose Hamiltonians explicitly depend on time. The solutions of such systems are one-parameter families of symplectic diffeomorphisms without a group property.

On the one hand, time-dependent Hamiltonian systems describe more general mechanical systems, *e.g.*, those with a time-dependent potential field. If this dependence is periodic in time, then the resulting dynamical systems lie wholly within the paradigm discussed in this chapter since the period map generates a discrete-time symplectic system in the phase space. But it also turns out that time-dependent (in particular periodic) Hamiltonians provide a powerful technical tool for symplectic geometry and the study of orbits of ordinary Hamiltonian systems. In particular they play a central role in the construction of certain symplectic capacities [**HZ**].

The variational principle in phase space, whether for time-dependent Hamiltonians or not, still requires a choice of coordinates that locally have the Darboux canonical form. In a somewhat more abstract form this amounts to choosing two transverse foliations whose leaves are Lagrangian submanifolds corresponding to "coordinates" and "momenta". There are of course cases where such a structure is given, *i.e.*, for an open subset of $\mathbb{R}^{2n}$ with the standard symplectic structure, or for the cotangent bundle Section 5.3c2. The difference between the least action principle of Hamilton and the action principle in the phase space is that the space of candidate orbits is much larger in the latter case. They are curves in the phase space whose initial and final coordidates are fixed, but there is no correlation between the derivatives of the coordinates and the momenta.

THEOREM 5.6.3 (Action principle in the phase space). *The solution of the system of Hamiltonian equations 5.1 with time-dependent Hamiltonian H for $a \le t \le b$. is a critical point of the* action functional

$$\int_a^b (p(t)q'(t) - H(t, p(t), q(t)))dt$$

defined on the space of smooth curves $c\colon [a,b] \to M$ *such that* $c(a) = x(a)$ *and* $c(b) = x(b)$*, where* $x(\cdot)$ *denotes the configuration space coordinate. Conversely, any critical point is a solution of* (5.1)*.*

Since this action functional is defined on a much larger space than the Lagrangian action functional in Section 5.6b, it is not *a priori* surprising that its critical points tend to have infinite index. Sophisticated topological methods have been developed, culminating in the Floer cohomology theory, to substitute for the more traditional Morse theory ("variational calculus at large" in a somewhat oldfashioned terminology) used in the Lagrangian context.

7. Holomorphic dynamics

We would like to emphasize that holomorphic dynamics is not covered in any serious way in the two volumes of the handbook to which this survey serves as an introduction. It is planned that at least one-dimensional holomorphic dynamics will be covered in [**DS2**]. Thus the comments below are not meant as an introduction or an overview of the material presented elsewhere but as a set of brief remarks about an extensive field that is naturally connected with some of the material of these volumes.

a. Conformal dynamics. (Se also [**MS, S-JS**].) The underlying structure of this branch of the theory of dynamical systems is a complex manifold (not necessarily compact, an open set in $\mathbb{C}^n$ is an example) and a holomorphic map defined in a neighborhood of a compact invariant set (the semilocal setting). The corresponding global situation is a holomorphic map of a compact complex manifold (*e.g.*, complex projective space $\mathbb{C}P(n)$) into itself. Holomorphic maps, both in one and several complex variables, possess a certain rigidity, manifested both locally (Taylor coefficients at a point define the map in an open set) and globally (Liouville Theorem, maximum modulus principle *etc.*). This sets holomorphic dynamics apart from general differentiable dynamics (where different locally defined maps can be easily glued together) and to a lesser extent Hamiltonian dynamics, where there are no local restrictions either, but there are some global ones. In this respect holomorphic dynamics is closer to the algebraic dynamics of translations and affine maps on homogeneous spaces (Section 3.3c) although the dynamical paradigms for the two areas tend to be quite different, *e.g.*, no nice invariant measure is usually present in the holomorphic case and dissipative behavior is quite common. One of the characteristic features of holomorphic dynamics is the important role played by singularities of holomorphic maps. Since the singular set has positive complex codimension and hence real codimension at least two the singularities tend to be more manageable than in the real, even the real-analytic, case.

Holomorphic dynamics in one variable is well developed. In fact, the classical works of Fatou, Julia and Montel appeared at a time when real differentiable dynamics, not to mention ergodic theory, was in its infancy. It rests on two pillars: Conformality and uniformization. The former is an *infinitesimal* property. It is a characteristic property of low-dimensional differentiable dynamics. From this point of view one can define the area of conformal dynamics, which essentially includes differentiable dynamics in real dimension one and holomorphic dynamics in complex dimension one. That this short list is exhaustive follows from the fact that any conformal map in real dimension two is holomorphic and that in higher dimension there are too few conformal maps (essentially only

higher-dimensional counterparts of fractional linear transformations). The main technical corollaries of conformality that are crucial for the analysis of a growing number of iterates of a map, are various kinds of bounded distortion estimates. Thus, the emphasis on conformality brings together one-dimensional real dynamics and one-dimensional complex holomorphic dynamics [**MS**].

On the other hand, uniformization, whose most elementary manifestation is the Riemann mapping theorem and a more advanced one the Koebe uniformization theorem, is an essentially one-dimensional *complex* phenomenon [**S-JS**].

For an introduction in one-dimensional complex dynamics see [**Bd**]; a more advanced source is [**CG**].

b. Holomorphic maps in higher dimension. Multidimensional complex dynamics is a much newer and less developed field. While neither conformality nor uniformization are available, there are other powerful tools from complex analysis that make it possible to understand the structure of certain classes of holomorphic maps (*e.g.*, polynomials) to a considerably greater degree than in the case of real differentiable dynamics. The basis of those tools are extremal properties of holomorphic maps which allow, *e.g.*, to prove a proper formula for the topological entropy. A useful observation is that in complex dimension two, hyperbolic behavior of invertible maps forces both stable and unstable manifolds to be one-dimensional complex submanifolds. Thus, the dynamics on these families of manifolds is conformal and some tools from one-dimensional complex analysis can be adapted to this situation.

There is not yet a comprehensive monograph on holomorphic dynamics in higher dimension. [**Si**] can be recommended as a thorough and extensive survey with an excellent list of references, [**MNTU**] contains an introduction and treatment of selected topics.

CHAPTER 6

Hyperbolic dynamics: Orbit instability and structural stability

1. Introduction

a. The hyperbolic paradigm. The dynamics of hyperbolic systems is dominated by exponential behavior of orbits relative to each other. While various distinct classes of dynamical systems belong to the hyperbolic category, two central aspects of behavior stand out as the main features.

The first is rich and pervasive recurrence of great complexity much like that found in transitive topological Markov chains. For certain classes of hyperbolic dynamical systems this correspondence is, in fact, quite precise (almost-isomorphism, Section 2.2f; see Section 6.7g). Accordingly, this recurrence is coupled with highly sensitive dependence on initial conditions and exponential behavior in all aspects of orbit growth. Furthermore, invariant measures abound.

The other central feature of hyperbolic systems is stability of the orbit structure as a whole. This is manifested as strong C^1 structural stability under uniformity assumptions, and as persistence of essential aspects of the orbit complexity in general.

Among the three main paradigms (hyperbolic, elliptic, parabolic) this is the one where one finds the linearization of a dynamical system to be most useful in studying the dynamics. The exponential behavior of the linearization directly translates into various kinds of exponential orbit behavior.

The combination of intricacy and robustness of the orbit structure on the one hand with the utility of linearization on the other hand has fueled great interest in hyperbolicity. In dealing with general structural questions, hyperbolic dynamics is by far the best developed and furthest advanced area in differentiable dynamics. Hyperbolicity is the leading and in some sense the only available paradigm that explains complicated or "stochastic" or "chaotic" behavior in differentiable dynamical systems of a general kind. In particular, all rigorous work related to "strange attractors" involves establishing some sort of hyperbolic behavior. Thus it is natural that hyperbolicity pervades a large number of the surveys in these volumes. The surveys [**S-H, S-C, S-P, S-BKP, S-W, S-K**] focus on various aspects of hyperbolic behavior, as do parts of the surveys [**S-FK, S-FM, S-KSS, S-JS**]. A systematic introduction to hyperbolic dynamics is given in [**KH**, Chapters 6,17–20].

b. Hyperbolic linear maps. In the case of linear maps, hyperbolicity is defined in terms of the spectrum: A continuous linear map $A\colon X \to X$ of a Banach space is said to be *hyperbolic* if its spectrum (or rather, that of its complexification), $\mathrm{Sp}(A)$, does not intersect the unit circle in $\mathbb{C}$. Unlike ellipticity or parabolicity, this is an open condition: There exist constants $0 < \lambda < 1 < \mu$ such that $\mathrm{Sp}(A) \cap \{z \in \mathbb{C} \mid \lambda < |z| < \mu\} = \varnothing$.

We then say that A is (λ, μ)-hyperbolic. This implies that there are subspaces E_s, E_u with $E = E_s \oplus E_u$, $A(E_s) \subset E_s$, $A(E_u) = E_u$, $\mathrm{Sp}(A_{\restriction E_s}) = \mathrm{Sp}(A) \cap \{|z| < 1\}$, and $\mathrm{Sp}(A_{\restriction E_u}) = \mathrm{Sp}(A) \cap \{|z| > 1\}$. Indeed, one can easily find an "adapted" norm $\|\cdot\|$ such that $\|A_{\restriction E_s}\| \le \lambda$, $\|(A_{\restriction E_u})^{-1}\| \le 1/\mu$ and $\|x_s + x_u\| = \max(\|x_s\|, \|x_u\|)$ for $x_s \in E_s$ and $x_u \in E_u$. It is the exponential behavior of orbits that dominates the dynamics and makes it so different from that of elliptic or parabolic maps.

A consequence of this behavior is structural stability, a primitive precursor of which is persistence of hyperbolic fixed points. We present a strong version of this result as an example, but also because it provides one of the principal technical tools in the development of the hyperbolic theory.

THEOREM 6.1.1 (Hyperbolic Fixed Point Theorem, [**Y3**]). *Suppose* $0 < \lambda < 1 < \mu$ *and* $A\colon E \to E$ *is* (λ, μ)*-hyperbolic. If* $f\colon E \to E$ *is a map with* $\epsilon := \mathrm{Lip}(f - A) < \epsilon_0 := \min(1 - \lambda, 1 - \mu^{-1})$ *(see* (2.5)*) then* f *has a unique fixed point* $p \in E$ *and* $\|p\| < \|f(0)\|/(\epsilon_0 - \epsilon)$.

This result is a fairly straightforward consequence of the Contraction Principle.

2. Main features of hyperbolic behavior

Here we describe those features that set hyperbolic dynamics apart from the rest of differentiable dynamics, in particular from elliptic and parabolic dynamics. Some of the descriptions given here are strictly correct only for the uniformly hyperbolic case but are equally distinctive when formulated more carefully in greater generality. Much of the core theory can be pushed to nonuniformly hyperbolic systems, and some aspects of partially hyperbolic dynamical systems can be understood with hyperbolic techniques as well. This is usually the case when the hyperbolic behavior so dominates the dynamics, that effects of subexponential order do not play a role in the properties one looks at. Stable ergodicity of partially hyperbolic systems is an example where this is a useful approach (Section 6.9c).

a. Growth of the orbit complexity. Hyperbolic systems exhibit exponential orbit growth any way one measures it. Periodic orbits are isolated, hence finite in number for any period, but grow at an exponential rate, if there is any nontrivial recurrence at all. This rate can, in fact, be determined with remarkable precision (Section 6.7c, [**S-H, S-P**]). Likewise, topological entropy is positive. The same goes for the homotopical and fundamental group entropies (Section 2.5m).

b. Relative behavior of orbits. Expansiveness (Section 2.4d) is the most straighforward consequence of uniform hyperbolicity and in fact one of the reasons for the effectiveness of symbolic representation for such systems. Various shadowing properties (see e.g. Section 6.6c) also fall into this category. These properties are crucial for understanding the nature of recurrence in hyperbolic systems.

c. Recurrence. In hyperbolic systems recurrence is far from uniform, but very pervasive and complex. Topological mixing is typically present, even though there are trivial exceptions (periodic components in the discrete time case and suspensions in the continuous time situation). Basically, hyperbolic dynamical systems can be decomposed into mixing pieces (Theorem 6.7.1). The most effective and precise description, however, of the orbit structure is that it corresponds precisely to a topological Markov chain. In fact, the

study of shifts was directly motivated by hyperbolic dynamics. Markov models arise (topologically or measurably) via the standard device of Markov partitions, which is outlined in Section 6.7g and explained in [**S-C**]. It decomposes the phase space of a hyperbolic dynamical system in such a way that the map corresponds to a shift on allowed sequences of partition elements. Up to negligible sets associated with the boundaries of the partition elements, the correspondence between points and sequences of partition elements (namely the *itineraries* of points) is exact, *i.e.,* it is an almost-isomorphism (Section 2.2f, Section 6.7g). Hyperbolic dynamical systems are often obtained as the restriction of a smooth system to an invariant set. Accordingly, the phase space may be a Cantor set and this coding may be a conjugacy.

d. Invariant measures. The measure-theoretic structure of hyperbolic dynamics is characterized by an abundance of invariant measures. Among these are evidently the many (ergodic) measures concentrated on periodic orbits. The weak* closure of these contains many more ergodic (indeed, mixing) measures, however. There is essentially one distinguished measure for each Hölder function on the phase space (Section 6.7c). See [**S-C**] for detailed treatment.

e. Stability. Among the properties specific to the smooth category, structural stability stands out. Hyperbolic dynamical systems are strongly C^1 structurally stable. To be more precise, the restriction to a hyperbolic set is always structurally stable, an Axiom A system (Section 6.4a) with a transversality condition on stable and unstable leaves is structurally stable, and a similar condition guarantees Ω-stability. Conversely, stability has been found to characterize hyperbolic dynamical systems: The sufficient conditions just mentioned are necessary as well. This is Mañé's and Hayashi's Stability Theorem [**S-H**], a high point in the development of smooth dynamics. Why hyperbolic dynamical systems are structurally stable is suggested by the Hyperbolic Fixed Point Theorem 6.1.1 [**S-H, Y3**], which, in fact, implies structural stability.

The abundance of periodic points immediately gives a large and intricate set of moduli of smooth conjugacy (Section 5.2a, Section 5.2c, Section 5.2e) that change nontrivially under perturbations. While there are often further invariants of smooth conjugacy, there are important situations where fixing these moduli determines a smooth equivalence class (Section 6.7h).

f. Prevalence of semilocal phenomena. Hyperbolicity on a part of phase space is a much more common phenomenon than hyperbolicity on the whole space. Accordingly, semilocal analysis (locally maximal or isolated, in particular *basic* hyperbolic sets, Definition 6.4.2, [**S-H, B1**]) plays a central role in hyperbolic theory. Since such hyperbolic sets often are totally disconnected this part of the theory does not involve any serious topological considerations. Global hyperbolic theory deals with Anosov systems (Definition 6.4.3). Global structures associated with such systems have nontrivial topology and impose various restriction on the topology of the phase space and the system itself. The problem of topological classification of Anosov systems has proved worth pursuing but is not well understood and even many simple-sounding special questions remain unanswered [**S-H**].

3. Stable manifolds

Because hyperbolicity can be defined with varying degrees of stringency, we first exhibit a fact central to most hyperbolic theory in order to illustrate what should be viewed as the essence of local hyperbolic behavior.

THEOREM 6.3.1 (The Hadamard–Perron Theorem, [**KH**]). *Let $\lambda < \mu$ and choose $0 < \gamma < \min\left(1, \sqrt{\mu/\lambda}-1\right)$ and $0 < \delta < \min\left(\dfrac{\mu-\lambda}{\gamma+2+1/\gamma},\ \dfrac{\mu-(1+\gamma)^2\lambda}{(1+\gamma)(\gamma^2+2\gamma+2)}\right)$. For $r \geq 1$ and for each $m \in \mathbb{Z}$ let $f_m\colon \mathbb{R}^n \to \mathbb{R}^n$ be a (surjective) C^r diffeomorphism such that*

$$f_m(x,y) = (A_m x + \alpha_m(x,y),\ B_m y + \beta_m(x,y))$$

for $(x,y) \in \mathbb{R}^k \oplus \mathbb{R}^{n-k}$, where $A_m\colon \mathbb{R}^k \to \mathbb{R}^k$ and $B_m\colon \mathbb{R}^{n-k} \to \mathbb{R}^{n-k}$ are linear maps with $\|A_m^{-1}\| \leq \mu^{-1}$, $\|B_m\| \leq \lambda$ and $\alpha_m(0) = 0$, $\beta_m(0) = 0$, $\|\alpha_m\|_{C^1} < \delta$, $\|\beta_m\|_{C^1} < \delta$. Then there is

(1) *a unique family $(W_m^+)_{m\in\mathbb{Z}}$ of k-dimensional C^1 manifolds*

$$W_m^+ = \{(x, \varphi_m^+(x)) \mid x \in \mathbb{R}^k\} = \operatorname{graph} \varphi_m^+$$

and

(2) *a unique family $(W_m^-)_{m\in\mathbb{Z}}$ of $(n-k)$-dimensional C^1 manifolds*

$$W_m^- = \{(\varphi_m^-(y), y) \mid y \in \mathbb{R}^{n-k}\} = \operatorname{graph} \varphi_m^-,$$

where $\varphi_m^+\colon \mathbb{R}^k \to \mathbb{R}^{n-k}$, $\varphi_m^-\colon \mathbb{R}^{n-k} \to \mathbb{R}^k$, $\sup_{m\in\mathbb{Z}} \|D\varphi_m^\pm\| < \gamma$, and the following properties hold:

(1) $f_m(W_m^-) = W_{m+1}^-$, $\quad f_m(W_m^+) = W_{m+1}^+$.

(2) *$\|f_m(z)\| < \lambda'\|z\|$ for $z \in W_m^-$, $\|f_{m-1}^{-1}(z)\| < (\mu')^{-1}\|z\|$ for $z \in W_m^+$, where $\lambda' := (1+\gamma)\,(\lambda + \delta(1+\gamma)) < \dfrac{\mu}{1+\gamma} - \delta =: \mu'$.*

(3) *Let $\lambda' < \nu < \mu'$. If $\|f_{m+L-1} \circ \cdots \circ f_m(z)\| < C\nu^L\|z\|$ for all $L \geq 0$ and some $C > 0$ then $z \in W_m^-$.*
Similarly, if $\|f_{m-L}^{-1} \circ \cdots \circ f_{m-1}^{-1}(z)\| \leq C\nu^{-L}\|z\|$ for all $L \geq 0$ and some $C > 0$ then $z \in W_m^+$.

Finally, in the hyperbolic case $\lambda < 1 < \mu$, the families $(W_m^+)_{m\in\mathbb{Z}}$ and $(W_m^-)_{m\in\mathbb{Z}}$ consist of C^r manifolds.

If one takes $\lambda < 1 < \mu$ from the start, then the hypotheses of the theorem describe a family of maps that are C^1 perturbations of a sequence of linear hyperbolic maps with expanding subspace $\{0\} \times \mathbb{R}^k$ and contracting subspace $\mathbb{R}^{n-k} \times \{0\}$. Note that hyperbolicity, or, more generally, the rate conditions, are cast in terms of asymptotic behavior rather than a spectral condition, although this is also possible (Section 6.4a).

The reason for considering families of maps lies in local analysis near a nonperiodic orbit. The f_m in the result above are local coordinate representations of the global map at the mth iterate.

Note also that the above result, while stated for a discrete family of maps, is also directly applicable to flows via time-one maps.

4. Definitions

The Hadamard–Perron Theorem suggests several possible definitions of hyperbolicity of a map. In each of these one requires the iterates of a map along orbits to be a family of the kind described in the theorem. However, choices are possible as to the degree of uniformity as well as whether to require strict hyperbolicity or partial hyperbolicity, *i.e.*, whether to admit nontrivial center manifolds.

a. Hyperbolic sets.

DEFINITION 6.4.1. Let M be a smooth manifold, $U \subset M$ a open, $f\colon U \to M$ a C^1 embedding. An f-invariant set Λ is said to be *hyperbolic* if the linear map defined on the space of bounded sections of $T_\Lambda M$ by $X \mapsto Df \circ X \circ f^{-1}$ is hyperbolic in the sense that the spectrum $\sigma(f)$ of the complexification of this map (sometimes called the *Mather spectrum*) is disjoint from the unit circle.

Equivalently, for some λ, μ with $lambda < 1 < \mu$ there is a decomposition $T_\Lambda M = E^+ \oplus E^-$ such that $Df_x E^\pm_x = E^\pm_{f(x)}$ and

$$\|Df^n{\restriction}_{E^-}\| \le C\lambda^n, \quad \|Df^{-n}{\restriction}_{E^+}\| \le C\mu^{-n} \text{ for } n \in \mathbb{N}.$$

This is called a (λ, μ)-*splitting* for $Df|T_\Lambda M$ and is often considered also in situations when λ and μ lie on the same side of 1.

Equivalently, following Alekseev [**Ax**], require that for some metric there exist $\lambda < 1 < \mu$ and $\gamma > 0$ such that for every $x \in \Lambda$ there is a decomposition $T_x M = S_x \oplus T_x$ with

$$Df_x H_x \subset \operatorname{Int} H_{f(x)} \text{ and } {Df_x}^{-1} V_{f(x)} \subset \operatorname{Int} V_x,$$

where

$$H_x := \{\xi + \eta \mid \xi \in S_x,\ \eta \in T_x,\ \|\eta\| \le \gamma\|\xi\|\},$$
$$V_x := \{\xi + \eta \mid \xi \in S_x,\ \eta \in T_x,\ \|\xi\| \le \gamma\|\eta\|\},$$

and if furthermore $\|Df_x\xi\| \ge \mu\|\xi\|$ for $\xi \in H_x$, and $\|{Df_x}^{-1}\xi\| \ge \lambda^{-1}\|\xi\|$ for $\xi \in V_{f(x)}$. H_x and V_x are called invariant horizontal and vertical *cone fields*.

An embedding $f\colon U \to M$ is said to satisfy *Axiom A* if $NW(f)$ is hyperbolic and periodic points are dense in it. (The latter is not automatic [**D**] except in dimension 2 [**NP**], although it is generic [**Pu1**].)

DEFINITION 6.4.2. A topologically transitive compact locally maximal (isolated) hyperbolic set is usually called a *basic set*.

This is motivated by the Spectral Decomposition Theorem 6.7.1, where such sets arise as building blocks.

It is quite easy to see that E^+_x and E^-_x have locally constant dimension and are continuous, hence uniformly transverse. They are, in fact, Hölder continuous. Under restrictive assumptions they may be differentiable, but they are C^2 only in special situations [**S-H**].

To apply the Hadamard–Perron Theorem note that this is equivalent to existence of a Riemannian metric (called a *Lyapunov metric*) on U such that for any $x \in \Lambda$ the sequence of differentials $(Df)_{f^n(x)}\colon T_{f^n(x)}M \to T_{f^{n+1}(x)}M$, $n \in \mathbb{Z}$, admits a (λ, μ)-splitting, *i.e.*,

there exist decompositions $T_{f^n(x)}M = E^+_{f^n(x)} \oplus E^-_{f^n(x)}$ such that $(Df)_{f^n(x)}E^\pm_{f^n(x)} = E^\pm_{f^{n+1}(x)}$ and

$$\|(Df)_{f^n(x)}\restriction_{E^-_{f^n(x)}}\| \le \lambda, \quad \|(Df)^{-1}_{f^n(x)}\restriction_{E^+_{f^{n+1}(x)}}\| \le \mu^{-1}.$$

Therefore, one can choose, for each $x \in \Lambda$, a local coordinate system mapping x to the origin and depending continuously on x, such that with respect to these coordinates f satisfies the hypotheses of the Hadamard–Perron Theorem with *uniform* $\lambda < 1 < \mu$.

DEFINITION 6.4.3. If for a diffeomorphism f of a compact manifold, the whole manifold is a hyperbolic set then f is said to be an *Anosov diffeomorphism*.

Compact hyperbolic sets, Anosov diffeomorphisms, and related issues are surveyed by [**S-H**].

b. Nonuniform hyperbolicity. Nonuniform hyperbolicity employs the same idea, namely separation of all directions into exponenetially expanding and exponentially contracting ones but does not require uniformity. Precise definitions are given in [**S-BKP**] and for the present discussion it suffices to think of it as hyperbolicity of orbits without uniform control. It should be noted that uniformity is abandoned both for the coefficients in front of the contraction and expansion rates as well as the angles between the stable and unstable subspaces.

c. Partial hyperbolicity. (See also [**S-Bu**].) Partial hyperbolicity requires a (λ, μ)-splitting with $\lambda < 1$ *or* $\mu > 1$, but not both. Often one requires $TM = E^+ \oplus E^0 \oplus E^-$ such that $Df^n\restriction_{E^0}$ has subexponential growth (usually) and $\|Df^{-n}\restriction_{E^+}\|\|Df^n\restriction_{E^0}\| \le \lambda < 1$ and $\|Df^{-n}\restriction_{E^0}\|\|Df^n\restriction_{E^-}\| \le \lambda$ for $n \in \mathbb{N}$, *i.e.*, E^+ expands more than anything in E^0 and E^- contracts more than anything in E^0. The "slow" distribution E^0 is not necessarily uniquely integrable. The extent to which these systems can be understood is limited by the fact that no restriction is imposed on the "subexponential part" of their behavior. For example, the product of any dynamical system with only subexponential expansion with a hyperbolic dynamical system is partially hyperbolic. Accordingly, hyperbolic methods may give some global insights but often do not help study the nonhyperbolic factor. But hyperbolic techniques may well resolve global issues that are dominated by the hyperbolic behavior. Stable ergodicity (Section 6.9c) is of this kind.

One can also consider nonuniformly partially hyperbolic systems [**BKP**].

d. Flows. A flow $\varphi^t\colon M \to M$ is said to be hyperbolic if $T_\Lambda M = E^0 \oplus E^+ \oplus E^-$ such that $E^0 = \langle\dot\varphi\rangle \ne \{0\}$, $D\varphi^t_x E^\pm_x = E^\pm_{f(x)}$ and $\|D\varphi^t\restriction_{E^-}\| \le C\lambda^t$, $\|D\varphi^{-t}\restriction_{E^+}\| \le C\mu^t$ for $t \ge 0$. $E^\pm$ are the *strong* subbundles, $E^0 \oplus E^\pm$ the weak ones.

Other natural examples of partially hyperbolic systems are isometric extensions of hyperbolic systems and elements of *transversely hyperbolic* actions of $\mathbb{R}^m$.

e. Hölder regularity. The category of Hölder continuous functions and maps plays an important role in hyperbolic dynamics. It is pervasive both in hypotheses and conclusions of statements. Its importance is related to the basic fact that if $d(x_n, y_n)$ is exponentially small in n (hence summable) then so is $d(f(x_n), f(y_n))$ whenever f is Hölder continuous. Here are a few specific examples of necessity. Hölder continuity of a function

makes it a member of the class C^f in Section 4.4g for which equilibrium states are defined. It is indispensable in the Livschitz Theorem. Absolute continuity of the invariant foliations (Section 6.7e) requires the Anosov diffeomorphism to be $C^{1+\alpha}$ [**RY**], even though some more basic results may hold with only a C^1 assumption.

Conversely, many of the objects naturally attached to hyperbolic dynamical systems are Hölder continuous: Invariant subbundles and foliations, conjugacies, *etc.* The Hölder exponent, by the way, is not determined primarily by that of the diffeomorphism, but rather by relations between contraction and expansion rates.

For nonuniformly hyperbolic systems the $C^{1+\alpha}$ hypothesis becomes entirely indispensable [**Pu2**].

5. Examples

Several standard examples display the principal features of hyperbolic behavior in a way that can be easily visualized.

FIGURE 6.1. A toral automorphism

a. Toral automorphisms. This example was mentioned before in Section 2.1c and Section 5.1h. Any integer $m \times m$ matrix A with determinant ± 1 induces a map on $\mathbb{R}^m/\mathbb{Z}^m$, which is hyperbolic whenever A is a hyperbolic linear map, *i.e.*, has no eigenvalue on the unit circle. Stable and unstable manifolds are simply translates of (projections of) eigenspaces and are dense. By structural stability (Section 6.7h) C^1-perturbations are also examples, but topologically equivalent. Figure 6.1 illustrates $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$.

b. The Smale horseshoe. Let $\mathcal{R}$ be a rectangle in $\mathbb{R}^2$ and $f\colon \mathcal{R} \to \mathbb{R}^2$ an embedding such that $\mathcal{R} \cap f(\mathcal{R})$ consists of two "horizontal" rectangles $\mathcal{R}_0$ and $\mathcal{R}_1$ and the restriction of f to the components $\mathcal{R}^i \subset f^{-1}(\mathcal{R})$, $i = 0, 1$, of $f^{-1}(\mathcal{R})$ is a hyperbolic affine map, contracting in the vertical direction and expanding in the horizontal direction. This implies

FIGURE 6.2. The horseshoe

that $\mathcal{R}^0$ and $\mathcal{R}^1$ are "vertical" rectangles. One of the simplest ways to achieve this effect is to bend $\mathcal{R}$ into the shape of a "horseshoe", or permanent magnet, or into a G- or "paper clip" shape. The maximal invariant subset of $\mathcal{R}$ is $\Lambda = \bigcap_{n=-\infty}^{\infty} f^{-n}(\mathcal{R})$, which is a Cantor set

FIGURE 6.3. The paper clip

with a natural product structure. The local stable and unstable manifolds are vertical and horizontal line segments.

c. The Smale attractor. (See also Section 5.2i.) On the solid torus $M = S^1 \times D^2$, where D^2 is the unit disk in $\mathbb{R}^2$, define coordinates (φ, x, y) such that $\varphi \in S^1$ and $x^2+y^2 \leq 1$. Let

$$f\colon M \to M, \quad f(\varphi, x, y) = \Big(2\varphi, \frac{1}{10}x + \frac{1}{2}\cos\varphi, \frac{1}{10}y + \frac{1}{2}\sin\varphi\Big).$$

If $C = \{\theta\} \times D^2$ is a cross section then $f(M) \cap C$ consists of two disjoint disks of radius $1/10$. $\Lambda := \bigcap_{l\in\mathbb{N}_0} f^l(M)$ is an attractor on which f is expanding. Locally it is the product of a Cantor set with an interval, but it is connected. The stable manifolds are the sections $C = \{\theta\} \times D^2$, the unstable manifold of each point is entirely contained in the attractor.

FIGURE 6.4. The Smale attractor

d. Suspensions. Any hyperbolic set for a diffeomorphism gives rise to a hyperbolic set for a flow via the suspension construction (Section 1.3j, Section 2.2j, Section 5.2j).

e. Geodesic flows. The geodesic flow (Section 5.4c)) on the unit tangent bundle of a negatively curved manifold is an Anosov flow [**S-K, KH, Kb**]. This is the primary example of an Anosov flow and is surveyed carefully by [**S-K**].

1. *Ergodicity.* Due to the Hamiltonian structure, volume (Liouville measure) is invariant. Volume is ergodic, indeed mixing and Bernoulli, because it is an equilibrium state, see Section 6.7c, Section 4.4g, [**S-H, Ld**], [**KH**, Theorem 20.4.1].

Ergodicity of volume was not initially obtained from the theory of equilibrium states, and it is good to see more directly how the hyperbolic structure produces ergodicity. We discuss this earlier approach due to Hopf and Anosov in Section 6.7e.

2. *The hyperbolic plane and surfaces of constant negative curvature.* A special case allows an interpretation in terms of homogeneous dynamics (Section 2.1b, [**KH**, Section 17.5]). Consider the hyperbolic or Poincaré upper half plane $\mathbb{H} = \{z \in \mathbb{C} \mid \operatorname{Im} z > 0\}$ with the hyperbolic Riemannian metric

$$\langle u + iv, u' + iv' \rangle_z := \operatorname{Re} \frac{(u + iv)(u' - iv')}{(\operatorname{Im} z)^2}.$$

Note that angles agree with the Euclidean ones. The geodesics for this metric are the lines $x + i\mathbb{R}$ and semicircles with real center. The group $\mathrm{PSL}(2, \mathbb{R})$ (obtained from $\mathrm{GL}(2, \mathbb{R})$ by identifying scalar multiples) acts isometrically by fractional linear transformations $(\begin{pmatrix} a & b \\ c & d \end{pmatrix}, z) \mapsto \dfrac{az + b}{cz + d}$, and this action induces a transitive action on unit tangent vectors.

In order to obtain a flow on a compact manifold one consider factors by cocompact lattices, *i.e.,* compact surfaces of constant negative curvature.

3. *Horocycles and stable manifolds.* The unit tangent bundle $S\mathbb{H}$ can be identified with $\mathrm{PSL}(2, \mathbb{R})$ by $A \mapsto A\mathfrak{i}$, where $\mathfrak{i}$ is the upward unit vector at i. In particular, $\begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{t/2} \end{pmatrix} \mathfrak{i}$ parametrizes an orbit of the geodesic flow (corresponding to the geodesic $t \mapsto ie^t$), so the geodesic flow is algebraically described by the translations of $\mathrm{PSL}(2, \mathbb{R})$ by $g^t = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{t/2} \end{pmatrix}$. The strong stable manifold of $\mathfrak{i}$ is the family of upward unit normals on

$\mathbb{R}+i$. The images of $\mathbb{R}+i$ under fractional linear transformations are called *horocycles*. They are horizontal lines and circles tangent to $\mathbb{R}$ and hence orthogonal to geodesics. $\mathbb{R}+i$ is parametrized (with unit speed) by $\begin{pmatrix}1 & t\\ 0 & 1\end{pmatrix}$, whose left translations define the *horocycle flow* h_t on $\mathrm{PSL}(2,\mathbb{R})$ (Section 4.3f). This flow is fairly clearly parabolic. Locally it has precisely the "triangular" orbit structure described in Section 8.1c (not to be confused with upper triangularity of the matrix). A simple way of seeing this is to note that $\begin{pmatrix}1 & t\\ 0 & 1\end{pmatrix}\begin{pmatrix}a & b\\ c & d\end{pmatrix} = \begin{pmatrix}a+tc & b+td\\ c & d\end{pmatrix}$, so these left translations act as a diagonal action of two shears on the space of 2×2 matrices, of which $\mathrm{PSL}(2,\mathbb{R})$ is the submanifold $\det^{-1}(\{1\})$.

FIGURE 6.5. Geodesics and horocycles

Geometrically, horocycles are limit circles (hence the name) in that $\mathbb{R}+i = \partial\big(\bigcup_{t>0} B(ie^t,t)\big)$. Generally, horocycles are obtained as boundaries of limit balls. This latter construction has the virtue of being independent of the algebraic structure, constant curvature, or dimension. On a simply connected Riemannian manifold with negative sectional curvature $\bigcup_{t>0} B(c(t),t)$ is convex for any geodesic c, and its boundary is called the *horosphere* for $\dot c(0)$. The "inward" unit normal vector field (the one containing $\dot c(0)$) is the strong stable manifold of $\dot c(0)$. The other unit vector field is the strong unstable manifold for $-\dot c(0)$ [**KH**].

Geodesic and horocycle flows satisfy the following commutation relation, which can be deduced from direct computation:

$$g_{-s}\circ h_t\circ g_s = h_{e^s t} \text{ for any } s,t\in\mathbb{R}.$$

It plays a central role in the study of dynamics of horocycle flows. It also allows to find by direct algebraic calculations many dynamical properties of geodesic flows that in general follow from hyperbolicity.

6. The core theory

Much of the theory of uniformly hyperbolic dynamical systems is described in [**S-H, S-C, S-P**]. The nonuniform case is discussed in [**S-BKP, S-W**]. We present the common mechanism behind the core results of the theory and then showcase a few basic and important results.

a. Applications of fixed point results. The Hyperbolic Fixed Point Theorem 6.1.1 (or the Contraction Principle) is the device that underlies most of the development of the core of at least the uniform hyperbolic theory. It can be applied in proofs by viewing the desired result as providing an object, such as a periodic point, an invariant manifold or a conjugacy, and restating the asserted qualities of this object as a fixed point property. This involves an appropriate, often infinite-dimensional, space of candidate objects on which one can construct an action derived from the dynamical system at hand. The successes of this approach are due to the notable fact that these derived actions inherit hyperbolicity from the underlying dynamics. While not unrelated to the cleverness one shows in setting

up the argument, this phenomenon is intrinsic to hyperbolicity in that it reflects the fact that the linearization provides direct and meaningful information about the dynamics.

This approach yields the core of the theory, *i.e.,* closing and shadowing results, local and global stability, as well as invariant foliations or, more generally laminations. With these tools in hand the development can proceed to finer investigations of the orbit structure (spectral decomposition, specification, Markov partitions) as well as further local and global properties, the study of invariant measures, smooth aspects, *etc.*

It is interesting that there can be quite different proofs of one result, both along these lines. The Hadamard–Perron Theorem is a case in point (as described also in [**S-H**]). It can be proved by the graph transform method of Hadamard, which obtains the unstable manifold as a fixed point of the action by the dynamics on candidate manifolds [**KH**]. But one can also follow Perron's approach, in which the local stable manifold is obtained as the set of bounded orbits [**Y3**]. In this approach the candidate objects are not orbits, by the way, and the fixed point property is not boundedness. Instead, an action is defined on bounded "things", and being an orbit is a fixed point property.

b. The Anosov Closing Lemma. Let (X, d) be a metric space, $U \subset M$ open and $f\colon U \to X$. For $a \in \mathbb{Z} \cup \{-\infty\}$ and $b \in \mathbb{Z} \cup \{\infty\}$ a sequence $(x_n)_{a<n<b}$ in U is called an *ϵ-orbit* or ϵ-pseudo-orbit for f if $d(x_{n+1}, f(x_n)) < \epsilon$ for all $a < n$, $n + 1 < b$.
It is said to be *periodic* if $-\infty < a < b < \infty$ and $d(f(x_b), x_a) < \epsilon$.
It is said to be *δ-shadowed* by the orbit $\mathcal{O}(x)$ of $x \in U$ if $d(x_n, f^n(x)) < \delta$ for all $a < n < b$.

THEOREM 6.6.1 ([**KH**, Theorem 6.4.15]). *Let Λ be a hyperbolic set for $f\colon U \to M$. Then there exists an open neighborhood $V \supset \Lambda$ and C, $\epsilon_0 > 0$ such that for $\epsilon < \epsilon_0$ and any periodic ϵ-orbit $(x_0, \dots, x_{m-1}) \subset V$ there is a point $y \in U$ such that $f^m(y) = y$ and $\operatorname{dist}(f^k(y), x_k) < C\epsilon$ for $k = 0, \dots, m-1$. In fact, $\operatorname{dist}(f^k(y), f^k(x)) < C\,\alpha^{\min(k,m-k)} \cdot (\operatorname{dist}(x, y) + \operatorname{dist}(f^m(x), f^m(y)))$.*

In particular, recurrent points are limits of periodic points. This is a strong statement about the abundance of periodic points. If Λ is locally maximal the periodic point y is in Λ.

SKETCH OF PROOF. f acts by $(x_0, \dots, x_{m-1}) \mapsto (f(x_{m-1}), f(x_1), \dots, f(x_{m-2}))$ on m-tuples of points [**KH**]. This action is easily seen to be hyperbolic and its fixed points are periodic orbits of f. Apply Theorem 6.1.1. □

c. The Shadowing Lemma. (See also [**KH**, Theorem 18.1.2].)

THEOREM 6.6.2. *Let M be a Riemannian manifold, $U \subset M$ open, $f\colon U \to M$ a diffeomorphism, and $\Lambda \subset U$ a compact hyperbolic set for f. Then there is a neighborhood $U(\Lambda) \supset \Lambda$ such that whenever $\delta > 0$ there is an $\epsilon > 0$ such that every ϵ-orbit in $U(\Lambda)$ is δ-shadowed by an orbit of f.*

A much more powerful counterpart of this result is the Shadowing Theorem [**KH**, Theorem 18.1.3] about coherent shadowing of families of orbits.

SKETCH OF PROOF. Assume (after possible extension) that the pseudo-orbit is bi-infinite. Then define an action as before: $(x_i)_{i\in\mathbb{Z}} \mapsto (f(x_{i-1}))_{i\in\mathbb{Z}}$ [**KH**]. Fixed points give orbits for f and hyperbolicity is easy to check. Apply Theorem 6.1.1. □

d. The Hartman–Grobman Theorem.

THEOREM 6.6.3. *If T is a bounded hyperbolic linear map in a Banach space E and f is sufficiently Lipschitz-close to T then there is a homeomorphism h such that $h \circ T = f \circ h$.*

This result is usually stated in a form more suitable for local analysis [**S-H, KH**], which is where its utility lies. Since the result is not used in this survey we chose this form in order to make it easier to outline the proof.

SKETCH OF PROOF. [**Y3**] Suppose both $f = T + \Delta F$ and $g = T + \Delta g$ are Lipschitz-perturbations of T and seek $h = \mathrm{Id} + \Delta h$ such that $f \circ h = h \circ g$. This is equivalent to $\mathcal{T}(\Delta h) + \Delta\mathcal{F}(\Delta h) = \Delta h$, where $\mathcal{T}(\Delta h) := T \circ \Delta h \circ g^{-1}$ and

$$\Delta\mathcal{F}(\Delta h) = \Delta f \circ (\mathrm{Id} + \Delta h) \circ g^{-1} + T \circ g^{-1} - \mathrm{Id}\,.$$

Check that T is hyperbolic and $\Delta\mathcal{F}$ is Lipschitz-small, then apply the Hyperbolic Fixed Point Theorem 6.1.1. □

The introduction of g is a device for showing invertibility of h by symmetry and uniqueness.

e. Structural stability of hyperbolic sets. Compact locally maximal hyperbolic sets are strongly C^1 structurally stable.

As mentioned earlier, this is one of the outstanding features of uniformly hyperbolic systems. This result can also be proved almost directly by applying the Hyperbolic Fixed Point Theorem 6.1.1 to a cleverly constructed action on candidate conjugacies and embeddings [**Y3**]. Alternatively, one can employ a shadowing theorem that gives coherent shadowing of entire families of orbits [**KH**, Theorems 18.1.3, 18.2.1].

By the way, with the invariant cone definition of hyperbolicity (Section 6.4a), it is immediate that a C^1 perturbation g of an embedding f with a hyperbolic set Λ in an isolating neighborhood U is hyperbolic on its maximal invariant set $K \subset U$, but structural stability requires that K is homeomorphic to Λ, and in particular nonempty.

This fundamental semilocal result is the starting point for the solution of global problems of C^1-structural stability and Ω-stability (Section 5.2f,Section 6.7i). Necessary conditions for the former were found by Palis and Smale [**PaS**] and for the latter essentially by Robbin [**R**], with the concluding step by Robinson [**Ro**]. In both cases, hyperbolicity on the nonwandering set is the principal condition, supplemented by a transversality condition on global stable and unstable manifolds (described in the next subsection).

f. Invariant laminations. The Hadamard–Perron Theorem (Theorem 6.3.1) gives, in each of these classes of hyperbolic dynamical systems, local stable and unstable manifolds. There are profound differences, however, in the structure of the resulting invariant laminations.

1. *Local leaves in the uniform case.* For a compact hyperbolic set one obtains local stable and unstable manifolds for every point, and these local leaves are compatible in that the intersection of any two of them is open in either one. These local leaves are also of uniform size, *i.e.*, each of them contains an ϵ-ball around their base point, where $\epsilon > 0$ is uniform:

Let Λ be a hyperbolic set for a C^1 embedding $f\colon V \to M$ such that Df on Λ admits a (λ, μ)-splitting with $\lambda < 1 < \mu$. Then for each $x \in \Lambda$ there is a pair of embedded C^1

discs $W^s(x)$, $W^u(x)$, called the *local stable manifold* and the *local unstable manifold* of x, respectively, such that

(1) $T_xW^s(x) = E^-_x$, $\quad T_xW^u(x) = E^+_x$;
(2) $f(W^s(x)) \subset W^s(f(x))$, $f^{-1}(W^u(x)) \subset W^u(f^{-1}(x))$;
(3) for every $\delta > 0$ there exists $C(\delta)$ such that for $n \in \mathbb{N}$

$$\operatorname{dist}(f^n(x), f^n(y)) < C(\delta)(\lambda + \delta)^n \operatorname{dist}(x, y) \text{ for } y \in W^s(x),$$
$$\operatorname{dist}(f^{-n}(x), f^{-n}(y)) < C(\delta)(\mu - \delta)^{-n} \operatorname{dist}(x, y) \text{ for } y \in W^u(x);$$

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

$$W^s(x) = \{y \mid f^n(y) \in O_{f^n(x)}, \qquad n = 0, 1, 2, \dots\},$$
$$W^u(x) = \{y \mid f^{-n}(y) \in O_{f^{-n}(x)}, \quad n = 0, 1, 2, \dots\}.$$

COROLLARY 6.6.4. *The restriction of a diffeomorphism or flow to a hyperbolic set is expansive (see Section 2.4d).*

For $d(x, y)$ sufficiently small the intersection $W^s(x) \cap W^u(y)$ consists of exactly one point $[x, y]$. Every point in its orbit is said to be *heteroclinic* to x and y. If the intersection is transverse and x, y are periodic then interesting dynamics arises [**KH**, Theorem 6.5.5].

The local stable and unstable manifolds are not unique but by 3 and 4 for any two local stable manifolds $W^s_1(x)$ and $W^s_2(x)$ satisfying the assertions of the theorem their intersection contains an open neighborhood of x on each of them. Equivalently, one can say that for some $n \geq 0$ one has $f^n(W^s_1(f^{-n}(x))) \subset W^s_2$ and $f^n(W^s_2(f^{-n}(x))) \subset W^s_1$. In fact such a number n can be chosen uniformly for all $x \in \Lambda$. The same is true for local unstable manifolds with n replaced by $-n$.

2. *Global manifolds.* This also implies that global stable and unstable manifolds

$$\widetilde{W}^s(x) = \bigcup_{n=0}^{\infty} f^{-n}(W^s(f^n(x))) \tag{6.1}$$
$$\widetilde{W}^u(x) = \bigcup_{n=0}^{\infty} f^n(W^u(f^{-n}(x)))$$

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

$$\widetilde{W}^s(x) = \{y \in U \mid \operatorname{dist}(f^n(x), f^n(y)) \to 0, \quad n \to \infty\},$$
$$\widetilde{W}^u(x) = \{y \in U \mid \operatorname{dist}(f^{-n}(x), f^{-n}(y)) \to 0, \quad n \to \infty\}.$$

These manifolds are injectively immersed Euclidean spaces, but by no means embedded. They are commonly dense.

For flows one analogously obtains *strong* and *weak* stable and unstable leaves tangent to the corresponding subbundles. Weak leaves are foliated by strong ones.

3. *Center manifolds.* (See also [**R, HPS, Kl**].) Note that taking $\lambda < 1 = \mu$ in the Hadamard–Perron Theorem one obtains *center-unstable* manifolds W^{cu} and taking $\lambda = 1 < \mu$ gives *center-stable* manifolds W^{cs}. Their intersection gives (possibly zero-dimensional) *center manifolds*, which are characterized by contraction/expansion that is slower than λ and μ, respectively.

DEFINITION 6.6.5. If $TM = E^+ \oplus E^0 \oplus E^-$ such that $\|Df^{-n}{\restriction}_{E^+}\|\|Df^n{\restriction}_{E^0}\| \leq \lambda < 1$ and $\|Df^{-n}{\restriction}_{E^0}\|\|Df^n{\restriction}_{E^-}\| \leq \lambda$ for $n \in \mathbb{N}$ then E^0 is called a center direction.

This expresses the fact that E^+ expands more than anything in E^0 and E^- contracts more than anything in E^0. If $Df^n{\restriction}_{E^0}$ has subexponential growth then this corresponds to the partially hyperbolic situation.

If $E^0 \neq \{0\}$ then there are nontrivial center manifolds tangent to E^0. In the case of a C^r diffeomorphism ($r \in \mathbb{N} \cup \{\infty\}$) the center-stable and center-unstable manifolds are C^k for any integer $k < r + 1$, at least in a neighborhood of the base point. (The weak-stable manifolds of a flow, however, are globally as smooth as the flow.) Unlike the stable and unstable manifolds, center manifolds are *not* uniquely defined, even though it may thus appear from the above theorem. The problem is that in concrete situations the above result is applied via a localization procedure, which is benign with respect to strong leaves, but not center leaves. For example, the flow on $\mathbb{R}^2$ generated by the vector field $(x^2, -y)$ fixes the origin (only), for which the y-axis is the stable manifold but any curve $Ce^{1/x}\chi_{(-\infty,0)}(x)$ is a C^∞ center stable manifold [**R**]. Local uniqueness of W^{cs} holds, however, when every neighborhood $0 \in U \subset W^{cs}(0)$ contains a neighborhood $0 \in V \subset U$ such that $\bigcup_{n\in\mathbb{N}} f^n(V) \subset U$ [**HPS**, Theorem 5A.3]. This assumption averts problems with the localization procedure is innocuous. Global uniqueness of center manifolds is a different issue and can sometimes be assured by cone field conditions similar to those in Section 6.4a.

A manifold is said to be *normally hyperbolic* if the rate conditions from Definition 6.6.5 hold for its tangent bundle in place of E^0. The theory of normally hyperbolic manifolds gives persistence results for perturbations of the map, and regularity results when sharper estimates of the rate differences are imposed [**HPS, R**].

7. Developments of the theory

This section is essentially an abbreviated version of parts of [**KH**] with some mention of subjects surveyed with slightly more detail in [**S-H**]. It is included for the sake of readers interested in a compact overview of hyperboic theory

a. Spectral decomposition. The structure of the set of periodic points of a hyperbolic set is rather intricate. This makes it an interesting object of study, but periodic points are also remarkably useful as a technical tool in the study of hyperbolic sets. This is due to their abundance both in the sense of exponential growth of the number of periodic points with the period and that of reflecting much of the nonperiodic dynamics in ways that made precise below. Density of periodic points in the nonwandering set (Anosov Closing Lemma, Theorem 6.6.1) together with that of stable and unstable manifolds implies that hyperbolic sets decompose into topologically transitive components:

THEOREM 6.7.1 (Spectral decomposition). *Let M be a Riemannian manifold, $U \subset M$ open, $f\colon U \to M$ a diffeomorphism, and $\Lambda \subset U$ a compact locally maximal (isolated) hyperbolic set for f (see Section 2.2e). Then there exist disjoint closed sets $\Lambda_1, \ldots, \Lambda_m$ and a permutation σ of $\{1, \ldots, m\}$ such that $NW(f{\restriction}_\Lambda) = \bigcup_{i=1}^m \Lambda_i$, $f(\Lambda_i) = \Lambda_{\sigma(i)}$, and when $\sigma^k(i) = i$ then $f^k{\restriction}_{\Lambda_i}$ is topologically mixing.*

This result is among the applications of the core theory (as opposed to being a direct consequence of a fixed point result). It uses the stable manifold theorem. The transitive components are obtained more or less constructively. Define an equivalence relation on $\mathrm{Per}(f_{\restriction\Lambda})$ by $x \sim y$ if and only if $W^u(x) \cap W^s(y) \neq \varnothing$ and $W^s(x) \cap W^u(y) \neq \varnothing$ with both intersections transverse at at least one point. Then each Λ_i is the closure of an equivalence class [**KH**, Theorem 18.3.1].

b. The Livschitz Theorem. (See also [**KH**, Theorem 19.2.1.].)

THEOREM 6.7.2. *If Λ is a transitive compact locally maximal hyperbolic set for an embedding $f\colon U \to M$ and $\varphi\colon \Lambda \to \mathbb{R}$ a Hölder continuous function such that $f^k(x) = x \implies \sum_{i=0}^{k-1} \varphi(f^i(x)) = 0$ then $\varphi = \psi \circ f - \psi$ for some Hölder continuous $\psi\colon \Lambda \to \mathbb{R}$,* i.e., *(the cocycle generated by) φ is a coboundary. The transfer function ψ is unique up to an additive constant.*

SKETCH OF PROOF. Pick a point x with dense orbit. Fix $\psi(x)$ and use $\varphi = \psi \circ f - \psi$ to define ψ on $\mathcal{O}(x)$. By the Anosov Closing Lemma (with exponential closeness) the assumption on periodic orbits and Hölder continuity imply uniform continuity (in fact, uniform Hölder continuity) of φ on $\mathcal{O}(x)$, hence the existence of a unique extension to $\mathcal{O}(x) = \Lambda$. □

This result is yet another manifestation of the abundance of periodic data: Periodic points determine cohomology completely.

There are additional smoothness results in this situation to the effect that ψ is as regular as φ [**S-H**].

c. Specification and equilibrium states. The spectral decomposition makes the following theorem of Bowen pertinent to any locally maximal hyperbolic set:

THEOREM 6.7.3. *A locally maximal topologically mixing hyperbolic set has the specification property (Section 4.4f).*

SKETCH OF PROOF [**KH**]. Use that stable and unstable leaves are uniformly dense in a mixing hyperbolic set. If x is the last point of the first orbit segment of the specification and y is the first point of the next segment let $z = W^u_\epsilon(x) \cap W^s(f^{-N}(y))$. Taking N large enough (this depends only on ϵ) ensures that $f^N(z)$ ϵ-shadows the second orbit segment, and the first orbit segment is ϵ-shadowed also. Continue with the shadowing point at the end of the second segment, connecting to the third in the same way. The changes to the earlier portions are exponentially small, so they settle down independently of the number of segments to be shadowed. □

Together with expansivity (Corollary 6.6.4) this implies that there is a unique equilibrium state (Section 4.4g) for every Hölder continuous function on a mixing compact locally maximal hyperbolic set. Invariant measures therefore abound.

Other consequences are positivity of topological entropy and an abundance of periodic orbits, in terms of number as well as the possibility of approximating orbit segments and invariant measures (Section 4.4).

As mentioned in Section 4.4g, this gives a fine growth asymptotic for periodic points: $c_1 e^{nh_{\text{top}}(f)} \leq P_n(f) \leq c_2 e^{nh_{\text{top}}(f)}$ [**KH**, Theorem 18.5.5], which, together with specification, in turn implies positive entropy if there is more than one point. Orbit growth estimates

as fine as those in Proposition 2.6.6 are possible via coding. A detailed discussion of equilibrium states can be found in [**S-C**].

d. Sinai–Ruelle–Bowen measure. A hyperbolic set Λ is an attractor if and only if for any point $x \in \Lambda$ the unstable manifold $W^u(x)$ lies in Λ.

Among the many invariant Borel probability measures for a dynamical system with a hyperbolic attractor, one is deemed especially noteworthy. This *Sinai–Ruelle–Bowen measure* [**S-H, S-C**] is characterized as the asymptotic distribution of *Lebesgue* a.e. point in a neighborhood of the set [**S-C**]. This suggests that computer pictures represent an attractor by approximating the Sinai–Ruelle–Bowen measure on it, or that this is the "physically observed" measure. It is obtained as the equilibrium state for $\log(J^u f)$, where $J^u f$ is the Jacobian of f on unstable leaves. Therefore it has all the expected stochastic complexity (Section 4.4g).

e. Absolutely continuous invariant measures for Anosov systems. For an Anosov system there are two Sinai–Ruelle–Bowen measures associated with positive and negative time asymptotics. They coincide if and only if there is an absolutely continuous invariant measure which then coincide with the Sinai–Ruelle–Bowen measure and is smooth. A useful criterion for the existence of such a measure follows from the Livschitz Theorem 6.7.2 once one uses the cohomological criterion for existence of an absolutely continuous invariant measure (Section 5.2m).

THEOREM 6.7.4. *A topologically transitive Anosov diffeomorphism f has an absolutely continuous invariant measure if and only if $f^n(x) = x \Rightarrow Jf^n(x) = 1$.*

f. Ergodicity of volume. Thus ergodicity of volume-preserving Anosov systems is a corollary of the theory of equilibrium states (once volume is identified as the Sinai–Ruelle–Bowen measure [**KH**, Theorem 20.4.1]).

However, the original approach retains its independent value because it can also be applied to various classes of partially hyperbolic and nonuniformly hyperbolic systems. It is based on the *Hopf argument*, first used by E. Hopf for geodesic flows on surfaces and extended by Anosov to general Anosov systems [**A**]. A contemporary rendering of the argument for geodesic flows is in [**Bm**], and [**KH**, Theorem 5.4.16] gives the argument in a simple situation. The central analytic ingredient is absolute continuity of the local holonomy maps of stable and unstable foliations (Section 5.1e). This produces essential openness of ergodic components. Then density of stable an unstable leaves leads to global ergodicity.

A good example of the applicability of the Hopf argument in the partially hyperbolic situation is the proof of ergodicity of the time-one map for a geodesic flow on a compact manifold of negative curvature. The additional structural feature is density of *strong* stable and unstable leaves. This follows from the fact that geodesic flows preserve a contact structure, which renders the strong foliations *completely nonintegrable* (Section 5.5a). This means that any two points in the phase space can be connected by a path of finitely many segments, each inside a strong leaf. Put differently, time is only locally meaningful, and one can achieve a change in time by a path that never has a time component. This *accessibility property* of the strong foliations plays a role in partially hyperbolic systems in connection with the stable ergodicity problem (Section 6.9c).

The proof of necessity of the periodic orbit condition in Theorem 6.7.4 is based on a modified version of the Hopf argument.

g. Local product structure, Markov partitions. We say that a hyperbolic set has *local product structure* if it is closed under the map $[\cdot,\cdot]$ (Section 6.6f). This is equivalent to local maximality.

To build Markov partitions define a *rectangle* to be a subset R of a compact locally maximal hyperbolic set Λ of small diameter that is closed under $[\cdot,\cdot]$ and the closure of its interior (in the topology of Λ). For $x \in R$ and $i = s, u$, let $W^i_R(x) := R \cap W^i_{\text{loc}}(x)$. A *Markov partition* is a finite cover of Λ by rectangles R_i with pairwise disjoint interiors such that if $x \in \operatorname{Int} R_i \cap f^{-1}(\operatorname{Int} R_j)$ then $W^u_{R_j}(f(x)) \subset f(W^u_{R_i}(x))$ and $f(W^s_{R_i}(x)) \subset W^s_{R_j}(f(x))$. Markov partitions of arbitrarily small diameter always exist for compact locally maximal hyperbolic sets [**S-C**], [**KH**, Section 18.7], and they provide a coding, *i.e.*, an almost-isomorphism (Section 2.2f) to a topological Markov chain.

This coding preserves entropy, and can easily be shown to be benign as far as periodic points go. One consequence is the following [**KH**]:

COROLLARY 6.7.5. *Let Λ be a compact locally maximal hyperbolic set for f. Given a Markov partition, $f_{\restriction\Lambda}$ is a factor (via an almost-conjugacy) of the topological Markov chain σ_A defined by allowed itineraries, which is topologically transitive (mixing) if and only if $f_{\restriction\Lambda}$ is, and has the same topological entropy as $f_{\restriction\Lambda}$. If Λ is totally disconnected then the factor map is a conjugacy. If $f_{\restriction\Lambda}$ is topologically mixing then $|P_n(f_{\restriction\Lambda}) - e^{nh_{top}(f_{\restriction\Lambda})}| < K\lambda^n$ for some $\lambda < e^{h_{top}(f_{\restriction\Lambda})}$, $K > 0$ (by Proposition 2.6.6).*

h. Stability, moduli and smooth classification. Some results related to structural stability are discussed in [**S-H**]. For Anosov systems, a remarkable extension of structural stability is that all known examples of Anosov diffeomorphisms have been classified: Each is topologically conjugate to a hyperbolic automorphism of an infranilmanifold, of which toral automorphisms are the prime example. At the same time, it remains unknown whether there may be further examples of Anosov diffeomorphisms. For Anosov flows, there is no classification, and the question of existence of yet unknown Anosov flows seems even more open [**S-H**]. A related issue is that all known Anosov diffeomorphisms are topologically transitive, but it is not known whether this is the case for all Anosov diffeomorphisms. For flows, there are nontransitive examples [**FrW**], but it is not clear just how exceptional this is.

There is the large and intricate array of moduli of smooth conjugacy provided by periodic points alone such as local normal forms around such points. Furthermore, ergodic invariant measures, which can be viewed as generalizations of periodic orbits, provide further invariants such as Lyapunov characteristic exponents. Still, there are some classes of hyperbolic systems in which a classification up to a smooth conjugacy is possible. One example is that smooth area-preserving Anosov diffeomorphisms on $\mathbb{T}^2$ are smoothly classified by their eigenvalue data as defined in Section 5.2e:

THEOREM 6.7.6 ([**KH**, Theorem 20.4.3]). *Suppose $f, g\colon \mathbb{T}^2 \to \mathbb{T}^2$ are C^2 area-preserving Anosov diffeomorphisms and $f \circ h = h \circ g$ for a homeomorphism h homotopic to the identity. Then f and g are C^1 conjugate if and only if their eigenvalues at corresponding periodic points p and $h(p)$ coincide.*

SKETCH OF PROOF. Both h and the logarithm $\varphi_f = \log(J^u f)$ of the unstable Jacobian are Hölder continuous [**S-H, KH**], so $\psi_f := \log(J^u f) \circ h$ has a unique equilibrium state μ_f. The hypothesis implies that ψ_f and $\varphi_g := \log J^u g$ have the same sums over periodic orbits. By the Livschitz Theorem they are cohomologous, which implies that they have the same equilibrium state, *i.e.*, $\mu_f =$ area. Equilibrium states are equivariant under Hölder homeomorphism by construction, *i.e.*, μ_f is the pullback of the equilibrium state $\nu_f =$ area for φ_f by h. This means that h preserves area.

In dimension 2 the stable and unstable foliations for f and g are C^1, *i.e.*, there are local C^1 coordinates for f and g in which the foliations are linear. The image of area under these coordinates is a measure with continuous density, which therefore induces continuous densities on every leaf. Since h preserves area it preserves these densities. This shows that on a leaf h is locally obtained by integration of a continuous density, hence is itself C^1. One can show that this implies C^1-smoothness of h. □

In fact, if $f, g \in C^\infty$ then so is h. Here are some interesting consequences of this result. The classification of Anosov diffeomorphisms on tori (or only $\mathbb{T}^2$) gives:

COROLLARY 6.7.7. *If $f\colon \mathbb{T}^2 \to \mathbb{T}^2$ is a C^2 area-preserving Anosov diffeomorphism and $\lambda \in \mathbb{R}$ such that for every $x \in \operatorname{Fix} f^n$ the expanding eigenvalue of $Df^n(x)$ is λ^n, then f is C^1 conjugate to a linear automorphism.*

The case of φ_f a coboundary takes the following form:

COROLLARY 6.7.8. *A C^2 area-preserving Anosov diffeomorphism $f\colon \mathbb{T}^2 \to \mathbb{T}^2$ with $h_\lambda(f) = h_{top}(f)$ is C^1 conjugate to a linear automorphism.*

i. The stability theorem. The celebrated Mañé–Hayashi Stability Theorem [**M3, Hy**] characterizes hyperbolicity as necessary for structural stability. Mañé showed that the sufficient conditions established by Robbin [**R**] and Robinson [**Ro**] (Section 6.2e) are necessary for C^1 structural stability. The precise statement of this result and the related NW-Stability (or Ω-Stability) Theorem are given in [**S-H**], but the essence is that structurally stable systems have hyperbolic nonwandering set with a transversality condition on invariant manifolds. Note that the "dissipative part" is included now, unlike in the above discussion of stability of hyperbolic sets. The continuous-time counterpart of Mañé's result yielded only much later. It was proved by Hayashi. A central ingredient of the proof is the Hayashi Connecting Lemma [**S-H, Hy**], which is based on the Pugh and Mañé closing lemmas (Section 5.2p).

The stability theorem is one of the high points in the development of smooth dynamics. A major component of the Smale program was the intent to pursue a classification of diffeomorphisms by topological type. The Stability Theorem identifies the open equivalence classes, *i.e.*, those systems, where such a scheme is feasible.

This, combined with the complete classification of known Anosov diffeomorphisms, also calls attention to a different interpretation of the term "roughness" for structural stability (which was coined by the seminal 1937 paper "Systèmes grossiers" by Andronov and Pontrjagin). It was intended to convey roughness in the sense of imperviousness to perturbation, but if one takes the classification results for Anosov diffeomorphisms to indicate a certain paucity of examples, one may think of "roughness" as expressing excessive strength of the hypothesis of uniform hyperbolicity and inviting the study of dynamical systems under "less rough" assumptions. Accordingly, the development of the theory of

nonuniformly hyperbolic systems was motivated to a large extent to meet the demand in applications (inside and outside of mathematics) to provide a theory adapted to systems with less than uniform hyperbolicity.

8. The theory of nonuniformly hyperbolic systems

A serious survey of the theory of nonuniform hyperbolicity is given by [**S-BKP**], and introduction can be found in [**KH**, Supplement] and in [**P**]; [**S-K**] discusses some aspects related to geometry. A comprehensive treatment is in preparation [**BKP**]. This is not the place to give the details of this extensive field. Nevertheless, the spirit of the work as well as some aspects of its present state can be outlined.

a. Contrast with the uniform case. Although this theory shares with the uniform one the use of linearization and other aspects of smoothness of the dynamical system, one pervasive distinction is that the heart of the approach is in invariant measures. This may be viewed as the intrinsically natural generalization, but is also closely connected to the Oseledets Multiplicative Ergodic Theorem [**S-BKP**, Oseledets Multiplicative Ergodic Theorem], [**Os, Rg, W2**]. Nevertheless, some results do not involve measure theory in their statements. A nice example is that a $C^{1+\alpha}$ diffeomorphism of a surface with positive topological entropy has a (hyperbolic) periodic point [**KH, P**].

Before describing the theory of nonuniform hyperbolicity, it is good to recall the collection of facts that embody the hyperbolic paradigm in the uniform case: Expansivity, closing and shadowing lemma, Livschitz theorem, spectral decomposition, Markov partitions, equilibrium states, absolute continuity of foliations, ergodicity of volume, the Bernoulli property of volume. After introducing the framework in which the theory is developed, we give the structural results aimed at recovering the features just listed for the uniform case. This leads to a useful comparison between the two situations.

b. Lyapunov exponents and tempering. A proper definition of nonuniformly hyperbolic systems has to be preceded by that of Lyapunov exponents. Even when one is not interested in the maximal possible generality, the natural setting is that of cocycles. However, for here we expressly consider the derivative extension which due to the triviality of skew products in the measurable category (Proposition 3.2.3) can be given by a cocycle. The Lyapunov exponents of an orbit $\mathcal{O}(x)$ are the exponential asymptotic growth rates of vectors under iteration of the differential, *i.e.*, $\chi^+(x,v) := \lim_{m\to\infty}(1/m)\log\|Df^n|_x v\|$. The Oseledets Multiplicative Ergodic Theorem [**S-BKP**, Oseledets Multiplicative Ergodic Theorem], [**Os, Rg, W2**] shows that with respect to an f-invariant Borel probability measure this is well-defined a.e. (on the *regular set*) and, at a given x, attains at most $\dim M$ different values. Furthermore, there is a Lyapunov decomposition into subspaces corresponding to the various Lyapunov exponents, whose dimension defines the multiplicity of the corresponding exponent.

An important related device is that of *tempering*, which introduces coordinate changes that bring the differential into a block form adapted to the Lyapunov decomposition and the Lyapunov exponents. The price is a distortion (of lengths and angles) that may grow exponentially, but at an arbitrarily slow rate.

c. Hyperbolic measures and Pesin sets. Hyperbolic measures are those for which all Lyapunov exponents are nonzero a.e. Note that up to this point there were no hyperbolicity assumptions of any kind. In fact, one of the strengths of the theory of nonuniformly hyperbolic systems is that it can make some interesting statements about dynamical systems without any such assumption. On the other hand, the theory has brought new insights into uniformly hyperbolic dynamics as well.

For hyperbolic measures, one has exponential behavior a.e., which is a much weaker assumption than uniform hyperbolicity. To extend the theory of uniformly hyperbolic dynamical systems to this situation one uses that for any given hyperbolicity estimate (with fixed constants) there is a (possibly empty) set, where this estimate holds, and that the union of these sets is the entire regular set. In other words, there are sets of arbitrarily large measure, called *Pesin sets*, on which one has uniformly hyperbolic conditions. One of the difficulties is that these are not usually invariant. Nevertheless, one obtains invariant laminations of a measurable kind. Often it is easier to work with approximations thereof (admissible manifolds [**KH**]).

d. Stable manifolds. The Hadamard–Perron Theorem applied in the nonuniform case also gives invariant laminations, but instead of uniformity in the size of leaves there is a measurable lower bound only. The same goes for the angle between stable and unstable leaves. In the uniform case the picture of stable leaves along an unstable one can be arranged (via local coordinates) as a horizontal line (unstable leaf) crossed by vertical ones. The nonuniform situation is best imagined as a horizontal line with a "fence" of vertical line segments, in the gaps of which there are somewhat crooked short line segments, between which there are much shorter line segments, some of them possibly quite close to horizontal, *etc.*

Their lack of regularity nonwithstanding, the invariant families of stable and unstable manifolds do retain absolute continuity. Among the consequences is that the ergodic decomposition of a hyperbolic measure consists of sets of positive measure, in particular, there are at most countably many components.

e. Structural theory. Remarkably, several of the central results of the uniform theory have counterparts in this setting [**KH**]. Among these are the Anosov Closing Lemma (which produces a hyperbolic periodic point), the Shadowing Lemma, the existence of Markov partitions (which here are approximate), and the Livschitz Theorem. There is also a spectral decomposition of a Pesin set for a hyperbolic measure into a finite union of orbit closures. While there is no structural stability, a vestige of it remains in certain stability properties of hyperbolic measures under perturbation: If μ is a hyperbolic measure for $\lim f_n$ then it is a weak limit of hyperbolic measures μ_n for the f_n.

1. *Entropy and horseshoes.* The theory also contains a beautiful result in line with our division into elliptic–parabolic and hyperbolic dynamical systems: The entropy of an ergodic hyperbolic measure, if positive, is approximated arbitrarily well by the topological entropies of horseshoes [**KH**]. In the case of surfaces, positive entropy of a measure implies hyperbolicity and hence by the Variational Principle the topological entropy is approximated by that of horseshoes. In other words, horseshoes are *the* mechanism for the production of exponential orbit growth.

For interval maps the same happens even without smoothness.

f. Sinai–Ruelle–Bowen measure. Because of its important role in the study of attractors, especially numerical experiments, there is great interest in producing a counterpart of the Sinai–Ruelle–Bowen measure outside the uniformly hyperbolic context. Simple examples suggest some difficulty. Smooth systems may fail to have a Sinai–Ruelle–Bowen measure even if hyperbolicity breaks down only in the most benign way. The example is a hyperbolic automorphism of $\mathbb{T}^2$ perturbed so as to remain hyperbolic except at the fixed point, where the derivative has an eigenvalue one and the other less than one [**HY**]. (This is also an example of nonuniqueness of equilibrium states [**K3**].) It is interesting that the introduction of benign singularities to the uniformly hyperbolic setting is not nearly as problematic [**C**]. When studying attractors, an essential problem is that sets of positive Lebesgue measure may have asymptotic distribution unrelated to the invariant measure of interest.

Nevertheless, there are some remarkable successes. First of all, the equivalence of the three characterizations of the Sinai–Ruelle–Bowen measure that constitute its main interest (equilibrium state, absolute continuity on unstable leaves, asymptotic distribution for Lebesgue-a.e. points [**S-H, S-BKP**]), have a useful counterpart in the nonuniform situation. A measure satisfying Pesin's entropy formula [**S-BKP**] (entropy is the integral of the positive Lyapunov exponents) is also absolutely continuous on unstable leaves and represents the asymptotic distribution of a set of points of positive Lebesgue measure [**Ld**]. Therefore it is clear what to look for, and such a measure is again called a Sinai–Ruelle–Bowen measure.

The other success is that for some important attractors of nonuniform type, a Sinai–Ruelle–Bowen measure has been found. The Hénon attractor (for appropriate parameters) is the most prominent example [**BY**].

g. Comparison. The list of structural results that transfer (with appropriate modification) from the uniform to the nonuniform situation is quite impressive, which can be taken as a testament to the basic robustness of the hyperbolic paradigm. Closing, shadowing, spectral decomposition, Markov partitions and absolute continuity remain valid with relatively moderate adjustment. Expansivity could be recovered in a substantially restated fashion that is hardly worthwhile. A partial counterpart of ergodicity of volume is positive measure of ergodic components for any absolutely continuous invariant measure [**S-BKP, P**].

More difficulties appear in conjunction with the theory of equilibrium states. Those with the Sinai–Ruelle–Bowen measure are a clear indication. Remarkably, uniqueness (and ergodicity) of the measure of maximal entropy was proved recently for the case of geodesic flows on rank 1 (weakly hyperbolic) manifolds [**S-K**].

9. Partial hyperbolicity

See[**S-Bu**] for a detailed account. As for nonuniformly hyperbolic systems, we can ask, how much of the uniformly hyperbolic theory works in the partially hyperbolic situation (Section 6.4c).

a. Structural results. As suggested earlier, there is little reason to expect much of the structural theory of the uniformly or nonuniformly hyperbolic situation to hold for all partially hyperbolic systems, because the effects of the subexponential component in a partially hyperbolic system can be substantial.

Of expansivity, for example, there remains sensitive dependence on initial conditions, *i.e.*, for any point there are nearby points whose orbit moves away (simply make sure to arrange for nontrivial distance in the hyperbolic direction). Likewise, product examples show that closing and shadowing cannot be expected. If the subexponential direction is integrable then one might hope for orbits that at least return to the same subexponential leaf, even if not close to the starting point. Such results were obtained for geodesic flows on nonpositively curved manifolds [**BBES**].

Ergodicity of volume or even ergodic components of positive measure can also not be expected, because this fails for products or time one maps of suspensions. However, ruling out situations of this kind does give results of some interest. There is a specific assumption on the invariant manifolds that carries much of the hyperbolic theory to this setting.

b. Invariant foliations. In the partially hyperbolic situation the distributions E^+ and E^- are uniquely integrable to invariant laminations W^u and W^s, which satisfy 1–3 of Section 6.6f1. The main difference to the hyperbolic case is that the dimensions of these leaves do not sum to that of the ambient manifold.

How to overcome this defect is best explained in the case of a dynamical system that is partially hyperbolic on a compact manifold M. In this case the invariant laminations are foliations. The model situation that illustrates how hyperbolic effects may dominate the dynamics, is that in which the distributions $E^{\pm}$ are smooth and $E := E^+ \oplus E^-$ is totally nonintegrable. This means that the closure of the space of vector fields tangent to E under the Lie bracket is TM. This happens in numerous homogeneous systems, such as time one maps of geodesic flows of compact locally symmetric spaces of rank 1 [**KH**, Section 17.7] or left translations of compact factors $\mathrm{GL}(n,\mathbb{R})/\Gamma$ by the one-parameter subgroup e^{tA} for A diagonal with distinct elements [**KSp**].

Such systems have many properties similar to hyperbolic systems: Topological transitivity, ergodicity and the Bernoulli property of the main invariant measures and exponential decay of correlations for smooth functions. However, they usually have no periodic points.

The smoothness assumption of this discussion is fragile under perturbation, but it is not essential. Without it, one can assume the accessibility property (Section 6.7e), which requires no differentiability and produces the same local effect of connecting any two nearby points by a path that is piecewise tangent to E. This is the key assumption for proving persistence of topological transitivity [**BP**].

c. Stable ergodicity. Volume-preserving Anosov systems are *stably ergodic*, *i.e.*, all volume-preserving C^2 perturbations are ergodic. This observation has led to the question of which volume-preserving C^2 diffeomorphisms have this property. Partially hyperbolic systems that do not have an obvious product-like structure seem like a good candidate and have been studied in this regard, beginning with time one maps for geodesic flows of negatively curved manifolds [**Wk**].

Again, the required property is the accessibility property of the invariant foliations (Section 6.7e). So far it is known that volume-preserving partially hyperbolic systems are stably ergodic if they have the accessibility property and are dynamically coherent (the center distribution is integrable to a foliation whose leaves foliate the stable and unstable manifold of each of its elements). It is not known whether these additional hypotheses can be dropped, but experts conjecture that stable ergodicity is generic in the partially

hyperbolic volume-preserving class [**GPS, PS**]. In other words, volume is "prevalently" ergodic.

Particularly substantial progress has been recently made in the case of partially hyperbolic dynamical systems in dimension three. In this case the central direction has to be one-dimensional and carries a single Lyapunov exponent. It was established in many situations that this exponent is prevalently nonvanishing and thus the conclusions of the previous section apply [**Do**].

One conjectures furthermore that any open set of ergodic volume-preserving diffeomorphisms has an open dense subset of Bernoulli diffeomorphisms.

CHAPTER 7

Elliptic dynamics: Stable recurrent behavior

1. Introduction

a. Main features. The elliptic paradigm revolves around two features at the opposite end of the orbit complexity scale from the exponential behavior captured by the hyperbolic paradigm. The first and most important is a remarkable persistence for fairly general classes of conservative dynamical systems of stable behavior (in certain parts of the phase space), which can be modeled on a translation of a torus. The other is is somewhat less precise. It can be roughly described as the appearance of exceptionally precise simultaneous return of many orbits close to their initial positions. In this case no identifiable complete set of models is available but certain typical features of both topological and measure-theoretical behavior can be identified. The interaction between the properties of the linearized and nonlinear systems is more subtle than in the hyperbolic case.

Both conceptually and technically, elliptic dynamics is related to hard analysis to a greater extent than hyperbolic dynamics, where geometric and probabilistic ideas and methods are very prominent. This is one of the reasons for the comparatively small role elliptic dynamics plays in this volume. The survey [**S-LL**] is mostly dedicated to some of the central issues in elliptic dynamics. Various questions related to elliptic dynamics are also discussed in [**S-JS, S-BK, S-KT**].

b. Linear elliptic maps. For a linear map $L\colon \mathbb{R}^n \to \mathbb{R}^n$ the absence of growth in both the positive and negative direction of time means that all eigenvalues of L have absolute value 1 and no nontrivial Jordan blocks are present. Thus, for a linear map ellipticity is equivalent to existence of an invariant scalar product or to conjugacy to a map of the form

$$\begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & 0 & & \\ & & & -1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & -1 & & & \\ & & & & & & R_{\varphi_1} & & \\ & & 0 & & & & & \ddots & \\ & & & & & & & & R_{\varphi_k} \end{pmatrix},$$

where $R_\varphi := \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}$ is the rotation by φ. Naturally, since $\mathrm{Id} = R_0$ and $-\mathrm{Id} = R_\pi$ we may assume that 1 and -1 appear at most once.

Any orbit closure S is a finite union of tori, and the restriction of the map to S is the product of a minimal translation on a torus and a cyclic permutation.

In the space of all linear maps, ellipticity is atypical and unstable: A typical small perturbation of an elliptic map makes it nonelliptic. For *symplectic* linear maps in $\mathbb{R}^{2n}$, however, ellipticity is an open property. For $n = 1$ this is a familiar fact about $SL(2, \mathbb{R})$.

If all eigenvalues have absolute value and the map has a nontrivial Jordan block present then the growth of the norm is at least linear. This is why by very slow growth in a nonlinear system we will usually mean growth of $\|Df^n\|$ that is slower than linear. The latter marks the transition to *parabolic* behavior discussed in the next chapter.

c. Isometries. The absence of growth in a nonlinear system in the topological or smooth setting is best represented by the case of an isometry. The interplay between the smooth structure and the topology generated by it produces several versions of the notion of isometry and equicontinuity for a dynamical system. Restricting to the discrete time case, we give definitions in increasing order of strength.

Assume $f\colon M \to M$ is a diffeomorphism of a compact connected differentiable manifold.

1. *Isometries via equicontinuity.* $\{f^n\}_{n\in\mathbb{Z}}$ are equicontinuous with respect to some metric d on X.

Alternatively, the f-invariant metric $d'(x, y) = \sup_{n\in\mathbb{Z}} d(f^n(x), f^n(y))$ is equivalent to d.

The structure of isometries in the topological setting was described in Section 4.3d. In particular, if f is topologically transitive then f is topologically conjugate to a minimal translation. Note, however, that the conjugacy may not be differentiable (see below). In fact, since such an f is uniquely ergodic (Corollary 4.3.6), the sole invariant measure is absolutely continuous if the conjugacy is smooth.

Even in the absence of topological transitivity, it is possible to show that the orbit closures are finite unions of tori (see Corollary 7.1.2 for the case of a smooth metric).

2. *Isometries via equicontinuous derivatives.* The sequence $\{Df^n\}_{n\in\mathbb{Z}}$ of derivatives is equicontinuous.

This implies the previous definition since $d(f(x), f(y)) \le \max \|Df\| d(x, y)$. In this case the norms $\|v\|_n = \max_{0\le|i|\le n} \|Df^i v\|$ on TM are all equivalent and $\|v\|_\infty = \sup \|v\|_n = \lim_{n\to\infty} \|v\|_n$ is Df-invariant. However, the $\|\cdot\|_n$ are not in general Riemannian metrics, but *Finsler* metrics. If f is topologically transitive (and hence minimal by Proposition 2.2.4), $\|\cdot\|_\infty$ is, in fact, continuous. This can be deduced by considering the norm on TM as a map that associates to a point $x \in M$ the unit ball and using the Hausdorff metric on the space of compact closed subsets of $\mathbb{R}^n$ to measure the distance between such maps. Since $\|\cdot\|_\infty$ is the pointwise limit of increasing continuous norms $\|\cdot\|_n$ the map described above possesses a natural semicontinuity. But then the set where the oscillation of this function is at least ϵ is closed and by minimality either M or empty. The first possibility easily gives a contradiction. Furthermore, by a similar argument, one can show existence of an invariant *continuous Riemannian* metric.

3. *Riemannian isometries.* f preserves a *smooth* Riemannian metric on M.

THEOREM 7.1.1. *Any diffeomorphism f of a compact manifold that preserves a smooth Riemannian metric can be included in a smooth action of the group $\mathbb{T}^k \times F$ by isometries, where F is a finite abelian group.*

PROOF. The group of isometries of a compact Riemannian manifold is a compact Lie group G in which $\{f^n\}_{n\in\mathbb{Z}}$ is embedded such that its closure is an abelian subgroup of G, hence as described. □

COROLLARY 7.1.2. *Every orbit closure of f is a product of a torus and a finite group and the action of f on it is smoothly equivalent to a translation.*

Thus, from the dynamical point of view, nonlinear isometries of manifolds do not look too different from linear isometries of $\mathbb{R}^n$.

d. Distinction between different classes of isometries. Even in the simplest situation, namely for minimal diffeomorphisms of the circle, the preceding definitions are distinct.

By the Denjoy Theorem 5.1.1 a C^2 circle diffeomorphism without periodic points is topologically conjugate to a rotation and hence preserves some metric. Thus it satisfies the first of the three definitions. However, the conjugacy may be singular (Theorem 7.5.8.1, [**KH**, Theorem 12.5.1]) and hence no continuous Riemannian metric is preserved, *i.e.*, the second definition is not satisfied. We call a diffeomorphism that is topologically but not smoothly conjugate to a translation T_γ of the torus a *nonstandard smooth realization* of T_γ in the topological category. Later (see Corollary 7.5.5) we will discuss a similar concept of nonstandard smooth realization in the measurable category, where the place of topological conjugacy is taken by metric isomorphism. In that case even the phase space may not be homeomorphic to the torus.

If the second definition is satisfied then a continuous Riemannian metric is preserved and in the circle case the conjugacy to a rotation is C^1. But there are examples of C^∞ circle diffeomorphisms for which the conjugacy is C^1 but not C^2 (Theorem 7.5.8.3, [**KH**, Theorem 12.6.1]), so no C^1 Riemannian metric is preserved and the third definition is not satisfied.

All these effects appear for irrational rotation numbers that are exceptionally well approximable by rationals (Liouville numbers). For other rotation numbers (Diophantine numbers, see Section 7.3b) the third definition is always satisfied (Theorem 7.3.5). This dichotomy between the Liouville and Diophantine paradigms plays a central role in the subsequent discussion.

e. Completely integrable systems. In the case of a general (not minimal) isometry one can generalize the structural description given by Theorem 7.1.1. Ignoring the finite part for the moment (*e.g.*, by passing to an iterate), one may ask which translation is conjugate to the restriction of a given isometry to an orbit closure. To make this question precise, one needs to choose proper cyclic coordinates on each torus (this can be done up to the action of $SL(n,\mathbb{Z})$). If one compares the translations thus obtained on different tori, the coordinates need to be chosen in agreement. This is certainly easy to do locally in a neighborhood of a regular orbit. Then clearly the translations are the same on all nearby tori.

A natural generalization of this situation is a diffeomorphism for which the phase space (or at least an open dense subset) splits into invariant tori, on each of which the diffeomorphism is smoothly conjugate to a translation, but not necessarily the same one for different tori. According to our definition, such a system is not elliptic on the whole phase space. In fact, it is parabolic on the union of those invariant tori (although hyperbolic

behavior may be present on the nowhere dense complement). However, on each torus, such a diffeomorphism is elliptic.

This situation, modified in an obvious way for flows, appears naturally in Hamiltonian dynamics. The Liouville–Arnold Theorem (Section 5.3h) tells us that this happens whenever a Hamiltonian flow on $2n$-dimensional symplectic manifold has n independent integrals *in involution, i.e.,* with pairwise vanishing Poisson brackets, whose joint level sets are compact.

Completely integrable systems occupy a distinguished place in dynamics. There are multiple compelling reasons for this. One is related to the symplectic action-angle variables (I, ω) for such a system, which are provided by the Liouville–Arnold Theorem (Section 5.3h). They bring it into a simple form where the Hamiltonian depends only on the action variable I, which therefore gives first integrals. These action-angle variables can be found *by quadrature, i.e.,* by integration of explicitly present data, the taking of square roots, and inversion of functions [**An1**]. Furthermore, in these coordinates the solutions can be written explicitly as

$$I_k(t) = I_0$$
$$\omega_k(t) = \omega_0 + \alpha(I_0)t,$$

where $\alpha_k(I_0) = \dfrac{\partial H}{\partial I_k}|_{I=I_0}$ is the frequency vector. Therefore, the Hamiltonian equations can be explicitly integrated (solved) in the original coordinates as well.

A completely integrable system is said to be *nondegenerate* at $I = I_0$ if I_0 is a regular point for α, *i.e.,*

$$\det(\frac{\partial \alpha_k}{\partial I_l}|_{I=I_0}) \neq 0.$$

To find explicit solutions of the equations of motion was, of course, the central goal of mechanics until Poincaré. Hence the search for and attention given to integrable cases in important problems involving parameters, such as the motion of a rigid body.

Beginning with Poincaré, it was understood that complete integrability is rather exceptional and that there are intrinsic obstructions to solving the equations explicitly. Still, completely integrable systems hold their place and even enjoyed a major revival as a research topic for two reasons.

First, and this is central for the present setting, many important problems in mechanics can be viewed as perturbations of completely integrable systems (see Section 7.4b). The motion of the solar system is a prime example with the Kepler model (no interaction between planets) being the completely integrable system in question [**Ms3, An3, Pc**], [**An1**, Appendix 8]. The second is the relation between complete integrability and the presence of symmetries in the system, which are of great interest to physics. The Noether Theorem (Section 5.3g) provides the correspondence between symmetries and integrals. Thus, finding integrals can be accomplished by finding symmetries. There are interesting situations, however, in which the situation is reversed in that system is shown to be integrable without having sufficiently many apparent symmetries, *i.e.,* there are "hidden symmetries", which are found by finding the corresponding integrals [**Ms3**]. An early example is the geodesic flow on the triaxial ellipsoid where the integrals were found by Jacobi.

2. The setting for elliptic dynamics

a. Perturbation problem and Diophantine conditions. As follows from the previous discussion, the basic paradigm of elliptic behavior in dynamics (no growth in the linearized system accompanied by recurrence in the phase space) is represented essentially by a single class of models: translations and linear flows on tori. The principal content of the investigation of elliptic phenomena in dynamics is thus the search for traces of behavior represented by these simple models within various classes of systems. The most general orbit behavior in differentiable dynamical systems is expected to be unstable and complicated (see Chapter 6, [**S-BKP, S-H**]). So, in order to look for stable behavior or its traces, one usually needs to have a point of departure where elliptic behavior is observed. Thus, elliptic dynamics deals mostly with *perturbations* of systems with recognized elliptic behavior.

For a perturbation of a single translation of a torus the natural question is: under what condition is such a perturbation differentiably conjugate to this or a nearby translation? This problem is local in the space of dynamical systems but global in the phase space. The answer is provided in terms of the rotation number and Diophantine conditions (Section 7.2c, Section 7.4a). In low-dimensional situations (diffeomorphisms of the circle and flows on the two-dimensional torus) the corresponding global problem (in the space of dynamical systems) can also be successfully investigated (see Section 7.3b).

However, the most significant part of the perturbation approach deals with perturbations of nontransitive systems with elliptic behavior, namely completely integrable Hamiltonian systems and their counterparts in the volume-preserving category. The question is to find some orbits that exhibit behavior similar to that on the invariant tori of the unperturbed system. It is not generally expected and is, in fact, unusual, that such orbits fill the whole phase space of the perturbed system or even an open subset of it. This subject is essentially semilocal in character, since the approach is to take a single invariant torus of the unperturbed system and to find certain motions for the perturbed ones that remain in a sufficiently small neighborhood of it. The most natural approach would be to look for an invariant torus for the perturbed system, such that the restriction of the perturbation to that torus is conjugate to the unperturbed system on the reference torus. Its success depends on certain nondegeneracy conditions for the unperturbed system and again on Diophantine conditions on the rotation vector for the original torus. Different variations of these conditions will be discussed in due time (Section 7.3c, Section 7.4c). This approach usually goes under the name of *KAM theory* after Kolmogorov, who discovered the principal phenomena, Arnold who provided proofs of Kolmogorov's results and Moser, who greatly extended the approach and gave it a convenient form. Further elaboration will be provided in the next two sections and detailed discussion can be found in [**S-LL**] and a different volume of this series [**DS2**].

Due to its semilocal character, this approach may be applied even when the system under consideration is not a small perturbation of a completely integrable system in the whole phase space. It is sufficient that in a part of the phase space, the system looks like such a perturbation. This happens naturally if the system is a Hamiltonian system with m degrees of freedom (*i.e.,* with $2m$-dimensional phase space) and has an isotropic invariant torus of dimension $k \leq m$. Certain conditions on the linearized system on the torus (essentially nonhyperbolicity) are needed to consider the system in a neighborhood of the torus as a small perturbation of a completely integrable system. Further conditions on

the higher jets on the torus ensure nondegeneracy. The most natural cases of this situation appear for

(1) $k = 0$: fixed point, which must be elliptic in order for the outlined method to apply, although if it is partially hyperbolic the analysis sometimes can be carried out within the *center manifold* of the point
(2) $k = 1$: elliptic periodic orbit; the same comment applies
(3) $k = m$: Lagrangian invariant torus.

Another aspect of the perturbation problem concerns perturbations of a pure nontransitive isometry. In this case, various pathologies are possible and one is concerned more with the possibility of obtaining a certain type of behavior (*e.g.*, topological transitivity, ergodicity, minimality or unique ergodicity). Among the questions that arise in this area are those of existence of nonstandard realizations of translations and other models, and genericity of certain types of behavior. This approach is discussed in Section 7.2f and Section 7.5d.

b. Circle diffeomorphisms and twist maps. In some low-dimensional situations, elliptic behavior or some traces of it may be found without a perturbation assumption. The two main cases are

(1) diffeomorphisms of the circle (see Theorem 5.1.1, Section 7.3b and [**S-JS**]), and
(2) some Hamiltonian systems with two degrees of freedom and similar systems. These include geodesic flows on the two-torus with an arbitrary Riemannian metric, billiards in smooth convex domains, and forced one-dimensional oscillations. The analysis of all these situations can be reduced to that of twist maps [**KH**, Section 9.3].

Detailed discussions of twist maps can be found in [**S-BK**], [**KH**, Section 9.3, Chapter 13]. Here we provide just a brief introduction.

Consider the open cylinder $C := S^1 \times (0,1)$. Its universal cover is the strip $S = \mathbb{R} \times (0,1)$ with the projection $\pi\colon S \to C$, $(x,y) \mapsto ([x],y)$. A lift of a map $f\colon C \to C$ is a map $F = (F_1, F_2)\colon S \to S$ such that $\pi \circ F = f \circ \pi$. Thus, F_1 commutes with integer shifts in the x-direction, while F_2 is periodic in the first variable.

DEFINITION 7.2.1. A (surjective) diffeomorphism $f\colon C \to C = S^1 \times (0,1)$ is said to be an *area-preserving twist map* if

(1) f preserves area,
(2) f preserves orientation,
(3) f "preserves boundary components", *i.e.*, there exists an $\epsilon > 0$ such that $f(S^1 \times (0,\epsilon)) \subset S^1 \times (0,1/2)$, and,
(4) if $F = (F_1, F_2)$ is a lift of f then $F_1(x,\cdot)$ is monotone increasing for each x.

An area-preserving twist diffeomorphism is said to be *uniform* if $\dfrac{\partial}{\partial y} F_1(x,y) \geq \lambda > 0$.

FIGURE 7.1. A twist map

Note that a twist map may not extend continuously to the closed cylinder. There are obvious modifications of these definitions for $S^1 \times I$, where I is any finite or infinite interval of the real line.

For $x \in \mathbb{R}$ let $T(x) = (\lim_{y\to 0} F_1(x,y), \lim_{y\to 1} F_1(x,y))$. Define the *twist interval* as the set of rotation numbers of those circle homeomorphisms that possess a lift H to $\mathbb{R}$ with the property that for some $x \in \mathbb{R}$ we have $H^{n+1}(x) \in T(H^n(x))$ for all $n \in \mathbb{Z}$. The twist interval is defined up to an integer translation.

Less formally, the twist interval consist of those rotation numbers for which some dynamics with this rotation number on the circle is compatible with the twist map.

The first main conclusion of the theory of twist maps is that for any number α from the twist interval there is a closed invariant minimal set O_α that projects injectively to the cyclic coordinate and such that the dynamics preserves the cyclic order of orbits on O_α. This implies that the rotation number of f on O_α can be defined, and it turns out to be α [**KH**, Theorem 13.2.6].

Obviously O_α is a periodic orbit when α is rational. Such orbits are often called *Birkhoff periodic orbits* [**KH**, Section 9.3]. For an irrational α the set O_α can be either an invariant curve or a Cantor set similar to the exceptional Denjoy minimal sets for homeomorphisms of the circle. These Cantor sets are called *Aubry–Mather sets* .

Which of these two possibilities holds for a particular map and a particular rotation number is the central question in the analysis of twist maps. In general terms, the answer depends on how far the map is from an integrable one and on the arithmetic properties of the number α.

A remarkable fact is that while for circle diffeomorphisms irrational rotation number and C^2 regularity guarantee topological conjugacy to the corresponding rotation (Theorem 5.1.1), Aubry–Mather sets necessarily appear in most smooth or real-analytic twist diffeomorphisms, since the sets of rotation numbers for which invariant circles exist are usually nowhere dense.

Twist maps provide an excellent setting for applications of variational methods in dynamics (Section 5.6). Due to the discrete time nature of the situation many technical difficulties related to describing the proper spaces of candidate orbits disappear. Variational methods produce Birkhoff periodic orbits, Aubry-Mather sets (either directly or as limits of Birkhoff periodic orbits), various heteroclinic orbits, orbits with prescribed asymptotic or oscillating behavior in regions of instsbility (Section 7.3d), *etc.*

c. Diophantine and Liouvillian behavior. (See also [**La**].) A vector $\gamma = (\gamma_1, \dots, \gamma_m) \in \mathbb{R}^m$ satisfies a *resonance relation* if its coordinates and 1 are rationally dependent, *i.e.,* if $\sum_{i=1}^m k_i\gamma_i = l$ for some $(k_1, \dots, k_m, l) \in \mathbb{Z}^{m+1} \setminus \{0\}$. The translation T_γ on $\mathbb{T}^m$ is minimal if and only if γ (which is, of course, defined only up to an addition of an integer vector) does not satisfy any resonance relation (Example 2.2.2, Proposition 2.2.4).

The dynamics of a minimal translation T_γ as well as the properties of its perturbations depend on the appearance of approximate resonances, *i.e.,* such integers $k_1, \dots, k_m, l$ that $|l - \sum_{i=1}^m k_i\gamma_i|$ is small. A convenient way to measure the quality of approximate resonances is to compare $|l - \sum_{i=1}^m k_i\gamma_i|$ with functions of $k_1, \dots, k_m, l$, *e.g.,* $(\max(k_1, \dots, k_m))^{-\alpha}$ for various $\alpha > 0$.

In general, the appearance of good "almost exact" approximate resonances leads to instability in the dynamics, whereas the presence of only moderate approximate resonances is related to stability.

DEFINITION 7.2.2. A vector $\gamma \in \mathbb{R}^m$ is said to be *Diophantine* if there exist $C, \alpha > 0$ such that

$$|\sum_{i=1}^{m} k_i\gamma_i - l| \geq C(\max(k_1, \ldots, k_m))^{-\alpha}$$

for any $(k_1, \ldots, k_m, l) \in \mathbb{Z}^{m+1} \smallsetminus \{0\}$ [**La**].

The translation T_γ of the torus $\mathbb{T}^m$ where γ is a Diophantine vector will be called a *Diophantine translation*.

The opposite type of arithmetic appears when the coordinates of the vector are simultaneously well approximable by rational numbers. Such vectors are sometimes called *Liouvillian*. The corresponding minimal translations on the torus, which we will also call Liouvillian, are exceptionally well approximable by periodic translations.

Between these two cases lies the situation, where very good approximate resonances appear, but no simultaneous approximation with the same denominators. The corresponding minimal translations are very well approximable by nonminimal nonperiodic translations, which may be Diophantine on lower-dimensional tori. This case tends to show more similarity with the Liouvillian than with the Diophantine situation.

The case $m = 1$ plays a special role in elliptic dynamics. In this case Diophantine and Liouvillian conditions reduce to restrictions on the speed of approximation of an irrational number by rationals from above and below, respectively, and no intermediate case appears (Definition 7.3.1).

The notion of Diophantine behavior for flows is slightly different. A vector $\gamma \in \mathbb{R}^m$ is said to be *weakly Diophantine* if there exist $C, \alpha > 0$ such that $|\sum_{i=1}^{m} k_i\gamma_i| \geq C(\max(k_1, \ldots, k_m))^{-\alpha}$ for any $(k_1, \ldots, k_m) \in \mathbb{Z}^m \smallsetminus \{0\}$. We will call a constant vector field and the corresponding linear flow on the torus *Diophantine* if its right-hand side is a weakly Diophantine vector.

d. The Newton method. The central role in establishing stability of elliptic behavior is played by a high-powered variant of the Newton method. This method, which lies at the heart of KAM theory, was discovered by Kolmogorov and developed systematically by Moser, who also used some earlier ideas of Nash.

We outline the application of this method following [**Ms2**] for a situation similar to the one above, where one seeks a smooth or analytic conjugacy.

The idea is to cast the conjugacy equation as an implicit-function problem rather than a fixed-point problem as in structural stability or in highly dissipative situations. Therefore the Newton method is suitable for finding smooth and analytic conjugacies and it works in many nonhyperbolic problems. However, its applicability is restricted to perturbation problems, *i.e.*, situations where f and g are close to each other.

Write the conjugacy equation as

$$g = \mathcal{F}(f, h) := h^{-1} \circ f \circ h.$$

The main feature of the operator $\mathcal{F}$ is the "group property":

$$\mathcal{F}(f, \varphi \circ \psi) = \mathcal{F}(\mathcal{F}(f, \varphi), \psi), \qquad \mathcal{F}(f, \mathrm{Id}) = f. \tag{7.1}$$

As in the elementary Newton method, we want to linearize the operator and hence we need to assume that there is a linear structure on a neighborhood of (g, Id) in the appropriate functional space and that f is close to g. Then one can linearize $\mathcal{F}$ on this neighborhood. Write $D_1\mathcal{F}$ and $D_2\mathcal{F}$ for the partial differentials with respect to f and h, respectively. To look for an "approximate solution" $h = \mathrm{Id} + w$ of the conjugacy equation linearized at (g, Id), write

$$\mathcal{F}(f,h) = \mathcal{F}(g,\mathrm{Id}) + D_1\mathcal{F}(g,\mathrm{Id})(f-g) + D_2\mathcal{F}(g,\mathrm{Id})(h-\mathrm{Id}) + \mathcal{R}(f,h),$$

where $\mathcal{R}(f,h)$ is of second order in $(f-g, h-\mathrm{Id})$. In other words, if h solves the linearized equation (obtained by dropping $\mathcal{R}$), then $w = h - \mathrm{Id}$ is a solution of

$$\mathcal{F}(g,\mathrm{Id}) + D_1\mathcal{F}(g,\mathrm{Id})(f-g) + D_2\mathcal{F}(g,\mathrm{Id})w = g.$$

Using that $D_1\mathcal{F}(g,\mathrm{Id}) = \mathrm{Id}$ (since $\mathcal{F}(\cdot,\mathrm{Id}) = \mathrm{Id}(\cdot)$ by (7.1)), this simplifies to

$$(f-g) + D_2\mathcal{F}(g,\mathrm{Id})w = 0.$$

If $D_2\mathcal{F}(g,\mathrm{Id})$ is invertible, then $w = -\left(D_2\mathcal{F}(g,\mathrm{Id})\right)^{-1} u$, where $u = f - g$. In this case, w is of the same order as u, and substituting $h = \mathrm{Id} + w$ into $\mathcal{F}(f,h)$ we obtain a function $f_1 = h^{-1} \circ f \circ h = \mathcal{F}(f,h) = g + \mathcal{R}(f,h)$, so the size of $u_1 = f_1 - g = \mathcal{R}(f,h)$ should be of second order in the size of $u = f - g$. To justify this, one needs to estimate the difference between $\mathcal{F}$ and its linearization near (g, Id).

Thus, consider an iterative process as follows. Assuming that $f_1, \ldots, f_n$ have been constructed, we solve the equation

$$f_n - g + D_2\mathcal{F}(g,\mathrm{Id})w_{n+1} = 0$$

and set

$$h_{n+1} = h_n \circ (\mathrm{Id} + w_{n+1}) \text{ and } f_{n+1} = (\mathrm{Id} + w_{n+1})^{-1} \circ f_n \circ (\mathrm{Id} + w_n).$$

The last step of the construction is the proof of convergence of the sequence h_n in an appropriate topology. It follows from the same estimates that provide the fast decrease of the size of the $f_n - g$.

Notice that at every step the linear part is inverted at (g, Id), rather than at the intermediate points as in the elementary Newton method. This is why the method can be applied in nonhyperbolic situations.

e. Fast periodic approximation in dynamics. The property of fast rational approximation for numbers or vectors has counterparts in dynamics both in the measurable and the smooth context.

One such property is rigidity for a measure-preserving transformation (Section 3.6e), which provides for a systematic uniform return of most initial conditions. It does not, however, imply ergodicity and hence is not a sufficiently adequate generalization of the properties of the best rational approximation of irrational numbers that arises from the continued fraction expansion. A dynamical interpretation of the approximation of an irrational number α by a rational p_n/q_n in lowest terms is that the rotation R_α is approximated by a cyclic permutation of q_n small intervals of equal size. The quality of this approximation depends on $|\alpha - (p_n/q_n)|$ and is particularly good for Liouvillian numbers. Thus, a proper measure-theoretic counterpart of this situation is the property of good periodic approximation as well as its refinement involving the speed of approximation (Section 3.6e, [**CFS**]).

Good periodic approximation implies a variety of properties: ergodicity, simple spectrum, rigidity and, hence, singularity of the maximal spectral type (Section 3.4q) and absence of mixing. While any ergodic translation on an abelian group allows good periodic approximation, this property is also compatible with weak mixing and a variety of more exotic properties [**S-KT, K7**]. Furthermore, approximation with sufficiently high speed implies that the maximal spectral type is concentrated on certain sets of Liouvillian numbers [**KS**], thus generalizing properties of eigenvalues for Liouvillian translations.

Fast periodic approximation and similar methods are used in two different ways: analyzing properties of given systems, and producing (or "designing") examples of systems within particular classes with various prescribed properties. The latter purpose requires a particular kind of convergence of approximations, and purely measure-theoretic concepts and methods are insufficient for constructing smooth examples. Thus, various concepts of fast periodic approximation in the smooth category have been developed. The most useful and successful is the conjugation-approximation method outlined in Section 7.2f and Section 7.5b. Among the applications of this method is the construction of nonstandard smooth realizations of translations of the torus on different manifolds (*i.e.*, a system with a smooth invariant measure that is measure-theoretically but not topologically conjugate to a toral translation), nonstandard smooth realizations of some other systems and smooth models of some systems, whose natural models are not smooth (*e.g.*, some translations on infinite-dimensional tori and solenoids). Some of these ideas are further developed in [**S-KT**].

f. The conjugation-approximation method. (See also [**AK**].) The general outline of this method is as follows.

Let $\{S_t\}_{t\in\mathbb{R}/\mathbb{Z}}$ be a smooth circle action on a compact manifold M, possibly with boundary. Choose $p_0/q_0 \in \mathbb{Q}$ and $f_0 := S_{p_0/q_0}$. Inductively define

$$\frac{p_{n+1}}{q_{n+1}} = \frac{p_n}{q_n} + \frac{1}{k_n l_n q_n^2} = \frac{p_n k_n l_n q_n + 1}{k_n l_n q_n^2}$$

and periodic maps

$$f_n = H_n \circ S_{p_n/q_n} \circ H_n^{-1} \tag{7.2}$$

with rapidly increasing periods, where

$$H_{n+1} = H_n \circ h_{n+1} \text{and} h_{n+1} \circ S_{1/q_n} = S_{1/q_n} \circ h_{n+1}.$$

Thus, at the nth step of the construction the parameters are the diffeomorphism h_n and the integers k_n and l_n. The roles played by of these parameters are quite different. First, one fixes k_n to make the orbits of the next step of the construction sufficiently long. The diffeomorphism h_n is then chosen to provide controlled behavior of the conjugacies of the finite group $(S_{k/(k_n q_n)})_{k\in\mathbb{N}}$ of isometries. Now H_n is likely to have large derivative, and so l_n is chosen large enough to guarantee closeness of f_{n+1} to f_n with sufficiently many derivatives to guarantee convergence of the sequence f_n in the C^∞ topology.

The power and flexibility of the method comes from the fact that in contrast to the Newton method this convergence is not connected with any convergence of the conjugating diffeomorphisms H_n. Without loss of generality we can assume that the circle action S preserves volume λ. Then the maps H_n may either preserve a volume λ, thus guaranteeing that the limit diffeomorphism is volume-preserving, or the weak limit points of the

sequence $(H_n)_*(\lambda)$ may be controlled, producing invariant measures with desired exotic properties. Furthermore, for any specified $k \geq 0$ the sequence H_n may converge in the C^k topology but not in the C^{k+1} topology (Section 7.5b), or diverge in C^0 but converge in probability (leading to nonstandard smooth realizations of the rotation $R_{\lim_{n\to\infty} p_n/q_n}$), or diverge in probability in a controlled fashion (producing required ergodic properties or even a measurable conjugacy of f with a transformation other than a rotation).

3. Diophantine phenomena with a single frequency

a. Linear stability of Diophantine behavior. The Diophantine–Liouvillian dichotomy is particularly clear and unambiguous in the case of a single frequency, *i.e.*, where the underlying isometry is a rotation of the circle or the linear flow on the two-torus. We describe this situation in some detail. In these cases the motion is determined by a single number α, the angle of rotation or the slope of the flow. Diophantine conditions correspond to restrictions on the speed of approximation of α by rationals.

DEFINITION 7.3.1. $\alpha \in \mathbb{R}$ is said to satisfy *the Diophantine condition with exponent* $\beta \geq 0$ *and constant* $C > 0$ if

$$|\alpha - (p/q)| \geq C/q^{2+\beta}$$

for all $p, q \in \mathbb{Z}$, $q \neq 0$ [**S-JS**, Definition 5.3]. Denote the set of all such numbers by $D_{\beta,C}$. The number α is said to be *Diophantine* if it satisfies a Diophantine condition for some $\beta, C > 0$, *Liouvillian* otherwise, *i.e.*, if there are sequences $p_n, q_n \in \mathbb{Z}$ such that

$$|\alpha - \frac{p_n}{q_n}| = o(q_n^{-\gamma}) \text{ for all } \gamma > 0. \tag{7.1}$$

Almost every number satisfies the Diophantine condition with any positive exponent β and a constant depending on β, *i.e.*, $\bigcap_{\beta>0} \bigcup_C D_{\beta,C}$ has full measure. Notice that each set $D_{\beta,C}$ is closed and nowhere dense. The set of Liouvillian numbers has Hausdorff dimension zero but is a dense G_δ, hence topologically significant.

Here is a simple example how this dichotomy is reflected in the dynamical properties of the rotation. It shows how Diophantine conditions help to deal with problems of *small denominators*.

PROPOSITION 7.3.2. *The number α is Diophantine if and only if every $\mathbb{R}$-valued C^∞ cocycle over the rotation R_α is C^∞ cohomologous to a constant.*

PROOF. $\varphi(x) = \sum_{n\in\mathbb{Z}} \varphi_n \exp 2\pi i n x$ is C^∞ if and only if $|\varphi_n| = o(|n|^{-\gamma})$ for all $\gamma > 0$. The cohomological equation takes the form $\varphi(x) - \varphi_0 = h(x+\alpha) - h(x)$, from where the Fourier coefficients $h_n = \varphi_n/(\exp 2\pi i n\alpha - 1)$ of h are uniquely determined for $n \neq 0$. Thus the Diophantine condition on α guarantees that the h_n decay faster than any negative power of n.

Conversely, suppose α is Liouvillian and choose sequences $p_n, q_n \in \mathbb{Z}$ satisfying (7.1). Define φ by choosing the Fourier coefficients $\varphi_{q_n} = |\alpha - (p_n/q_n)|^{1/2}$ and $\varphi_m = 0$ otherwise. By (7.1) φ is C^∞, but the Fourier coefficients of h are $h_{q_n} = |\alpha - (p_n/q_n)|^{1/2}/(\exp 2\pi i q_n\alpha - 1) \approx |\alpha - (p_n/q_n)|^{-1/2} \to \infty$, which precludes the existence of even an L^1 solution. □

COROLLARY 7.3.3. *If R_α is a rotation by a Diophantine angle and φ is a C^∞ function then $\sum_{n=0}^{N-1} \varphi \circ R_\alpha - \int \varphi = O(1/N)$ in any C^r-norm.*

Using the connection between cocycles and time changes (Section 1.3m), one immediately obtains the following result due to Kolmogorov.

PROPOSITION 7.3.4. *Any C^∞ time change of a linear flow on $\mathbb{T}^2$ with Diophantine slope is C^∞ conjugate to a linear flow.*

b. Smooth linearization of circle diffeomorphisms. (See also [**S-JS**].) A much deeper nonlinear counterpart of Proposition 7.3.2 is the following theorem of Yoccoz [**Y1**], which followed and completed the pioneering work of Herman [**Hm1**].

THEOREM 7.3.5. *The number α is Diophantine if and only if every C^∞ diffeomorphism of the circle with rotation number α is C^∞ conjugate to the rotation R_α.*

There is a similar result giving necessary and sufficient conditions for analytic conjugacy of analytic diffeomorphisms to a rotation [**Y2**], [**S-JS**, Theorem 6.5]. As might be expected, the condition on the rotation number (Brjuno condition) is weaker than the Diophantine requirement but it is still expressed in terms of the speed of rational approximation.

On the other hand, more specific Diophantine conditions guarantee finitely smooth conjugacy of C^r diffeomorphisms to a rotation (with some loss of regularity) [**KO1**], [**S-JS**, Theorem 5.2]. At the opposite end of the scale are results for the most "robust" Diophantine numbers, *i.e.*, those with exponent 0. Such numbers are said to be of *constant type* because the inequality $|\alpha - (p/q)| > C/q^2$ for all $p, q \in \mathbb{Z}$, $q \neq 0$ is equivalent to boundedness of all coefficients a_m in the continued fraction expansion $\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \dots}}$. For diffeomorphisms with rotation numbers of constant type, sufficient regularity conditions for existence of an absolutely continuous conjugacy to a rotation are below C^2 and are remarkably close to the condition required for continuous conjugacy [**KO2**], [**S-JS**, Proposition 11].

Note that the set of numbers of constant type has measure zero but Hausdorff dimension one.

c. The invariant curve theorem. Consider the following semilocal situation. Let M be a two-dimensional differentiable manifold with an area form ω, $U \subset M$ an open set, $\Gamma \subset U$ a C^∞ closed curve $f : U \to M$ a C^∞ embedding such that $f(\Gamma) = \Gamma$. Assume the rotation number of $f_{\restriction \Gamma}$ is a Diophantine number α. Then by Theorem 7.3.5 $f_{\restriction \Gamma}$ is C^∞ conjugate to the rotation R_α. Using the normal forms approach (Section 5.2c, Section 5.2g) it is possible to show that there are C^∞ coordinates (x, y) in a neighborhood of Γ where x is a cyclic coordinate, such that $\omega = dx \wedge dy$ and

$$f(x, y) = (x + \alpha + h_1(x, y), y + cx^k + h_2(x, y))$$

where k is a natural number, c a constant and the functions h_1 and h_2 vanish with all their derivatives of all orders at $y = 0$. We call the curve Γ *nonflat* if $c \neq 0$ and *nondegenerate* if in addition $k = 1$.

The following theorem, which we reproduce in a slightly imprecise form, is the prototype result of KAM theory and is proved using the method of Section 7.2d.

THEOREM 7.3.6 ([**Rm**]). *Assume that the curve Γ is nonflat. Let $g = f + r$ be a C^∞ map sufficiently close to f; the number of derivatives of r that are required to be small and the estimates on the size of these derivatives depend on β and C in the Diophantine condition, and k and c above. Then g has a unique smooth invariant curve Γ' close to Γ, given by the equation $y = \phi(x)$, with rotation number α.*

Applying Theorem 7.3.6 to Diophantine numbers sufficiently close to α one obtains

COROLLARY 7.3.7. *If the curve Γ in nondegenerate then Γ' is the limit of other smooth invariant curves for g from both sides.*

d. Twist maps; nondegenerate case.

1. *Integrable twist maps.* Any map of the cylinder of the form $f(x, y) = (x+g(y), y)$, where g is a monotone increasing differentiable function, is an area-preserving twist diffeomorphism (Definition 7.2.1). Such maps are called *integrable twist maps*. If $g' > 0$ then this map is a uniform twist diffeomorphism, and it is said to be an *integrable uniform twist diffeomorphism*.

Integrable twist maps are simple examples of completely integrable discrete time systems because all circles $y = \mathrm{const}$ are invariant. The nondegeneracy assumption ensures that the set of rotation numbers of the invariant circles contains an interval.

2. *Regular invariant circles.* Description of what remains of this foliation by invariant circles when a nondegenerate integrable twist map is perturbed (in the area-preserving category) is a special case of KAM-theory which deals with perturbations of completely integrable systems (Section 7.4c, [**S-LL**]) . Precise statements require rather careful formulation, but the picture is easy to describe. For a twist map f we will call any invariant closed curve of the form $\mathrm{graph}\,\phi$ where $\phi : S^1 \to \mathbb{R}$ an *invariant circle* for f.

Perturbations in a sufficiently fine topology generically have only a reduced collection of invariant circles.

The nonminimal circles with rational rotation numbers and most of the circles with Liouvillean rotation numbers are usually the first to break up, with finite sets of Birkhoff periodic orbits taking the place of the former and Aubry–Mather sets (Section 7.2b) taking the place of the latter. Each Aubry–Mather set is embeddable in a (noninvariant) circle.

Circles with Diophantine rotation number are the most robust, and they make up an invariant set of positive measure. Thus, a perturbation of a nondegenerate integrable twist has a set of large measure of invariant Diophantine circles. In fact, the measure of the complement to the set of invariant circles goes to zero as the perturbations approach the integrable system. Preservation of these circles is proved using Theorem 7.3.6 but careful estimates are needed to make sure that all curves corresponding to Diophantine numbers with fixed β and C remain for any sufficiently small perturbation. Let us call the invariant circles whose existence follows from Theorem 7.3.6 *regular*. In particular, given any proper closed subinterval I of the twist interval and a perturbation which is sufficently small to a given sufficiently high order $r \geq 4$ of derivatives, there are regular invariant circles with all rotation numbers from the set $I \cap D_{\beta,C}$.

3. *Global structure of invariant circles.* In order to understand the global orbit structure of the perturbed map the following general fact about twist maps is useful.

PROPOSITION 7.3.8. *All invariant circles of a given area-preserving twist map ϕ are Lipschitz graphs with the fixed Lipschitz constant. The limit of a sequence of invariant*

circles is an invariant circle. There is at most one invariant circle with a given rotation number and rotation number is continuous on the set of invariant circles.

COROLLARY 7.3.9. *The set $\mathcal{R}(f)$ of rotation numbers of invariant circles of a twist map f is closed.*

4. *Regions of instability.* Between invariant circles there are *regions of instability*. These regions correspond to complementary intervals to the set $\mathcal{R}(f)$.

The boundary components of any region of instability are invariant circles with rotation numbers corresponding to the endpoints of these intervals, but they are not regular since by Corollary 7.3.7 any regular invariant circle is a limit of other regular circles from both sides.

The dynamics in the regions of instability is complicated. In particular they contain Aubry–Mather sets with rotation numbers from the corresponding complementary interval. Each region of instability also contains orbits traveling from one boundary component to the other in either direction as well as orbits oscillating between the boundary components. They also typically contain periodic orbits with remarkable semilocal behavior. There are "necklaces" of alternating hyperbolic and elliptic periodic points, with typically heteroclinic tangles between the hyperbolic ones. This means that there is an abundance of horseshoes (Section 6.5b, [**KH**, Theorem 6.5.5]) in this ring, hence exponentially complex dynamics (positive topological entropy, *etc.*). The elliptic points also contribute to the complexity of the picture, because a neighborhood of each of them looks much like the entire picture (Section 7.3e): Invariant circles separating regions of instability in which hyperbolic and elliptic periodic points alternate... Despite the great complexity of this orbit structure, a certain basic stability has been achieved by confining each orbit to an invariant circle or the narrow region between two invariant circles.

Several changes occur when the perturbation becomes larger. A quantitative change is that the regions of instability widen. A qualitative change is that increasingly more invariant circles break up. The order in which they do so is affected greatly by arithmetic. The disappearance of the Diophantine invariant circles progresses according to the strength of the Diophantine condition on the rotation number. Those with the smallest exponent, such as the one with the golden mean as rotation number, typically survive the longest, even though macroscopic perturbations eventually break up even these.

In fact, for any given Diophantine rotation number in the twist interval a corresponding invariant circle will persist under sufficiently small perturbations. For any Liouvillian invariant circle one can produce arbitrarily small perturbations that destroy it, but no perturbation will destroy all of them immediately (see Section 7.5c).

e. Neighborhood of an elliptic fixed point. This is a situation very similar to the one which appeared in Theorem 7.3.6 for the nonflat but degenerate case. Using polar coordinates, a smooth area-preserving map f of the plane for which 0 is an elliptic fixed point, *i.e.*, Df has eigenvalues $e^{\pm 2\pi i\alpha}$ for some $\alpha \in \mathbb{R}$, can be brought (up to terms that vanish with all derivatives at the origin) into the *Birkhoff normal form* (Section 5.2c) $(\theta, r) \mapsto (\theta + \omega(r^2), r)$ near 0, where ω is given by a (in general formal) power series. For technical reasons a finite number of *rational* α's have to be excluded. Then, unless $\omega = \text{const}$, we can separate the leading nonconstant term in ω to produce an integrable twist so that the map f can be viewed as its small perturbation. This twist is degenerate since $\omega'(0) = 0$. But it is nondegenerate for sufficiently small r away from zero. One can

extend this to a smooth area-preserving twist in $S^1 \times \mathbb{R}$ and then apply perturbation results like the ones above to obtain persistence of some invariant curves.

Therefore one obtains a picture just like the one described for area-preserving perturbations of nondegenerate integrable twists. Even among analytic perturbations those with such complicated dynamics are generic in a strong sense, *i.e.,* for any $k \in \mathbb{N}$, even among those perturbations whose k-jet at the fixed point agrees with the original map, the complicated ones are generic [**Gn**].

f. Caustics in convex billiards and related problems. (See also [**Lz, Zh**].) A convex billiard is the flow of free unit speed motion in a convex region, with "optical" reflection at the boundary (Figure 7.1), *i.e.,* angle of incidence=angle of reflection. A natural

FIGURE 7.1. A convex billiard

section for this flow is given by inward vectors on the boundary. Since their length is normalized, one can parametrize it by $S^1 \times (0, \pi)$ using a parameter on the boundary and the reflection angle. Thus, the phase space is a cylinder (or annulus), and the section map is a twist map (Figure 7.1, p. 160). After slight reparametrization, it is area-preserving, and strict convexity makes it nondegenerate. When the region is an ellipse, the billiard map is an integrable twist. This is closely related to the abundance of *caustics*, *i.e.,* curves whose tangents are reflected into tangents to the same curve. Generally, the existence of (convex) caustics is closely related to the existence of invariant circles for the billiard map. Analytic geometry shows that all confocal ellipses and hyperbolae are caustics of the ellipse (for the hyperbolae tangencies alternate between branches). The integrability is related to

FIGURE 7.2. Elliptic billiard with confocal ellipses and hyperbolas

that of the geodesic flow on the ellipsoid, which was discovered by Jacobi. Caustics are often envelopes of families of trajectories, and are therefore easily visible in a reflective metal ring. A question, unanswered since Birkhoff posed it, is whether the elliptic billiard

FIGURE 7.3. An envelope

table is the only one with an open set of caustics. The issue is that, while one can easily perturb a twist map in such a way that a selected set of invariant circles persists (by supporting the perturbation away from this set), one cannot localize perturbations arising from deformations of the billiard table in a like fashion because a boundary point corresponds to a curve connecting the boundary components of the cylinder. An extreme example of breakdown of any semblance of integrability is the impossibility of convex caustics when there is a flat boundary point ([**KH**, Proposition 13.5.3], this is not so hard to see using the mirror equation of geometric optics). Furthermore, there are also delicate estimates that confine all convex caustics to small neighborhoods of the boundary as the minimal boundary curvature decreases to 0 [**GkK**].

g. Preservation of Diophantine circles without twist. While the twist condition pervades the theory of invariant circles and arises naturally from the mechanical systems that motivated this work, there is recent remarkable work showing that to some extent arithmetic conditions alone can provide for the presence of invariant circles. Specifically, Herman proved that for an area-preserving map with a smooth invariant closed curve Γ with Diophantine rotation number α has invariant curves arbitrary close to Γ. This gives new information compared to Theorem 7.3.6 when the curve Γ is flat. Naturally in this case there may be many curves with the same rotation number. In the case of Liouvillean rotation number on the boundary there may be no invariant circles inside altogether (Section 7.5d).

4. Diophantine phenomena with several frequencies

Due to the greater technical difficulty of this subject, including the need for elaborate notations, we restrict ourselves to a brief outline without presenting precise formulations, let alone sketches of proofs. For a detailed overview see [**S-LL**]. We will emphasize both similarities and differences with the previous case.

a. Perturbation of linear maps and vector fields on the torus. (See also [**Ms2**].) The global conjugacy results like Theorem 7.3.5 do not hold in higher dimension, *i.e.*, for maps of the torus $\mathbb{T}^n$, $n \geq 2$ and for flows on $\mathbb{T}^n$, $n \geq 3$. Aside from obvious reasons (no topological conjugacy to a rotation in general) it is not known whether any diffeomorphism of $\mathbb{T}^n$, $n \geq 2$ that is topologically conjugate to a Diophantine rotation is smoothly conjugate to it. However, for small perturbations of rotations and linear flows the conjugacy holds. A convenient way to formulate such a perturbation result is to consider an n-parameter family of maps or vector fields close to the family of translations or vector fields near a given Diophantine translation (corr. vector field). Then under proper assumptions on size and regularity such a family contains a unique element differentiably conjugate to the given translation (corr. vector field). An example of such a family of diffeomorphisms is

$$f_\lambda(x) = x + \alpha + h(x) + \lambda,$$

where α is a Diophantine vector and h is small with sufficiently many derivaives. Similarly, for vector fields one may consider

$$\frac{dx}{dt} = \alpha + h(x) + \lambda$$

for a weakly Diophantine α.

b. Stability problem in celestial mechanics. Poincaré's prize-winning memoir on the question of the stability of the solar system showed that the difficulty of this question had been seriously underestimated. Generically, there are no integrals other than the classical ones, so attempts to recognize integrability, for example, have little promise. On the other hand, because of the smallness of the planetary masses compared to that of the sun, our solar system is relatively close to a superposition of two-body problems, which is integrable.

The traditional approach in celestial mechanics concentrated on establishig stability of motions for such a perturbed system for a finite but sufficiently long length of time, via normal forms, asymptotic series and the like. Neither Poincaré, who pioneered the

qualitative approach to celestical mechanics, nor Birkhoff, who was the leading figure in the field in the first half of the twentieth century, seem to have realized that stability for infinite time may take place in substantial portions of the phase space. One can speculate that if they had had computers at their disposal they would have stumbled upon the invariant tori by computer simulation.

The crucial insight into the problem was achieved by Kolmogorov in the fifties. Kolmogorov discovered that under certain nondegeneracy conditions a large set of invariant tori of an integrable Hamiltonian system survive under any sufficiently small perturbation. This produces a large invariant set with genuine elliptic behavior.

c. The Kolmogorov theorem in the nondegenerate Hamiltonian case. (See also [**An2, Ms2**].) The core result of KAM theory is a multidimensional version of Theorem 7.3.6 in the nondegenerate case. It deals with a perturbation of a Hamiltonian system with n degrees of freedom defined in a neighborhood of a Lagrangian (hence n-dimensional) invariant torus T. It is assumed that the Hamiltonian vector field restricted to T is conjugate to a constant Diophantine vector field. Assuming furthermore a nondegeneracy condition for the Hessian of the Hamiltonian in the direction transverse to the torus T one deduces existence and uniqueness of an invariant torus T' for the perturbed Hamiltonian system near T such that the perturbed system on T' is differentiably conjugate to the unperturbed system on T.

For systems with two degrees of freedom this result can be derived from Theorem 7.3.6 by restricting the Hamiltonian flow to the constant energy surface and taking a section map near the torus T.

This basic result is then applied to a perturbation of a globally defined completely integrable mechanical system. By the Liouville–Arnold Theorem (Section 5.3h) a nondegenerate completely integrable mechanical system (Section 7.1e) decomposes (in regions with compact level sets) into a foliation by tori on each of which the dynamics is that of a linear flow, and the set of whose rotation vectors has nonempty interior. From the point of view of mechanics, this is the natural generalization of nondegenerate twist maps. Symplectic perturbations of such a system have a set of invariant tori of almost full measure, and the order of their breakup is related to the arithmetic properties of their rotation vector. There are also regions of instability containing periodic orbits around which a similar scaled picture appears (a nice picture is in [**AM**]).

Unlike in the two-dimensional situation, however, the n-dimensional invariant tori do not separate the $2n$-dimensional phase space. Therefore, stability in the sense of a priori boundedness of orbits is not an automatic consequence in the same way as for twist maps. Indeed, there is a phenomenon discovered by Arnold and called *Arnold diffusion*: Some orbits of an arbitrarily small perturbation of a completely integrable system near an invariant torus may drift a finite distance away from the torus or even leave any compact part of the phase space.

An important qualitative consequence of the Kolmogorov theorem is nonergodicity (with respect to volume) of open sets of Hamiltonian systems on the level surfaces of the Hamiltonian. This contrasts with ergodicity of open sets of Hamiltonian systems that are obtained by perturbing geodesic flows on compact manifolds of negative curvature (Section 6.5e, Section 6.7f).

d. Degenerate case and stability of the solar system. Kolmogorov's result is not directly applicable to the solar system because the nondegeneracy assumption does not hold. This question was treated by Arnold [**An3**]. See [**S-LL**] for a description of this important and technically complicated subject.

e. Frequency locking for special symplectic structures. (See also [**HZ**].) M. Herman [**Hm2, Hm3**] noticed that the stability results mentioned in Section 7.4a can be directly applied to obtain instances of differentiable stability (differentiable conjugacy of all nearby systems) for Hamiltonian systems for some special symplectic structures on the torus. Specifically, consider the symplectic form

$$dx_2 \wedge dx_1 + dx_4 \wedge dx_3 + \alpha_1 dx_3 \wedge dx_2 + \alpha_2 dx_1 \wedge dx_3$$

on $\mathbb{T}^4$, where (α_1, α_2) is a Diophantine vector. The Hamiltonian vector field with the Hamiltonian $H_0(x) = x_4$ is actually the constant vector field $(\alpha_1, \alpha_2, 1, 0)$ and hence it is Diophantine on every three-dimensional torus $H_0 = \text{const}$. Perturbing the Hamiltonian one obtains the one-parameter family of perturbed flows on the level surfaces. Each of these flows can be made conjugate to the original linear flow by adding a constant. As it turns out, due to the Hamiltonian structure of the flows, these constants must vanish.

f. Preservation of tori in the volume-preserving category. As we pointed out there is a substantial difference between the conclusions of KAM theory in the case of one and several frequencies. In the former case *all* orbits of the perturbed system stay forever near invariant tori of the unperturbed system while in the latter there is a diffusion for some orbits away from these tori. The reason of course is that for area-preserving maps of surfaces invariant circles locally divide the phase space (and the invariant tori for systems with two degrees of freedom locally divide the manifolds of constant energy).

Using proper modifications of the KAM techniques Herman, Xia and others [**X**] found a generalization of the low-dimensional results in the *volume-preserving category*. The semilocal situation for these results in the discrete time case is as follows.

The unperturbed system is defined on $\mathbb{T}^n \times [-\epsilon, \epsilon]$ and has the form $f(x,t) = (x + \alpha(t), t)$ where the function $\alpha : [-\epsilon, \epsilon] \to \mathbb{R}^n$ satisfies a Diophantine type nondegeneracy condition. One considers volume-preserving perturbations that preserve the torus $t = 0$. Then if the perturbation is sufficiently small in theC^r topology, where r depends on the Diophantine conditions, most of $\mathbb{T}^n \times [-\epsilon, \epsilon]$ is filled by Diophantine invariant tori close to the tori $t = \text{const}$. Notice that the rotation numbers on the tori of the perturbed map will in general be different from those for the unperturbed one. Under global conditions on the unperturbed map this leads to global results for an open set of volume-preserving diffeomorphisms, which imply in particular nonergodicity with respect to volume and confinement of orbits near the tori of the unperturbed system.

5. Liouvillian phenomena

a. Linear instability of Liouvillian behavior. While we already know that Liouvillian rotations do not possess stability of smooth cocycles (Proposition 7.3.2), the behavior of smooth cocycles over some Liouvillian rotations exhibits various more specialized types of exotic behavior than simply failing to be a smooth coboundary. We first describe a phenomenon that produces examples of minimal real analytic nonergodic transformations as

well as real analytic time changes in a linear flow on $\mathbb{T}^2$ with highly discontinuous eigenfunctions. Namely, there is an analytic function φ such that $\varphi(x) = \Phi(x+\alpha) - \Phi(x)$ for a very discontinuous Φ. A strong notion of discontinuity is the following:

DEFINITION 7.5.1. Let X, Y be topological spaces and μ a measure on X. A measurable map $f\colon X \to Y$ is said to be *metrically dense* with respect to μ if $\mu(U \cap f^{-1}(V)) > 0$ for all nonempty open $U \subset X$, $V \subset Y$.

PROPOSITION 7.5.2. *There exists $\alpha \in \mathbb{R}$ and an analytic function $\varphi\colon S^1 \to \mathbb{R}$ such that $\varphi(x) = \Phi(x+\alpha) - \Phi(x)$ with $\Phi\colon S^1 \to \mathbb{R}$ measurable and metrically dense with respect to Lebesgue measure.*

SKETCH OF PROOF. It is convenient to switch to multiplicative notation on the circle by considering it as the unit circle in $\mathbb{C}$. Then $x \mapsto x+\alpha$ becomes $z \mapsto \lambda z$ with $\lambda = e^{2\pi i\alpha}$. Take the *Dirichlet kernel*

$$D_{q,m}\colon S^1 \to \mathbb{R}, \quad D_{q,m}(z) = \sum_{j=1}^{m-1} (z^{jq} + z^{-jq}).$$

Its density is concentrated around the qth roots of unity. Inductively define $\Phi_n = \sum_{k=1}^{n} C_k D_{q_k,m_k}(z)$, choosing the parameters of the construction such that Φ_n converges in probability to a metrically dense function Φ, $p_n/q_n \to \alpha \in \mathbb{R} \smallsetminus \mathbb{Q}$ very rapidly, and $\Phi_n \circ \mathbb{R}_{p_n/q_n} - \Phi_n$ converges in the real-analytic topology to a function φ, so $\Phi \circ R_\alpha - \Phi = \varphi$. □

To produce minimal nonergodic examples we use:

PROPOSITION 7.5.3 ([**KH**, Propositions 4.2.5,6]). *Consider the torus $\mathbb{T}^2$, a function $\varphi\colon S^1 \to \mathbb{R}$, and a map $f\colon (x,y) \mapsto (x+\alpha, y+\varphi(x))$ of $\mathbb{T}^2$ with $\alpha \in \mathbb{R} \smallsetminus \mathbb{Q}$.*

1. *If $\varphi(x) = \Phi(x+\alpha) - \Phi(x)$ for some Lebesgue measurable function $\Phi\colon S^1 \to \mathbb{R}$ then for any ergodic invariant measure, f is metrically isomorphic to the rotation R_α and there are uncountably many different ergodic invariant measures.*
2. *Either $\varphi(x) = \Phi(x+\alpha) - \Phi(x) + r_1\alpha + r_2$ for some continuous $\Phi\colon S^1 \to \mathbb{R}$ and $r_1, r_2 \in \mathbb{Q}$ or f is minimal.*

The first conclusion is due to the observation that $h^{-1} \circ f \circ h(x,y) = (x+\alpha, y)$ for $h(x,y) = (x, y+\Phi(x))$, which implies that any invariant measure for f projects to Lebesgue measure, so the invariant ergodic measures for f are exactly the measures induced from measures on circles, and the graph of $\Phi + c$ for any $c \in \mathbb{R}$ supports such a measure.

COROLLARY 7.5.4. *There exist analytic minimal nonergodic diffeomorphisms of $\mathbb{T}^2$.*

PROOF. Let $f\colon \mathbb{T}^2 \to \mathbb{T}^2$, $(x,y) \mapsto (x+\alpha, y+\varphi(x))$, where φ is as in Proposition 7.5.2. By Proposition 7.5.3 f has distinct ergodic invariant measures, hence it has nonergodic invariant measures. If f were not minimal then by Proposition 7.5.3 we would have $\varphi(x) = \psi(x+\alpha) - \psi(x) + r_1\alpha + r_2$ for some continuous $\psi\colon S^1 \to \mathbb{R}$ and $r_1, r_2 \in \mathbb{Q}$. Suppose $r = r_1\alpha + r_2 < 0$ (the case $r < 0$ is similar). Then $F = \psi - \Phi$ satisfies $F(x+\alpha) = F(x) - r > F(x)$ for all $x \in S^1$, whereas there is a set $F^{-1}(-\infty, c)$ of positive measure, contradicting the Poincaré Recurrence Section 3.4c. □

Using Section 1.3m as in the Diophantine case, one obtains

COROLLARY 7.5.5. *There exists a real-analytic time-change of an irrational linear flow on $\mathbb{T}^2$ that is metrically isomorphic to the linear flow but whose nonconstant eigenfunctions are metrically dense as maps from the torus to the unit circle*

This is an example of a *nonstandard smooth realization* (Section 7.1d) of the linear flow as well as the translations that comprise it.

Existence of an eigenfunction for the special flow over a measure-preserving transformation T with roof function φ implies that $\exp i(a + b\varphi)$ is a coboundary for some $a, b \in \mathbb{R}$. Another phenomenon related to cocycles over Liouvillian rotations leads to the absence of solutions not only of the ordinary additive cohomological equation but of the above multiplicative cohomological equations as well.

THEOREM 7.5.6 ([**K7**]). *Let $h(z) = \sum_{n\neq 0} h_n z^n$ be a C^2 real valued function on S^1 with zero average. Let R_α be a rotation on S^1,*

$$\frac{|\alpha - p_n/q_n| q_n}{\sum_{k=1}^{\infty} |h_{kq_n}|} \to 0 \quad \textit{and} \quad \frac{|h_{q_n}|}{\sum_{k=1}^{\infty} |h_{kq_n}|} > c > 0$$

for some $p_n/q_n \in \mathbb{Q}$. Then $\exp ir(h_0 + h(z))$ is not a coboundary for any h_0 and r and consequently the special flow over R_α built under the function $h_0 + h(z)$ is weakly mixing for all h_0.

Since the conditions of the last theorem are satisfied for many real analytic functions h and numbers α one can pass from special flows to time changes as before to get

COROLLARY 7.5.7. *There exists a real-analytic time-change of an irrational linear flow on $\mathbb{T}^2$ that is weakly mixing.*

b. Circle diffeomorphisms with Liouvillian rotation numbers. For smooth and analytic circle diffeomorphisms with extremely well approximable rotation number, the conjugacy to a rotation and hence the invariant measure tend to be singular. Arnold's theorem [**KH**, Theorem 12.5.1] exposes singularity of the conjugacies as a generic phenomenon in typical one-parameters families of real-analytic maps.

On the other hand, the properties of conjugacies for specially constructed diffeomorphisms can be carefully controlled by the conjugation-approximation method (Section 7.2f), an inductive construction that produces C^∞ diffeomorphisms with irrational (but very well approximable) rotation number with virtually every possible degree of regularity of the conjugacy to the rotation. Specifically, we can make the conjugacy singular, absolutely continuous without being Lipschitz, and C^r without being C^{r+1} for any $r \in \mathbb{N}$.

The phenomenon of having a singular conjugacy to a rotation is generic in the following sense: In the C^∞-closure of the set of C^∞ circle diffeomorphisms C^∞ conjugate to a rotation there is a residual set of diffeomorphisms that are conjugate to a rotation by a singular homeomorphism. This implies that diffeomorphisms that are conjugate to a circle rotation via an absolutely continuous homeomorphism, in particular a Lipschitz continuous or smooth one, form a set of first category.

THEOREM 7.5.8 ([**KH**, Theorem 12.6.1]). *Given any neighborhood U of 0 in $C^\infty(S^1)$ and $i \in \{1, 2, 3\}$ there exist α and a C^∞ diffeomorphism f of S^1 such that $f - R_\alpha \in U$ and f is conjugate to R_α via a conjugacy h that has the ith of the following three properties:*

(1) *h is singular.*

(2) *h is absolutely continuous but not Lipschitz continuous.*
(3) *h is C^r but not C^{r+1}, where $r \in \mathbb{N}$ is arbitrary.*

SKETCH OF PROOF. Take α to be the limit of $\alpha_n = p_n/q_n \in \mathbb{Q}$, where $\alpha_{n+1} = \alpha_n + \beta_n$ and $\beta_n = 1/K_n q_n$ with K_n chosen below so as to ensure smoothness of f. For each n take $A_n = \mathrm{Id} + a_n \colon S^1 \to S^1$ with $a_n(x + (1/q_{n-1})) = a_n(x)$ and set $h_n = A_1 \circ \cdots \circ A_n$, $f_n := h_n \circ R_{p_n/q_n} \circ h_n^{-1}$. To find out how to choose K_n so that $f_n \to f$ in the C^∞ topology, set

$$\begin{aligned} f_{n+1,K} &= h_{n+1} \circ R_{p_n/q_n} \circ R_{1/Kq_n} \circ h_{n+1}^{-1} \\ &= h_n \circ A_{n+1} \circ R_{p_n/q_n} \circ R_{1/Kq_n} \circ A_{n+1}^{-1} \circ h_n^{-1} \\ &= h_n \circ R_{p_n/q_n} \circ A_{n+1} \circ R_{1/Kq_n} \circ A_{n+1}^{-1} \circ h_n^{-1} \end{aligned}$$

(since A_{n+1} commutes with R_{1/q_n}). Then $f_{n+1,K} \to f_n$ in the C^∞ topology as $K \to \infty$, and one can take K_n large enough so that the sequence defined by $f_{n+1} = f_{n+1,K_n}$ converges in the C^∞ topology to a function f with $f - R_\alpha \in U$, as desired.

Choose the C^0 norm of a_n such that h_{n+1} is so C^0-close to h_n that $h_n \to h$ uniformly (and the same holds for the inverses), where $h \colon S^1 \to S^1$ is a monotone surjective map. Since $f_n \circ h_n = h_n \circ R_{p_n/q_n}$ we have $f \circ h = h \circ R_\alpha$. h is a homeomorphism because if h maps an interval to a point then it maps its image under R_α to a point as well, and since finitely many such images cover S^1 this is impossible by surjectivity.

There is enough freedom left in the choice of the a_n to produce any one of the three properties in the statement □

c. Destruction and preservation of Liouvillian circles for twist maps. An individual circle with irrational but very well approximable (*e.g.*, Liouvillian) rotation number disappears under a typical arbitrarily small perturbation of an integrable twist map. Instead of this circle, an Aubry-Mather set appears Section 7.2b. However, there are always "unexpected" invariant circles.

THEOREM 7.5.9. *Any sufficiently small perturbation of an integrable twist always has, in addition to the Diophantine invariant circles, uncountably many invariant circles with Liouvillian rotation numbers.*

SKETCH OF PROOF. We will use Corollary 7.3.9. The set of invariant circles is closed and intersects any vertical segment in a closed set. If this set contains an interval then there are countably many invariant circles with rational rotation number and uncountably many invariant circles with Liouvillean rotation number. Thus we may assume that the set $\mathcal{R}(f)$ is nowhere dense. We need to show that $\mathcal{R}(f)$ contains uncountably many Liouvillean numbers. By the Baire category theorem this will follow if we show that $\mathcal{R}(f)$ is not contained in a set $D_{\beta,C}$ (Definition 7.3.1). But this follows from the fact that the boundary circles of the regions of instability are not regular (Section 7.3d), hence their rotation numbers cannot all belong to a single set $D_{\beta,C}$. Since such boundary circles are dense in the set of invariant circles, the statement follows. □

d. Perturbations of isometries in higher dimension. (See also [**AK, FH, GuK**].) The general method outlined in Section 7.2f can be used to construct perturbations of isometries that arise from nontransitive circle actions. The main idea is that the space

of orbits of such an action is sufficiently large to allow enough freedom in choosing the conjugating maps h_n in the inductive step of the construction.

There are two versions of the method. One of them aims at establishing ergodic properties of the perturbed system with respect to a particular invariant measure, usually a volume. In this case almost every orbit of the limit map has to exhibit the desired properties, and this is achieved by controlling orbit segments of rapidly growing length away from sets of measure decreasing to zero in the inductive steps. This allows to disregard such things as the presence of boundaries, of singular orbits, nontriviality of certain bundles, *etc.*

In the other version such properties as minimality, unique ergodicity or a description of all invariant measures are produced. In this situation *all* orbits of the limit map need to be controlled. Aside from stronger assumptions such as free (rather than nontrivial) action of S^1, the constructions in these cases tend to be more subtle and elaborate.

We make some comments on the constructions of the first kind. Let M be an m-dimensional manifold, not necessarily compact and possibly with boundary. An action of S^1 preserves a smooth Riemannian metric (by taking any metric and averaging it with respect to the action), and hence a volume. Furthermore, there is an open dense set of orbits that have the same period, which without loss of generality may be assumed equal to one. Let N be the space of orbits of the action, N_q the space of orbits of the subgroup $q^{-1}\mathbb{Z}/\mathbb{Z} \subset S^1$ of order q. The action projects into N_q. Naturally, an open dense set of orbits of the projected action has period $1/q$. There is an S^1-invariant closed set A of volume zero such that $M \setminus A$ with the S^1 action is diffeomorphic to a direct product of an open ball B^{m-1} with S^1 with the fiberwise action. This allows to construct a coherent sequence of fundamental domains F_q for the actions of $q^{-1}\mathbb{Z}/\mathbb{Z}$. A particular way to construct the conjugation map h_n is to take a compactly supported diffeomorphism of F_{q_n} and extend it to a diffeomorphism f of M that commutes with the action of $q_n^{-1}\mathbb{Z}/\mathbb{Z}$. In particular, one can pick a sufficiently large number k_n and constuct h_n in the above manner so that most orbits of $(k_n q_n)^{-1}\mathbb{Z}/\mathbb{Z}$ will be almost uniformly distributed in N_{q_n}. This will produce ergodicity of the limit diffeomorphism.

Another source of maps commuting with the action of S^1 and hence $q_n^{-1}\mathbb{Z}/\mathbb{Z}$ are nonconstant shifts along the orbits.

Combining these two types of constructions produces a great variety of properties, both purely ergodic and related to the differentiable structure of the limit map. The first category includes a prescribed number of ergodic components for the volume, weak mixing, nonstandardness in the sense of Kakutani equivalence, *etc.* An example of a property of the second kind is the existence of a measurable discontinuous invariant Riemannian metric, which may coexist with weak mixing [**GuK**].

CHAPTER 8

Parabolic dynamics: A special case of intermediate orbit growth

1. Introduction

a. Systems with intermediate orbit growth. The two preceding chapters presented an overview of the two extremes among the widespread types of phenomena in differentiable dynamics. These may be termed "stable" (elliptic) and "random" (hyperbolic and partially hyperbolic) motions, as in the title of a classical exposition by Moser [**Ms1**].

What remains is a grey middle ground of subexponential behavior with zero entropy that cannot be covered by the elliptic paradigms of stability and fast periodic approximation. At present, there is no way to attempt a classification of the characteristic phenomena for this kind of behavior. This is to a large extent due to the problem that the linearization is in general not well structured and not sufficiently representative of the nonlinear behavior. Furthermore, the outer boundary of the usefulness of the elliptic paradigm is not sufficiently well-defined. For example, its applicability to arbitrary interval exchange transformations (Section 4.3g) is open to question. However, there is a type of behavior within this intermediate class that is modeled well enough by the infinitesimal and local "shear" orbit structure of unipotent linear maps. It is this kind of behavior that we call parabolic, and whose typical features we aim to identify.

The survey [**S-MT**] deals with parabolic dynamics, as does a substantial part of [**S-KSS**].

b. Parabolic linear paradigm: Jordan blocks and polynomial growth. From the dynamical point of view, a parabolic linear map, *i.e.,* one that has only eigenvalues of absolute value one and possesses some nontrivial Jordan blocks, is characterized by the presence of some isometric directions (corresponding to the eigenspaces) and polynomial growth in both the positive and negative direction of time for all other vectors. Of course, this growth is achieved by shifts relative to each other of invariant affine subspaces parallel to the sum of the eigenspaces. In every such space the map acts isometrically.

Parabolicity is an exceptional and hence unstable phenomenon in all natural spaces of linear maps. For example, in $\mathrm{SL}(2,\mathbb{R})$ parabolic matrices, characterized by the condition $|\mathrm{tr}A| = 2$, separate the open sets of elliptic and hyperbolic matrices. Accordingly, one may expect parabolic behavior of nonlinear systems to be exceptional and generally unstable under perturbations. There are, however, special situations, where parabolic behavior is the norm due to low-dimensionality or special properties of the dynamical systems under consideration.

c. Nonlinear systems with parabolic linear part: Local shear. Parabolic phenomena in nonlinear systems appear in several ways. One is the uniform parabolicity best

represented by affine and homogeneous unipotent examples (Section 8.3a, Section 8.3b). Various classes of skew-products such as smooth distal systems (Section 2.4a) also belong to that class. In this case, the relative behavior of orbits throughout the phase space resembles that of parabolic linear systems. Another way for parabolicity to appear for typical orbits is the presence of nonuniformities in the system. Such nonuniformities may have the form of outright singularities, as in billiards and mechanical systems involving collisions (Section 8.5), or special features, such as hyperbolic fixed points, which are exceptional by themselves, but have a long-term influence on typical orbits. Smooth flows on surfaces of higher genus are the most typical examples of the latter effect (Section 8.4).

Parabolic systems are characterized by triangular (or block-triangular) derivative at a typical orbit, corresponding to the corner of which there is usually an integrable distribution with isometric behavior. Then there are invariant filtrations for the linearized system, which are usually integrable. The relative behavior of leaves of an invariant foliation within the next one is also usually isometric. Thus, separation of nearby orbits is achieved via a "shear", rather than by outright expansion of distances.

d. Parabolic systems with singularities. Smooth dynamical systems with singularities are important objects of study, because singularities may be an essential intrinsic feature of the system (such as collisions in mechanical systems) or may appear as a result of applying constructions such as section maps for flows. Singularities introduce an essentially new feature compared to smooth or even topological systems. By encountering a discontinuity, orbits may instantaneously diverge. The usual mechanisms of divergence (infinitesimal and local growth) help to make these events more likely. Thus the complexity of orbit behavior is produced by a combination of "cutting" at singularities and local expansion, which may become extreme near singularities due to unbounded derivatives.

In assessing the influence of singularities and, in particular, in deciding whether behavior in such a system should be classified as parabolic, one needs to modify the criteria put forward for the smooth case. A good indicator of parabolicity is nontrivial polynomial (or subexponential) growth of the complexity of the orbit structure. Example 2.6.9 shows that different ways of measuring this complexity may lead to classifying the same system as elliptic or parabolic.

It is interesting to note that when cutting is not accompanied by expansion or where expansion is slow outside of the singularities, the system tends to display at least some features of parabolic behavior. A prototypical class of examples is that of interval exchange maps (Section 4.3g), which are piecewise isometries. In other words, systems with singularities that are locally elliptic or parabolic, tend to be parabolic.

2. Main features of parabolic behavior

Unlike in the hyperbolic and elliptic situation there are no general theorems that identify parabolic behavior in broad classes of dynamical systems. Therefore we take a "botanical" approach in this chapter. After going through the list of characteristic common features of parabolic systems, we describe known representatives of each species individually.

a. Growth of the orbit complexity. Parabolic systems exhibit a distinctly subexponential pattern of the *global* orbit behavior. This should not be confused with a requirement of subexponential infinitesimal or local orbit growth. For example, hyperbolic saddles are an essential feature of area-preserving flows on surfaces of higher genus (Section 8.4),

which are quite typical parabolic systems. A characteristic feature of parabolic systems is the coincidence of the upper and lower topological power entropies and that the (thereby well-defined) topological power entropy $\mathrm{ent}_p^{\mathrm{top}}$ (Section 2.5i) is finite, but not zero. See also Section 3.7l. The cases where $\mathrm{ent}_p = 0$ (see Section 2.5h) and $\mathrm{ent}_p^{\mathrm{top}} > 0$ (Example 2.6.9) represent the border area between elliptic and parabolic behavior.

b. Relative behavior of orbits. The "shear" structure of the linearized system leads to the following crucial phenomenon: Once two orbits come close to each other, they stay close for a certain time, which is usually commensurate to a negative power of the distance. Thereafter, for a much longer time they maintain a distance that is moderate, but bounded from below.

Thus, each close encounter is followed by a controlled period of separation. This is in stark contrast with typical hyperbolic behavior and is a root cause of the comparative uniformity of the orbit structure in parabolic systems.

c. Recurrence. In terms of recurrence, parabolic behavior is associated with some uniformity of recurrence and as such is closer to elliptic than to (partially) hyperbolic dynamics. In the elliptic situation, minimality is characteristic for the topologically transitive case, and a general elliptic system typically decomposes into disjoint regular minimal orbit closures. In the parabolic case, minimality is still among the prevalent types of recurrence, especially for systems of algebraic origin (Section 8.3), but it is not automatic for topologically transitive systems. Nevertheless, orbit closures are typically regular (essentially submanifolds of the phase space), and in the topologically transitive case there is usually an open subset of dense orbits, a phenomenon called *quasi-minimality*. This is prevalent for area-preserving flow on surfaces (Section 8.4). The *exceptional set* of nondense orbits is often a finite union of orbits.

d. Invariant measures. Analogously to the way recurrence typically presents itself in parabolic systems, the structure of the space of invariant measures in a parabolic system is much closer to that in the elliptic situation than to the hyperbolic or partially hyperbolic scenarios.

1. *Essential unique ergodicity.* Unique ergodicity is still the most characteristic behavior in the minimal case. In the quasi-minimal situation there likewise is usually only one "dominant" invariant measure that is positive on open sets, accompanied by an (often finite) collection of invariant measures, which are often quite regular. In this "quasi uniquely ergodic" case the totality of invariant measures usually has a simple structure that is well-understood. Note, however, that deviations from uniform distribution with respect to the dominant invariant measure are often inevitable due to the presence of semiorbits asymptotic to the exceptional set. Such is the case for flows on surfaces. For algebraic systems due to a remarkable result of Ratner [**S-KSS**] each orbit is uniformly distributed according to a certain invariant measure.

2. *Essential finiteness of the set of invariant measures.* Another possible feature, which also appears in elliptic systems with Liouvillian recurrence, is the presence of several ergodic measures that share the dominant component. Their number is often finite and related to the geometry of the system (Theorem 8.4.5). An essential difference from the elliptic case is that in natural finite-parameter families of systems minimality and unique ergodicity appear in distinct ways, even though both are still prevalent. Nonminimality only

appears, when the parameters satisfy some algebraic relations and is hence restricted to a countable union of positive-codimension submanifolds of the parameter space (Corollary 8.4.4). In contrast, unique ergodicity holds for a set of parameters of full measure, whose complement has a more complicated structure (Section 8.4d).

e. Mixing properties. For mixing behavior in parabolic systems it is important that the local triangular, or shear, structure provides locally for an invariant isometric factor. This factor may or may not be globally integrable. If it is not, then the system is topologically mixing.

In the integrable case there is, globally, an isometric factor. In extreme cases it may be the identity, such as for completely integrable systems. However, one can still define mixing relative to this factor. Topological mixing relative to an isometric factor means that whenever A is an open set that projects onto the isometric factor, then

$$(\text{for all open } B\,)(\exists N \in \mathbb{N})(\forall n \geq N) \quad f^n(A) \cap B \neq \varnothing. \tag{8.1}$$

Topological mixing is one area where the boundary between the elliptic and parabolic situations cannot be clearly drawn. While the leading elliptic paradigm, equicontinuity of iterates and smooth rigidity associated with fast periodic approximation (Section 7.3b), are incompatible with (8.1) for any proper open subset A, it is probably possible to have, for example, topologically mixing nonstandard smooth realizations of some rotations or translations on the torus.

On the other hand, for parabolic systems without an isometric factor, the behavior with respect to an invariant measure supported on the dominant set is characterized by the presence of moderate mixing. The weakest such property, which holds in all known cases, is mild mixing (Section 3.6g). The possibility of mixing (Section 3.6h) is much less clear. It seems that both mixing and its absence are fully compatible with parabolic behavior. Uniformly parabolic systems are mixing, unless they have an isometric factor, but otherwise mixing is dependent on rather subtle estimates of the power of local "shearing" (Section 8.4e).

For systems with an isometric factor, mixing appears for functions orthogonal to all eigenfunctions or, equivalently, for sets independent of all sets from the isometric σ-algebra. Furthermore, uniformly parabolic systems have countable Lebesgue spectrum in the orthogonal complement to eigenfunctions and are multiply mixing (Section 3.6h) whenever they are mixing. On the other hand, maybe some systems with nonuniform parabolicity are mixing without being multiply mixing, or have absolutely continuous spectrum of finite multiplicity. Neither of these situations is known to hold for any measure-preserving transformation.

f. Decay of correlations. As mentioned in Section 3.1, the speed of correlation decay with respect to an invariant measure is not an invariant, while the convergence of correlations to 0 is simply equivalent to mixing. For mixing smooth dynamical systems, an interesting and important problem is to find the rate of correlation decay for particular classes of functions, primarily those of smooth or Hölder continuous functions. For parabolic systems, not much about this is known in general. However, for systems of an algebraic nature and others that exhibit uniformly parabolic behavior, polynomial decay of correlations seems to be typical.

g. Invariant distributions. An interesting feature of parabolic systems, which sets it apart from both elliptic and hyperbolic systems, is the presence of invariant distributions (Section 5.2n) that are not determined by measures (Example 5.2.3). There are two typical paradigms:

(1) The presence of infinitely many independent distributions of a particular order (Section 8.3b6).
(2) The existence of a finite (but growing) number of invariant distributions for any finite order (Section 8.4f).

h. Speed of convergence of ergodic averages. There is a connection between invariant distributions and the speed of convergence of ergodic averages for some natural classes of functions such as smooth or characteristic functions of nice sets. For parabolic systems the speed is typically given by a negative power of time. The estimates improve when more and more invariant distributions vanish [**FFo**].

i. Rigidity of the measurable orbit structure. The measurable structure with respect to the dominant invariant measure tends to be fairly rigid and to determine the differentiable structure. Specifically this means that within a given class of parabolic systems (such as horocycle flows [**Ra1**]) any measurable isomorphism is smooth and in fact of a very special kind. Furthermore, these properties include full descriptions of measurable centralizers, factors, and joinings of the systems within a given class in algebraic terms.

All of these present a sharp contrast with the hyperbolic situation, where all interesting invariant measures are Bernoulli. In this situation the centralizer of a system is huge and consequently, once two systems are isomorphic there is a huge variety of different isomorphisms. Thus, the measurable structure carries very little information about the differentiable one.

3. Parabolic systems with uniform structure

Natural classes of parabolic systems are provided by algebraic dynamics, *i.e.,* homogeneous and affine maps and flows on homogeneous spaces of Lie groups [**S-KSS**, Chapter 4].

a. Affine maps on the torus. Let $L \in \mathrm{SL}(m,\mathbb{Z})$ be a quasi-unipotent matrix, *i.e.,* all eigenvalues are roots of unity, and assume L has some nontrivial Jordan blocks. For $v \in \mathbb{R}^n$ let $A_{L,v}\colon \mathbb{T}^m \to \mathbb{T}^m$, $x \mapsto Ax + v \pmod 1$. The simplest prototype examples correspond to $L = \begin{pmatrix} 1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & 1 \\ & & & 1 \end{pmatrix}$ and $v = (\alpha, 0, \dots, 0)$ with $\alpha \in \mathbb{R} \smallsetminus \mathbb{Q}$, and were denoted by $A_{n,\alpha}$ in Section 4.3e. Let E be the sum of the eigenspaces of L. It is always rational and hence projects to a rational torus in $\mathbb{T}^m$. Thus, there is a natural factor of $A_{L,v}$ on a lower-dimensional torus, with isometric behavior in the fibers. It is also a quasi-unipotent affine map. Continuing inductively, we arrive at a pure translation.

PROPOSITION 8.3.1. *A map $A_{L,v}$ is is always distal (Section 2.4a). It is minimal if and only if the translation factor is. In this case, $A_{L,v}$ is uniquely ergodic.*

Further properties of these maps fit well into the general scheme outlined in the previous section

(1) If $A_{L,v}$ is minimal then $L^2(\mathbb{T}^m, \lambda) = H_0 \oplus E_L$, where H_0 is the invariant space spanned by eigenfunctions of the unitary operator $U_{A_{L,v}}$, and $U_{A_{L,v}}$ has countable Lebesgue spectrum in the space E_L.
(2) There are infinitely many independent invariant distributions for $A_{L,v}$ on the space of functions with absolutely convergent Fourier series, and these span the space of invariant distributions.
(3) $\mathrm{ent}_p(A_{L,v}) = k - 1$, where k is the maximal size of a Jordan block of L. In particular, $\mathrm{ent}_p(A_{n,\alpha}) = n$.

b. Homogeneous dynamics for nonabelian groups. (See also [**S-KSS**].) Parabolic behavior also appears in homogeneous maps and flows on homogeneous spaces of nonabelian Lie groups. Recall that an element g of a Lie group G is said to be *unipotent* if $(\mathrm{Ad}_g - \mathrm{Id})^k = 0$ for some $k \in \mathbb{N}$ (*i.e.*, all eigenvalues of Ad_g are 1). It is said to be *quasi-unipotent* if all eigenvalues of Ad_g are on the unit circle. The action of the left translations L_g on the space of right-invariant vector fields on G is given by the linear operator Ad_g. Thus, for any nonabelian Lie group G, any discrete subgroup $\Gamma < G$ and any quasi-unipotent $g \in G$, the left rotation L_g on G/Γ is parabolic.

Two special examples of this kind merit particular attention.

EXAMPLE 8.3.2. Nilflows: Here G is a simply connected nonabelian nilpotent Lie group and $g \in G$ does not belong to the center of G. Nilflows are distal (Section 2.4a).

EXAMPLE 8.3.3. Unipotent translations on semisimple groups, in particular the horocycle flow (Section 4.3f), where $G = \mathrm{SL}(2, \mathbb{R})$ and $h^t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$. These systems are not distal since they do not have isometric factors.

The dynamical and ergodic properties of these homogeneous maps and flows is the most complete illustration of the characteristic features of parabolic dynamics we just described.

(1) Every orbit closure is a homogeneous submanifold (Ratner's topological theorem [**Ra3**]).
(2) Every ergodic invariant measure is Haar measure on a homogeneous submanifold, in particular, minimality implies unique ergodicity (Ratner's measure-theoretic theorem [**Ra4**]).
(3) In general, Haar measure has a mixture of discrete and countable Lebesgue spectrum. Nilflows provide an example where both components are present and their properties are close to those of affine unipotent maps on the torus.
(4) In the semisimple case and many others, Haar measure has countable Lebesgue spectrum. In this case, there is mixing of all orders.
(5) Smooth functions have polynomial decay of correlations.
(6) For $G = \mathrm{SL}(2, \mathbb{R})$ if only principal and discrete series appear in the representation of G on $L^2(C/\Gamma)$, then there are infinitely many invariant distributions of fixed order, and they determine all invariant distributions. If there are also representations from the complementary series, infinitely many distributions of growing orders also appear.

(7) If $g \in G$ is quasi-unipotent, Γ is a cocompact lattice and L_g the induced translation on G/Γ then $\text{ent}_p(L_g) = k - 1$, where k is the size of the largest Jordan block of Ad_g.

c. Extensions. Parabolic behavior is inherited by compact group extensions and isometric extensions (Section 2.2i). Furthermore as we already saw, some extensions of isometries and other systems with elliptic behavior are parabolic. In general, the apperance of parabolic behavior in an extension depends on the "shear" produced by the cocycle generating the extension.

Naturally, in order to discuss parabolic behavior properly, one needs to assume a smooth structure throughout the construction. It is interesting that, while no transitive translations on compact nonabelian groups exist, even irrational rotations of the circle have transitive and uniquely ergodic compact group extensions with nonabelian fiber.

d. Time changes. Time changes are another construction that preserves the uniformly parabolic behavior while (in general) destroying the homogeneous structure. Ratner [**Ra2**] showed that many typical properties of horocycle flows including rigidity of measurable orbit strucute extend to a broad class of time changes of such flows.

4. Flows on surfaces

This subject is developed in [**KH**, Chapter 14].

a. Section maps. Smooth flows on compact surfaces that preserve a measure whose support is the whole surface, are examples of parabolicity in a setting that is not algebraic or connected to isometric behavior. The leading case is the area-preserving one. Locally, the triangular structure near any reference orbit arises naturally by way of the orbit foliation, *i.e.,* the invariance of the generating vector field. Interesting effects may be produced already by slowing down the velocity in a fixed point free flow on the torus to produce an isolated fixed point. This case is a useful model where explicit calculations are greatly simplified.

However, the really essential case appears on surfaces of genus greater than one, where the presence of fixed points is unavoidable due to the Poincaré-Hopf index formula. Assuming that all fixed points are isolated, the invariant measure condition guarantees that all fixed points are either centers or saddles (possibly multiple ones). Then there are only finitely many fixed points, while by area-preservation and the Poincaré Recurrence Theorem (Section 3.4c) recurrent points are dense. Therefore [**KH**, Proposition 14.1.4], for any transversal to the flow, the return map is defined and continuous except possibly at finitely many points (the last intersections of the incoming separatrices of the saddles with the transversal), and at those points both one-sided limits exist. In fact, one can consider the flux of the invariant measure through the transversal and thus find a reparametrization such that the induced map is an interval exchange map (Section 4.3g). Furthermore, in the area-preserving case, this reparametrization is smooth. In summary we have:

THEOREM 8.4.1. *For area-preserving flows with only isolated singularities the return map to any transversal is smoothly conjugate to an interval exchange.*

Thus, the transverse aspects of the dynamics of the flow are competely reducible to those of interval exchanges. As we mentioned above, interval exchange transformations

represent a borderline situation between elliptic and parabolic behavior. To be more precise, aperiodic interval exchanges do exhibit linear orbit growth, but in certain situations this is just an artifact, such as for the exchange of two intervals, which is semiconjugate to a rotation via identification of the endpoints of the interval (*cf.* Example 2.6.9). Furthermore, even genuine interval exchange maps, *e.g.*, weakly mixing exchanges of three intervals [**KS**], often allow good periodic approximation (Section 3.6e), which is typical for elliptic rather than parabolic behavior.

However, these distinctions are irrelevant, when one considers not only the transverse dynamics of flows with fixed points, but the full dynamics including time-change aspects. Parabolicity appears due to "shearing" near the fixed points, which slows various orbits down according to their distance from the incoming separatrices. Thus, for example, even the time change of an irrational linear flow on the torus with zero velocity at some point displays parabolic behavior, despite the fact that it has rotations as natural section maps.

It is interesting that the presence of hyperbolic fixed points in this situation does not produce hyperbolicity in the return map. The mechanism of neutralizing the hyperbolic effect is that the slowdown of orbits near the saddles concentrates the splitting of orbits into a singularity of the return map (at which the return times diverge). Another interesting observation is that, contrary to superficial intuition, nondegenerate saddles, *i.e.,* hyperbolic fixed points, produce less "shearing" effects than nonhyperbolic degenerate saddles. In some cases this may result in such significant distinctions as absence versus presence of mixing (Section 8.4e).

b. Measured foliations. A generalization of the notion of an area-preserving flow on a surface is that of a *measured foliation* [**FLP**]. It turned out to be extremely fruitful for the Teichmüller theory [**S-MT**] as well as for the Nielsen–Thurston theory of surface homeomorphisms [**S-FM**]. A measured foliation on a surface (possibly with boundary) is a foliation with one-dimensional leaves that has a transverse holonomy-invariant measure that is positive on open sets and whose singularities are isolated and of the "generalized saddle" type. A "generalized saddle" is a fixed point with $N \neq 2$ saddle-like sectors bounded by prongs as separatrices. For $N = 2k$ it is topologically an ordinary saddle of multiplicity $k - 1$, for odd N it is locally nonorientable. The index of such a saddle is $(2 - N)/2$ and the usual Poincaré-Hopf formula holds. The basic cases are $N = 1$ (a return point) with index 1/2 and $N = 3$ (a half-saddle) with index $-1/2$. Measured foliations provide another interesting example of dynamics without time (Section 1.2c). Ordinary foliations with one-dimensional leaves generated by nonvanishing line fields can be made into flows by lifting to a double cover. Similarly a measured foliation can be made orientable by passing to an appropriate finite *branched cover* that branches over all singularities with odd number of prongs.

As an interesting example, on the sphere there are measured foliations with nontrivial recurrence and dense leaves with four one-prong saddles. These appear as stable and unstable "foliations" for pseudo-Anosov type diffeomorphisms, which provided the first examples of area-preserving sphere maps with stochastic behavior [**K3**]. Among the foliations with half saddles are the stable and unstable "foliations" of generic pseudo-Anosov diffeomorphisms on surfaces of genus at least two [**S-FM, FLP, GK**].

c. Topological properties. *A saddle connection* is an orbit that is positively and negatively asymptotic to (not necessarily distinct) saddles.

THEOREM 8.4.2 ([**KH**, Theorem 14.6.3]). *Let M be a surface of genus g and φ a flow with finitely many fixed points that preserves a measure positive on open sets. Then M decomposes invariantly as $M = \bigcup_{i=1}^{k} P_i \cup \bigcup_{j=1}^{l} T_j \cup C$, where the P_i are periodic components (open sets of periodic orbits), the T_j are transitive components (open with every semiorbit dense except incoming separatrices of fixed points), and C is a finite union of fixed points and saddle connections. Furthermore, $l \leq g$.*

A flow on a surface (possibly with boundary) is said to be *quasiminimal* if it has finitely many fixed points and every semiorbit other than a fixed point or a separatrix of a saddle is dense. Thus, the closure of any transitive component is a surface with a quasiminimal flow.

COROLLARY 8.4.3. *Transitivity, topological mixing and quasiminimality are equivalent for flows on compact surfaces that have finitely many fixed points and preserve a measure positive on open sets.*

Periodic components can appear in two different ways. If such a component contains an orbit homologous to zero (and hence consists of only such orbits) then this component persists under small perturbations in the space of area-preserving vector fields. We call such periodic components *stable*. Stable periodic components appear around centers but they may also be present in flows whose only fixed points are saddles. An example is a flow on the orientable surface of genus two obtained by taking two flows on the torus with a hole formed by a homoclinic loop of a single saddle and connecting these by a collar filled with closed orbits. Equivalently, glue two toral flows with a single center each along closed orbits around the centers (deleting neighborhoods of the centers).

An area-preserving flow on a surface (possibly with a boundary) is said to be *irreducible* if it has no stable periodic components. The statement of Theorem 8.4.2 can be refined in an obvious way by taking out stable periodic components and dividing the rest into irreducible components. This situation differs only in minor ways from the leading case of an irreducible flow on a closed compact surface. There are open sets of area-preserving irreducible flows on the torus and on all closed compact surfaces (orientable or not) with Euler characteristic ≤ -2. Any area-preserving flow on the sphere, the projective plane or the Klein bottle has an open dense set of stable periodic orbits. A somewhat less familiar fact is that on the nonorientable surface of genus -1 an open dense set of area-preserving flows have this property, although irreducible flows also exist. Flows in the interior of the space of irreducible flows are said to be *stably irreducible*. To study flows with multiple saddles one defines the notions of irreducibility and stable irreducibility subject to restrictions on the structure of the saddles.

Thus, for further study of nontrivial dynamical behavior the quasiminimal case is the main one. Furthermore, it is prevalent among irreducible flows, as will be shown later.

COROLLARY 8.4.4. *If there are no saddle connections then the flow is quasiminimal.*

d. Invariant measures and smooth orbit classification. It is quite remarkable that the restriction on the number of quasiminimal components has a counterpart in the restriction on the number of ergodic invariant measures supported by these components.

THEOREM 8.4.5 ([**St**], [**KH**, Theorem 14.7.6]). *For any area-preserving flow on a compact surface of genus g there are at most g ergodic nonatomic invariant measures.*

Furthermore, for any k with $1 \leq k \leq g$ there exists a quasiminimal area-preserving flow that has exactly k nonatomic ergodic invariant measures.

SKETCH OF PROOF. The transverse smooth structure for area-preserving flows is locally (in the space of vector fields) determined by finitely many parameters [**KH**, Section 14.7c]. First among these are the number of saddles and their indices. If each saddle is generic among those with given index, continuous parameters may be defined as fluxes through a basis of the first homology group $H_1(M, F, \mathbb{R})$ relative to the set F of fixed points. These are defined up to a common scalar multiple and fix the *fundamental class* of the flow in the cohomology group $H^1(M, F, \mathbb{R})$. Thus, orbit equivalence is determined by the projectivization of the fundamental class. Restriction of the fundamental class to the group $H^1(M, \mathbb{R})$, which is naturally embedded into $H^1(M, F, \mathbb{R})$, gives an element which is Poincaré dual to the asymptotic cycle (Section 2.3f) with respect to the invariant area measure. The orbit equivalence between nearby vector fields with the same fundamental class is Lipschitz continuous everywhere and smooth outside of F. The total number of parameters is maximal for flows with nondegenerate saddles, for which it is $4g - 4$.

The estimate on the number of nonatomic ergodic invariant measures is obtained in three steps. First one notices that such a measure is defined by its fluxes through a basis in $H_1(M, F, \mathbb{R})$, thus defining an element of $H^1(M, F, \mathbb{R})$. Then one sees that, in fact, this element is determined by its restriction to absolute cycles, *i.e.*, the asymptotic cycle, thus decreasing the dimension of the space of ergodic measures to no more than $2g$. Finally, from the fact that different orbits of the flow do not intersect and using typical orbits for various invariant measures, one deduces that asymptotic cycles of various measures are in involution with respect to the intersection form, which is a symplectic form. This gives the final estimate. □

Notice that a constant time change multiplies the fundamental class by a scalar. Thus, both minimality and unique ergodicity depend only on the projectivization of the fundamental class. Both minimality and unique ergodicity are, in fact, prevalent among irreducible volume-preserving flows. For minimality, this follows immediately from Corollary 8.4.4 because the presence of a saddle connection means that the flux through some relative cycle vanishes. Thus we have

THEOREM 8.4.6. *For values of the projectivized fundamental class outside of a countable union of codimension one submanifolds, the corresponding irreducible flows are quasiminimal.*

There are similar statements for flows with multiple saddles and for flows with stable periodic components.

Prevalence of unique ergodicity follows from the corresponding result for interval exchange transformations (Section 4.3g).

THEOREM 8.4.7. *There is a set of full measure in the space of projectivized fundamental classes for which the corresponding irreducible area-preserving flows are uniquely ergodic.*

The value of the fundamental class does not determine the topological type of even a quasiminimal flow globally in the space of vector fields [**Lv**]. An alternative parametrization for the topological types of area-preserving flows follows from the Thurston parametrization of the boundary of the Teichmüller space [**S-MT, FLP**] and, in fact, gives a global

parametrization up to a topological conjugacy homotopic to the identity modulo some rearrangement of saddle connections. Most of the boundary is identified with generic (and hence nonorientable) measured foliations (Section 8.4b), so in order to include flows, one needs a careful analysis of the lower-dimensional strata of the boundary. While Thurston's parametrization brings a better view of the global geometry of a general area-preserving flow, it does not give any significant new dynamical information, either in terms of topological recurrence or of invariant measures.

The analysis of this and the previous subsection extends more or less straightforwardly from the case of flows to general measured foliations. In particular the stable and unstable "foliations " of pseudo-Anosov and generalized pseudo-Anosov maps are uniquely ergodic [**FLP**]. For prevalence of unique ergodicity among measured foliations see [**Ms2**].

e. Mixing and return-time singularities. A conjugacy between the section map of an area-preserving flow on a compact surface and an interval exchange transformation T produces a metric conjucacy between the flow and a special flow over T. The roof function for the special flow is the return time for the section map and can be expressed through the flux parameter. It is convenient to assume that the transversal itself is a cycle relative to F, *i.e.,* that is it is either closed or has endpoints in F. The roof function is positive and differentiable everywhere except for the last points of intersection of the incoming separatrices of the saddles with the transversal, where it goes to infinity on both sides. The set S of these points contains all discontinuity points of T and $\operatorname{card} S \leq 4g$. In the case of a closed transversal the conjugacy with the special flow is, in fact, differentiable outside of F, where it is undefined. The singularities of the return function produce a strong shear, which may cause mixing.

The following result is not directly applicable to the case of area-preserving flows, but it demonstrates that the transverse behavior is not sufficient for mixing when augmented by only a moderate shear.

THEOREM 8.4.8 ([**CFS, K5**]). *Let $T\colon I \to I$ be an interval exchange transformation. There exists $N \in \mathbb{N}$ such that for any measurable $A \subset I$ and any $n \in \mathbb{N}$ there are $n_1, n_2, \dots, n_N > n$ with $A \subset \bigcup_{i=1}^{N} T^{n_i}(A)$.*

Let $\tau\colon I \to \mathbb{R}$ be of bounded variation and φ the special flow over T with roof function τ. Then there exist $N \in \mathbb{N}$ and $V \in \mathbb{R}$ such that for any measurable set A and any $n \in \mathbb{N}$ there are $n_1, n_2, \dots, n_N > n$ with $A \subset \bigcup_{i=1}^{N} \bigcup_{t=0}^{V} \varphi^{n_i+t} A$.

Since these conclusions give arbitrarily late returns to A with relative measure at least $1/N$ (rather than $\lambda(A)$), they preclude mixing.

COROLLARY 8.4.9. *Neither an interval exchange transformation nor a special flow over an interval exchange transformation with roof function of bounded variation is mixing with respect to any invariant measure.*

This result implies absence of mixing for billiards in rational polygons on their ergodic components (Corollary 8.5.2).

Mild mixing (Section 3.6g) is possible already for an exchange of three intervals, although most of such interval exchanges are rigid [**KS**].

Thus, unboundedness of the return function is essential for mixing behavior. As it turns out, nondegenerate hyperbolic saddles produce logarithmic singularities, while degenerate saddles produce stronger power singularities. In both cases, singularities are

symmetric up to a bounded function. As was shown by Kocergin [**Kc1**], logarithmic singularities are not sufficient by themselves to produce mixing for the special flow. Specifically, he proves absence of mixing for special flows over a rotation with only symmetric logarithmic singularities. Thus, there are area-preserving flows with nondegenerate singularities that are not mixing because a rotation appears as a section map.

On the other hand, Kocergin [**Kc2**] shows that an ergodic area-preserving flow is mixing if there is at least one degenerate saddle. Thus, power singularities (symmetric or not) are sufficient to produce mixing in the special flow over an interval exchange transformation, while symmetric logarithmic singularities are not. It is not known whether for a flow with only nondegenerate saddles the combination of transversal dispersion and weak shear can cause mixing, even though neither phenomenon is sufficient by itself.

Interestingly, *nonsymmetric* logarithmic singularities do produce mixing [**Kc2, KhS**], although such singularities do not appear from smooth area-preserving flows.

f. Invariant distributions and smooth classification. While a classification of area-preserving flows up to smooth orbit equivalence is essentially given by the projectivized fundamental class (Section 8.4d), a classification up to flow equivalence (smooth or topological) has to take time changes into account. This involves a classification of cocycles (Section 1.3m). We know already from the simpler situation on the torus that the Diophantine and Liouvillian situations produce strikingly different pictures with respect to cocycles and time changes (Proposition 7.3.2, Proposition 7.5.2), and one should expect similar problems in the higher genus case.

However, the basis for any kind of classification of smooth cocycles is to determine the set of invariant distributions, because their vanishing determines the closure of the space of coboundaries (Section 5.2n). Some invariant distributions are related to the behavior near the singularities in that they are determined by the normal form at the singularities, which, in general, depends on infinitely many paramenters.

To concentrate on the essential dynamical phenomena, one should consider time changes that are trivial near the singular set, and the cocycles corresponding to them. Forni [**Fo**] found a complete classification of such cocycles. The picture here is quite different both from isometries (no nontrivial invariant distributions) and from uniformly parabolic systems of transverse dimension greater than one (infinitely many independent distributions of a certain order). Forni found that for area-preserving flows there are finitely many nontrivial invariant distributions of any finite order r but their number goes to infinity with r.

Furthermore, there is a counterpart of the cocycle rigidity for Diophantine rotations and translations of the torus: For almost every value of the fundamental class, a sufficiently smooth cocyle in the kernel of the proper set of invariant distributions is a coboundary (with a transfer function of lower regularity).

5. Billiards in polygons and polyhedra and related systems

This subject is systematically developed in the survey [**S-MT**]. See also [**Ta**].

a. Billiard flow and the section map. Polygonal billiards are dynamical systems that display both a moderate expansion of parabolic type and cutting due to singularities. Certain subclasses of polygonal billiards clearly fall within the parabolic paradigm and have, in fact, been analyzed quite comprehensively. But the general case remains somewhat

elusive, and, while certain features of parabolic behavior are unmistakably present, no good hold on the global complexity of the orbit structure has been achieved yet.

Consider a connected polygonal domain P in the plane, which need not be convex or even simply connected. As in Section 7.3f, the *billiard flow* in P is defined on the space M of unit tangent vectors with foot points inside P or on the boundary and pointing inside. A tangent vector moves along its axis with unit speed and upon reaching the boundary instantaneously reflects from it, *i.e.,* changes its direction according to the rule of optics "the angle of incidence is equal to the angle of reflection". Ambiguities appear for an orbit hitting a vertex. We are not concerned with such orbits *per se*, but, of course, the presence of vertices is responsible for discontinuities, which are essential features of polygonal billiard flows.

Locally, the billiard flow is essentially linear, so its infinitesimal and local orbit structures coincide. They are, in fact, unipotent: There are two isometric directions, that of the flow itself and one corresponding to parallel translation of the reference tangent vector perpendicularly to its axis. There is one direction of linear growth corresponding to the rotation with fixed foot point. Thus, locally there is a double eigenvalue one and one Jordan block of size two. The billiard flow clearly preserves the phase volume.

Reflections do not change this pattern. The commonly used device of "unfolding" replaces the reflection of an orbit by reflection of the polygon, continuing the orbit along the same straight line.

FIGURE 8.1. Unfolding

Clearly, any orbit not hitting a vertex is completely determined by its encounters (point and direction) with the boundary. Thus the set S of vectors with foot points on the boundary provides a natural section for the billiard flow. It can be viewed as the cylinder (for a simply connected polygon) or as a union of several cylinders. Vectors with foot points on a given side form a rectangle parametrized by the length parameter l and the angle φ with the side. The corresponding section map, which is defined for vectors that do not hit any side of any of these rectangles, is called *the billiard map*. It is piecewise linear in the coordinates (l, φ) and has a local triangular structure with a single Jordan block of size two inherited from the billiard flow.

The decomposition of the boundary into sides provides a natural procedure for *coding* of the billiard map. The resulting symbolic systems are certainly not of finite type and their properties are not easily identifiable. Still, it is a useful tool in studying orbit growth in billiards [**K8**].

b. Integrable billiards. One might expect that the local triangular structure described above sometimes integrates and produces a one-parameter family of invariant manifolds. In the most straightforward way, this only happens for a limited collection of polygons, namely those which tile the plane by the group generated by reflections in their sides, or, equivalently, have all angles of the form π/p, $p = 2, 3, 4, 6$. This list includes rectangles, the equilateral triangle and two right triangles. Such billiards are said to be *integrable*. For integrable billiards a *complete unfolding* is produced by the tiling and hence the billiard flow is reduced to an invariant subset of the completely integrable geodesic flow on the flat torus, which splits into invariant tori with linear flows. Notice in particular that in this case

FIGURE 8.2. Complete unfolding

FIGURE 8.3. Fundamental domains

orbits hitting the vertices can be uniquely extended through the vertices.

c. Rational billiards and quadratic differentials. The situation is much more interesting, when the integrable structure is allowed to possess some singularities.

1. *Rational polygons and the angle integral.*

DEFINITION 8.5.1. A polygon is said to be *rational* if the angle between any two of its sides is a rational multiple of π.

Equivalently, a polygon is rational if the linear part (*i.e.,* the factor Λ consisting of rotations and reflections) of the group generated by reflections in its sides is finite. Thus, for a rational billiard, the direction of a vector in the phase space may take only finitely many values under the flow evolution, namely, the values corresponding to the orbit of the initial direction under the finite group Λ. In other words, the billiard flow has a first integral called *the angle integral*, and each level surface of the angle integral is the union of finitely many copies of P.

Correspondingly, for the billiard map, there are invariant sets, each of which is a union of finitely many intervals. Since the section map preserves the area element $\sin\varphi\, d\varphi\, dl$, each invariant set can be reparametrized as the unit interval in such a way that the billiard map acts as an interval exchange transformation (all arguments above neglect the ambiguities appearing in the finite discontinuity sets). Thus, the billiard map for a rational polygon splits into a one-parameter family of interval exchange transformations. It is easy to see that for each interval exchange the return time is a piecewise linear function. Thus, the billiard flow splits into one-parameter family of invariant sets on each of which it is isomorphic to the special flow over an interval exchange transformation with piecewise linear roof function. Since piecewise linear functions have bounded variation, Corollary 8.4.9 implies:

COROLLARY 8.5.2. *The billiard flow in a rational billiard is not mixing on any level surface of the angle integral.*

We also obtain a reasonably complete description of ergodic invariant measures for rational billiards: every such measure is concentrated on a single level surface of the angle integral and every level surface can carry only a fixed number (depending on P) of different nonatomic ergodic measures. Furthermore, the support of any such measure is a domain on the level surface. Atomic measures appear in continuous families and only for countably many values of the angle integral. The number of periodic families for each level is again bounded by the same constant.

2. *The associated Riemann surface.* There is a more geometric and altogether more elegant way of describing the structure of billiards in rational polygons. The survey [**S-MT**] extensively discusses this, so we give only a brief outline (see also [**Gk2**]).

Copies of the polygon corresponding to the values of the angle coordinate for a fixed value of the angle integral (*i.e.,* an orbit of the finite group Λ) are glued together according

to the unfolding of the trajectories. The result is a compact topological surface with an obvious smooth structure outside of finitely many points corresponding to the vertices of P. In fact, the surface carries not only a smooth structure, but a complex analytic structure, and is thus a Riemann surface with a linear flow, which has singularities at the vertices. The resulting surface is the same for all regular values of the angle integral, so there is, in fact, a single Riemann surface with a one-parameter family of linear flows, which is generated by vector fields obtained from each other by rotations. This is easily identified with a *quadratic differential* on the Riemann surface. Notice that in the simple integrable case the singularities of the quadratic differential and hence of the corresponding linear flows are, in fact, removable. In the general case, the linear flows may be slowed down at their singularities to produce area-preserving smooth flows, which were discussed in the previous section.

Not all quadratic differentials appear from the billiard construction because those that do have a certain symmetry and are thus rather special. However, any quadratic differential on a Riemann surface generates a one-parameter family of line fields and, subject to an orientability condition, those line fields generate a one-parameter family of flows with the same behavior as described above for the billiards.

For an interesting class of billiards, called *almost integrable* billiards, which is intermediate to general rational billiards and integrable ones, see [**S-MT**, Section 1.5], [**Gk1**].

d. Prevalence of minimality and unique ergodicity in rational billiards. As we pointed out, a rational billiard defines a one-parameter family of interval exchange transformations. Recall that both minimality and unique ergodicity are typical properties of interval exchange transformations, albeit in a different way: minimality holds outside of a countable union of finite codimension submanifolds in the parameter space and unique ergodicity holds on a somewhat smaller set of full measure. One may expect the one-parameter families appearing in rational billiards to be typical and hence to intersect the nonminimal set in a countable subset and the set on nonuniquely ergodic interval exchanges in a set of Lebesgue measure zero. This is indeed the case [**S-MT, KZ, KMS**]. While the result for minimality is a fairly simple corollary of the no-saddle-connections criterion (Corollary 8.4.4), the original proof of prevalence of unique ergodicity uses the powerful machinery of Teichmüller theory [**KMS**]. A more elementary proof was later found by Boshernitzan [**Bs**].

The conclusion is thus that any rational billiard is (quasi)-minimal for all but countably many level surfaces of the angle integral and is uniquely ergodic (modulo a convention about the orbits hitting vertices) for a set of values of the angle integral that is of full Lebesgue measure.

e. Topologically transitive and ergodic irrational billiards. As the denominators of the angles of a rational billiard grow, each level surface of the angle integral becomes more and more dense and more and more uniformly distributed in the phase space M of the billiard flow. This simple observation, combined with the prevalence of minimality and ergodicity of most level surfaces for rational billiards, makes it possible to apply a rather general approximation construction to obtain irrational billiards that are topologically transitive and ergodic with respect to the invariant phase volume in the *whole* phase space

[**KMS**]. In fact, the set of such billiards is residual with respect to the natural parametrization of these spaces. These ergodic billiards are counterparts of interval exchange transformation with extreme Liouvillian behavior, an exceptionally good simultaneous approximation of the vector of lengths defining the interval exchange. Unfortunately, these billiards may not be typical for irrational behavior. It is not known whether these or any irrational billiards are mixing.

Nothing is known about the possible nature of a singular invariant measure in general polygonal billiards.

f. Subexponential behavior in polygonal billiards. There are various characteristics of the global orbit complexity for an arbitrary polygonal billiard. These characteristics can be divided into topological ones, which deal with the total complexity, and those dealing with behavior with respect to Lebesgue measure.

Topological characteristics include asymptotic growth of the number of *generalized diagonals*, *i.e.*, orbits connecting vertices, growth of the number of different codes, topological a-entropy for the corresponding symbolic systems *etc.*

The only known general results in this direction assert subexponentiality of all these asymptotics without any specific estimates on the growth rate [**K8**]. This implies in particular that the topological entropy of the symbolic system associated to a polygonal billiard is zero and that the entropy of the billiard map and the billiard flow with respect to any invariant measure is also equal to zero.

On the other hand, the growth characteristics with respect to Lebesgue measure are known to be polynomial. More specifically, the natural coding of the billiard map produces a partition, which is not a topological generator (because the coding does not distinguish periodic orbits within a parallel family) but is a generator with respect to Lebesgue measure as well as any other nonatomic ergodic invariant measure [**K8**]. Then the entropy of the corresponding iterated partition with respect to Lebesgue measure grows like $O(\log n)$. This, of course, implies finiteness of the power metric entropy (Section 3.7l).

Furthermore, one can extend the calculation of the growth of generalized diagonals by fixing any two points p and q inside the billiard table or on its boundary and by counting the asymptotic growth of the number of orbit segments connecting these points. No estimate better than subexponential is available for any fixed pair of points, but if one averages the number of segments of length at most T over all p and q then the growth of the average is quadratic in T. Notice that for rational billiards the number of generalized diagonals grows quadratically with multipcicative bounds that depend on the arithmetic of the angles and thus cannot be extended to other billiards [**S-MT**].

g. Geodesic flows on locally flat surfaces. A class of systems that is close to polygonal billiards and can be analyzed by the same methods with similar degree of success, consists of geodesic flows on compact surfaces with a Riemannian metric that is flat except for a finite number of singular points. The Riemann surfaces that appear in the construction associated with rational billiards are of that kind (Section 8.5c). The locally flat Riemannian metric in question is obtained from flat metrics inside the copies of the billiard table making up the surface. These metrics are glued seamlessly across the sides but produce singularities around the vertices making the total angle around each vertex a multiple of 2π. The latter fact is due to the rationality of the billiard.

A more general construction includes a simplicial decomposition of a surface with a Euclidean metric in each triangle. As in the case of billiards, the orbits of the geodesic flow are defined as long as they do not hit singular points. If all angles of all triangles are rational multiples of π, the conclusions of the theory of rational billiards extend to the geodesic flow on such a surface.

See [**S-MT**] for a discussion of certain typical properties of flows on flat surfaces which are not available for billiards.

h. Billiards in polyhedra. A natural generalization of the constructions of billiards and locally flat geodesic flows is to higher dimension. In particular, consider a polyhedron P in the m-dimensional euclidean space. The *billiard flow* is defined on the space of tangent vectors with footpoints inside P or on the boundary and pointing inside. Similarly to the two-dimensional case the flow is defined as long as an orbit does not hit a face of codimension greater than one. The *billiard map* is again defined as the first return map on the boundary.

The billiard flow and billiard map in a polyhedron are parabolic, the latter having locally $m-1$ Jordan blocks of size two. Thus the local orbit growth is linear. The vanishing of topological entropy and other subexponentiality results from Section 8.5f extend to this case [**GkH**]. One can also define and fully analyze integrable polygonal billiards as those for which the "table" P tiles the space. However, there is no good counterpart to the notion of a rational polygonal billiard. While the latter form dense subsets in the natural spaces of polygonal billiards, polyhedral billiards that possess enough first integrals are exceedingly rare. Symmetry is not of sufficient help. For example, even the dynamics of the billiard inside the regular tetrahedron is not well understood!

Bibliography

SURVEYS IN THIS AND THE COMPANION VOLUME

[S-BK] Victor Bangert, Anatole Katok: *Variational methods II: Twist maps, Lagrangian systems, and closed geodesics*

[S-BKP] Luís Barreira, Anatole Katok, Yakov Pesin: *Introduction to smooth ergodic theory and nonuniformly hyperbolic dynamics*

[S-B] Vitaly Bergelson: *Ergodic theorems and combinatorial ergodic theory*

[S-Bu] Keith Burns: *Partially hyperbolic dynamical systems*

[S-C] Nikolai Chernov: *Invariant measures for hyperbolic dynamical systems*

[S-FK] Renato Feres, Anatole Katok: *Ergodic theory and dynamics of G-spaces*

[S-FM] John Franks, Michał Misiurewicz: *Topological methods in dynamics*

[S-F] Alexander Furman: *Random dynamics*

[S-H] Boris Hasselblatt: *Hyperbolic dynamical systems*

[S-HZ] Helmut Hofer, Eduard Zehnder: *Symplectic methods*

[S-JS] Michael Jakobson, Gregorz Świątek: *One-dimensional maps*

[S-KT] Anatole Katok, Jean-Paul Thouvenot: *Ergodic theory: Spectral theory and combinatorial constructions*

[S-KSS] Dmitry Kleinbock, Nimish Shah, Alexander Starkov: *Homogeneous dynamics*

[S-K] Gerhard Knieper: *Hyperbolic dynamics and Riemannian geometry*

[S-LL] Mark Levi, Rafael de la Llave: *Classical mechanics and KAM theory*

[S-LS] Douglas Lind, Klaus Schmidt: *Symbolic Dynamics and automorphisms of compact groups*

[S-MT] Howard Masur, Serge Tabachnikov: *Dynamics, Teichmüller theory and billiards*

[S-P] Mark Pollicott: *Distribution of periodic orbits and zeta-functions*

[S-R] Paul Rabinowitz: *Variational methods for Hamiltonian systems*

[S-T] Jean-Paul Thouvenot: *Ergodic theory: Entropy, isomorphism and Kakutani equivalence*

[S-W] Maciej Wojtkowski: *Nonuniformly hyperbolic systems: Applications*

MAJOR SOURCES

[AM] Ralph Abraham, Jerrold Marsden: *Foundations of mechanics. Second edition, revised and enlarged*, with the assistance of Tudor Raţiu and Richard Cushman. Benjamin/Cummings Publishing Co., Reading, MA, 1978

[Ax] Vladimir M. Alekseev: *Quasirandom dynamical systems. I. Quasirandom diffeomorphisms*, Mat. Sb. (N.S.) **76 (118)** (1968), 72–134; *Invariant Markov subsets of diffeomorphisms*, Uspehi Mat. Nauk **23** (1968), no. 2 (140) 209–210

[ALM] Lluís Alsedà, Jaume Llibre, Michał Misiurewicz: *Combinatorial dynamics and entropy in dimension one*, Advanced Series in Nonlinear Dynamics, 5. World Scientific Publishing Co., Inc., River Edge, NJ, 1993

[A] Dmitry V. Anosov: *Geodesic flows on Riemannian manifolds with negative curvature*, Proceedings of the Steklov Institute of Mathematics **90** American Mathematical Society, Providence, RI, 1967

[An1] Vladimir Igorevich Arnold: *Mathematical methods of classical mechanics*, Graduate Texts in Mathematics **60**, Springer Verlag Berlin, New York, 1978, 1989

[BKP] Luís Barreira, Anatole Katok, Yakov Pesin: *Nonuniformly hyperbolic dynamical systems*, in preparation

[Bi] Birkhoff, George David: *Dynamical systems*, Colloquium Publications **9**, American Mathematical Society, Providence, RI, 1927

[B1] Rufus Bowen: *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Springer Lecture Notes in Mathematics **470**, 1975

[CG] Lennart Carleson, Theodore W. Gamelin: *Complex dynamics*, Universitext: Tracts in Mathematics. Springer-Verlag, New York, 1993

[CFS] Isaak P. Cornfeld, Sergei V. Fomin, Yakov G. Sinai: *Ergodic theory*, Grundlehren der mathematischen Wissenschaften **245**, Springer-Verlag, 1982

[DS2] *Dynamical systems, vol 3., Bifurcations*, Handbooks in Mathematics, Elsevier

[DS3] *Dynamical systems, vol 4., Applications*, Handbooks in Mathematics, Elsevier

[F1] Hillel Furstenberg: *Recurrence in ergodic theory and combinatorial number theory*, Princeton University Press 1981

[H1] Paul R. Halmos: *Lectures on ergodic theory*, Chelsea Publishing Co., New York, 1960

[H2] Paul R. Halmos: *Measure theory*, D. van Nostrand Co., New York, 1950

[HZ] Helmut Hofer, Eduard Zehnder: *Symplectic invariants and Hamiltonian dynamics*, Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser Verlag, Basel, 1994

[KH] Anatole Katok, Boris Hasselblatt: *Introduction to the Modern Theory of Dynamical Systems*, Encyclopedia of Mathematics and its Applications, 54. Cambridge University Press, 1995

[Kb] Wilhelm Klingenberg: *Riemannian geometry*, de Gruyter Studies in Mathematics, 1. Walter de Gruyter & Co., Berlin-New York, 1982

[Kg] Ulrich Krengel: *Ergodic theorems*, de Gruyter Studies in Mathematics, 6. Walter de Gruyter & Co., Berlin-New York, 1985

[LM] Douglas Lind, Lawrence Marcus: *An introduction to symbolic dynamics and coding*, Cambridge University Press, 1995

[M1] Ricardo Mañé: *Ergodic theory and differentiable dynamics*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 8. Springer-Verlag, Berlin-New York, 1987

[MS] Welington de Melo, Sebastian van Strien: *One-dimensional dynamics*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 25. Springer-Verlag, Berlin-New York, 1993

[Ms1] Jürgen Moser: *Stable and random motions in dynamical systems (with special emphasis on celestial mechanics)*, Hermann Weyl Lectures, the Institute for Advanced Study, Annals of Mathematics Studies **77**, Princeton University Press, 1973

[Na] Mahendra G. Nadkarni: *Spectral theory of dynamical systems*, Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser Verlag, Basel, 1998

[O] Donald Ornstein: *Ergodic theory, randomness, and dynamical systems*, James K. Whittemore Lectures in Mathematics given at Yale University. Yale Mathematical Monographs **5**. Yale University Press, New Haven, CT-London, 1974

[Pa1] William Parry: *Entropy and generators in ergodic theory*, W. A. Benjamin, Inc., New York–Amsterdam, 1969

[Pa2] William Parry: *Topics in ergodic theory*, Cambridge tracts in Mathematics, 75. Cambridge University Press, 1981

[PP] William Parry, Mark Pollicott: *Zeta functions and periodic orbit structure of hyperbolic dynamics*, Astérisque **187–188**, Société Mathématique de France, Paris, 1990

[Pt] Karl Petersen: *Ergodic theory*, Cambridge Studies in Advanced Mathematics, 2. Cambridge University Press, Cambridge-New York, 1983

[R] Clark Robinson: *Dynamical systems. Stability, symbolic dynamics and chaos*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1995

[Ru1] David Ruelle: *Thermodynamic formalism, The mathematical structure of classical equilibrium statistical mechanics*, Encyclopedia of Mathematics and its Applications, 5. Addison-Wesley Publishing Co., Reading, MA, 1978

[Ru2] David Ruelle: *Statistical mechanics: Rigorous results*, W. A. Benjamin, Inc., New York-Amsterdam, 1969

[S] Klaus Schmidt: *Dynamical systems of algebraic origin*, Progress in Mathematics **128**, Birkhäuser Verlag, Basel, Boston, 1995

[W1] Peter Walters: *An introduction to ergodic theory*, Graduate Texts in Mathematics, 79. Springer-Verlag, New-York-Berlin, 1982

[Z] Robert Zimmer: *Ergodic theory and semisimple groups*, Monographs in Mathematics, 81. Birkhäuser Verlag, Basel-Boston, MA, 1984

OTHER SOURCES

[AdKM] Roy L. Adler, Alan G. Konheim, M. Harry McAndrew: *Topological entropy*, Transactions of the American Mathematical Society **114** (1965) 309–319

[AdM] Roy Adler, Brian Marcus: *Topological entropy and equivalence of dynamical systems*, Memoirs of the American Mathematical Society **20** (1979), no. 219

[AmK] Warren Ambrose, Shizuo Kakutani: *Structure and continuity of measurable flows*, Duke Mathematical Journal **9** (1942) 25–42

[AK] Dmitry V. Anosov, Anatole B. Katok: *New examples in smooth ergodic theory. Ergodic diffeomorphisms*, Trudy Moskovskogo Matematičeskogo Obščestva **23** (1970) 3–36; Transactions of the Moscow Mathematical Society **23** (1970), American Mathematical Society, 1972

[An2] Vladimir Igorevich Arnold: *Proof of a theorem of A. N. Kolmogorov on the preservation of conditionally periodic motions under a small perturbation of the Hamiltonian*, Uspehi Matematičeskih Nauk **18** (1963), no. 5, (113) 13–40

[An3] Vladimir Igorevich Arnold: *Small denominators and problems of stability of motion in classical and celestial mechanics*, Uspehi Matematičeskih Nauk **18** (1963), no. 6, (114) 91–192

[Bm] Werner Ballmann, *Lectures on spaces of nonpositive curvature. With an appendix by Michael Brin*, Birkhäuser Verlag, Basel, 1995

[BBES] Werner Ballmann, Michael Brin, Patrick Eberlein: *Structure of manifolds of nonpositive curvature. I.*, Annals of Mathematics (2) **122** (1985), no. 1, 171–203; Werner Ballmann, Michael Brin, Ralf Spatzier: *Structure of manifolds of nonpositive curvature. II.*, Annals of Mathematics (2) **122** (1985), no. 2, 205–235

[By] Augustin Banyaga: *Contact geometry*, preprint

[Bd] Alan Beardon: *Iterations of rational functions: Complex analytic dynamical systems*, Graduate Texts in Mathematics, 132. Springer-Verlag, New York, 1991

[Bl] Grigoriĭ R. Belitskiĭ: *Equivalence and normal forms of germs of smooth mappings*, Russian Mathematical Surveys **31** (1978), no. 1, 107–177

[BF] Alexandra Bellow, Hillel Furstenberg: *An application of number theory to ergodic theory and the construction of uniquely ergodic models*, Israel Journal of Mathematics **33** (1979), no. 3–4, 231–240

[BY] Michael Benedicks, Lai-Sang Young: *Sinai–Bowen–Ruelle measures for certain Hénon maps*, Inventiones Mathematicae **112** (1993), no. 3, 541–576

[Bi] Patrick Billingsley: *Ergodic theory and information*, John Wiley & Sons, Inc., New York-London-Sydney, 1965; Robert E. Krieger Publishing Co., Huntington, NY, 1978

[Blu] Frank Blume: *Minimal rates of entropy convergence for completely ergodic systems*, Israel Journal of Mathematics **108** (1998) 1–12; *Possible rates of entropy convergence*, Ergodic Theory and Dynamical Systems **17** (1997), no. 1, 45–70

[B2] Rufus Bowen: *Entropy for group endomorphisms and homogeneous spaces*, Transactions of the American Mathematical Society **153** (1971) 401–414

[B3] Rufus Bowen: *Some systems with unique equilibrium states*, Mathematical Systems Theory **8** (1975) no. 3, 193–202

[Bs] Michael Boshernitzan: *A condition for minimal interval exchange maps to be uniquely ergodic*, Duke Mathematics Journal **52** (1985), no. 3, 723–752

[BP] Michael I. Brin, Yakov B. Pesin: *Partially hyperbolic dynamical systems*, Uspehi Matematičeskih Nauk **28** (1973), no. 3 (171), 169–170; Izvestija Akademii Nauk SSSR. Seriya Matematicheskaya **38** (1974) 170–212

[Br] Alexander D. Brjuno: *The analytical form of differential equations*, Transactions of the Moscow Mathematical Society **25** (1971) 131–288; **26** (1972) 199–238

[Bn] Kenneth S. Brown: *Cohomology of groups*, Graduate Texts in Mathematics **87**, Springer Verlag New York, Berlin, 1982

[BK] Dmitry Burago, Bruce Kleiner: *Separated nets in Euclidean space and Jacobians of bi-Lipschitz maps*, Geometric and Functional Analysis **8** (1998), no. 2, 173–182

[CO] Rafael V. Chacon, Donald Ornstein: *A general ergodic theorem*, Illinois Journal of Mathematics **4** (1960) 153–160

[C] Nikolai Chernov: *Statistical properties of piecewise smooth hyperbolic systems*, Discrete and Continuous Dynamical Systems **5**, no. 2, 425–448

[Cn] Alain Connes: *Une classification des facteurs de type III*, Annales Scientifiques de l'École Normale Supérieure (4) **6** (1973), 133–252

[CFW] Alain Connes, Jack Feldman, Benjamin Weiss: *An amenable equivalence relation is generated by a single transformation*, Ergodic Theory and Dynamical Systems **1** (1981), 431–450

[D] Alan Dankner: *On Smale's Axiom A dynamical systems*, Annals of Mathematics (2) **107** (1978), no. 3, 517–553

[Do] Dmitry Dolgopyat: *On mixing properties of compact group extensions of hyperbolic systems*, preprint

[Dy] Henry Abel Dye: *On groups of measure preserving transformations, I*, American Journal of Mathematics **81** (1959), 119–159; *II*, American Journal of Mathematics **85** (1963), 551–576

[El] Robert Ellis: *Lectures on topological dynamics* Benjamin, N.Y, 1969

[FH] Albert Fathi, Michael Robert Herman: *Existence de difféomorphismes minimaux*, Dynamical Systems, Vol. 1—Warsaw, Astérisque bf 49, Société Mathematique de France, Paris, 1977, 37–59

[FLP] Albert Fathi, Francois Laudenbach, Valentin Poenaru: *Travaux de Thurston sur les surfaces*, Séminaire Orsay. Astérisque **66–67**, Société Mathématique de France, Paris, 1979

[FM] Jacob Feldman, Calvin C. Moore: *Ergodic equivalence, cohomology, and von Neumann algebras, I, II*, Transactions of the American Mathematical Society **234** (1977), no. 2, 289–324, 325–359

[Fe1] Sebastian Ferenczi: *Measure-theoretic complexity of ergodic systems*, Israel Journal of Mathematics **100**, (1997) 189–207

[Fe2] Sebastian Ferenczi: *Complexity of sequences and dynamical systems*, Combinatorics and number theory (Tiruchirapalli, 1996). Discrete Mathematics **206** (1999), no. 1–3, 145–154

[FFo] Livio Flaminio, Giovanni Forni: *On the cohomological equation for horocycle flows*, in preparation

[Fo] Giovanni Forni: *Solutions of the cohomological equation for area-preserving flows on compact surfaces of higher genus*, Annals of Mathematics (2) **146** (1997), no. 2, 295–344

[FrW] John Franks, Robert Williams: *Anomalous Anosov flows*, in: Global theory of dynamical systems, Zbigniew Nitecki and Clark Robinson, eds., Lecture Notes in Mathematics **819**, Springer Verlag Berlin, New York, 1980, 158–174

[Fr] David Fried: *Rationality for isolated expansive sets*, Adv. in Math. **65** (1987), 35–38

[Fm] Alexander Furman: *Orbit equivalence rigidity*, Annals of Mathematics (2) **150** (1999), no. 3, 1083–1108

[F2] Hillel Furstenberg: *The structure of distal flows*, American Journal of Mathematics **85** (1963), 477–515

[F3] Hillel Furstenberg: *The unique ergodicity of the horocycle flow*, Springer Lecture Notes in Mathematics **318** (1973), 95–115

[F4] Harry Furstenberg: *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, Journal d'Analyse Mathématique **31** (1977), 24–256; Hillel Furstenberg, Yitzhak Katznelson, Donald Ornstein: *The ergodic theoretical proof of Szemerédi's theorem*, Bulletin of the American Mathematical Society (N.S.) **7** (1982), no. 3, 527–552

[FW] Hillel Furstenberg, Benjamin Weiss: *The finite multipliers of infinite ergodic transformations*, The structure of attractors in dynamical systems, Proc. Conf. North Dakota State University, Fargo, ND, 1977, 127–132, Springer Lecture Notes in Mathematics **668** (1978)

[Gn] Chr. Genecand: *Transversal homoclinic orbits near fixed points of area-preserving diffeomorphisms of the plane*, Dynamics reported (N.S.) **2** 1–30, Springer-Verlag, 1993

[GK] Marlies Gerber, Anatole B. Katok: *Smooth models of Thurston's pseudo-Anosov maps*, Annales Scientifiques de l'École Normale Supérieure (4) **15** (1982), no. 1, 173–204

[GPS] Matthew Grayson, Charles Pugh, Michael Shub, *Stably ergodic diffeomorphisms*, Annals of Mathematics(2) **140** (1994), no. 2, 295–329

[Gl] Frederick P. Greenleaf: *Invariant means on topological groups and their applications*, Van Nostrand Mathematical Studies, **16**. Van Nostrand Reinhold Co., New York-Toronto-London, 1969

[GuK] Roland Gunesch, Anatole Katok: *Construction of weakly mixing diffeomorphisms preserving measurable Riemannian metric and smooth measure*, Discrete and Continuous Dynamical Systems **6** (2000), no. 1, The Millenium Issue, 61–88

[Gt] Carlos Gutierrez: *A counterexample to a C^2 closing lemma*, Ergodic Theory and Dynamical Systems **7** (1987) no. 4, 509–530

[Gk1] Eugene Gutkin: *Billiards on almost integrable polyhedral surfaces*, Ergodic Theory and Dynamical Systems **4** (1984), no. 4, 569–584

[Gk2] Eugene Gutkin: *Billiards in polygons*, Physica D **19** (1986), no. 3, 311–333

[GkH] Eugene Gutkin, Nicolai T. A. Haydn: *Topological entropy of generalized polygon exchanges*, Bulletin of the American Mathematical Society (N.S.) **32** (1995), no. 1, 50–56; *Topological entropy of polygon exchange transformations and polygonal billiards*, Ergodic Theory and Dynamical Systems **17** (1997), no. 4, 849–867

[GkK] Eugene Gutkin, Anatole Katok: *Caustics in inner and outer billiards*, Communications in Mathematical Physics **173** (1995) 101-133

[Hy] Shuhei Hayashi: *Connecting invariant manifolds and the solution of the C^1 stability and Ω-stability conjectures for flows*, Ann. of Math. (2) **145** (1997), no. 1, 81–137; *The stability and Ω-stability conjectures and a lemma connecting stable and unstable manifolds*, Comptes Rendus de l'Académie des Sciences - Séries 1 - Mathematics **322** (1996), no. 2, 159–163; *The C^1 connecting lemma and the stability conjecture for flows*, Sūrikaisekikenkyūsho Kōkyūroku No. 938 (1996), 114–122

[Hm1] Michael Robert Herman: *Sur la conjugation différentiable des difféomorphismes du cercle à des rotations*, Publications Mathématiques de l'Institut des Hautes Études Scientifiques **49** (1979) 5–233

[Hm2] Michael Robert Herman: *Exemple de flots Hamiltoniens dont aucune perturbation en topologie C^∞ n'a d'orbites periodiques sur un ouvert de surfaces d'energies*, Comptes Rendus de l'Académie des Sciences - Séries 1 - Mathematics **312** (1991), no. 13, 989–994

[Hm3] Michael Robert Herman: *Differentiabilité optimale et contre-exemples à la fermeture en topologie C^∞ des orbites recurrentes de flots Hamiltoniens*, Comptes Rendus de l'Académie des Sciences - Séries 1 - Mathematics **313** (1991), no. 1, 49–51

[Hm4] Michael Robert Herman: *Construction d'un difféomorphisme minimal d'entropie topologique non nulle*, Ergodic Theory and Dynamical Systems **1** (1981), no. 1, 65–76

[HPS] Morris W. Hirsch, Charles C. Pugh, Michael Shub: *Invariant manifolds*, Springer Lecture Notes in Mathematics **583**, 1977

[Ho] Eberhard Hopf: *Ergodentheorie*, Ergebnisse der Mathematik und ihrer Grenzgebiete 5. Band, Springer Verlag, Berlin, New York, 1937

[Hos] Bernard Host: *Mixing of all orders and pairwise independent joining of systems with singular spectrum*, Israel Journal of Mathematics **76** (1991), no. 3, 289–298

[HY] Hu Yi Hu, Lai Sang Young: *Nonexistence of SBR measures for some diffeomorphisms that are "almost Anosov"*, Ergodic Theory and Dynamical Systems **15** (1995), no. 1, 67–76

[HK] Steven Hurder, Anatole Katok: *Differentiability, rigidity and Godbillon–Vey classes for Anosov flows*, Publications Mathématiques de l'Institut des Hautes Études Scientifiques **72** (1990) 5–61

[J] Robert I. Jewett: *The prevalence of uniquely ergodic systems*, J. Math. Mech. **19** (1969/1970) 717-729

[Ji] Boju Jiang: *Lectures on Nielsen fixed point theory*, Contemporary Mathematics **14**, American Mathematical Society, Providence, RI, 1983

[Kl] Stephen Arthur Kalikow: *Twofold mixing implies threefold mixing for rank one transformations*, Ergodic Theory and Dynamical Systems **4** (1984), no. 2, 237–259

[K1] Anatole B. Katok: *Monotone equivalence in ergodic theory*, Izvestija Akademii Nauk SSSR. Seriya Matematicheskaya **41** (1977), no. 1, 104–157, 231

[K2] Anatole B. Katok: *The special representation theorem for multi-dimensional group actions*, Dynamical systems, Vol. I—Warsaw, Astérisque **49**, Société Mathematique de France, Paris, 1977, 117–140

[K3] Anatole B. Katok: *Bernoulli diffeomorphisms on surfaces*, Annals of Mathematics (2) **110** (1979), no. 3, 529–547

[K4] Anatole B. Katok: *Lyapunov exponents, entropy and periodic orbits for diffeomorphisms*, Publications Mathématiques de l'Institut des Hautes Études Scientifiques **51** (1980) 137–173

[K5] Anatole B. Katok: *Interval exchange transformations and some special flows are not mixing*, Israel Journal of Mathematics **35** (1980), no. 4, 301–310

[K6] Anatole B. Katok in collaboration with E. A. Robinson: *Constructions in ergodic theory*, manuscript, first part published as *Combinatorial constructions in Ergodic Theory: I. Approximation and Genericity*, Proceedings of the Steklov Institute, "Anosov Seminars" Volume, 2001;

[K7] Anatole B. Katok in collaboration with E. A. Robinson: *Constructions in ergodic theory*, manuscript, second part published as *Cocycles, cohomology and combinatorial constructions in Ergodic Theory*,

Smooth Ergodic Theory. American Mathematical Society Proceedings of Symposia in Pure Mathematics, Summer Research Institute Seattle, WA, 1999, to appear 2001

[K8] Anatole B. Katok: *The growth rate for the number of singular and periodic orbits for a polygonal billiard*, Communications in Mathematical Physics **111** (1987), no. 1, 151–160

[KSp] Anatole B. Katok, Ralf Jürgen Spatzier: *First cohomology of Anosov actions of higher rank abelian grouos and applications to rigidity*, Publications Mathématiques de l'Institut des Hautes Études Scientifiques **79** (1994) 131–156

[KS] Anatole B. Katok, Anatole M. Stepin: *Approximations in ergodic theory*, Uspehi Matematičeskih Nauk **22** (1967) no. 5 (137), 81–106

[KT] Anatole Katok, Jean-Paul Thouvenot: *Slow entropy type invariants and smooth realization of commuting measure preserving transformations*, Ann. Inst. Henri Poincaré Prob. Statist. **33** (1997), no. 3, 323–338

[KZ] Anatole Katok, A. N. Zemljakov: *Topological transitivity of billiards in polygons*, Mat. Zametki **18** (1975), no. 2, 291–300; Math. Notes **18** (1975), no. 1–2, 760–764 (1976)

[KO1] Yitzhak Katznelson, Donald Ornstein: *The differentiability of conjugation of certain diffeomorphisms of the circle*, Ergodic Theory and Dynamical Systems **9** (1989), 643–680

[KO2] Yitzhak Katznelson, Donald Ornstein: *The absolute continuity of the conjugation of certain diffeomorphisms of the circle*, Ergodic Theory and Dynamical Systems **9** (1989) 681–691

[KW] Yitzhak Katznelson, Benjamin Weiss: *The classification of nonsingular actions, revisited*, Ergodic Theory and Dynamical Systems **11** (1991), no. 2, 333–348

[Kl] Al Kelley: *The stable, center-stable, center, center-unstable, unstable manifolds*, Journal of Differential Equations **3** (1967) 546–570

[KMS] Stephen Kerckhoff, Howard Masur, John Smillie: *Ergodicity of billiard flows and quadratic differentials*, Annals of Mathematics **124** (1986) 293–311

[KhS] Konstantin M. Khanin, Yakov G. Sinai: *Mixing of some classes of special flows over rotations of the circle*, Funktional. Anal. i Prilozhen. **26** (1992), no. 3, 1–21 (Functional Analysis and its Applications **26** (1992), no. 3, 155-169)

[Kc1] A. V. Kočergin: *Nondegenerate saddles and the absence of mixing*, Mat. Zametki **19** (1976), no. 3, 453–468; Math. Notes **19** 1976), no. 3, 277–286

[Kc2] A. V. Kočergin: *Mixing in special flows over an interval exchange and in smooth flows on surfaces*, Mat. Sbornik (N.S.) **96 (138** (1975) 471–502, 504; Math. USSR-Sbornik **25** (1975), no. 3, 441–469

[Kr1] Wolfgang Krieger: *On ergodic flows and the isomorphism of factors*, Mathematische Annalen **223** (1976), no. 1, 19–70

[Kr2] Wolfgang Krieger: *On entropy and generators of measure preserving transformations*, Transactions of the American Mathematical Society **119** (1970), no. 2, 98–119

[Kr3] Wolfgang Krieger: *On unique ergodicity*, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (University of California, Berkeley 1970/1971), Vol. II: Probability theory, University of California Press, 1972, 327–346

[KB] Nicolas Kryloff, Nicolas Bogoliouboff: *La théorie générale de la mesure dans son application àl'étude des systèmes dynamiques de la mécanique non linéaire*, Annals of Math. **38** (1937) no. 1, 65–113

[Ks] Anatoly G. Kushnirenko: *An upperbound for the entropy of a classical dynamical system*, Doklady Akademii Nauk SSSR **161** (1965) 37–38

[La] Serge Lang: *Introduction to Diophantine approximations*, Addison-Wesley Publishing Co., Reading, MA-London-Don Mills, Ont., 1966; Second edition, Springer-Verlag, New York, 1995

[Lz] Vladimir F. Lazutkin: *The existence of caustics for a billiard problem in a convex domain*, Math. USSR, Isvestija **7** (1973), 185–214

[Ld] François Ledrappier: *Propriétés ergodiques des mesures de Sinaï*, Publications Mathématiques de l'Institut des Hautes Études Scientifiques **59** (1984) 163–188

[Lv] Gilbert Levitt: *Flots topologiquement transitifs sur les surfaces compactes sans bord: contrexemples à une conjecture de Katok*, Ergodic Theory and Dynamical Systems **3** (1983), no. 2, 241–249

[L] Jorge Lewowicz: *Expansive homeomorphisms of surfaces*, Bol. Soc. Brasil. Mat. (N.S.) **20** (1989), no. 1, 113–133

[Li] Elon Lindenstrauss: *Pointwise theorems for amenable groups*, Electronic Research Announcements, American Mathematical Society **5** (1999), 82–90

[M2] Ricardo Mañé: *An ergodic closing lemma*, Annals of Mathematics (2) **116** (1982), no. 3, 503–540

[M3] Ricardo Mañé: *A proof of the C^1 stability conjecture*, Publications Mathématiques de l'Institut des Hautes Études Scientifiques **66** (1987), 161–210

[Mt] John Mather: *Action minimizing invariant measures for positive definite Lagrangian systems*, Math. Zeitschrift **207** (1991), no. 2, 169–207

[MN] Joseph Mathew, Mahendra Nadkarni: *A measure preserving transformation whose spectrum has Lebesgue component of multiplicity two*, Bulletin of the London Mathematical Society **16** (1984), no. 4, 402–406

[Ms1] Howard Masur: *Interval exchange transformations and measured foliations*, Annals of Math. **115** (1982), 169–200

[Ms2] Howard Masur: *Hausdorff dimension of the set of nonergodic foliations of a quadratic differential*, Duke Mathematical Journal **66** (1992), no. 3, 387–442

[MM] Curtis McMullen: *Lipschitz maps and nets in Euclidean space*, Geometric and Functional Analysis **8** (1998), no. 2, 304–314

[Mi] Michał Misiurewicz: *A short proof of the variational principle for a $\mathbb{Z}_+^n$ action on a compact space*, Astérisque **40**, Société Mathématique de France, Paris, 1976, 147–157

[Mi2] Michał Misiurewicz, *Topological conditional entropy,* Studia Math. **55** (1976), 175–200, **66** (1992), no. 3, 387–442

[MNTU] Shunsuke Morosawa, Yasuihiro Nishimura, Masachiko Taniguchi, Tetsuo Ueda: *Holomorphic dynamics*, Cambridge Studies in Advanced Mathematics, Cambridge University Press, 2000

[Ms2] Jürgen Moser: *A rapidly convergent iteration method and nonlinear partial differential equations. I.*, Ann. Scuola Norm. Sup. Pisa (3) **20** (1966) 265–315; *A rapidly convergent iteration method and nonlinear differential equations. II.*, Ann. Scuola Norm. Sup. Pisa (3) **20** (1966) 499–535

[Ms3] Jürgen Moser: *Dynamical systems—past and present*, Proceedings of the International Congress of Mathematicians, Vol. I (Berlin, 1998), Documenta Math. **1998**, Extra Vol. I, 381–402 (electronic)

[Mz] Shahar Mozes: *Mixing of all orders of Lie group actions*, Inventiones Mathematicae **107** (1992), no. 2, 235–241; *Erratum*, Inventiones Mathematicae **119** (1995), no. 2, 399;

[N1] Sheldon Newhouse: *Continuity properties of entropy*, Annals of Mathematics (2) **129** (1989), no. 2, 215–235; *Corrections to "Continuity properties of entropy"*, Annals of Mathematics (2) **131** (1990), no. 2, 409–410

[N2] Sheldon Newhouse: *Entropy and volume*, Ergodic Theory and Dynamical Systems **8*** (Conley Memorial Issue, 1988) 283–300

[NP] Sheldon Newhouse, Jacob Palis: *Hyperbolic nonwandering sets on two-dimensional manifolds*, Dynamical systems (Proc. Sympos. Univ. of Bahia, Salvador, 1971), pp. 293–301, Academic Press, NY, 1973

[NZ] Igor Nikolaev, Evgeny Zhuzhoma: *Flows on 2-dimensional manifolds. An overview*, Springer Lecture Notes in Mathematics, **1705** (1999)

[ORW] Donald Ornstein, Daniel Rudolph, Benjamin Weiss: *Equivalence of measure preserving transformations*, Memoirs of the American Mathematical Society **37** (1982), no. 262

[OS] Donald Ornstein, Meir Smorodinsky: *Ergodic flows of positive entropy can be time changed to become K-flows*, Israel Journal of Mathematics **26** (1977), no. 1, 75–83

[OW] Donald Ornstein, Benjamin Weiss: *Entropy and isomorphism theorems for actions of amenable groups*, Journal Analyse Math. **48** (1987) 1–141; *Ergodic theory of amenable group actions. I. The Rohlin lemma*, Bulletin of the American Mathematical Society (N.S.) **2** (1980), no. 1, 161–164

[Os] Valeriĭ I. Oseledets: *A multiplicative ergodic theorem. Liapunov characteristic numbers for dynamical systems* Trudy Moskovskogo Matematičeskogo Obščestva **19** (1968) 179–210; Transactions of the Moscow Mathematical Society **19** (1968) 197–221

[Ox] John Oxtoby: *Ergodic sets*, Bulletin of the American Mathematical Society **58** (1952) 116–136

[PaS] Jacob Palis, Steven Smale: *Structural stability theorems*, in: Global Analysis, Proceedings of Symposia in Pure Mathematics, **14**, American Mathematical Society, Providence, RI 1970, 223–231

[P] Yakov Pesin: *Dimension theory in dynamical systems. Contemporary view and applications*, Chicago Lectures in Mathematics, The University of Chicago Press, Chicago and London, 1997

[PPi] Yakov Pesin, B. S. Pitskel: *Topological pressure and the variational principle for noncompact sets*, Funktsionalnyĭ Analiz i ego Prilozheniya **18** (1984) no. 4, 50–63

[Ph] Robert R. Phelps: *Lectures on Choquet's Theorem*, D. Van Nostrand Co., Inc., Princeton-Toronto-London, 1966.

[PiS] B. S. Pitskel, Anatole M. Stepin: *The property of the entropy equidistribution of commutative groups fo metric automorphisms*, Doklady Akademii Nauk SSSR **198** (1971), 1021–1024

[Pc] Henri Jules Poincaré: *Les méthodes nouvelles de la mécanique celeste*, Gauthier-Villars, Paris, 1892 vol. I, 1893 vol. II, 1899 vol. III; *New methods of celestial mechanics*, National Aeronautics and Space Administration, Clearinghouse for Federal Scientific and Technical Information, 1967; History of Modern Physics and Astronomy **13**, American Institute of Physics, 1993

[P] Mark Pollicott: *Lectures on ergodic theory and Pesin theory on compact manifolds*, London Mathematical Society Lecture Note Series **180**, Cambridge University Press, 1993

[PY] Mark Pollicott, Michiko Yuri: *Dynamical systems and ergodic theory*, London Mathematical Society Student Texts **40**, Cambridge University Press, 1998

[Pu1] Charles Pugh: *The closing lemma; An improved closing lemma and a general density theorem*, American Journal of Mathematics **89** (1967), 956–1021

[Pu2] Charles Pugh: *The $C^{1+\alpha}$ hypothesis in Pesin theory*, Publications Mathématiques de l'Institut des Hautes Études Scientifiques **59** (1984) 143–161

[PS] Charles Pugh, Michael Shub: *Stable ergodicity and partial hyperbolicity*, in "International Conference on Dynamical Systems (Montevideo, 1995)", 182–187, Longman, Harlow, 1996; *Stably ergodic dynamical systems and partial hyperbolicity*, Journal of Complexity 13 (1997), no. 1, 125–179

[Rg] Madabusi S. Raghunathan: *A proof of Oseledec's multiplicative ergodic theorem*, Israel Journal of Mathematics **32** (1979), no. 4, 356–362

[Ra1] Marina Ratner: *Rigidity of horocycle flows*, Annals of Mathematics (2) **115** (1982), no. 3, 597–614

[Ra2] Marina Ratner: *Rigidity of time changes for horocycle flows*, Acta Math. **156** (1986), no. 1–2, 1–32

[Ra3] Marina Ratner: *Ragunathan's topological conjecture and distributions of unipotent flows*, Duke Mathematical Journal **63** (1991), no. 1, 235–280

[Ra4] Marina Ratner: *On Ragunathan's measure conjecture*, Annals of Mathematics (2) **134** (1991), no. 3, 545–607

[R] Joel Robbin: *A structural stability theorem*, Annals of Mathematics (2) **94** (1971), 447–493

[Ro] Clark Robinson: *Structural stability of C^1 diffeomorphisms*, Journal of Differential Equations **22** (1976), no. 1, 28–73

[RY] Clark Robinson, Lai Sang Young: *Nonabsolutely continuous foliations for an Anosov diffeomorphism*, Inventiones Math. **61** (1980), no. 2, 159–176

[Rk1] Vladimir Abramovich Rokhlin: *On the fundamental ideas of measure theory*, Translations of the American Mathematical Society (1) **10** (1962)

[Rk2] Vladimir Abramovich Rokhlin: *Lectures on the entropy theory of transformations with invariant measure*, Uspehi Matematičeskih Nauk **22** (1967), no. 5 (137), 3–56; Russ. Math. Surveys **22** (1967)

[Rm] Helmut Rüssmann: *Kleine Nenner I. Über invariante Kurven diferenzierbarer Abbildungen eines Kreisringes*, Nachr. d. Akad. Wiss. Göttingen Math.-Phys. Klasse II **1970**(1970) 67–105

[Ru3] David Ruelle: *Statistical mechanics on a compact set with $\mathbb{Z}^v$ action satisfying expansiveness and specification*, Transactions of the American Mathematical Society **187** (1973) 237–251

[Ry] V. V. Ryzhikov: *Stochastic intertwinings and multiple mixing of dynamical systems*, Journal of Dynamical and Control Systems **2** (1996), no. 1, 1–19

[St] E. A. Sataev: *The number of invariant measures for flows on orientable surfaces*, Izvestija Akademii Nauk SSSR. Seriya Matematicheskaya **39** (1975), no. 4, 860–878

[Sh] Michael Shub: *Dynamical systems, filtrations and entropy*, Bulletin of the American Mathematical Society **80** (1974) 27–41

[Si] Nessim Sibony: *Dynamique des applicationes rationelles de P^k*, Dynamique et géometrie complexes (Lyon, 1997), ix–xii, 97–185, Panoramas et Synthèses **8**, Société Mathématique de France, Paris, 1999

[Sm] Stephen Smale: *Differentiable dynamical systems*, Bulletin of the American Mathematical Society **73** (1967), 747–817

[Ta] Serge Tabachnikov: *Billiards*, Panoramas et Synthèses **1**, Société Mathématique de France, Paris, 1995

[Tp] Arkady Tempelman: *Ergodic theorems for group actions. Informational and thermodynamical aspects*, Mathematics and its Applications **78**, Kluwer Academic Publishers Group, Dordrecht, 1992

[Va] Veeravalli S. Varadarajan: *Lie Groups, Lie Algebras, and Their Representation Theory*, Prentice-Hall Series in Modern Analysis, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1974; Graduate Texts in Mathematics **102**, Springer-Verlag, New York-Berlin, 1984.

[V1] William Veech: *Periodic points and invariant pseudomeasures for toral endomorphisms*, Ergodic Theory and Dynamical Systems **6** (1986), no. 3, 449–473

[V2] William Veech: *Gauss measures for transformations on the space of interval exchange maps*, Annals of Mathematics **115** (1982), 201–242

[Vs] Anatoly M. Vershik: *A measurable realization of continuous groups of automorphisms of a unitary ring*, Izvestija Akademii Nauk SSSR. Seriya Matematicheskaya **29** (1965) 127–136

[W2] Peter Walters: *A dynamical proof of the multiplicative ergodic theorem*, Transactions of the American Mathematical Society **335** (1993), no. 1, 245–257

[Wk] Amie Wilkinson: *Stable ergodicity of the time-one map of a geodesic flow*, Ergodic Theory and Dynamical Systems **18** (1998), no. 6, 1545–1587

[Wl] Robert Williams: *Classification of one dimensional attractors*, Global Analysis, Proceedings of Symposia in Pure Mathematics **14**, 341–361, American Mathematical Society, Providence, RI, 1970

[X] Zhihong Xia: *Existence of invariant tori for certain nonsymplectic diffeomorphisms*, Hamiltonian dynamical systems (Cincinnati, OH, 1992), 373–385, IMA Vol. Math. Appl., 63, Springer-Verlag, New York, 1995

[Y1] Jean-Christophe Yoccoz: *Conjugaison différentiable des difféomorphismes du cercle dont le nombre de rotation vérifie une condition Diophantienne*, Annales Scientifiques de l'École Normale Supérieure **17** (1984), 333–361

[Y2] Jean-Christophe Yoccoz: *Théorème de Siegel, nombres de Bruno et polynômes quadratiques. Petits diviseurs en dimension 1*, Astérisque **231**, Société Mathématique de France, Paris, 1995, 3–88; *Linéarization des germes de difféomorphismes holomorphes de* $(\mathbb{C}, 0)$, Comptes Rendus de l'Académie des Sciences - Séries 1 - Mathematics **306** (1988), no. 1, 55–58

[Y3] Jean-Christophe Yoccoz: *Introduction to hyperbolic dynamics*, Real and complex dynamical systems (Hillerød 1993), 265–291, NATO Advanced Sciences Institutes Series C: Mathematical and Physical Sciences **464**, Kluwer Academic Publishers Group, Dordrecht, 1995

[Yd] Yosif Yomdin: *Volume growth and entropy*, Israel Journal of Mathematics **57** (1987), 285–300

[Zh] Vadim Zharnitsky: *Invariant tori in the systems of billiard type under the slow periodic variation of the Hamiltonian*, preprint