Introduction to the Ergodic Theory of Chaotic Billiards

Nikolai Chernov Roberto Markarian

Contents

Preface

In the last three decades of the twentieth century, chaotic billiards became one of the most active and popular research areas in statistical mechanics. This started with a seminal paper by Ya. Sinai in 1970 [Si2], where he developed a mathematical apparatus for the study of hyperbolic and ergodic properties for a large class of plane billiards. He also obtained an exact formula for the entropy of billiards. Sinai's theory led to an outburst of papers in mathematics and physics journals devoted to various types of billiards on plane and space of any dimension. The remarkable progress of Sinai's theory culminated in a solution, in some form, of a classical hypothesis by L. Boltzmann (stated back in the 1880's) on the ergodicity of gases of hard balls. The advances in the study of billiards also penetrated nonequilibrium statistical mechanics and some other sciences.

The goal of this book is to introduce the reader to the up-to-date theory of chaotic billiards. It addresses graduate students and young researchers, both physicists and mathematicians. We assume basic knowledge of calculus, elements of Riemannian geometry and some familiarity with probability. For the reader's convenience, the book provides necessary background in measure theory.

This book is a revised, updated and improved version of the monograph by R. M. published in 1993 [Ma5]. It has the same general structure and consists of two major parts.

The first one comprises Chapters I, II and III. It contains a brief exposition of the most important elements of ergodic theory. Chapter I is largely a transcription of the first section of the notes by Ricardo Mañe that he prepared for his course at the ICTP, Trieste in 1988. (He kindly gave one of us, R. M., his permission to include this Chapter in the original version of the book. Regrettably, Mañe passed away in 1995.) Chapter II contains a proof of Ergodic Theorem and is also inspired by the text of

Mañe. Chapter III focuses on smooth hyperbolic dynamics and covers the main results of the Pesin theory.

The second part of the book is formed by Chapter IV and is fully devoted to billiards. It starts with elementary properties of planar billiards, then goes on to multidimensional models, including Lorentz gases and hard ball systems. It covers mean free path formulas and bounds on the number of reflections. Sinai's theory of planar dispersing billiards is outlined. Then other hyperbolic billiards are described in detail. Formulas for the entropy are derived. Overall, the material of Chapter IV constitutes a basic course in billiards, after which one should be able to read main research papers in the area.

R. M. would like to thank Instituto de Matemática Pura e Aplicada (IMPA, Rio de Janeiro), Department of Mathematics of the University of Alabama at Birmingham, and Instituto de Matemáticas (IM-UNAM), sede Cuernavaca, México, where important parts of the book have been prepared. R. M. was partially supported by CSIC (Universidad de la República) and CONICYT (Uruguay). N. Ch. was partially supported by NSF grant DMS-9732728.

N. Chernov
R. Markarian
July 2001

Chapter I

Measure Preserving Maps

I.1 Introduction

Let $U \subset \mathbb{R}^n$ be an open set and $\mathbf{v} : U \to \mathbb{R}^n$ a C^r vector field $(r \geq 1)$. Consider the differential equation

$$\dot{\mathbf{x}} = \mathbf{v}(\mathbf{x}) \tag{I.1.1}$$

Suppose that for every $\mathbf{p} \in U$ there exists a (unique) solution $\mathbf{x} : \mathbb{R} \to U$ of (I.1.1) with initial condition $\mathbf{p}$, i.e. $\mathbf{x}(t)$ satisfies $\dot{\mathbf{x}}(t) = \mathbf{v}(\mathbf{x}(t))$ for all $t \in \mathbb{R}$ and $\mathbf{x}(0) = \mathbf{p}$.

Theorem I.1.1 *If U has finite volume and the divergence of $\mathbf{v}$ is zero, then, for almost every $\mathbf{p} \in U$, the solution $\mathbf{x}(t)$ of* (I.1.1) *with initial condition $\mathbf{p}$, is recurrent, i.e.*

$$\liminf_{t \to +\infty} \operatorname{dist}(\mathbf{x}(t), \mathbf{p}) = 0. \tag{I.1.2}$$

The words "almost every" mean that the set S of points $\mathbf{p} \in U$ for which the property (I.1.2) fails has measure zero. The divergence $\operatorname{div} \mathbf{v}$ of $\mathbf{v}$ is the scalar function defined by

$$\operatorname{div} \mathbf{v}(\mathbf{x}) = \frac{\partial v_1}{\partial x_1}(\mathbf{x}) + \cdots + \frac{\partial v_n}{\partial x_n}(\mathbf{x})$$

The $\operatorname{div}\mathbf{v} \equiv 0$ hypothesis holds in the important case of Hamiltonian vector fields. The latter are vector fields $\mathbf{v} : U \to \mathbb{R}^n$, where $n = 2m$, for which there exists a scalar function $H : U \to \mathbb{R}$ (called the Hamiltonian of $\mathbf{v}$) such that, denoting points in $\mathbb{R}^n$ as $(q_1, \cdots, q_m, p_1, \cdots, p_m)$, satisfies

$$\mathbf{v} = \left(\frac{\partial H}{\partial p_1}, \cdots, \frac{\partial H}{\partial p_m}, \quad -\frac{\partial H}{\partial q_1}, \cdots, -\frac{\partial H}{\partial q_m}\right)$$

Theorem I.1.1, proved by Poincaré in his ”Les Nouvelles Méthodes de la Mécanique Céleste” [Po], can nowadays be regarded as a minor corollary of deeper and more accurate results; but, due to the simplicity of its statement, it still serves as a beautiful introductory example conveying the flavor of ergodic theory. Its proof relies on the fact that, due to our hypothesis on the existence of solutions defined on all of R for every initial condition $\mathbf{p}$, our vector field $\mathbf{v}$ defines a flow of diffeomorphisms $\varphi_t : U \to U$, $t \in \mathbb{R}$. This flow, due to the zero divergence hypothesis, is *volume preserving*. This, together with the crucial assumption on the finiteness of the volume of U (without which the the theorem fails as the vector field in $\mathbb{R}^n$ given by $\mathbf{v} = (1, 0, \ldots, 0)$ shows) is what makes the proof work.

For every $t \in \mathbb{R}$ define a map $\varphi_t : U \to U$ by $\varphi_t(\mathbf{p}) = \mathbf{x}(t)$ where $\mathbf{x} : \mathbb{R} \to U$ is the solution of (I.1.1) with initial condition $\mathbf{x}(0) = \mathbf{p}$. Basic results of ordinary differential equations show that, for every t, the map φ_t is a C^r diffeomorphism, and the family of maps $\varphi_t : U \to U$, $t \in \mathbb{R}$ makes a one-parameter group (a flow), i.e. $\varphi_0 =$identity and $\varphi_{t+s} = \varphi_t \circ \varphi_s$ for all $t, s \in \mathbb{R}$. Moreover, by Liouville's formula

$$\det(\varphi_t'(\mathbf{p})) = \exp\left[\int_0^t \operatorname{div}\mathbf{v}(\varphi_s(\mathbf{p}))\, ds\right]$$

for all $\mathbf{p}$ and t. Hence, due to the assumption $\operatorname{div}\mathbf{v} = 0$, we have, $\det(\varphi_t'(\mathbf{p})) = 1$ for all t and $\mathbf{p}$. Theorem I.1.1 now follows from its “discrete version”:

Theorem I.1.2 *Let $U \subset \mathbb{R}^n$ be an open set with finite volume and let $f : U \to U$ be a C^1 diffeomorphism with $|\det(f'(\mathbf{p}))| = 1$ for all $\mathbf{p} \in U$. Then almost every $\mathbf{p} \in U$ is recurrent, i.e.* $\liminf_{n\to+\infty} \operatorname{dist}(\mathbf{p}, f^n(\mathbf{p})) = 0$.

Theorem I.1.1 follows easily from Theorem I.1.2 applied to $f = \varphi_1$, observing that if $\mathbf{x} : \mathbb{R} \to U$ is the solution of (I.1.1) with initial condition $\mathbf{p}$, then $\mathbf{x}(n) = \varphi_1^n(\mathbf{p})$ for all $n \in \mathbb{Z}$. In turn, Theorem I.1.2, whose proof in a much more general framework will be given in Section I.5, works because the hypothesis on the determinant implies that f is volume preserving, i.e.

$$\mathrm{Vol}\, f(V) = \mathrm{Vol}(V)$$

for every open set $V \subset U$. The surprising conclusion of the Poincaré theorem is obtained by studying the iterates (the dynamics) of the map and exploiting the fact that the map leaves volume invariant. This exemplifies the objective of ergodic theory: to study the dynamics of a map with the help of its invariant measure. The volume, in this context, is an invariant measure. Naturally, ergodic theory relies on measure theory. Therefore we shall begin by reviewing the basic definitions and facts of measure theory.

I.2 Measures

Let X be a set. A family $\mathcal{O}$ of subsets of X is called an *algebra* if

$$\begin{aligned} X &\in \mathcal{O} \\ A &\in \mathcal{O} \Rightarrow A^c \in \mathcal{O} \\ A, B &\in \mathcal{O} \Rightarrow A \cup B \in \mathcal{O} \end{aligned}$$

It easily follows that

$$\begin{aligned} A, B &\in \mathcal{O} \Rightarrow A \cap B = (A^c \cup B^c)^c = \in \mathcal{O} \\ A, B &\in \mathcal{O} \Rightarrow A \setminus B = A \cap B^c \in \mathcal{O} \end{aligned}$$

The family $\mathcal{O}$ is called a $\sigma-$*algebra* if it is an algebra and satisfies

$$A_i \in \mathcal{O},\ i = 1, 2, \dots \quad \Rightarrow \quad \cup_i A_i \in \mathcal{O}.$$

If $\mathcal{O}_0$ is a family of subsets of X, the $\sigma-$algebra generated by $\mathcal{O}_0$ is the smallest $\sigma-$algebra containing $\mathcal{O}_0$: this is the $\sigma-$algebra $\mathcal{O}$ such that $\mathcal{O} \supset \mathcal{O}_0$ and $\mathcal{O} \subset \mathcal{O}_1$ for every $\sigma-$algebra $\mathcal{O}_1$ that contains $\mathcal{O}_0$.

If $\mathcal{O}$ is an algebra of subsets of X, we say that a function $\mu : \mathcal{O} \to [0, \infty]$ is a *measure on O* if $\mu(\emptyset) = 0$ and for every finite or countable collection of disjoint sets $A_i \in \mathcal{O},\ i = 1, 2, \dots,\ \cup_i A_i \in \mathcal{O}$, we have

$$\mu\left(\cup_i A_i\right) = \sum_i \mu(A_i)$$

A measure is $\sigma-$*finite* if X can be decomposed into a countable union of sets with finite measure.

Can we always extend a measure on an algebra to a bigger $\sigma-$algebra? The answer is contained in the following Extension Theorem whose proof can be found in [Ha] or [F].

Theorem I.2.1 (Hahn-Kolmogorov) *Let* $\mathcal{O}_0$ *be an algebra of sets of* X *and* ν *a measure defined on* $\mathcal{O}_0$ *. Then, there exist a* $\sigma-$*algebra* $\mathcal{O}$*, and a measure* μ *on* $\mathcal{O}$ *such that* $\mathcal{O}_0 \subset \mathcal{O}$ *and* $\nu(A) = \mu(A)$ *for every* $A \in \mathcal{O}_0$ *(*μ *is an extension of* ν *on* $\mathcal{O}$*). If* ν *is* $\sigma-$*finite, the extension on the* $\sigma-$*algebra generated by* $\mathcal{O}_0$ *is unique.*

If X is a topological space, we define the *Borel* $\sigma-$*algebra* of X to be the $\sigma-$algebra generated by the closed subsets of X. Since the complements of open subsets are closed, it coincides with the $\sigma-$algebra generated by the open subsets of X. Sets in the Borel $\sigma-$algebra are called *Borel sets.*

The most standard example of a Borel $\sigma-$algebra with a measure is the Lebesgue measure in Euclidean space $\mathbb{R}^n$. A *cube* in $\mathbb{R}^n$ is a set Q of the form $Q = J_1 \times \cdots \times J_n$ where $J_1, \ldots, J_n$ are bounded intervals of $\mathbb{R}$. Its volume $\mathrm{Vol}(Q)$ is defined by the product

$$\mathrm{Vol}(Q) = |J_1| \cdot |J_2| \cdots |J_n|$$

Theorem I.2.2 *There exists a unique measure* λ *on the Borel* $\sigma-$*algebra of* $\mathbb{R}^n$ *such that* $\lambda(Q) = \mathrm{Vol}(Q)$ *for every cube* $Q \subset \mathbf{R}^n$.

This measure is called the Lebesgue measure. To prove Theorem I.2.2 we introduce the notion of outer measure. Let X be a metric space and denote by $\mathcal{P}(X)$ the family of all its subsets. A function $\mu^* : \mathcal{P}(X) \to [0, +\infty]$ is called an *outer measure* on X if it satisfies the following properties:

(a) $\mu^*(\emptyset) = 0$

(b) $A_i \subset X,\ i = 1, 2, \cdots \quad \Rightarrow \quad \mu^*\left(\cup_i A_i\right) \le \sum_i \mu^*(A_i)$

(c) If $A \subset X,\ B \subset X$ and $\mathrm{dist}(A, B) > 0$, then $\mu^*(A \cup B) = \mu^*(A) + \mu^*(B)$.

Here $\mathrm{dist}(A, B) \stackrel{def}{=} \inf\{\, \mathrm{dist}(x, y) : x \in A,\ y \in B\}$.

Theorem I.2.3 *If X is a separable metric space and μ^* is an outer measure on X, then μ^* is a measure on the Borel $\sigma-$algebra of X.*

For a proof of this theorem see [Ha].

Now we consider the Lebesgue outer measure μ^* on $\mathbb{R}^n$ defined by

$$\mu^*(S) = \inf \sum_i \mathrm{Vol}(Q_i)$$

where the infimum is taken over all finite or countable coverings of S by cubes $\{Q_i\}$. It is easy to check that μ^* is an outer measure. By Theorem I.2.3, μ^* is a measure on the Borel $\sigma-$algebra in $\mathbb{R}^n$. Then, to obtain the existence of a measure λ on the Borel $\sigma-$algebra of $\mathbb{R}^n$ such that $\lambda(Q) = \mathrm{Vol}(Q)$ for cubes Q, we only have to show that $\mu^*(Q) = \mathrm{Vol}(Q)$ when Q is a cube and then take λ as the restriction of μ^* to the Borel $\sigma-$algebra. We leave the verification of $\mu^*(Q) = \lambda(Q)$ for cubes as an exercise (see Ex. I.2.1.). To prove the uniqueness of λ suppose that λ' is another measure on the Borel $\sigma-$algebra of $\mathbb{R}^n$ such that $\lambda'(Q) = \mathrm{Vol}(Q)$ for every cube Q. Cover $\mathbf{R}^n$ by disjoint cubes $Q_1, Q_2, \ldots$ and let $\mathcal{O}_0$ be the family of Borel sets A such that $\lambda(A \cap Q_i) = \lambda'(A \cap Q_i)$ for every i. When Q is a cube, so is $Q \cap Q_i$. Hence $\lambda(Q \cap Q_i) = \mathrm{Vol}(Q \cap Q_i) = \lambda'(Q \cap Q_i)$. Then $\mathcal{O}_0$ contains every cube. Moreover it is easy to check that $\mathcal{O}_0$ is a $\sigma-$algebra. Since it contains all the cubes and it is a $\sigma-$algebra itself, $\mathcal{O}_0$ must contain all the countable unions of cubes. But every open set is a countable union of cubes. Hence $\mathcal{O}_0$ contains all the open sets. Since the family of open sets generates the Borel $\sigma-$algebra, $\mathcal{O}_0$ contains the Borel $\sigma-$algebra. Hence $\lambda(A) = \lambda'(A)$ for every Borel set. $\Box$

Definition. A *measure space* is a triple $(X, \mathcal{O}, \mu)$, where X is a set, $\mathcal{O}$ a $\sigma-$algebra of subsets of X, and $\mu : \mathcal{O} \to [0, \infty]$ a measure. If $\mu(X) = 1$, we say that $(X, \mathcal{O}, \mu)$ is a *probability space* and μ a *probability measure*, or a *probability.*

The Lebesgue probability on the Borel $\sigma-$algebra of the torus $\mathbb{T}^n$ is defined by a method similar to that used for $\mathbb{R}^n$. The torus $\mathbb{T}^n$ is the cartesian product $S^1 \times \cdots \times S^1$, where S^1 is the unit circle $S^1 = \{z \in \mathbb{C} : |z| = 1\}$. A cube in $\mathbb{T}^n$ is a set Q of the form $Q = J_1 \times \cdots \times J_n$ where $J_1, \cdots, J_n$ are intervals in S^1. The volume

$\mathrm{Vol}(Q)$ of a cube $Q \subset \mathbb{T}^n$ is defined by

$$\mathrm{Vol}(Q) = \prod_{i=1}^{n} \frac{|J_i|}{2\pi}.$$

Theorem I.2.4 *There exists a unique measure λ on the Borel $\sigma-$algebra of $\mathbb{T}^n$ such that $\lambda(Q) = \mathrm{Vol}(Q)$ for every cube $Q \subset \mathbb{T}^n$. This measure is a probability.*

The measure λ is called the *Lebesgue probability* on $\mathbb{T}^n$.

Let $(X, \mathcal{O}, \mu)$ be a measure space. A set $A \subset X$ is called a *zero measure set* (or a *null set*) if there exists $\hat{A} \in \mathcal{O}$, $A \subset \hat{A}$ and $\mu(\hat{A}) = 0$. Two sets $A_1, A_2 \subset X$ are said to be *equivalent (mod 0)* if their symmetric difference $A_1 \triangle A_2 = (A_1 \setminus A_2) \cup (A_2 \setminus A_1)$ is a zero measure set. If $\mathcal{S}$ is a family of subsets of X we say that $A \in \mathcal{S}$ (mod 0) if there exists $A_1 \in \mathcal{S}$, which is equivalent (mod 0) to A. We say that $\mathcal{S}_1 = \mathcal{S}_2$ (mod 0) if for every $A_1 \in \mathcal{S}_1$ and $A_2 \in \mathcal{S}_2$ we have $A_1 \in \mathcal{S}_2$ (mod 0) and $A_2 \in \mathcal{S}_1$ (mod 0). Finally, $\mathcal{S}$ *(mod 0)-generates* $\mathcal{O}$ if $\mathcal{O} = \bar{\mathcal{S}}$ (mod 0) where $\bar{\mathcal{S}}$ is the $\sigma-$algebra generated by $\mathcal{S}$.

Theorem I.2.5 *Let $(X, \mathcal{O}, \mu)$ be a measure space and $\mathcal{O}_0 \subset \mathcal{O}$ an algebra that (mod 0)-generates $\mathcal{O}$. Then for every $A \in \mathcal{O}$ and $\varepsilon > 0$ there exists $A_0 \in \mathcal{O}_0$ such that $\mu(A \triangle A_0) \le \varepsilon$. When $(X, \mathcal{O}, \mu)$ is a probability space, the converse is also true.*

Proof. Let $\hat{\mathcal{O}}_0$ be the family of sets $A \in \mathcal{O}$ such that for every $\varepsilon > 0$ there exists $A_0 \in \mathcal{O}_0$ satisfying $\mu(A \triangle A_0) \le \varepsilon$. We claim that $\hat{\mathcal{O}}_0$ is a $\sigma-$algebra. If $A \in \hat{\mathcal{O}}_0$ and $A \in \mathcal{O}$ satisfies $\mu(A \triangle A_0) \le \varepsilon$, then $\mu(A^c \triangle A_0^c) \le \varepsilon$ because $A^c \triangle A_0^c = A \triangle A_0$. Hence $A \in \hat{\mathcal{O}}_0$ implies $A^c \in \hat{\mathcal{O}}_0$. Moreover if $A_1; A_2, \dots$ are in $\hat{\mathcal{O}}_0$, then, given $\varepsilon > 0$, we take sets $S_1, S_2, \dots$ in $\mathcal{O}_0$ satisfying $\mu(A_i \triangle S_i) \le \varepsilon / 4^i$. Clearly, for $N \ge 1$

$$(\cup_i A_i) \triangle \left(\cup_{i=1}^N S_i\right) \subset \left(\cup_{i=1}^N (A_i \triangle S_i)\right) \cup \left((\cup_i A_i) \setminus \left(\cup_{i=1}^N A_i\right)\right)$$

Then

$$\mu\left[(\cup_i A_i) \triangle \left(\cup_{i=1}^N S_i\right)\right] \le \sum_{i=1}^{N} \mu(A_i \triangle S_i) + \mu\left[(\cup_i A_i) \setminus \left(\cup_{i=1}^N A_i\right)\right]$$

Now,

$$\cap_N \left((\cup_i A_i) \setminus \left(\cup_{i=1}^N A_i\right)\right) = \emptyset$$

and so

$$\mu\left[(\cup_i A_i)\setminus\left(\cup_{i=1}^N A_i\right)\right]\le \varepsilon/2$$

for a sufficiently large N. Hence

$$\mu\left[(\cup_i A_i)\bigtriangleup\left(\cup_{i=1}^N S_i\right)\right]\le \varepsilon$$

if N is sufficiently large. This proves that $\mathcal{O}_0$ is a $\sigma-$algebra. But $\hat{\mathcal{O}}_0\supset\mathcal{O}_0$. Then $\hat{\mathcal{O}}_0\supset\tilde{\mathcal{O}}_0$ where $\tilde{\mathcal{O}}_0$ is the $\sigma-$algebra generated by $\mathcal{O}_0$. Therefore $\hat{\mathcal{O}}_0\supset\mathcal{O}$ (mod 0) and so every set $A\in\mathcal{O}$ (mod 0) (in particular, sets in $\mathcal{O}$) has the approximation property. Conversely if $A\in\mathcal{O}$ and there exist sets $A_0\in\mathcal{O}_0$ with $\mu(A\triangle A_0)$ arbitrarily small, we can take a sequence $A_n\in\mathcal{O}_0,\ n=1,2,\dots$ such that

$$\sum_n \mu(A_n\triangle A)<+\infty.$$

This implies (see Ex. I.2.4.) that

$$A=\cup_{m=1}^{\infty}\cap_{n\ge m}A_n \qquad (\text{mod}0)$$

Hence A is in the $\sigma-$algebra (mod 0)-generated by $\mathcal{O}_0$.

Theorem I.2.6 *If μ is a probability on the Borel $\sigma-$algebra of a separable metric space, then for every Borel set A and every $\varepsilon>0$ there exists a closed set $A_0\subset A$ and an open set $A_1\supset A$ such that $\mu(A\setminus A_0)\le\varepsilon$ and $\mu(A_1\setminus A)\le\varepsilon$.*

Proof. Let $\hat{\mathcal{O}}$ be the family of Borel sets such that for all $\varepsilon>0$ there exist a closed set $A_0\subset A$ and an open set $A_1\supset A$ such that $\mu(A\setminus A_0)\le\varepsilon$ and $\mu(A_1\setminus A)\le\varepsilon$. Then $\hat{\mathcal{O}}$ is a $\sigma-$algebra because if $A\in\hat{\mathcal{O}}$ and $A_0\subset A\subset A_1$ satisfy the above property for a certain $\varepsilon>0$, then $A_1^c\subset A^c\subset A_0^c$ satisfy $\mu(A^c\setminus A_1^c)=\mu(A\setminus A_1)\le\varepsilon$ and $\mu(A_0^c\setminus A^c)=\mu(A\setminus A_0)\le\varepsilon$. Moreover, if $A_1,A_2,\dots$ belong to $\hat{\mathcal{O}}$, then, given $\varepsilon>0$ we can take closed sets $C_i\subset A_i$ and open sets $U_i\supset A_i$ such that $\mu(A_i\setminus C_i)\le\varepsilon/4^i$ and $\mu(U_i\setminus A_i)\le\varepsilon/2^i$ for i. Then

$$\begin{aligned}\mu\left((\cup_i U_i)\setminus(\cup_i A_i)\right) &\le \mu\left(\cup_i(U_i\setminus A_i)\right)\\ &\le \sum_i \mu(U_i\setminus A_i)\le\varepsilon\end{aligned}$$

Moreover, for all $N > 0$,

$$(\cup_i A_i) \setminus \left(\cup_{i=1}^N C_i\right) \subset \left(\cup_{i=1}^N (A_i \setminus C_i)\right) \cup \left(\cup_i A_i \setminus \cup_{i=1}^N A_i\right)$$

Hence

$$\begin{aligned}
\mu\left[(\cup_i A_i) \setminus \left(\cup_{i=1}^N C_i\right)\right] &\leq \sum_{i=1}^N \mu(A_i \setminus C_i) + \mu\left[(\cup_i A_i) \setminus \left(\cup_{i=1}^N A_i\right)\right] \\
&\leq \sum_i \mu(A_i \setminus C_i) + \mu\left[(\cup_i A_i) \setminus \left(\cup_{i=1}^N A_i\right)\right] \\
&\leq \frac{\varepsilon}{2} + \mu\left[(\cup_i A_i) \setminus \left(\cup_{i=1}^N A_i\right)\right]
\end{aligned}$$

But

$$\cap_{N=1}^\infty \left[(\cup_i A_i) \setminus \left(\cup_{i=1}^N A_i\right)\right] = \emptyset$$

This implies

$$\lim_{N \to +\infty} \mu\left[(\cup_i A_i) \setminus \left(\cup_{i=1}^N A_i\right)\right] = 0$$

Hence, if N is large enough

$$\mu\left[(\cup_i A_i) \setminus \left(\cup_{i=1}^N A_i\right)\right] \leq \varepsilon/2$$

Then the open set $\cup_i U_i$ and the closed set $\cup_{i=1}^N C_i$ satisfy the required property showing that $\cup_i A_i \in \mathcal{O}$ and proving that $\hat{\mathcal{O}}$ is a $\sigma-$algebra. Moreover $\hat{\mathcal{O}}$ contains all the open sets. To prove this observe that if A is an open set, we can write $A = \cup_i K_i$ where $K_1, K_2, \ldots$ are closed sets (using the fact that X is a separable metric space). Then, given $\varepsilon > 0$, we take, as in the previous argument, N so large that $\mu\left[(\cup_i K_i) \setminus (\cup_{i \leq N} K_i)\right] < \varepsilon$. Then

$$\mu\left(A \setminus \cup_{i=1}^N K_i\right) = \mu\left[(\cup_i K_i) \setminus (\cup_{i \leq N} K_i)\right] < \varepsilon$$

Hence the closed set $\cup_{i=1}^N K_i \subset A$ satisfies one of the approximation requirements, and the open set A itself obviously satisfies the other one. Hence the $\sigma-$algebra $\hat{\mathcal{O}}$ contains every open set. Then $\hat{\mathcal{O}}$ contains every Borel set, and the theorem is proved. □

Theorem I.2.7 *If X is a complete separable metric space and μ is a probability on its Borel $\sigma-$algebra, then for every Borel set A and every $\varepsilon > 0$ there exists a compact set $K \subset A$ such that $\mu(A \setminus K) \leq \varepsilon$.*

Proof. Given a Borel set A and $\varepsilon > 0$ take a closed set $A_0 \subset A$ such that $\mu(A \setminus A_0) \leq \varepsilon/2$. Now it suffices to find a compact set $K \subset A_0$ with $\mu(A_0 \setminus K) \leq \varepsilon/2$. For each $n \in \mathbb{Z}^+$ take a covering of A_0 by closed balls $B_1^{(n)}, B_2^{(n)}, \ldots$ of radius $1/n$. Take N_n such that

$$\mu \left[A_0 \setminus \cup_{i=1}^{N_n} (B_i^{(n)} \cap A_0) \right] \leq \frac{\varepsilon}{4^n}$$

Such N_n exists because

$$A_0 = \cup_i (B_i^{(n)} \cap A_0)$$

Define

$$K = \cap_{n=1}^{\infty} \cup_{i=1}^{N_n} (B_i^{(n)} \cap A_0)$$

Clearly K is closed. Moreover for every $\delta > 0$ there exists a finite cover of K by balls of radius $\leq \delta$ (namely, the balls $B_1^{(n)}, \ldots, B_{N_n}^{(n)}$, with n so large that $1/n < \delta$). Since X is complete, these properties imply that K is compact. Moreover

$$\begin{aligned}
\mu(A_0 \setminus K) &= \mu \left[A_0 \setminus \cap_{n=1}^{\infty} \cup_{i=1}^{N_n} (B_i^{(n)} \cap A_0) \right] \\
&= \mu \left[\cup_{n=1}^{\infty} \left(A_0 \setminus \cup_{i=1}^{N_n} (B_i^{(n)} \cap A_0) \right) \right] \\
&\leq \sum_{n=1}^{\infty} \mu \left(A_0 \setminus \cup_{i=1}^{N_n} (B_i^{(n)} \cap A_0) \right) \\
&\leq \sum_{n=1}^{\infty} \frac{\varepsilon}{4^n} < \frac{\varepsilon}{2}.
\end{aligned}$$

The theorem is proved. □

Exercises:

I.2.1.

(a) If $Q, Q_1, \ldots, Q_m$ are cubes and $Q \subset \cup_{i=1}^m Q_i$ prove that $\mathrm{Vol}(Q) \leq \sum_{i=1}^m \mathrm{Vol}(Q_i)$.

(b) If Q and Q_i, $i = 1, 2, \ldots$ are cubes and $Q \subset \cup_i Q_i$, prove that $\mathrm{Vol}(Q) \leq \sum_i \mathrm{Vol}(Q_i)$. Hint: Take a compact cube $Q_0 \subset Q$ with $\mathrm{Vol}(Q_0) \geq \mathrm{Vol}(Q) - \varepsilon$ and open cubes $\hat{Q}_i \supset Q_i$ with $\mathrm{Vol}(\hat{Q}_i) \leq \mathrm{Vol}(Q_i) + \varepsilon/2^i$. Observe that Q_0 is covered by some finite collection $\hat{Q}_{i_1}, \ldots, \hat{Q}_{i_m}$ and apply (a).

I.2.2. Let $(X, \mathcal{O}, \mu)$ be a probability space and $A_i \in \mathcal{O}$, $i = 1, 2, \ldots$ Prove that

$$\mu\left(\cap_n \cup_{i \geq n} A_i\right) = \lim_{n \to +\infty} \mu\left(\cup_{i \geq n} A_i\right) \geq \limsup_{n \to +\infty} \mu(A_n)$$

I.2.3. Let $(X, \mathcal{O}, \mu)$ be a probability space. Suppose that $A \in \mathcal{O}$ and that a sequence $A_i \in \mathcal{O}$, $i = 1, 2, \ldots$ satisfies $\sum_i \mu(A \triangle A_i) < +\infty$. Prove that for almost every $x \in A$ there exists n such that $x \in A_i$ for all $i \geq n$. Hint: The set of points x in A such that $x \notin A_n$ for infinitely many values of n coincides with $\cap_{n=1}^{\infty} \cup_{i \geq n} (A \setminus A_i)$ and $\mu\left(\cap_{n=1}^{\infty} \cup_{i \geq n} (A \setminus A_i)\right) \leq \sum_{i \geq n} \mu(A \setminus A_i)$ for all n.

I.2.4. Let $(X, \mathcal{O}, \mu)$ be a probability space. We say that a sequence of sets $A_i \in \mathcal{O}$, $i \geq 1$, *converges pointwise* to a set $A \in \mathcal{O}$ if for almost every $x \in A$ there exists $n = n(x) \in \mathbb{Z}^+$ such that $x \in A_i$ for all $i \geq n$ and for almost every $x \notin A$ there exists $n = n(x)$ such that $x \notin A_i$ for all $i \geq n(x)$.

(a) Prove that $A_i \to A$ pointwise if and only if

$$A = \cup_{n=1}^{\infty} \cap_{i \geq n} A_i \qquad (\text{mod } 0)$$

and

$$A^c = \cup_{n=1}^{\infty} \cap_{i \geq n} A_i^c \qquad (\text{mod } 0)$$

(b) Prove that $A_i \to A$ pointwise if and only if

$$A = \cup_{n=1}^{\infty} \cap_{i \geq n} A_i \qquad (\text{mod } 0)$$

and

$$A = \cap_{n=1}^{\infty} \cup_{i \geq n} A_i \qquad (\text{mod } 0)$$

(c) Prove that $A_i \to A$ pointwise implies

$$\lim_{i \to +\infty} \mu(A_i \triangle A) = 0$$

and

$$\lim_{i \to +\infty} \mu(A_i) = \mu(A).$$

Hint: Observe that, by Exercise I.2.2,

$$\limsup_{i \to +\infty} \mu(A \setminus A_i) \leq \mu\left(\cup_{n=1}^{\infty} \cap_{i \geq n} (A \setminus A_i)\right)$$

and

$$\limsup_{i\to+\infty} \mu(A_i \setminus A) \le \mu\left(\cap_{n=1}^{\infty} \cup_{i\ge n} (A_i \setminus A)\right)$$

and that $A_i \to A$ pointwise implies that the sets on the right hand side have measure zero.

(d) Prove that if $\sum_n \mu(A_n \triangle A) < +\infty$ then $A_n \to A$ pointwise. Hint: Use Exercise I.2.1.

(e) Show that $\lim_{n\to+\infty} \mu(A_n \triangle A) = 0$ does not imply $A_n \to A$ pointwise. Hint: Let $p_1/q_1, p_2/q_2, \cdots$ be all the rational numbers in $[0,1]$ including reducible fractions, arbitrarily ordered, and $A_n = [0,1] \setminus [(p_n-1)/q_n, (p_n+1)/q_n]$. Prove that $A_n \to [0,1]$ is false. In fact, *every* point $x \in [0,1]$ misses infinitely many of A_i's.

I.2.5. Prove that if X is a complete metric separable space and μ is a probability on its Borel $\sigma-$algebra then every Borel set of X is the union of a Borel set N with $\mu(N) = 0$ and a countable collection of disjoint compact sets.

I.2.6. Let X be a metric space. Given $t \ge 0$, the Hausdorff outer measure h_t on X is defined by

$$h_t(S) = \lim_{\rho\to 0} \inf \sum_i (\operatorname{diam} B_i)^t$$

where the infimum is taken over all the coverings of X by balls B_i, $i \ge 0$, with $\operatorname{diam} B_i \le \rho$ for all i.

(a) Prove that if $X = \mathbb{R}^n$ there exists $C > 0$ such that if μ^* is the Lebesgue outer measure then

$$C^{-1}\mu^*(S) \le h_n(S) \le C\,\mu^*(S)$$

for every set $S \subset X$.

(b) Prove that if μ is a probability on a metric space X such that there exist $\delta > 0$ and $C > 0$ satisfying

$$\mu(B) \le C\,(\operatorname{diam} B)^\delta$$

for every ball $B \subset X$, then

$$\mu(S) \le C\,h_\delta(S)$$

for every Borel set $S \subset X$.

(c) Prove that if X is a Borel subset of $\mathbb{R}^n$ and there exists a probability μ on X for which there exist $\delta > 0$ and $R > 0$ satisfying

$$\mu(B) \geq C^{-1}(\operatorname{diam} B)^{\delta}$$

for every ball B in X with $\operatorname{diam}(B) \leq R$, then there exists $K > 0$ such that

$$\mu(S) \geq K\, h_{\delta}(S)$$

for every Borel set $S \subset X$. Hint: Take an open set $U \supset S$ with $\mu(U \cap X) \leq \mu(S) + \varepsilon$. Decompose U into a disjoint union of cubes $Q_1, Q_2, \ldots$ such that all the sides of each Q_i have same length d_i. Then

$$\begin{aligned} \mu(S) + \varepsilon &\geq \sum_i \mu(Q_i \cap X) \geq \sum_i C^{-1} d_i^{\delta} \\ &= \sum_i C^{-1}(n^{-1/2} \operatorname{diam} Q_i)^{\delta} \end{aligned}$$

(d) If X is a metric space, prove that

$$\begin{array}{rcll} h_t(X) < +\infty & \Rightarrow & h_{t_2}(X) = 0 & \forall t_2 > t \\ h_t(X) > 0 & \Rightarrow & h_{t_1}(X) = +\infty & \forall t_1 < t \end{array}$$

The number d such that $h_t(X) = 0$ for $t > d$ and $h_t(X) = +\infty$ for $t < d$ is called the *Hausdorff dimension* of X and denoted by $HD(X)$.

(e) Let X and Y be metric spaces. Prove that if there exists a Lipschitz continuous homeomorphism between X and Y (i.e. a map $h : X \to Y$ such that there exists $C > 0$ satisfying $C^{-1}\operatorname{dist}(x, y) \leq \operatorname{dist}(h(x), h(y)) \leq C \operatorname{dist}(x, y)$ for all $x, y \in X$), then $HD(X) = HD(Y)$.

I.2.7. If $A \subset \mathbb{R}^n$ is an open set with $\lambda^*(\partial A) = 0$, then the characteristic function f_A of A (defined by $f_A(x) = 1$ if $x \in A$ and by $f_A(x) = 0$ if $x \notin A$) is integrable in the sense of Riemann and

$$\lambda(A) = \int f_A(x)\, dx$$

I.3 Measurable Maps

Let X, Y be two sets, and let $\mathcal{O}$ and $\mathcal{S}$ be $\sigma-$algebras of subsets of X and Y, respectively. A map $f : X \to Y$ is said to be *measurable* with respect to $\mathcal{O}$ and $\mathcal{S}$ if $f^{-1}(B) \in \mathcal{O}$ for every $B \in \mathcal{S}$. If X and Y are topological spaces, we say that $f : X \to Y$ is measurable if it is measurable with respect to the Borel $\sigma-$algebras of X and Y.

Proposition I.3.1 *If X and Y are topological spaces and $f : X \to Y$ is a map such that $f^{-1}(A)$ is a Borel subset of X for every open subset of Y, then f is measurable.*

Proof. Let $\mathcal{S}$ be the family of subsets $A \subset Y$ such that $f^{-1}(A)$ is a Borel subset of X. It is easy to check that $\mathcal{S}$ is a $\sigma-$algebra. By the hypothesis it contains all the open sets of Y. Then it contains the $\sigma-$algebra generated by the open sets that is exactly the Borel $\sigma-$algebra of Y. Hence every Borel subset A of Y belongs in $\mathcal{S}$, and then $f^{-1}(A)$ is a Borel subset of X. $\Box$.

Corollary I.3.2 *If X and Y are topological spaces, every continuous map $f : X \to Y$ is measurable.*

If X is a set and Y a topological space we say that a sequence of maps $f_n : X \to Y$, $n \geq 1$, converges (pointwise) to a map $f : X \to Y$ if

$$\lim_{n\to\infty} f_n(x) = f(x)$$

for every $x \in X$.

Corollary I.3.3 *If X is a set with a $\sigma-$algebra $\mathcal{O}$, Y a metric space with Borel σ-algebra, and $f_n : X \to Y$, $n \geq 1$, a sequence of measurable maps that converges pointwise to a map $f : X \to Y$, then f is measurable*

Proof. By Corollary I.3.2 we only have to show that $f^{-1}(A)$ is a Borel set whenever $A \subset Y$ is an open set. If A is an open set, then $A = \cup_{i=1}^{\infty} B_i$, where B_i are open sets such that $\text{dist}(B_i, Y \setminus A) > 0$ for every $i \geq 1$ (for instance, take $B_i = \{y \in Y : \text{dist}(y, Y \setminus A) > 1/i\}$). Then

$$f^{-1}(A) = \cup_{i=1}^{\infty} \cup_{m=1}^{\infty} \cap_{n\geq m} f_n^{-1}(B_i)$$

Since $f_n^{-1}(B_i) \in \mathcal{O}$ for all n and i (because every map f_n is measurable) then the set on the right hand side belongs in $\mathcal{O}$, thus proving that $f^{-1}(A)$ is measurable. □

If $(X, \mathcal{O}, \mu)$ is a measure space and Y is a topological space, we say that a sequence of maps $f_n : X \to Y$ *converges almost everywhere* to a map $f : X \to Y$ if the set of points $x \in X$ for which the convergence $\lim_{n\to\infty} f_n(x) = f(x)$ fails is a zero measure set.

Theorem I.3.4 (Egorov's theorem) *Let $(X, \mathcal{O}, \mu)$ be a probability space and Y a metric space. Let $f_n : X \to Y$ be a sequence of measurable maps that converges almost everywhere to a map $f : X \to Y$. Then for every $\varepsilon > 0$ there exists a set $A \in \mathcal{O}$ with $\mu(A) \geq 1 - \varepsilon$ such that $f_n|_A$ converges uniformly to $f|_A$.*

Proof. Let $X_0 \in \mathcal{O}$ be a set such that $\mu(X_0) = 1$ and $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in X_0$. Then $(X_0, \mathcal{O}_0, \mu)$ is a probability space, where $\mathcal{O}_0$ consists of all measurable subsets of X_0. The maps $f_n : X_0 \to Y$ are measurable and converge (pointwise) to $f : X_0 \to Y$. Hence $f : X_0 \to Y$ is a measurable map. Then the functions $X_0 \to \mathbb{R}$ defined by $x \mapsto \operatorname{dist}(f(x), f_n(x))$ are measurable, and (cf. Exercise I.3.1) so are the functions $X_0 \to \mathbb{R}$ defined by

$$x \mapsto \sup_{n \geq N} \operatorname{dist}(f(x), f_n(x))$$

Then the sets

$$S_N(\delta) = \{x \in X_0 | \sup_{n \geq N} \operatorname{dist}(f(x), f_n(x)) \geq \delta\}$$

are measurable. Moreover

$$S_1(\delta) \supset S_2(\delta) \supset \cdots \supset S_N(\delta) \supset S_{N+1}(\delta) \supset \cdots$$

and, since the maps $f_n : X_0 \to Y$ converge to $f : X_0 \to Y$, we have

$$\cap_N S_N(\delta) = \emptyset$$

for every $\delta > 0$. Indeed, if this intersection contains a point x, then $\operatorname{dist}(f_{n_i}(x), f(x))$ would be $\geq \delta$ for infinitely many values of n_i, contradicting $\lim_{n\to\infty} f_n(x) = f(x)$. Hence

$$\lim_{N\to\infty} \mu(S_N(\delta)) = 0$$

for all $\delta > 0$. Given $\varepsilon > 0$ choose, for each $n \in \mathbb{Z}^+$, an integer N_n such that

$$\mu(S_{N_n}(1/n)) \leq \varepsilon/2^n$$

Define

$$S = \cup_{n=1}^{\infty} S_{N_n}(1/n)$$

Then

$$\mu(S) \leq \sum_{n=1}^{\infty} \mu(S_{N_n}(1/n)) \leq \varepsilon$$

Let us prove that $f_n|_{X_0 \setminus S}$ converges uniformly to $f|_{X_0 \setminus S}$. Given $\varepsilon_0 > 0$ take n such that $1/n \leq \varepsilon_0$. Observe that $x \in X_0 \setminus S$ implies $x \in X_0 \setminus S_{N_n}(1/n)$. This means that

$$\sup_{i \geq N_n} \operatorname{dist}(f_i(x), f(x)) \leq \frac{1}{n} \leq \varepsilon_0$$

Hence

$$\operatorname{dist}(f_i(x), f(x)) \leq \varepsilon_0$$

for all $x \in X_0 \setminus S$ and $i \geq N_n$. This proves the uniform convergence of $f_n|_{X_0 \setminus S}$ to $f|_{X_0 \setminus S}$. □

Theorem I.3.5 (Lusin's theorem) *Let $(X, \mathcal{O}, \mu)$ be a probability space where X is a complete metric separable space, Y a separable metric space and $f : X \to Y$ a measurable map. Then, for every $\varepsilon > 0$ there exists a compact set $K \subset X$ such that $\mu(K) > 1 - \varepsilon$ and the map $f|_K$ is continuous.*

Proof. The function f is the limit of a sequence of simple functions $f_n : X \to Y$, i.e. measurable maps whose images are finite sets (see Exercise I.3.2). For these maps the theorem holds as a corollary of Theorem I.2.7. Given $\varepsilon > 0$ take a set $X_0 \in \mathcal{O}$ such that $\mu(X_0) > 1 - \varepsilon/2$ and $f_n|_{X_0} \to f|_{X_0}$ uniformly. For each n take a compact set $K_n \subset X_0$ such that $\mu(K_n) \leq \mu(X_0) - \varepsilon/4^n$ and $f_n|_{K_n}$ is continuous. Define K by $K = \cap_n K_n$. Then K is compact and

$$\begin{aligned}
\mu(K) &= \mu(X_0) - \mu(X_0 \setminus K) \\
&= \mu(X_0) - \mu(\cup_n (X_0 \setminus K_n)) \\
&\geq \mu(X_0) - \sum_n \varepsilon/4^n
\end{aligned}$$

$$\begin{aligned} &\geq \mu(X_0) - \varepsilon/2 \\ &\geq 1 - \varepsilon. \end{aligned}$$

Moreover, since $K \subset K_n$ for all n, the maps $f_n|_K : K \to Y$ are continuous for all n. Then $f|_K : K \to Y$, being the uniform limit of this sequence as $n \to \infty$, is also continuous. □

Exercises:

I.3.1. Let X be a topological space. Let $f_\alpha : X \to \mathbb{R}$ $\alpha \in \mathbb{N}$, be a countable family of maps. Define the function $\sup_\alpha f_\alpha : X \to \mathbb{R}$ by $(\sup_\alpha f_\alpha)(x) = \sup_\alpha f_\alpha(x)$. Prove that

$$(\sup_\alpha f_\alpha)^{-1}((a,b]) = \cap_\alpha f_\alpha^{-1}((-\infty,b]) \cap (\cup_\alpha f_\alpha^{-1}((a,+\infty)))$$

and that $\sup_\alpha f_\alpha$ is measurable.

I.3.2. Let $(X, \mathcal{O}, \mu)$ be a probability space and Y a separable metric space. A map $f : X \to Y$ is *simple* if it is measurable and $f(X)$ is a finite set. Prove that every measurable map is the limit of a sequence of simple maps. Hint: Show that for each $n \geq 1$ there exists a covering of Y by disjoint Borel sets $P_1^{(n)}, P_2^{(n)}, \ldots$ with $\operatorname{diam} P_i^{(n)} \leq 1/n$ for all i. For each n take N_n such that:

$$\mu(f^{-1}(\cup_{i>N_n} P_i^{(n)})) \leq 1/4^n.$$

Take a point $a \in Y$ and arbitrary points $p_i^{(n)} \in P_i^{(n)}$. Define $f_n : X \to Y$ by $f_n(x) = p_i^{(n)}$ if $x \in f^{-1}(P_i^{(n)})$, $1 \leq i \leq N_n$, and $f_n(x) = a$ if

$$x \in f^{-1}(\cup_{i>N_n} P_i^{(n)})$$

I.3.3. Let $(X, \mathcal{O}, \mu)$ be a probability space, $\mathcal{O}_0 \subset \mathcal{O}$ a subalgebra that generates $\mathcal{O}$ and $\mu' : \mathcal{O} \to [0,1]$ a probability. If $\mu|_{\mathcal{O}_0} = \mu'|_{\mathcal{O}_0}$, then $\mu = \mu'$. Hint: Prove that the family $\mathcal{O}'$ of sets $A \in \mathcal{O}$ such that $\mu(A) = \mu'(A)$ is a $\sigma-$algebra that contains $\mathcal{O}_0$.

I.4 Measure Preserving Maps

If $(X, \mathcal{O}, \mu)$ and $(Y, \mathcal{S}, \nu)$ are measure spaces, we say that a map $T : X \to Y$ is *measure preserving* if

$$B \in \mathcal{S} \quad \Rightarrow \quad T^{-1}(B) \in \mathcal{O} \ \text{ and } \ \mu(T^{-1}(B)) = \nu(B).$$

The following is a simple but useful method of checking that a map is measure preserving.

Proposition I.4.1 *Let $(X, \mathcal{O}, \mu)$ and $(Y, \mathcal{S}, \nu)$ be probability spaces and $\mathcal{S}_0 \subset \mathcal{S}$ an algebra that generates $\mathcal{S}$. If $T^{-1}(B) \in \mathcal{O}$ and $\mu(T^{-1}(B)) = \nu(B)$ for all $B \in \mathcal{S}_0$, then T is measure preserving.*

Proof. Let $\hat{\mathcal{S}}$ be the family of sets $B \in \mathcal{S}$ such that $T^{-1}(B) \in \mathcal{O}$ and $\mu(T^{-1}(B)) = \nu(B)$. It is easy to see that $\hat{\mathcal{S}}$ is a $\sigma-$algebra, and it obviously contains $\mathcal{S}_0$. Then $\hat{\mathcal{S}} \supset \mathcal{S}$. Hence $B \in \mathcal{S}$ implies $T^{-1}(B) \in \mathcal{O}$ and $\mu(T^{-1}(B)) = \nu(B)$. $\Box$

Examples:

1. Translations of $\mathbb{T}^n$. For any $k = (k_1, \ldots, k_n) \in \mathbb{T}^n$ define the translation $L_k : \mathbb{T}^n \to \mathbb{T}^n$ by

$$L_k(x_1, \ldots, x_n) = (k_1 x_1, \ldots, k_n x_n)$$

Proposition I.4.2 *The Lebesgue probability on $\mathbb{T}^n$ is invariant under all translations. Moreover, it is the unique probability on the Borel $\sigma-$algebra of $\mathbb{T}^n$ with this property.*

Proof. Let L_k be a translation. If $Q \subset \mathbb{T}^n$ is a cube, it is clear that

$$\lambda(L_k^{-1}(Q)) = \operatorname{Vol} L_k^{-1}(Q) = \operatorname{Vol} Q = \lambda(Q).$$

Then, applying Proposition I.4.1 to the algebra of disjoint unions of cubes gives that λ is L_k-invariant. Conversely, suppose that λ' is another probability on the Borel $\sigma-$algebra of $\mathbb{T}^n$ invariant under all translations. Let $\{Q_1, \ldots, Q_m\}$ be a finite collection of disjoint cubes covering $\mathbb{T}^n$, which are all translations of the same cube. Then $\lambda'(Q_1) = \cdots = \lambda'(Q_m)$ and $\lambda(Q_1) = \cdots = \lambda(Q_m)$. Therefore, if $\lambda'(Q_i) \neq \lambda(Q_i)$ for some i, say $\lambda'(Q_i) > \lambda(Q_i)$, it would immediately imply that $\lambda'(Q_j) > \lambda(Q_j)$ for all

$1 \leq j \leq m$. This would imply $1 = \lambda'(\mathbb{T}^n) = \sum_j \lambda'(Q_j) > \sum_j \lambda(Q_j) = \lambda(\mathbb{T}^n) = 1$, a contradiction. Hence, $\lambda(Q_i) = \lambda'(Q_i)$ for all i. This shows that λ and λ' coincide on cubes belonging to this type of partition. But it is easy to see that every cube $Q \subset \mathbb{T}^n$ is a (disjoint) union of cubes belonging to such partition. Then λ and λ' agree on the family of all cubes, and so they agree on the algebra of disjoint unions of cubes. Hence $\lambda = \lambda'$ on the Borel $\sigma-$algebra of $\mathbb{T}^n$. $\square$

2. Linear maps of $\mathbb{T}^n$. Define a map $\pi : \mathbb{R}^n \to \mathbb{T}^n = S^1 \times \cdots \times S^1$ by

$$\pi(t_1, \ldots, t_n) = (e^{2\pi i t_1}, \ldots, e^{2\pi i t_n}).$$

Clearly

$$\pi(x) = \pi(y) \quad \Longleftrightarrow \quad x - y \in \mathbb{Z}^n$$

and $\pi(\mathbb{Z}^n) = \{(1, \ldots, 1)\}$. Given a linear isomorphism $\hat{T} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\hat{T}(\mathbb{Z}^n) \subset \mathbb{Z}^n$, or, equivalently, whose matrix in the canonical basis has integral components, there exists $T : \mathbb{T}^n \to \mathbb{T}^n$ such that $\pi \circ \hat{T} = T \circ \pi$. The map T is defined by taking, for each $p \in \mathbb{T}^n$, a point $x \in \pi^{-1}(p)$ and setting

$$T(p) = \pi[\hat{T}(x)]$$

This definition is consistent because if $y \in \pi^{-1}(p)$ is another point, then $x - y \in \mathbb{Z}^n$ hence $\hat{T}(x - y) \in \mathbb{Z}^n$, and so

$$\pi[\hat{T}(y)] = \pi[\hat{T}(x)]$$

Maps constructed as T above are called linear maps of $\mathbb{T}^n$, and $\hat{T}$ is called the linear lifting of T to the covering space $\mathbb{R}^n$.

Proposition I.4.3 *The Lebesgue probability on $\mathbb{T}^n$ is invariant under every linear map of $\mathbb{T}^n$.*

Proof. Let $T : \mathbb{T}^n \to \mathbb{T}^n$ be a linear map. Define a probability ν on the Borel $\sigma-$algebra of $\mathbb{T}^n$ by setting $\nu(A) = \lambda(T^{-1}(A))$. If $L_k : \mathbb{T}^n \to \mathbb{T}^n$ is a translation and $Tx = k$, we have for every Borel set A:

$$\nu(L_k^{-1}(A)) = \lambda(T^{-1}L_k^{-1}(A)) = \lambda(L_x^{-1}T^{-1}(A)) = \lambda(T^{-1}(A)) = \nu(A)$$

Hence ν is invariant under all translations. Then $\lambda = \nu$. This proves that λ is invariant under T.

3. Bernoulli Shifts. Let X be a compact metric space. Let $B(X)$ denote the space of double-sided sequences $\theta : \mathbb{Z} \to X$ endowed with metric
$$d(\alpha, \beta) = \sum_{n=-\infty}^{\infty} \frac{1}{k^{|n|}}\, d_0(\alpha(n), \beta(n))$$
where $k > 1$ is a constant and d_0 the metric on X. Observe that, in this metric, a sequence of sequences $\{\theta_n\} \in B(X)$ converges to a sequence $\theta \in B(X)$ if and only if it converges componentwise, i.e.
$$\lim_{n\to\infty} \theta_n(j) = \theta(j)$$
for all $j \in \mathbb{Z}$. Hence the convergence is independent of the constant $k > 1$ used to define the metric d. The *shift* $\sigma : B(X) \to B(X)$ is defined by
$$(\sigma\theta)(n) = \theta(n+1)$$
Clearly σ is a homeomorphism. When X is a finite set, $X = \{1, \ldots, m\}$, then we denote $B(\{1, \ldots, m\})$ simply by $B(m)$.

Given Borel sets $A_0, \ldots, A_m$ in X and $j \in \mathbb{Z}$ we define a cylinder $C(j, A_0, \ldots, A_m)$ by
$$C(j, A_0, \ldots, A_m) = \{\theta \in B(X) |\ \theta(j+i) \in A_i, \quad 0 \le i \le m\}$$
Finite disjoint unions of cylinders make an algebra that generates the Borel $\sigma-$algebra of $B(X)$. Moreover, given a probability μ_0 on the Borel $\sigma-$algebra of X, there exists a unique probability μ on the Borel $\sigma-$algebra of $B(X)$ (called the product measure associated with μ_0) such that for every cylinder:
$$\mu(C(j, A_0, \ldots, A_m)) = \prod_{i=0}^{m} \mu_0(A_i) \tag{I.4.1}$$
The existence and uniqueness of μ can be deduced from Theorem I.2.3 following a construction similar to that of Lebesgue measure on $\mathbb{R}^n$ or $\mathbb{T}^n$. Moreover, μ is invariant under σ. This follows from the fact $\mu(\sigma^{-1}(C)) = \mu(C)$ for every cylinder C, as it can be checked by using the above formula, and from the fact that finite disjoint unions of cylinders make an algebra that generates the Borel $\sigma-$algebra of $B(X)$. Denote

by $B_\mu(X)$ the space $B(X)$ endowed with the probability μ. The shift $\sigma : B_\mu(X) \to B_\mu(X)$ is called a *Bernoulli shift.*

When X is a finite set, $X = \{1, \ldots, m\}$, the probability μ_0 is determined by the numbers $p_i = \mu_0(\{i\})$ and in this case $B_\mu\ (\{1, \ldots, m\})$ is simply denoted by $B(p_1, \ldots, p_m)$.

In a similar way we define $B^+(X)$, the space of one-sided sequences $\theta : \mathbb{Z}^+ \to X$ endowed with the metric

$$d(\alpha, \beta) = \sum_{n=0}^{\infty} \frac{1}{k^n}\, d_0(\alpha(n), \beta(n))$$

where $k > 1$. The *one sided shift* $\sigma : B^+(X) \to B^+(X)$ is defined also by the formula $(\sigma\theta)(n) = \theta(n+1)$, but now $n \geq 0$. The map σ on $B^+(X)$ is only a continuous surjective map, and not a homeomorphism. If μ_0 is a probability on the Borel $\sigma-$algebra of X, a product measure μ on the Borel $\sigma-$algebra of $B^+(X)$ is defined by the same formula for cylinders $C(j, A_0, \cdots, A_m)$ with $j \geq 0$. Again μ turns out to be $\sigma-$invariant . Then $B_\mu^+(X)$ and $B^+(p_1, \ldots, p_m)$ are defined as in the previous case.

4. Volume Preserving Diffeomorphisms. Let U and V be open sets of $\mathbb{R}^n$. We say that a diffeomorphism $f : U \to V$ is *volume preserving* if $|\det f'(x)| = 1$ for all $x \in U$. Then

$$\lambda(f^{-1}(A)) = \lambda(A)$$

for every Borel subset $A \subset V$. To prove this property first observe that if $\mu^*(\cdot)$ denotes the outer Lebesgue measure, then

$$\lambda(A) = \int \chi_A(x)\, dx$$

for every open set $A \subset \mathbb{R}^n$ with $\mu^*(\partial A) = 0$. Here the integral is the Riemann integral and χ_A is the characteristic function (indicator) of A (i.e. $\chi_A(x) = 0$ if $x \notin A$ and $\chi_A(x) = 1$ if $x \in A$, see also Ex. I.2.7). But $\mu^*(\partial A) = 0$ implies $\mu^*(\partial f^{-1}(A)) = 0$. Hence

$$\lambda(f^{-1}(A)) = \int \chi_{f^{-1}(A)}(x)\, dx.$$

Note that $\chi_{f^{-1}(A)} = \chi_A \circ f$. Therefore

$$\lambda(f^{-1}(A)) \quad = \quad \int \chi_{f^{-1}(A)}(x)\, dx \ = \ \int (\chi_A \circ f)\, dx$$

$$
\begin{aligned}
&= \int (\chi_A \circ f) |\det f'| \, dx \\
&= \int \chi_A(x) \, dx \ = \ \lambda(A)
\end{aligned}
$$

where the fourth identity is obtained by change of variables. This shows that the desired formula holds when A is an open set with $\mu^*(\partial A) = 0$, in particular, when A is a cube. Then we take a covering of V by disjoint cubes $Q_1, Q_2, \cdots$ and define the $\sigma-$algebra $\hat{\mathcal{O}}$ of all the Borel sets $A \subset V$ such that

$$
\lambda(f^{-1}(A \cap Q_i)) = \lambda(A \cap Q_i)
$$

for all i. This is a $\sigma-$algebra that contains the subalgebra of disjoint unions of cubes. Hence $\hat{\mathcal{O}}$ is the Borel $\sigma-$algebra and the formula is proved.

Exercises

I.4.1. An interval exchange transformation of S^1 is a map $T : S^1 \to S^1$ such that there exists a finite family of disjoint open intervals $J_1, \dots, J_n$, whose closures cover S^1 and such that $T|_{J_i} : J_i \to T(J_i)$ is an isometry, and the intervals $T(J_i)$ are disjoint, too. Prove that T preserves the Lebesgue probability.

I.4.2. Let $J_i = (a_i, b_i)$, $i \geq 1$ be a countable family of disjoint open intervals contained in $(0, 1)$ such that

$$
\sum_{i=1}^{\infty} (b_i - a_i) = 1
$$

Let $T : [0, 1] \to [0, 1]$ be a map satisfying:

$$
T(x) = \sigma_i \, \frac{x - a_i}{b_i - a_i} + \frac{1 - \sigma_i}{2}
$$

whenever $x \in J_i$. Here σ_i can be $+1$ or -1. Prove that T preserves the Lebesgue probability.

I.4.3. Prove that the set of periodic points of the shift $\sigma : B(X) \to B(X)$ is dense in $B(X)$.

I.4.4. Consider $B(m)$ endowed with metric

$$d(\alpha, \beta) = \sum_n \frac{1}{k^{|n|}} |\alpha(n) - \beta(n)|$$

where $k > 1$. Denote by $B_r(\theta)$ the closed ball of radius r centered at θ and put

$$S_N(\theta) = \{\alpha \in B(m) |\ \alpha(n) = \theta(n) \qquad \text{for} \qquad |n| \leq N\}$$

(a) Prove that $S_N(\theta) \subset B_r(\theta)$ if

$$\frac{1}{k^N} \cdot \frac{2m}{k-1} \leq r$$

(b) Prove that $S_N(\theta) \supset B_r(\theta)$ if

$$N \leq \frac{\log(1/r)}{\log k}$$

(c) Prove, by using (a) and (b), that if μ is the product measure associated with a uniform probability μ_0 on $\{1, \ldots, m\}$ given by $\mu_0(\{i\}) = 1/m$, then there exists $C > 0$ such that

$$C^{-1} r^\delta \leq \mu(B_r(\theta)) \leq C r^\delta$$

where

$$\delta = \frac{2 \log m}{\log k}.$$

Deduce that $2 \log m / \log k$ equals the Hausdorff dimension of $B(m)$.

I.5 Poincaré Recurrence Theorem

We shall now prove the Poincaré recurrence theorem. We deal first with a probabilistic version, which makes no reference to topology.

Theorem I.5.1 *Let T be a measure-preserving map of a probability space $(X, \mathcal{O}, \mu)$. Given $A \in \mathcal{O}$, let A_0 be the set of points $x \in A$ such that $T^n(x) \in A$ for infinitely many $n \geq 0$. Then A_0 belongs to $\mathcal{O}$, and $\mu(A_0) = \mu(A)$.*

Proof. Let $C_n := \{x \in A |\ T^j(x) \notin A \text{ for all } j \geq n\}$. It is clear that

$$A_0 = A \setminus \cup_{n=1}^{\infty} C_n$$

Thus, the theorem will be proved if we show that $C_n \in \mathcal{O}$ and $\mu(C_n) = 0$ for every $n \geq 1$. Observe that

$$C_n = A \setminus \cup_{j \geq n} T^{-i}(A).$$

which shows that $C_n \in \mathcal{O}$, and implies that

$$C_n \subset \cup_{j \geq 0} T^{-j}(A) \setminus \cup_{j \geq n} T^{-j}(A)$$

But since

$$\cup_{j \geq n} T^{-j}(A) = T^{-n}(\cup_{j \geq 0} T^{-j}(A))$$

we obtain that

$$\mu(\cup_{j \geq n} T^{-j}(A)) = \mu(\cup_{j \geq 0} T^{-j}(A))$$

This implies $\mu(C_n) = 0$. □

In order to state the topological version of the recurrence theorem, we need the notion of the $\omega-$limit set of a point under a map. Let X be a topological space and $T : X \to X$ a map. We define the $\omega-$limit set of a point $x \in X$ as the set of points $y \in X$ such that for every neighborhood U of y the relation $T^n(x) \in U$ holds for infinitely many positive values of n. If X is a metric space, this is equivalent to saying that

$$\liminf_{n \to \infty} \operatorname{dist}(T^n(x), y) = 0$$

Theorem I.5.2 *Let X be a separable metric space and $T : X \to X$ a Borel-measurable map. Let μ be a T-invariant probability measure on the Borel $\sigma-$algebra of X. Then $\mu(\{x :\ x \notin \omega(x)\}) = 0$. In other words, almost every point is recurrent.*

Proof. Let $\{U_n\}_{n=0}^{\infty}$ be a basis of open sets such that

$$\lim_{n \to \infty} \operatorname{diam} U_n = 0$$

and

$$\cup_{n\geq m} U_n = X$$

for every $m \geq 0$. Let $\tilde{U}_n := \{x \in U_n |\ T^j(x) \in U_n$ for infinitely many positive values of $j\}$. From the preceding theorem

$$\mu(U_n \setminus \tilde{U}_n) = 0.$$

Put

$$\tilde{X} := \cap_{m=0}^{\infty} \cup_{n\geq m} \tilde{U}_n$$

It follows that

$$\begin{aligned} \mu(X \setminus \tilde{X}) &= \mu(\cup_{m=0}^{\infty}(X \setminus \cup_{n\geq m}\tilde{U}_n)) \\ &= \mu(\cup_{m=0}^{\infty}(\cup_{n\geq m} U_n \setminus \cup_{n\geq m}\tilde{U}_n)) \\ &\leq \mu(\cup_{m=0}^{\infty} \cup_{n\geq m} (U_n \setminus \tilde{U}_n)) \\ &= 0 \end{aligned}$$

Thus we only have to show that $x \in \tilde{X}$ implies $x \in \omega(x)$. Let $r > 0$. Choose m such that $\operatorname{diam} U_n \leq r/3$ if $n \geq m$. Since $x \in \tilde{X}$, it follows that $x \in \cup_{n\geq m}\tilde{U}_n$. Thus there exists $n \geq m$ such that $x \in \tilde{U}_n$. Since $\operatorname{diam} U_n \leq r/3$, it follows that $U_n \subset B_r(x)$, which implies that $T^j(x) \in B_r(x)$ if $T^j(x) \in U_n$. But since $x \in \tilde{U}_n$, $T^j(x) \in U_n$ for infinitely many values of j, showing that $x \in \omega(x)$. $\Box$

Exercises:

I.5.1. Let T be a measure preserving map of a probability space $(X, \mathcal{O}, \mu)$. Given $A \in \mathcal{O}$, define $\tilde{A}$ as the set of points $x \in A$ such that $T^n(x) \in A$ for infinitely many $n > 0$. Define $N : \tilde{A} \to \mathbb{Z}$ by setting $N(x) = m$ if $m > 0$, $T^m(x) \in A$ and $T^n(x) \notin A$ for $0 < n < m$. The function $N(x)$ is called the *first return time*.

(a) Prove that $\tilde{A} \in \mathcal{O}$, $\mu(A \setminus \tilde{A}) = 0$, and $N : \tilde{A} \to \mathbb{Z}$ is a measurable function;

(b) Define $\tilde{T} : \tilde{A} \to X$ by

$$\tilde{T}(x) = T^{N(x)}(x)$$

Prove that $\tilde{T}(\tilde{A}) \subset \tilde{A}$ and that if T is invertible (i.e. if there exists a measure preserving map S of $(X, \mathcal{O}, \mu)$ such that $TS(x) = ST(x) = x =$ for a.e. x), then $\tilde{T} : \tilde{A} \to \tilde{A}$ leaves invariant the measure $\mu|_{\tilde{A}}$. Hint: Put $\tilde{A}_n = N^{-1}(\{n\})$. Prove that $\mu(\tilde{T}(C)) = \mu(C)$ for all $C \in \mathcal{O}$ such that $C \subset \tilde{A}_n$.

I.5.2. Consider the map $T : [0,1] \to [0,1]$ defined by $T(x) = x/2$ for $0 < x \leq 1$ and $T(0) = 1$. Prove that T is measurable and that there is no T-invariant probability on the Borel $\sigma-$algebra of $[0,1]$.

I.6 Integration

Let $(X, \mathcal{O}, \mu)$ be a measure space. Given a set $A \in \mathcal{O}$, its characteristic function χ_A is defined by $\chi_A(x) = 0$ if $x \notin A$ and $\chi_A(x) = 1$ if $x \in A$. A function $f : X \to \mathbb{C}$ is said to be *simple* if it can be written as $f = \sum_{i=1}^n \lambda_i \chi_{A_i}$, where $A_1, \ldots, A_n$ are in $\mathcal{O}$ and $\mu(A_i) < +\infty$ when $\lambda_i \neq 0$. The integral of f is defined by

$$\int_X f \, d\mu = \sum_{\lambda_i \neq 0} \lambda_i \, \mu(A_i)$$

if the series converges absolutely.

We say that $f : X \to \mathbf{C}$ is *integrable* if there exists a sequence of simple functions $f_n : X \to \mathbb{C}$ such that

$$\lim_{n\to\infty} f_n(x) = f(x) \quad \text{for a.e. } x$$

and

$$\lim_{n,m\to\infty} \int_X |f_n - f_m| \, d\mu = 0 \tag{I.6.1}$$

The *integral* of f is defined by

$$\int_X f \, d\mu = \lim_{n\to\infty} \int_X f_n \, d\mu \tag{I.6.2}$$

The existence of the limit on the right hand side follows from (I.6.1), because it implies

$$\limsup_{n,m\to\infty} \left| \int_X f_n \, d\mu - \int_X f_m \, d\mu \right| \leq \lim_{n,m\to\infty} \int_X |f_n - f_m| \, d\mu = 0$$

thus showing that the sequence in (I.6.2) is a Cauchy sequence of complex numbers. It is more difficult, but necessary, to show that the limit in (I.6.2) is independent of the sequence $\{f_n\}$. This is clearly guaranteed by the following property: If $g_n : X \to \mathbb{C}$, $n \geq 1$, is a sequence of simple functions such that $\lim_{n\to\infty} g_n(x) = 0$ a. e. and

$$\lim_{n,m\to\infty} \int_X |g_n - g_m| \, d\mu = 0$$

then

$$\lim_{n\to\infty} \int_X g_n \, d\mu = 0$$

The purpose of this section is to survey the basic properties of integration. The proofs of the results can be found in, for instance, [Rd, Ha] or [F].

Integrable functions do not have to be measurable (i.e. $f^{-1}(A) \in \mathcal{O}$ whenever $A \subset \mathbb{C}$ is a Borel set), but it follows from the definition that any integrable function f coincides with a measurable function on a set $X_0 \subset X$ with $\mu(X_0^c) = 0$. Hence, $f^{-1}(A) \in \mathcal{O} \pmod 0$, if $A \subset \mathbb{C}$ is a Borel set. It is also clear that given two functions $f_i : X \to \mathbb{C}$, $i = 1, 2$, which coincide almost everywhere, f_1 is integrable if and only if so is f_2, and in this case their integrals are equal. Given $f : X \to \mathbb{C}$, we say that f is integrable on $A \subset X$ if $f \cdot \chi_A$ is integrable, and we put

$$\int_A f \, d\mu = \int_X f \cdot \chi_A \, d\mu.$$

It is easy to see that if f is integrable, then it is integrable on every $A \in \mathcal{O}$. Moreover, observe that f is integrable if and only if $|f|$ is.

When dealing with applications, an often important problem is to decide if a function $f : X \to \mathbb{C}$, which is the limit a.e. of a sequence of integrable functions $f_n : X \to \mathbb{C}$, is integrable and if its integral is the limit of the integrals of f_n as $n \to \infty$. Obviously, affirmative answer will require additional hypotheses. The following three results are fundamental tools to handle this problem.

Theorem I.6.1 (Fatou's Lemma) *Let $f_n : X \to \mathbb{R}$ be a sequence of positive integrable functions such that*

$$\liminf_{n\to\infty} \int_X f_n \, d\mu < +\infty$$

and converging a.e. to $f : X \to \mathbb{R}$. *Then* f *is integrable and*

$$\liminf_{n\to\infty} \int_X f_n \, d\mu \geq \int_X f \, d\mu$$

Theorem I.6.2 (Monotone Convergence Theorem) *Let* $f_n : X \to \mathbb{R}$ *be a sequence of integrable functions such that for a.e.* x *the sequence* $\{f_n(x)\}$ *is monotonically increasing and*

$$\sup_n \int_X f_n d\mu < +\infty$$

Then the function $f(x) = \lim_{n\to\infty} f_n(x)$ *is integrable and*

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X f \, d\mu$$

Theorem I.6.3 (Dominated Convergence Theorem) *Let* $f_n : X \to \mathbb{C}$ *be a sequence of integrable functions* dominated *by an integrable function* $f : X \to \mathbb{R}$, *i.e.* $|f_n(x)| \leq f(x)$ *for all* n *and a.e.* x. *Then, if the sequence* $f_n(x)$, *converges for a.e.* x, *the limit function* $f(x) = \lim_{n\to\infty} f_n(x)$ *satisfies*

$$\int f \, d\mu = \lim_{n\to\infty} \int f_n \, d\mu$$

Given $p \geq 1$, denote by $\mathcal{L}^p(X, \mathcal{O}, \mu)$ the space of functions $f : X \to \mathbb{C}$ such that $|f|^p$ is integrable.

Theorem I.6.4 $\mathcal{L}^p(X, \mathcal{O}, \mu)$ *is a vector space of functions (i.e.* $f, g \in \mathcal{L}^p(X, \mathcal{O}, \mu) \Rightarrow \lambda f + \gamma g \in \mathcal{L}^p(X, \mathcal{O}, \mu)$ *for all* $\lambda, \gamma \in \mathbb{C}$*) and it becomes a Banach space when endowed with the norm*

$$\|f\|_p = \left(\int |f|^p \, d\mu \right)^{1/p}$$

In fact, $\|\cdot\|_p$ is not exactly a norm because $\|f\|_p = 0$ only implies $f(x) = 0$ for a. e. x. This “normal” problem is solved by considering, when necessary, $\mathcal{L}^p(X, \mathcal{O}, \mu)$ as the quotient space of the space of functions

f such that $|f|^p$ is integrable over the space of functions that are zero a.e.

A function $f : X \to \mathbb{C}$ is said to be $\mathcal{L}^\infty$ if it is measurable and there exists $K \geq 0$ such that $|f(x)| \leq K$ for a.e. x. Let $\|f\|_\infty$ denote the infimum of the constants K satisfying this property. Denote by $\mathcal{L}^\infty(X, \mathcal{O}, \mu)$ the set of $\mathcal{L}^\infty$ functions. Clearly, it is a vector space of functions, and on its quotient space over the subspace of functions that are zero a.e., $\|\cdot\|_\infty$ is a Banach norm.

Let X be a set and $\mathcal{O}$ a $\sigma-$algebra of subsets of X. If $\mu : \mathcal{O} \to [0, +\infty]$ and $\nu : \mathcal{O} \to [0, +\infty]$ are measures, we say that μ is *absolutely continuous* with respect to ν, and we write $\mu \ll \nu$, if $\nu(A) = 0$ implies $\mu(A) = 0$.

Theorem I.6.5 (Radon-Nikodym) *Let $(X, \mathcal{O}, \mu)$ be a measure space and $\nu : \mathcal{O} \to [0, +\infty)$ a measure satisfying $\mu \ll \nu$. If $(X, \mathcal{O}, \mu)$ is $\sigma-$finite (i.e. if there exists a countable covering of X by sets in $\mathcal{O}$ of finite μ-measure), then there exists a ν-integrable function $f : X \to \mathbb{R}^+$ such that for every $A \in \mathcal{O}$*

$$\mu(A) = \int_A f \, d\nu.$$

Moreover, a function $g : X \to \mathbb{C}$ is in $\mathcal{L}^1(X, \mathcal{O}, \mu)$ if and only if $gf \in \mathcal{L}^1(X, \mathcal{O}, \nu)$, and then

$$\int_X g \, d\mu = \int_X f g \, d\nu.$$

It is clear that the function f is essentially unique: if $f_1 : X \to \mathbb{R}$ is another function satisfying the requirements of the theorem, then $f_1 = f$ almost everywhere. The function f is called the *Radon-Nikodym derivative* of μ with respect to ν and denoted by $d\mu/d\nu$. When $\mu \ll \nu$ and $\nu \ll \mu$, we say that μ is *equivalent* to ν. In this case $d\mu/d\nu = (d\nu/d\mu)^{-1}$. a.e.

I.7 Existence of Invariant Measures

Let X be a compact metric space and $\mathcal{O}$ its Borel $\sigma-$algebra. In this section we shall first prove that the set $m(X)$ of probabilities $\mu : \mathcal{O} \to$

$[0,1]$ admits a unique topology that is metrizable and such that $\mu_n \to \mu$ if and only if

$$\int \varphi \, d\mu_n \to \int \varphi \, d\mu$$

for every continuous $\varphi : X \to \mathbb{R}$. Then we shall prove that with this topology $m(X)$ becomes a compact space and using this property we shall show that every continuous map $T : X \to X$ has at least one T-invariant probability, i.e. a probability $\mu \in m(X)$ such that $\mu(T^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{O}$. In fact the most frequent situation is when T has infinitely many invariant probabilities. When T has only finitely many invariant probabilities it is clear that is has only one (because if it has two, μ_1 and μ_2, then all the linear combinations $\lambda\mu_1 + (1-\lambda)\mu_2$, $0 \le \lambda \le 1$, would be invariant probabilities), and then T has extremely strong ergodic properties, as we shall see below.

Given a continuous map T, denote by $m_T(X)$ the set of T-invariant probabilities.

Proposition I.7.1 *$m_T(X)$ is non-empty.*

It will be necessary to introduce a topology in $m(X)$, defined by the following neighborhood basis:

$$V_{\varepsilon,\phi}(\mu) = \left\{ \nu \in m(X) : \left| \int_X \phi \, d\nu - \int_X \phi \, d\mu \right| \le \varepsilon \right\}$$

where $\varepsilon > 0$ and $\phi : X \to \mathbf{R}$ is a continuous function.

The following lemmas are simple but important:

Lemma I.7.2 *$m(X)$ is a compact metric space.*

Let $C^0(x)$ be the vector space of continuous functions $f : X \to \mathbb{R}$, endowed with the norm

$$\|f\|_0 = \sup_{x \in X} |f(x)|$$

Since X is a compact metric space, there exists a countable subset $\{g_i\}_{i>0}$ of $C^0(X)$ which is dense in the unit ball $B := \{f \in C^0(X) : \|f\|_0 \le 1\}$. Consider in $m(X)$ the metric

$$d(\mu, \nu) = \sum_{j=1}^{\infty} \frac{1}{2^j} \left| \int_X g_j \, d\mu - \int_X g_j \, d\nu \right|$$

We leave it to the reader to verify that $d(\cdot,\cdot)$ is indeed a metric. The proof of Lemma I.7.2 is based on the following fact:

Lemma I.7.3 *The following properties of a sequence* $\mu_n \in m(X)$ *are equivalent:*

(a) $\lim_{n\to\infty} d(\mu_n,\mu) = 0$;

(b) $\lim_{n\to\infty} \int_X g_j\, d\mu_n = \int_X g_j\, d\mu$ *for every* $j \geq 1$.

(c) $\lim_{n\to\infty} \int_X g\, d\mu = \int_X g\, d\mu$ *for every* $g \in C^0(X)$.

Proof. The implication (a) $\Rightarrow$ (b) follows from the inequality

$$\left|\int_X g_j\, d\mu_n - \int_X g_j\, d\mu\right| \leq 2^j d(\mu_n,\mu)$$

We now prove (b) $\Rightarrow$ (c). Given $g \in C^0(X)$ and $\varepsilon > 0$, let g_j be such that

$$\left\| g_j - \frac{g}{\|g\|}\right\| \leq \frac{\varepsilon}{3\|g\|}$$

the case $g = 0$ being trivial. Let n_0 be such that $n \geq n_0$ implies that

$$\left|\int_X g_j\, d\mu_n - \int_X g_j\, d\mu\right| \leq \frac{\varepsilon}{3\|g\|}$$

Then, for $n \geq n_0$

$$\begin{aligned}
\left|\int_X g\, d\mu_n - \int_X g\, d\mu\right| &\leq \|g\| \left|\int_X \frac{g}{\|g\|} - g_j\, d\mu_n\right| \\
&\quad + \|g\| \left|\int_X g_j\, d\mu_n - \int_X g_j\, d\mu\right| \\
&\quad + \|g\| \left|\int_X g_j - \frac{g}{\|g\|}\, d\mu\right| \\
&\leq \|g\| \frac{\varepsilon}{3\|g\|} + \|g\| \cdot \frac{\varepsilon}{3\|g\|} + \|g\| \frac{\varepsilon}{3\|g\|} \\
&= \varepsilon
\end{aligned}$$

We now prove (c) $\Rightarrow$ (a). Let j_0 be such that

$$\sum_{j=j_0+1}^{\infty} \frac{1}{2^j} \leq \frac{\varepsilon}{4}$$

Then

$$\begin{aligned} d(\mu_n, \mu) &\leq \sum_{j=1}^{j_0} \frac{1}{2^j} \left| \int_X g_j \, d\mu_n - \int_X g_j \, d\mu \right| \\ &\quad + \sum_{j=j_0+1}^{\infty} \frac{1}{2^j} \left(\int_X |g_j| \, d\mu_n + \int_X |g_j| \, d\mu \right) \\ &\leq \sum_{j=1}^{j_0} \frac{1}{2^j} \left| \int_X g_j \, d\mu_n - \int_X g_j \, d\mu \right| + 2 \cdot \frac{\varepsilon}{4} \end{aligned}$$

Since the sequence $\{\mu_n\}$ satisfies (c), we can find n_0 such that $n \geq n_0$ implies

$$\left| \int_X g_j \, d\mu_n - \int_X g_j \, d\mu \right| \leq \frac{\varepsilon}{2}$$

for every $1 \leq j \leq j_0$. This implies that for every $n \geq n_0$ we have

$$d(\mu_n, \mu) \leq \frac{\varepsilon}{2} + \sum_{j=1}^{j_0} \frac{1}{2^j} \left| \int_X g_j \, d\mu_n - \int_X g_j \, d\mu \right| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

The lemma is proved. □

Proof of Lemma I.7.2. We prove that $m(X)$ is metrizable by showing that the metric $d(\cdot, \cdot)$ generates the topology of $m(X)$. Let μ_n be a sequence such that $d(\mu_n, \mu)$ approaches 0; we will show that $\mu_n \to \mu$ in the topology of $m(X)$. Let $V_{\varepsilon,f}(\mu)$ be a neighborhood of μ defined as above. From Lemma I.7.3

$$\lim_{n \to \infty} \int_X f \, d\mu_n = \int_X f \, d\mu,$$

which implies that $\mu_n \in V_{\varepsilon,f}(\mu)$ for large values of n. This proves that every open set of $m(X)$ is open in the topology given by d.

Conversely, given $U \subset m(X)$ open with respect to d, we show that it is also open in the topology of $m(X)$ by finding for any $\mu \in U$ and $\varepsilon > 0$ a finite sequence $\{f_j\}_{1 \leq j \leq m}$, $f_j \in C^0(X)$, such that $\cap_{j=1}^{m} V_{\varepsilon, f_j}(\mu) \subset U$. Since U is open in the metric d, there exists $r > 0$ such that $d(\nu, \mu) \leq r$ implies $\nu \in U$. Let m be such that

$$\sum_{j=m+1}^{\infty} \frac{1}{2^j} \leq \frac{r}{4}$$

If we take $f_j = g_j$ and $\varepsilon = r/2$, we are done, since

$$\nu \in \cap_{j=1}^{m} V_{\varepsilon, g_j}(\mu)$$

implies that

$$\begin{aligned} d(\nu, \mu) &\le \sum_{j=1}^{m} \frac{1}{2^j} \left| \int_X g_j \, d\mu - \int_X g_j \, d\nu \right| \\ &+ \sum_{j=m+1}^{\infty} \frac{1}{2^j} \left(\int_X |g_j| \, d\mu + \int_X |g_j| \, d\nu \right) \\ &\le \sum_{j=1}^{m} \frac{1}{2^j} \cdot \frac{r}{2} + \sum_{j=m+1}^{\infty} \frac{1}{2^j} \cdot 2 \\ &\le \frac{r}{2} + 2 \cdot \frac{r}{4} = r \end{aligned}$$

and hence that $\nu \in U$.

The proof of compactness requires the following theorem (see [Rd] for a proof):

Theorem I.7.4 (Riesz Representation Theorem) *Let* $\phi : C^0(X) \to \mathbb{R}$ *be a positive linear functional, i.e. a linear map such that* $\phi(f) \ge 0$ *if* $f \ge 0$ *and* $\phi(1) = 1$, *where* $1 \in C^0(X)$ *is the constant function equal to 1. Then there exists a unique* $\mu \in m(X)$ *such that*

$$\int_X f \, d\mu = \phi(f)$$

for all $f \in C^0(X)$.

Since $m(X)$ is metrizable, proving compactness is equivalent to proving that every sequence $\{\mu_n\}_{n \ge 1}$, $\mu_n \in m(X)$, has a convergent subsequence. We associate to each μ_n a sequence of numbers $\{\mu_n(j)\}_{j \ge 1}$ defined by

$$\mu_n(j) = \int_X g_j \, d\mu_n$$

Note that $\mu_n(j) \in [-1, 1]$. By a standard diagonal argument, there exists a subsequence $\{\mu_{n_m}\}_{m \ge 1}$ such that $\{\mu_{n_m}(j)\}_{m \ge 1}$, converges for every j, i.e. for every j the sequence

$$\left\{ \int_X g_j \, d\mu_{n_m} \right\}_{m \ge 1}$$

converges. Using the same reasoning as in the proof of the implication (b) $\Rightarrow$ (c) of Lemma I.7.3, we conclude that

$$\left\{ \int_X g \, d\mu_{n_m} \right\}_{m \geq 1}$$

converges for every $g \in C^0(X)$. Let $\phi : C^0(X) \to \mathbb{R}$ be defined by

$$\phi(g) = \lim_{m \to \infty} \int_X g \, d\mu_{n_m}$$

It is easy to see that ϕ is a positive linear functional. Then, by Theorem I.7.4, there exists $\nu \in m(X)$ such that, for every $g \in C^0(X)$,

$$\phi(g) = \int_X g \, d\nu$$

Lemma I.7.2 is proved. $\Box$

Proof of Proposition I.7.1. Given a continuous map $T : X \to X$, we define $T^* : m(X) \to m(X)$ by

$$(T^* \mu)(A) = \mu(T^{-1}(A))$$

for every Borel set $A \subset X$; this is a continuous map. The proposition will be proven if we find $\mu \in m(X)$ such that $T^* \mu = \mu$. Take any $\mu_0 \in m(X)$, and consider the sequence $\{\mu_n\}_{n \geq 0}$ defined by

$$\mu_n = \frac{1}{n+1} \sum_{m=0}^{n} (T^*)^m \mu_0$$

By Lemma I.7.2, we can find a convergent subsequence $\{\mu_{n_j}\}_{j \geq 1}$ and take $\mu = \lim_{j \to \infty} \mu_{n_j}$. Then

$$\begin{aligned} T^* \mu_{n_j} &= \frac{1}{n_j + 1} \sum_{m=0}^{n_j} (T^*)^{m+1} \mu_0 \\ &= \frac{1}{n_j + 1} \sum_{m=0}^{n_j} (T^*)^m \mu_0 - \frac{1}{n_j + 1} \mu_0 + \frac{1}{n_j + 1} (T^*)^{n_j + 1} \mu \end{aligned}$$

The last two terms converge to 0 as $j \to \infty$. Thus

$$T^* \mu = \lim_{j \to \infty} T^* \mu_{n_j} = \lim_{j \to \infty} \frac{1}{n_j + 1} \sum_{m=0}^{n_j} (T^*)^m \mu_0 = \lim_{j \to \infty} \mu_{n_j} = \mu$$

The proposition is proved. □

It may happen that a continuous map of a compact metric space has exactly one invariant probability. Such maps are called *uniquely ergodic.* The class of uniquely ergodic maps is very restricted. The standard examples are certain translations of $\mathbb{T}^n$.

Theorem I.7.5 *Let f be a translation of $\mathbb{T}^n$. Then the following properties are equivalent*

(a) *Denoting $e = (1, \ldots, 1)$, the orbit $\{f^n(e) | n \in \mathbb{Z}\}$ is dense in $\mathbb{T}^n$*

(b) $\{f^n(x) | n \in \mathbb{Z}\}$ *is dense in $\mathbb{T}^n$ for every $x \in \mathbb{T}^n$*

(c) f *is uniquely ergodic.*

Proof. The implication (b) $\Rightarrow$ (a) is trivial.

(a) $\Rightarrow$ (c) Let $\mu \in m_f(\mathbb{T}^n)$. We shall prove that μ is invariant under all the translations of $\mathbb{T}^n$. This property, as proved in Section I.3, implies that μ is the Lebesgue probability. Let g be any translation. Take a sequence of positive integers $n_j,\ j \geq 1$, such that

$$\lim_{j \to \infty} f^{n_j}(e) = g(e)$$

Such a sequence exists by (a). But, since f and g are translations

$$\begin{aligned} d(f^{n_j}(x), g(x)) &= d(x \cdot f^{n_j}(e), x \cdot g(e)) \\ &= d(f^{n_j}(e), g(e)) \end{aligned}$$

Hence $f^{n_j} \to g$ uniformly. Now it is easy to see that $(f^{n_j})^* \nu \to g^* \nu$ for all $\nu \in m(\mathbb{T}^n)$. Hence

$$g^* \mu = \lim_{j \to \infty} (f^{n_j})^* \, \mu = \lim_{j \to \infty} (f^*)^{n_j} \mu = \mu$$

(c) $\Rightarrow$ (b). Let f be uniquely ergodic. We shall show that if X denotes the closure of $\{f^n(x) : n \in \mathbb{Z}\}$, for some $x \in \mathbb{T}^n$, then $X = \mathbb{T}^n$. Observe first that $f(X) = X$, so we may consider $f|_X : X \to X$. On the other hand, X is compact, and so it follows from Proposition I.7.1 that there exists an $(f|_X)-$invariant invariant probability ν defined on the Borel subsets of X. We extend ν to a probability μ on the Borel $\sigma-$algebra of $\mathbb{T}^n$ by defining

$$\mu(A) = \nu(A \cap X),$$

for every Borel subset $A \subset \mathbb{T}^n$. By the invariance of X and ν we have

$$\begin{aligned}\mu(f^{-1}(A)) &= \nu(f^{-1}(A) \cap X) \\ &= \nu((f|_X)^{-1}(A \cap X)) \\ &= \nu(A \cap X) \\ &= \mu(A)\end{aligned}$$

so μ is f-invariant. Since we assume f to be uniquely ergodic, $\mu = \lambda$, the Lebesgue probability on $\mathbb{T}^n$. Then

$$\lambda(X) = \mu(X) = \nu(X) = 1 = \lambda(\mathbb{T}^n)$$

and so $\lambda(\mathbb{T}^n \setminus X) = 0$. Since λ is positive on non-empty open sets, we obtain $\mathbb{T}^n \setminus X = \emptyset$, i.e. $X = \mathbb{T}^n$. □

The following theorem characterizes uniquely ergodic maps in terms of orbital averages of continuous functions.

Theorem I.7.6 *Let X be a compact metric space and $T : X \to X$ a continuous map. The following properties are equivalent:*

(a) T *is uniquely ergodic;*

(b) *For every $f \in C^0(X)$ the limit*

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{j=0}^{n} f(T^j(x))$$

exists for every $x \in X$ and does not depend on x.

(c) *For every $f \in C^0(X)$ the sequence of functions*

$$\frac{1}{n+1} \sum_{j=0}^{n} f \circ T^j \tag{I.7.1}$$

converges uniformly to a constant.

Proof. (a) ⇒ (c). If (c) does not hold, there exists $f \in C^0(X)$ such that the sequence (I.7.1) does not converge uniformly to

$$\int_X f \, d\mu$$

where μ is the only element of $m_T(X)$. Then there exist $\varepsilon > 0$, a sequence of integers $\{n_i\}_{i\geq 1}$ and a sequence $\{x_i\}_{i\geq 1}$ of points of X such that

$$\left| \frac{1}{n_i+1} \sum_{j=0}^{n_i} f(T^j(x_i)) - \int_X f\, d\mu \right| \geq \varepsilon$$

for every i. Let $\mu_{n_i} \in m(X)$ be such that

$$\int_X g\, d\mu_{n_i} = \frac{1}{n_i+1} \sum_{j=0}^{n_i} g(T^j(x_i))$$

for every $g \in C^0(X)$; the existence of such a measure is guaranteed by the Riesz representation Theorem I.7.4. Since $m(X)$ is compact by Lemma I.7.2, we can assume that the sequence $\{\mu_{n_i}\}$ converges to a measure $\nu \in m(X)$. We now prove that $\nu \in m_T(X)$. Let $g \in C^0(X)$; then

$$\begin{aligned}
\int_X (g \circ T)\, d\nu &= \lim_{n_i\to\infty} \int_X (g\circ T)\, d\mu_{n_i} \\
&= \lim_{n_i\to\infty} \frac{1}{n_i+1} \sum_{j=0}^{n_i} g(T^{j+1}(x_i)) \\
&= \lim_{n_i\to\infty} \int_X g\, d\mu_{n_i} - \lim_{n_i\to\infty} \frac{g(x_i)}{n_i+1} + \lim_{n_i\to\infty} \frac{g(T^{n_i+1}(x_i))}{n_i+1} \\
&= \int_X g\, d\nu,
\end{aligned}$$

so that $\nu \in m_T(X)$. But

$$\begin{aligned}
\left| \int_X f\, d\nu - \int_X f\, d\mu \right| &= \lim_{n_i\to\infty} \left| \int_X f\, d\mu_{n_i} - \int_X f\, d\mu \right| \\
&= \lim_{n_i\to\infty} \left| \frac{1}{n_i+1} \sum_{j=0}^{n_i} f(T^j(x_i)) - \int_X f\, d\mu \right| \\
&\geq \varepsilon
\end{aligned}$$

and we get $\nu \neq \mu$, contradicting that T is uniquely ergodic.

(c) $\Rightarrow$ (b). This is trivial.

(b) $\Rightarrow$ (a). Let $\phi : C^0(X) \to \mathbb{R}$ be the functional defined by

$$\phi(f) = \lim_{n\to\infty} \frac{1}{n+1} \sum_{j=0}^{n} f(T^j(x))$$

Then for $\mu \in m_T(X)$,

$$\int_X f\, d\mu = \frac{1}{n+1} \sum_{j=0}^{n} \int_X (f \circ T^j), d\mu$$

because

$$\int_X f\, d\mu = \int_X (f \circ T^j)\, d\mu$$

for every j. Since the sequence $\frac{1}{n+1}\sum_{j=0}^{n} f(T^j(x))$ is bounded by $\|f\|_0$, it follows from the dominated convergence theorem that

$$\begin{aligned}
\int_X f\, d\mu &= \lim_{n\to\infty} \frac{1}{n+1} \sum_{j=0}^{n} \int_X (f \circ T^j)\, d\mu \\
&= \int_X \lim_{n\to\infty} \frac{1}{n+1} \sum_{j=0}^{n} (f \circ T^j)\, d\mu \\
&= \int_X \phi(f)\, d\mu \\
&= \phi(f)
\end{aligned}$$

Thus, the only element of $m_T(X)$ is the measure associated with the positive linear functional ϕ. $\Box$

Exercises:

I.7.1 Let $f : S^1 \to S^1$ be a homeomorphism with fixed points. Prove that a point of S^1 is recurrent if and only if it is a fixed point of f^n for some $n \in \mathbb{Z}$. Deduce that if f has a finite set of recurrent points, then every probability measure invariant under f is a convex linear combination of Dirac delta measures concentrated at fixed points of f^n, $n \in \mathbb{Z}$.

I.7.2. Let $f : [0,1] \times S^1 \to S^1$ be a continuous map such that for every $t \in [0,1]$ the map $f_t(\cdot) = f(t,\cdot)$ is a homeomorphism $S^1 \to S^1$. Is there a continuous map $\mu : [0,1] \to m(S^1)$ such that $\mu(t)$ is f-invariant for every $0 \le t \le 1$?

I.7.3. Let X be a compact metric space and $T : X \to X$ a continuous map. If A is a subset of X, we put

$$\tau(x,A) := \limsup_{n\to\infty} \frac{1}{n} \#\{0 \le j < n|\ T^j(x) \in A\}$$

Prove that for every compact set $U \subset X$ and every $x \in X$ there exists $\mu \in m_T(X)$ such that

$$\mu(U) \ge \tau(x,U)$$

Hint: Let $\{n_i\}_{i\ge1}$ be a sequence of integers such that

$$\lim_{i\to\infty} \frac{1}{n_i} \#\{0 \le j < n_i|\ T^j(x) \in A\} = \tau(x,U)$$

Prove that one can assume that the sequence

$$\frac{1}{n_i} \sum_{j=0}^{n_i-1} \delta_{T^j(x)}$$

converges in $m(X)$. Prove that its limit, call it μ, is T-invariant and that

$$\int_X \phi\, d\mu \ge \tau(x,U)$$

for every continuous function $\phi : X \to [0,1]$ which takes value 1 on U.

I.7.4. Let X be a compact metric space and μ a Borel probability on X.

(a) If $U \subset X$ is open and $\mu(\partial U) = 0$, then for every $\varepsilon > 0$ there exists a function $f \in C^0(X)$ such that $f \ge \chi_U$ and $\int_X f\, d\mu \le \mu(U) + \varepsilon$.

(b) If $f \in C^0(X)$ and $\varepsilon > 0$, then there exists $g = \sum_{i=1}^{\infty} \lambda_i \chi_{U_i}$, where U_i are open sets with $\mu(\partial U_i) = 0$, such that $\int_X |f - g|\, d\mu < \varepsilon$.

(c) A sequence $\mu_n \in m(X)$, converges to μ if and only if $\mu_n(U)$ converges to $\mu(U)$ for every open set $U \subset X$ with $\mu(\partial U) = 0$.

I.7.5. If X is a compact metric space and $T : X \to X$ is a continuous map, prove that

$$\sup_{\mu \in m_T(X)} \int_X \phi \, d\mu = \sup_{x \in X} \limsup_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} \phi(T^j(x))$$

for every continuous function $\phi : X \to \mathbb{R}$. Hint: Observe that

$$\int_X \phi \, d\mu = \int_X \frac{1}{n} \sum_{j=0}^{n-1} \phi(T^j(x)) \, d\mu$$

I.8 The Equivalence Problem

The natural notion of equivalence between two measure preserving maps is given by the following definition.

Definition. We say that two measure preserving maps T_i, $i = 1, 2$, of two measure spaces $(X_i, \mathcal{O}_i, \mu_i)$, $i = 1, 2$, respectively, are *equivalent* if there exists a measure preserving map F taking $(X_1, \mathcal{O}_1(\text{mod } 0), \mu_1)$ into $(X_2, \mathcal{O}_2(\text{mod } 0), \mu_2)$ satisfying

(a) F is invertible, i.e. there exists a measurable map $G : X_2 \to X_1$ such that $GF(x) = x$ for a.e. $x \in X_1$ and $FG(y) = y$ for a.e. $y \in X_2$.

(b) F preserves measure, i.e. $\mu_1(F^{-1}(A)) = \mu_2(A)$ (mod 0) for every Borel $A \subset X_2$.

(c) $T_2 F = F T_1$ for a.e. $x \in X_1$.

Observe that (a)-(b) imply that G is a measure preserving map of $(X_2, \mathcal{O}_2, \mu_2)$ into $(X_1, \mathcal{O}_1, \mu_1)$, and, by (c), $GT_2 = T_1 G$ almost everywhere.

Hence, the equivalence is symmetric. Clearly it is transitive and reflexive, so it is a true equivalence relation.

One of the aims of ergodic theory is to classify measure preserving maps modulo this equivalence relation. One of the methods for this analysis consists in associating with a measure preserving map $T : (X, \mathcal{O}, \mu) \to (Y, \mathcal{S}, \nu)$ a linear operator $U_T : \mathcal{L}^2(Y) \to \mathcal{L}^2(X)$ defined by

$$U_T f = f \circ T.$$

The fact that T preserves measure implies that U_T is a unitary operator, i.e., denoting by $\langle \cdot, \cdot \rangle$ the inner product in $\mathcal{L}^2$ we have

$$\langle U_T f, U_T g \rangle = \langle f, g \rangle$$

for every $f, g \in \mathcal{L}^2(Y)$.

Definition. Let $(X_i, \mathcal{O}_i, \mu_i)$, $i = 1, 2$, be measure spaces and $T_i : X_i \to X_i$ measure-preserving maps, with the associated linear operators U_{T_i}. We say that T_1 and T_2 are *spectrally equivalent* if there exists an invertible isometry

$$L : \mathcal{L}^2(X_2) \to \mathcal{L}^2(X_1)$$

such that $LU_{T_2} = U_{T_1} L$.

If T_1 and T_2 are equivalent, they are spectrally equivalent, since the map $F : X_1 \to X_2$ given in the definition of equivalence gives rise to an isometry $U_F : \mathcal{L}^2(X_2) \to \mathcal{L}^2(X_1)$ which satisfies the condition $U_F U_{T_2} = U_{T_1} U_F$ (just take $U_F f = f \circ F$).

In general, however, spectrally equivalent maps are not necessarily equivalent; for example, all Bernoulli shifts are spectrally equivalent, but Kolmogorov proved in 1958 that they are not all equivalent. He did this by associating to each measure-preserving map $T : X \to X$ a real number $h(T) \in [0, +\infty]$, called the entropy of T which is an invariant under equivalence (meaning that all equivalent maps have the same entropy). The entropy of a Bernoulli shift $B(p_1, \ldots, p_n)$ is equal to $-\sum_{i=1}^n p_i \log p_i$. Thus, in particular, the shifts $B(1/2, 1/2), B(1/3, 1/3, 1/3)$, and $B(1/4, 1/4, 1/2)$ have different entropies and so cannot be equivalent.

In the case of Bernoulli shifts, the converse also holds: Two Bernoulli shifts with the same entropy are equivalent. This remarkable result was proved by Ornstein in 1970. Thus we have the following.

Theorem I.8.1 *Two Bernoulli shifts are equivalent if and only if they have same entropy.*

Exercises:

I.8.1. Prove that if Bernoulli shifts $B^+_\nu(p_1, \ldots, p_m)$ and $B^+_\mu(q_1, \ldots, q_\ell)$ are equivalent, then $m = \ell$ and $\{p_1, \ldots, p_m\}$ can be indexed so that $p_i = q_i$ for all $1 \leq i \leq m = \ell$. Hint: Put $A_i = \{\theta \in B^+(p_1, \ldots, p_m) | \, \theta(0) =$

$i\}$ and $C_j = \{\theta \in B^+(q_1, \ldots, q_\ell) |\, \theta(0) = j\}$. (a) Prove that the set A_i has the following property: $\sigma|_{A_i}$ is a bijection between A_i and $B^+(p_1, \ldots, p_m)$ and $\nu(\sigma(S)) = p_i^{-1}\nu(S)$ for all Borel subsets $S \subset A_i$. (b) Then, if $T : B^+_\nu(p_1, \ldots, p_m) \to B^+_\mu(q_1, \ldots, q_\ell)$ realizes the equivalence between the two shifts, show that $T(A_i)$ contains a Borel subset D such that $\sigma|_D$ is a bijection between D and a full measure subset of $B^+(q_1, \ldots, q_\ell)$ and $\mu(\sigma(S)) = p_i^{-1}\mu(S)$ for all $S \subset D$. (c) Deduce that there exists j such that $q_j = p_i$ and $T(A_i) = C_j$ (mod 0). Assume first that the q_j's are all distinct.

I.8.2. Let $f : [0, 1] \to [0, 1]$ be the map given in Exercise I.4.2. Prove that f is equivalent to $\sigma : B^+(p_1, p_2, \ldots) \to B^+(p_1, p_2, \ldots)$ with $p_i = b_i - a_i$. Hint: Define a map $T : [0, 1] \to B^+(p_1, p_2, \ldots)$ setting $(Tx)(n) = j$ if $f^n(x) \in (a_j, b_j)$.

I.9 Entropy

The entropy is an important numerical characteristic of any measure preserving map. The notion of entropy has some long history, it came from other branches of physics and sciences. Originally, R. Klausius introduced the entropy in 1864 to describe the transformation of heat energy into kinetic energy. Nowadays the entropy in physics is commonly regarded as a measure of complexity of systems with a large number of components and possible configurations.

In 1948, C. Shannon introduced the entropy into information theory as a quantitative characteristic of uncertainty (or information) for random events. Let $(X, \mathcal{A}, \mu)$ be a probability space. We want to measure the amount of information added by knowing that an event $A \in \mathcal{A}$ has actually occurred. Clearly such a measure should be a function, i, of the probability $\mu(A)$, and

a) be nonnegative, i.e. $i : [0, 1] \to [0, +\infty))$,
b) be zero if the event has probability one, i.e. $i(1) = 0$,
c) increase when $\mu(A)$ decreases, and
d) satisfy the following independence relation: if $A, B \in \mathcal{A}$ are independent events (i.e., $\mu(A \cap B) = \mu(A)\mu(B)$), the information added by knowing that both A and B occur (i.e., $A \cap B$ occurs) is equal to the sum of information given by knowing that A and B have occurred.

It can be proved that there is a unique function that satisfies these conditions: $i(p) = -k \log p$, where $k > 0$ is a constant, see Exercise I.9.1. If $k = 1/\log 2$, the unit of information is called "bit" (it is used in information theory).

Let now $A_1, \ldots, A_n$ be random events of which one and only one can actually occur (so, they are mutually exclusive). Let $p_i = \mu(A_i)$, $1 \le i \le n$, be their probabilities, and of course, $p_1 + \cdots + p_n = 1$. Then the amount of information added by knowing which one of these events actually occurs is the mean (expected value) of the information defined above:

$$h(p_1, \ldots, p_n) = -p_1 \log p_1 - \cdots - p_n \log p_n \tag{I.9.1}$$

Now this quantity is called the *entropy* of the probability distribution $\{p_1, \ldots, p_n\}$. We note that $h \ge 0$. If n is fixed, h attains its maximum value $h = \log n$ at the uniform distribution $p_1 = \cdots = p_n = 1/n$, see Exercise I.9.2. In this case the number of events, n, can be expressed by $n = e^h$.

What is the meaning of h in a general case, when n is very large and the distribution is very nonuniform, so that $h \ll \log n$? The answer is given by the classical Shannon-McMillan-Breiman theorem in information theory, which we describe somewhat loosely. It says that we can divide the set of events $\{A_1, \ldots, A_n\}$ into two groups: essential events, whose probabilities are large enough, and negligible events (the others), so that (i) the negligible events can be removed from the expression (I.9.1) without a significant loss for the value of h, and (ii) the number of essential events is approximately

$$n_{\text{ess}} \approx e^h \tag{I.9.2}$$

This is the basic meaning of the entropy in information theory: e^h gives, essentially, the number of possible events.

In 1958, based on Shannon's ideas, A. N. Kolmogorov introduced the entropy into the theory of dynamical systems. His version of entropy, developed jointly with Ya. G. Sinai, is now referred to as the measure-theoretic entropy or the Kolmogorov-Sinai entropy.

Let $T : X \to X$ be a map preserving a probability measure μ. We are going to measure the complexity of the map T^N as $N \to \infty$. What is the complexity? Let $X = A_1 \cup \cdots \cup A_n$ be a partition of X into disjoint measurable subsets, $A_i \cap A_j = \emptyset$ as $i \neq j$. If we label each set A_i by i, then a phase point x can be coded by a label i if $x \in A_i$. The complexity of the partition $\xi = \{A_1, \dots, A_n\}$ is measured by its entropy defined by Shannon's formula, i.e.

$$h(\xi) = -\mu(A_1) \log \mu(A_1) - \cdots - \mu(A_n) \log \mu(A_n)$$

Now the map T enters the construction. Given a point $x \in A_{i_0}$ (with a label i_0), the point $T^k x$ for a $k \geq 1$ can be labeled by i_k if $T^k x \in A_{i_k}$. Hence, the orbit $\{x, Tx, \dots, T^N x\}$ can be labeled by a string $\{i_0, i_1, \dots, i_N\}$. One can easily see that this string will actually label all the points in the set $A_{i_0} \cap T^{-1} A_{i_1} \cap \cdots \cap T^{-N} A_{i_N}$. All such sets (with a fixed N) make a partition of X that we denote by ξ_N. Its complexity is again measured by the entropy $h(\xi_N)$ defined by the same Shannon's formula.

To understand what the partition ξ_N looks like, for each $j \geq 1$ consider a partition of X given by $T^{-j}\xi = \{T^{-j}A_1, \dots, T^{-j}A_n\}$. Due to the invariance of μ under T we obviously have $h(T^{-j}\xi) = h(\xi)$ for all $j \geq 1$. Now ξ_N is obtained by taking all possible intersections of the elements of $T^{-j}\xi$, $0 \leq j \leq N$. This construction is called the *product of partitions* and denoted by $\bigvee_{j=0}^{N} T^{-j}\xi$. Note that if $N \geq M$, each set in ξ_N is contained in ξ_M and each element of ξ_M is the union of elements of ξ_N; so, ξ_N is a *refinement* of ξ_M.

The partition ξ_N consists of n^{N+1} sets, called *elements* or *atoms* (some of them may be empty, in this case we put 0 instead of $0 \log 0$). So, its maximal entropy cannot exceed $h(\xi_N) \leq \log n^{N+1} = (N+1) \log n$, i.e. $h(\xi_N)$ grows at most linearly in time N. In fact, this estimate is rather optimistic, as for many maps most of the elements of ξ_N will be just empty, and $h(\xi_N)$ will grow slowly, if at all, see Exercise III.9.3. But for many other maps there will be a linear growth of $h(\xi_N)$, indeed, see Exercise I.9.4.

The quantity

$$h(T, \xi) = \lim_{N \to \infty} \frac{1}{N+1} h(\xi_N)$$

is called the *entropy of the map T with respect to the partition ξ*. This limit always exists and is equal to $\inf_{N \geq 1} (N+1)^{-1} h(\xi_N)$, because the sequence $(N+1)^{-1} h(\xi_N)$ monotonically decreases with N, but we omit a proof of that. In particular, for every N we have $h(T, \xi) \leq (N+1)^{-1} h(\xi_N) \leq h(\xi)$.

The value $h(T, \xi)$ admits the following interpretation in terms of information theory. We want to code the information about the state of the system (X, T, μ) at every future iteration j by using the partition ξ. The itinerary $\{i_0, i_1, \dots, i_N\}$ of a given point x defined above codes an atom of the partition ξ_N that contains x. Since $h(\xi_N)$ is the (expected) amount of information given by knowing the itinerary of the point x during the first $N+1$ iterations, the value $h(\xi_N)/(N+1)$ is the mean speed of data transmission required to code the evolution of the system. Hence, $h(T,\xi)$ is the asymptotic speed of transmission of data as $N \to \infty$. More specifically, the amount of bits transferred per unit time is $h(T,\xi)/\log 2$. If this quantity is positive, the system is complex enough so that the description of its evolution during time N requires an amount of computer memory that must grow linearly with N. No finite amount of information about this system can describe its evolution in the entire future. Practically, this means that the future cannot be predicted, i.e. such a system is *deterministic* but *unpredictable*.

As it follows from our previous estimates, $h(T,\xi) \leq h(\xi)$, i.e. $h(T,\xi)$ is always finite, but may be unbounded as a function of ξ.The quantity

$$h_\mu(T) = \sup_\xi h(T,\xi)$$

where the supremum is taken over all finite measurable partitions of X is called the *measure-theoretic entropy* of the map T. This is the Kolmogorov-Sinai entropy.

We emphasize the meaning of the entropy. By the formula (I.9.2), the entropy $h(T,\xi)$ measures the exponential rate of growth of the number of essential elements in the partition ξ_N as $N \to \infty$ (those elements that make essential contribution to the quantity $h(\xi_N)$). One can think of $h(T)$ then as the exponential rate of growth of the complexity of the map T^N as $N \to \infty$.

There is a practically useful theorem by Kolmogorov and Sinai that simplifies the calculation of entropy. We need some definitions. If $\{\mathcal{P}_n\}_{n\geq 1}$ is a sequence of partitions, we denote by $\bigvee_{n\geq 1} \mathcal{P}_n$ the minimum σ-algebra that contains all the atoms of all these partitions. In other words, it is σ-algebra generated by $\bigcup_{n\geq 1} \mathcal{P}_n$. If T is a measurable map on the measure space $(X, \mathcal{A}, \mu)$, we say that $\mathcal{P}$ is a *generating partition* if it satisfies one of the following conditions:

i) either $\bigvee_{j=0}^{\infty} T^{-j}\mathcal{P} = \mathcal{A}$ (up to subsets of measure zero)

ii) or $\bigvee_{j=0}^{\infty} T^{-j}\mathcal{P} \neq \mathcal{A}$, but T is invertible with a measurable inverse and $\bigvee_{j=-\infty}^{\infty} T^{-j}\mathcal{P} = \mathcal{A}$

Then, the Kolmogorov-Sinai theorem states that if $\mathcal{P}$ is a generating partition with $h(\mathcal{P}) < \infty$, then

$$h_\mu(T) = h(T, \mathcal{P}).$$

Exercises:

I.9.1. Prove that the only function $i : [0,1] \to [0, +\infty)$ that satisfies (a)-(d) in the beginning of this section is $i(p) = -k \log p$, with a constant $k > 0$. Note that the condition (d) means $i(pq) = i(p) + i(q)$ for all $0 \le p, q \le 1$. Hint: change variable $x = -\log p$, then the function $f(x) = i(e^{-x})$ is increasing and satisfies $f(0) = 0$ and $f(x+y) = f(x) + f(y)$ for all $x, y > 0$. Denote $f(1) = a$ and show that $f(m/n) = am/n$ for all $m, n \in \mathbb{N}$. Then use the monotonicity of f to prove that $f(x) = ax$ for all $x > 0$.

I.9.2. Show that $h(p_1, \dots, p_n) \le \log n$ and the equality takes place if and only if $p_1 = \cdots = p_n = 1/n$. Hint: the function $g(x) = -x \log x - (a - x) \log(a - x)$, where $a > 0$ is given, takes its sole maximum at $x = a/2$.

I.9.3. Let $T : S^1 \to S^1$ be a rotation of the circle S^1 through a fixed angle $\theta > 0$. Show that $h(T, \xi) = 0$ for any finite partition of S^1 into arcs. Hint: let n be the number of arcs. Verify that ξ_N is a partition of S^1 into no more than $n(N + 1)$ arcs.

I.9.4. Consider a Bernoulli shift $B_\mu(p_1, \dots, p_m)$ and a partition ξ into the sets $A_i = \{\theta \in B(m) \colon \theta(0) = i\}$, for $1 \le i \le m$. Show that

$$h(\sigma, \xi) = h(p_1, \dots, p_m) = -p_1 \log p_1 - \cdots - p_m \log p_m$$

Hint: verify by induction that $h(\xi_N) = N h(p_1, \dots, p_m)$ for all $N \ge 1$.

Chapter II

Ergodicity

II.1 Ergodic Theorem of Birkhoff-Khinchin

Birkhoff-Khinchin theorem deals with the distribution of the orbits of a measure preserving map T of a probability space $(X, \mathcal{O}, \mu)$. In order to study how an orbit $\{x, T(x), T^2(x), \ldots\}$ is asymptotically distributed in X we introduce the *sojourn time* of x in a set $A \in \mathcal{O}$ by

$$\tau(x, A) = \lim_{n \to \infty} \frac{1}{n} \#\{0 \leq m < n | \, T^m(x) \in A\}$$

Birkhoff's Theorem states that this limit exists for a.e. x and that $\tau(x, A)$ is an integrable function of x whose integral is given by

$$\int_X \tau(x, A) \, d\mu(x) = \mu(A)$$

Moreover, one can easily check that, as a function of x, τ is T-invariant, i.e.

$$\tau(x, A) = \tau(T(x), A) \qquad \text{a.e.}$$

This motivates the following definition: we say that T is ergodic if all the T-invariant functions are constant a.e. Then τ must be constant and its integral becomes just its value a.e. Then, for a.e. x we have:

$$\tau(x, A) = \mu(A)$$

This is a remarkable conclusion and poses the problem of developing methods to decide when a map T is ergodic. Ergodicity is a strong property and many important transformations, as for instance those arising in Hamiltonian Mechanics, are frequently not ergodic. On the other hand, as we shall see in this chapter, there are plenty of important classes of ergodic transformations. Moreover, when X is a compact metric space and $T : X \to X$ is a continuous map, there always exist ergodic T-invariant probabilities on the Borel $\sigma-$algebra of X. They are important in the analysis of the dynamics of T.

An interesting example of an ergodic measure preserving map is the map $T : [0,1] \to [0,1]$ given by $T(x) = 10x - [10x]$ (sometimes denoted by $\{10x\}$, where $\{\cdot\}$ stands for the fractional part of a number). This map preserves the Lebesgue probability on [0,1] and is ergodic (for a proof see Ch. III.1. in [Mn]). An direct consequence of its ergodicity is the following important fact in number theory. Write $x \in [0,1]$ in decimal representation $x = 0.a_0a_1a_2\dots$ and let $N_n(x,j)$ be the number of times that the digit $0 \le j \le 9$ appears in the string $[a_0 \dots a_{n-1}]$. Then, for a.e. $x \in [0,1]$

$$\lim_{n\to\infty} \frac{1}{n} N_n(x,j) = \frac{1}{10}$$

This is a consequence of the ergodicity of T because $a_m = j$ if and only if $T^m(x) \in [j/10, (j+1)/10)$. Then, for a.e. x :

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} N_n(x,j) &= \tau(x, [j/10, (j+1)/10)) \\
&= \lambda([j/10, (j+1)/10)) \\
&= 1/10.
\end{aligned}$$

More subtle is a similar property for continued fractions. Every irrational number $x \in (0,1)$ can be written in a unique way as a continued fraction

$$x = \cfrac{1}{a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots}}}}$$

where $a_0, a_1, \dots$ are positive integers. Let $P_n(x,k)$ be the number of times that k appears among $a_0, \dots, a_{n-1}$. Then, for a.e. $x \in (0,1)$

$$\lim_{n\to\infty} \frac{1}{n} P_n(x,k) = \frac{1}{\log 2} \log(1 + \frac{1}{k(k+2)})$$

The proof of this property requires first transforming $P_n(x,k)$ into a sojourn time. This is done with the help of the Gauss map $T : [0,1] \to [0,1]$ defined by

$$T(x) = \begin{cases} \frac{1}{x} - \left[\frac{1}{x}\right] & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Observe that

$$\frac{1}{x} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots}}}$$

where

$$a_0 = \left[\frac{1}{x}\right]$$

and

$$T(x) = \frac{1}{x} - \left[\frac{1}{x}\right] = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots}}}$$

Repeating this argument gives

$$T^n(x) = \cfrac{1}{a_n + \cfrac{1}{a_{n+1} + \cfrac{1}{\ddots}}}$$

Then

$$\frac{1}{T^n(x)} = a_n + \cfrac{1}{a_{n+1} + \cfrac{1}{\ddots}}$$

thus implying

$$a_n = \left[\frac{1}{T^n(x)}\right]$$

Observe that

$$a_n = k \iff T^n(x) \in \left(\frac{1}{k+1}, \frac{1}{k}\right]$$

Hence, if the Lebesgue probability were T-invariant, we would have

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} P_n(x,k) &= \lim_{n\to\infty} \frac{1}{n} \# \left\{ 0 \le j < n | \, T^j(x) \in \left(\frac{1}{k+1}, \frac{1}{k}\right] \right\} \\ &= \tau\left(x, \left(\frac{1}{k+1}, \frac{1}{k}\right]\right) \end{aligned}$$

for almost every x. However, the Lebesgue probability is **not** T-invariant. Fortunately, there is another probability on the Borel $\sigma-$algebra, discovered by Gauss in 1799, that **is** T-invariant. It is defined by

$$\mu(A) = \frac{1}{\log 2} \int_A \frac{dx}{1+x}$$

for every Borel set A. The probability μ is T-invariant as the reader can check easily by proving first that $\mu(A) = \mu(T^{-1}(A))$ when A is an interval and then using the fact that finite unions of intervals make a generating subalgebra for the Borel $\sigma-$algebra. Moreover, it can be proved that T is ergodic with respect to the probability μ. In fact it belongs to a very well understood class of maps of the interval called Markov maps. Then, for μ a.e. x we have

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} P_n(x,k) &= \tau\left(x, (1/(k+1), 1/k]\right) \\ &= \frac{1}{\log 2} \int_{1/(k+1)}^{1/k} \frac{dx}{1+x} \\ &= \frac{1}{\log 2} \log\left(1 + \frac{1}{k(k+2)}\right) \end{aligned}$$

Since the measure μ is equivalent to the Lebesgue measure, we conclude that the above property holds for a.e. x with respect to the Lebesgue measure.

Given a measure preserving map T of a probability space $(X, \mathcal{O}, \mu)$ we say that a function $f : X \to \mathbb{R}$ is *invariant* (or T-invariant) if $f(T(x)) = f(x)$ for a.e. $x \in X$.

Theorem II.1.1 (Birkhoff-Khinchin) *Let $(X, \mathcal{O}, \mu)$ be a probability space and $T : X \to X$ a measure preserving map. If $f : X \to \mathbb{R}$ is an integrable function, the limit*

$$\tilde{f}(x) := \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j(x)) \qquad \text{(II.1.1)}$$

exists for a.e. $x \in X$, and the function $\tilde{f}$ is T-invariant, integrable and

$$\int_X \tilde{f}\, d\mu = \int_X f\, d\mu$$

$$\lim_{n\to\infty}\int_X \left|\tilde{f} - \frac{1}{n}\sum_{j=0}^{n-1} f\circ T^j\right| d\mu = 0$$

The function $\tilde{f}$ is called the *time average* of f. Sojourn times are time averages of characteristic functions because

$$\#\{0 \le j < n : T^j(x) \in A\} = \sum_{j=0}^{n-1} \chi_A(T^j(x))$$

and then

$$\tau(x, A) = \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} \chi_A(T^j(x)) = \tilde{\chi}_A(x)$$

Hence we have

$$\int \tau(x, A)\, d\mu(x) = \int \tilde{\chi}_A\, d\mu = \int \chi_A\, d\mu = \mu(A)$$

Proof. Suppose that we have proved the existence of the limit (II.1.1) that defines $\tilde{f}$. The integrability of $\tilde{f}$ would follow from the integrability of $|\tilde{f}|$ (because $|\int \tilde{f}\, d\mu| \le \int |\tilde{f}|\, d\mu$). Next, if the limit (II.1.1) exists, we have

$$|\tilde{f}(x)| \le \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} |f(T^j(x))|$$

for almost every x. So it is enough to prove that the above limit is integrable (the existence of this limit follows from our assumption that (II.1.1) exists, applied to $|f|$). As T^j preserves the measure μ,

$$\int |f(T^j(x))|\, d\mu = \int |f\circ T^j|\, d\mu = \int |f|\, d\mu < \infty$$

and so

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} \int |f(T^j(x))|\, d\mu = \|f\|_1 < \infty$$

Then, Fatou's lemma (Theorem I.6.1) implies that $|\tilde{f}|$ is integrable and that $\|\tilde{f}\|_1 \le \|f\|_1$.

Now, $\tilde{f}$ is T-invariant because

$$\begin{aligned}
\tilde{f}(T(x)) &= \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} f(T^{j+1}(x)) \\
&= \lim_{n\to\infty} \frac{n+1}{n}\frac{1}{n+1}\sum_{j=0}^{n} f(T^{j}(x)) - \lim_{n\to\infty}\frac{1}{n} f(x) \\
&= \lim_{n\to\infty} \frac{1}{n+1}\sum_{j=0}^{n} f(T^{j}(x)) \\
&= \tilde{f}(x)
\end{aligned}$$

for almost every x.

In order to prove the last inequality in Theorem II.1.1 we first assume that $f \in \mathcal{L}^\infty(X,\mathcal{O},\mu)$. The definition of $\tilde{f}$ implies that the following sequence converges to zero almost everywhere

$$\left| \tilde{f} - \frac{1}{n}\sum_{j=0}^{n-1} f\circ T^j \right| \tag{II.1.2}$$

We have also that for almost every x

$$|\tilde{f}(x)| \le \lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}|f(T^j(x))| \le \lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}\|f\|_\infty = \|f\|_\infty$$

Hence,

$$\left| \tilde{f}(x) - \frac{1}{n}\sum_{j=0}^{n-1} f\circ T^j(x) \right| \le \|f\|_\infty + \frac{1}{n}\sum_{j=0}^{n-1}|f\circ T^j(x)| \le 2\|f\|_\infty$$

and so the sequence (II.1.2) is dominated by a constant. Then, we apply the Dominated Convergence Theorem I.6.3.

If f is only integrable, given any $\varepsilon > 0$ there exists $\phi \in \mathcal{L}^\infty\ (X,\mathcal{O},\mu)$, and $N > 0$ such that $\|f-\phi\|_1 < \varepsilon/3$ (actually, ϕ may be chosen as a simple function) and

$$\left\| \tilde{\phi} - \frac{1}{n}\sum_{j=0}^{n-1}\phi\circ T^j \right\|_1 \le \varepsilon/3$$

for $n \geq N$. So

$$\left\| \tilde{f} - \frac{1}{n}\sum_{j=0}^{n-1} f \circ T^j \right\|_1 \leq \|\tilde{f} - \tilde{\phi}\|_1 + \left\| \frac{1}{n}\sum_{j=0}^{n-1} \phi \circ T^j - \tilde{\phi} \right\|_1 + \left\| \frac{1}{n}\sum_{j=0}^{n-1} \phi \circ T^j - \frac{1}{n}\sum_{j=0}^{n-1} f \circ T^j \right\|_1$$

But

$$\|\tilde{f} - \tilde{\phi}\|_1 = \|(f - \phi)^{\sim}\|_1 \leq \|f - \phi\|_1 \leq \varepsilon/3$$

and

$$\left\| \frac{1}{n}\sum_{j=0}^{n-1} (\phi - f) \circ T^j \right\|_1 \leq \frac{1}{n}\sum_{j=0}^{n-1} \|\phi - f\|_1 = \|\phi - f\|_1 \leq \varepsilon/3$$

It follows that

$$\left\| \tilde{f} - \frac{1}{n}\sum_{j=0}^{n-1} f \circ T^j \right\|_1 \leq \varepsilon$$

for all $n \geq N$, and this finishes the proof of the last equality in Theorem II.1.1.

The space averages (integrals) of f and $\tilde{f}$ are equal because of the convergence in $\mathcal{L}^1$ that we have just proved and the $\mu-$invariance of T:

$$\int \tilde{f}\, d\mu = \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} \int f \circ T^j\, d\mu = \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} \int f\, d\mu = \int f\, d\mu$$

So, it "only" remains to prove the existence of $\tilde{f}$. This will follow from the next theorem:

Theorem II.1.2 (Maximal Ergodic Theorem) *If* $f \in \mathcal{L}^1(X, \mathcal{O}, \mu)$ *and*

$$E(f) = \left\{ x : \sup_{n \geq 0} \sum_{j=0}^{n} f(T^j(x)) > 0 \right\}$$

then

$$\int_{E(f)} f\,d\mu \geq 0$$

Proof. Consider the increasing sequence

$$f_n(x) = \max\left\{ f(x), f(x) + f(T(x)), \ldots, \sum_{j=0}^{n} f(T^j(x)) \right\}$$

Since

$$E(f) = \cup_{n=0}^{\infty} \{x : f_n(x) > 0\}$$

it is sufficient to prove

$$\int_{\{x:f_n(x)>0\}} f\,d\mu \geq 0$$

for every $n \geq 0$.

If $f_n \circ T(x) \geq 0$, we have $f_n \circ T(x) + f(x) \geq f(x)$ and so,

$$\begin{aligned}
f_n \circ T(x) + f(x) &= \max\{f(T(x)), \ldots, \sum_{j=1}^{n+1} f(T^j(x))\} + f(x) \\
&= \max\{f(T(x)) + f(x), \ldots, \sum_{j=0}^{n+1} f(T^j(x))\} \\
&= \max\{f(x), f(x) + f(T(x)), \ldots, \sum_{j=0}^{n+1} f(T^j(x))\} \\
&= f_{n+1}(x) \geq f_n(x)
\end{aligned}$$

We now consider the following decomposition

$$\begin{aligned}
\int_{\{f_n \geq 0\}} f\,d\mu &= \int_{\{f_n \geq 0\} \cap \{f_n \circ T < 0\}} f\,d\mu + \int_{\{f_n \geq 0\} \cap \{f_n \circ T \geq 0\}} f\,d\mu \\
&\geq \int_{\{f_n \geq 0\} \cap \{f_n \circ T < 0\}} f\,d\mu + \int_{\{f_n \geq 0\} \cap \{f_n \circ T \geq 0\}} f_n\,d\mu \\
&\quad - \int_{\{f_n \geq 0\} \cap \{f_n \circ T \geq 0\}} f_n \circ T\,d\mu
\end{aligned}$$

On the domain of integration in the first integral of the last sum we have $f(x) \geq f_n(x)$. Indeed, from the definitions of f_n and $f_n \circ T$ and the positivity of the first one, we deduce that for some $0 \leq \bar{m} \leq n$

$$f_n(x) = \sum_{j=0}^{\bar{m}} f(T^j(x)) \geq 0$$

and from the negativity of $f_n \circ T$ we deduce that

$$\sum_{j=1}^{m+1} f(T^j(x)) < 0$$

for every $0 \leq m \leq n$. If $\bar{m} = 0$, then $f_n(x) = f(x)$. If $\bar{m} > 0$, then

$$f_n(x) = f(x) + \sum_{j=1}^{\bar{m}} f(T^j(x))$$

and $f_n(x) < f(x)$.

Now

$$\begin{aligned}
\int_{\{f_n \geq 0\}} f \, d\mu &\geq \int_{\{f_n \geq 0\} \cap \{f_n \circ T < 0\}} f_n \, d\mu \\
&\quad + \int_{\{f_n \geq 0\} \cap \{f_n \circ T \geq 0\}} f_n \, d\mu - \int_{\{f_n \geq 0\} \cap \{f_n \circ T \geq 0\}} f_n \circ T \, d\mu \\
&= \int_{\{f_n \geq 0\}} f_n \, d\mu - \int_{\{f_n \geq 0\} \cap \{f_n \circ T \geq 0\}} f_n \circ T \, d\mu \\
&= \int_{T^{-1}(\{f^n \geq 0\})} f_n \circ T \, d\mu - \int_{\{f^n \geq 0\} \cap \{f_n \circ T \geq 0\}} f_n \circ T \, d\mu
\end{aligned}$$

Note that the domain of integration in the first integral on the right is

$$\begin{aligned}
\{T^{-1}(x) \quad &: \quad \max\{f(x), \ldots, \sum_{j=0}^{n} f(T^j(x))\} \geq 0\} \\
&= \{y : \max\{f(Ty), \ldots, \sum_{j=1}^{n+1} f(T^j(y))\} \geq 0\} \\
&= \{f_n \circ T \geq 0\}
\end{aligned}$$

which contains the domain of integration of the second integral. So, the subtraction of the integrals gives the integral of $f_n \circ T$ over a subset of the set $\{x : f_n \circ T(x) \geq 0\}$, so that integral is ≥ 0. Theorem II.1.2 is proved. $\square$

Corollary II.1.3 *If A is a measurable set, $A \subset E(f)$, and $T^{-1}(A) = A$, then* $\int_A f\, d\mu \geq 0$.

Proof. If χ_A is the characteristic function of A, from the definition of $E(f\chi_A)$ we have that $A = E(f\chi_A)$. So,

$$0 \leq \int_{E(f\chi_A)} f\chi_A\, d\mu = \int_A f\chi_A\, d\mu = \int_A f\, d\mu$$

$\square$

Now we finish the proof of Theorem II.1.1. Let $f \in \mathcal{L}^1(X, \mathcal{O}, \mu)$ and define

$$E^+_\alpha(f) = \{x : \limsup_{n\to\infty} \frac{\sum_{j=0}^n f(T^j(x))}{n+1} > \alpha\}$$

$$E^-_\alpha(f) = \{x : \liminf_{n\to\infty} \frac{\sum_{j=0}^n f(T^j(x))}{n+1} < \alpha\}$$

Observe that $E^-_\alpha(f) = E^+_{-\alpha}(-f)$ and $T^{-1}(E^+_\alpha(f)) = E^+_\alpha(f) \subset E(f)$.

First, we will prove that if $A \subset E^+_\alpha(f)$ is a measurable set such that $T^{-1}(A) = A$, then

$$\int_A f\, d\mu \geq \alpha\mu(A) \tag{II.1.3}$$

Indeed,

$$\int_A f\, d\mu = \int_A (f - \alpha)\, d\mu + \alpha\mu(A) \geq \alpha\mu(A)$$

since $A \subset E^+_\alpha(f) = E^+_0(f - \alpha) \subset E(f - \alpha)$ and by Corollary II.1.3 the last integral is not negative. By the same method we can prove that if $A \subset E^-_\beta(f)$ and $T^{-1}(A) = A$ then

$$\int_A f\, d\mu \leq \beta\mu(A) \tag{II.1.4}$$

From the inequalities (II.1.3) and (II.1.4), with $A = E^+_\alpha(f) \cap E^-_\beta(f), \quad \alpha > \beta$, we obtain

$$\mu(E^+_\alpha(f) \cap E^-_\beta(f)) = 0 \tag{II.1.5}$$

Finally, if $\alpha_n,\ n \geq 1$, is a dense sequence in $\mathbb{R}$ (rational numbers, for example), it results that

$$\left\{ x : \limsup_{n\to\infty} \frac{\sum_{j=0}^n f(T^j(x))}{n+1} > \liminf_{n\to\infty} \frac{\sum_{j=0}^n f(T^j(x))}{n+1} \right\}$$

$$= \cup_{\alpha_n > \alpha_m} (E^+_{\alpha_n}(f) \cap E^-_{\alpha_m}(f))$$

This set has zero measure by (II.1.5) and so, on a set of full measure both limits coincide and define the function $\tilde{f}$ for $f \in \mathcal{L}^1(X)$. The proof of Birkhoff-Khinchin Theorem is completed. □

Corollary II.1.4 *If* T *is invertible and* $f \in \mathcal{L}^1(X, \mathcal{O}, \mu)$, *then* $\tilde{f}(x) = \tilde{f}^-(x)$ *a.e., where*

$$\tilde{f}^-(x) := \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^{-j}(x))$$

and hence

$$\tilde{f}(x) = \lim_{n\to\infty} \frac{1}{2n+1} \sum_{j=-n}^{n} f(T^j(x))$$

Proof. If $f_n \to f$ in $\mathcal{L}^1$, the inequality

$$\int |f_n - f|\, d\mu = \int |f_n - f|^\sim\, d\mu \geq \int |\tilde{f}_n - \tilde{f}|\, d\mu$$

shows that $\tilde{f}_n \to \tilde{f}$ in $\mathcal{L}^1$, and the same is true for $\tilde{f}^-_n$ and $\tilde{f}^-$. On a probability space X, square integrable functions $f \in \mathcal{L}^2(X)$ are dense in $\mathcal{L}^1(X)$ (actually, any function in $\mathcal{L}^1$ can be approximated by simple functions). Hence, it is enough to prove the corollary for $f \in \mathcal{L}^2$.

Let $F \subset \mathcal{L}^2$ be the subspace of T-invariant functions. It is a closed subspace of $\mathcal{L}^2$. If $\pi : \mathcal{L}^2 \to F$ is the orthogonal projection, the equality $\tilde{f} = \tilde{f}^-$ will follow from $\pi f = \tilde{f}$ for every $f \in \mathcal{L}^2$, since F is also the space of T^{-1}-invariant functions (i.e., such that $g \circ T^{-1} = g$ a.e.). But for any $g \in F$

$$\int (f - \tilde{f}) g \, d\mu = \int (fg - \tilde{f}g)^{\sim} \, d\mu = \int (fg)^{\sim} \, d\mu - \int (\tilde{f}g)^{\sim} \, d\mu = 0$$

because $(fg)^{\sim} = \tilde{f}g = (\tilde{f}g)^{\sim}$, as a consequence of $\tilde{f}$ and g being T-invariant. □

Remark. Let T^t be a continuous group of automorphisms on a probability space $(X, \mathcal{O}, \mu)$ (i.e., parameter $t \in \mathbb{R}$). This means that every T^t is invertible, $T^0 = \mathrm{Id}$, $T^{t+s}(x) = T^t(T^s(x))$, and $\mu(T^t(A)) = \mu(A)$ for every $t, s \in \mathbb{R}$, $x \in X$, $A \in \mathcal{O}$. We say that T^t is a flow if for every measurable function $g : X \to \mathbb{R}$, $g \circ T^t$ is measurable on $X \times \mathbb{R}$.

The corresponding version of the Birkhoff-Khinchin theorem for flows says that for every $f \in \mathcal{L}^1(X, \mathcal{O}, \mu)$

$$\begin{aligned} \tilde{f}(x) &= \lim_{t \to \infty} \frac{1}{t} \int_0^t f(T^s(x)) \, ds \\ &= \lim_{t \to \infty} \frac{1}{t} \int_0^t f(T^{-s}(x)) \, ds \\ &= \lim_{t \to \infty} \frac{1}{2t} \int_{-t}^t f(T^s(x)) \, ds \end{aligned}$$

where the limits exist for almost every $x \in X$, and $\tilde{f} \in \mathcal{L}^1(X, \mathcal{O}, \mu)$ and $\int \tilde{f} \, d\mu = \int f \, d\mu$.

Remark. In the same way as before, we can prove that if $f \in \mathcal{L}^p(X, \mathcal{O}, \mu)$ for some $1 \leq p < \infty$, then $\tilde{f} \in \mathcal{L}^p(X, \mathcal{O}, \mu)$, and

$$\lim_{n \to \infty} \left\| \tilde{f} - \frac{1}{n} \sum_{j=0}^{n-1} f \circ T^j \right\|_p = 0$$

Corollary II.1.5 *For every $A, B \in \mathcal{O}$ the limit*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} \mu(T^{-j}(A) \cap B)$$

exists.

Proof. We have

$$\mu(T^{-n}(A) \cap B) = \int_X \chi_{T^{-n}(A)} \chi_B \, d\mu = \int_X (\chi_A \circ T^n) \chi_B \, d\mu$$

Since $\chi_A \in \mathcal{L}^2(X)$, we can apply the Birkhoff-Khinchin theorem to conclude that the sequence

$$\frac{1}{n} \sum_{j=0}^{n-1} \chi_A \circ T^j$$

converges in $\mathcal{L}^2$. Thus

$$\frac{1}{n} \sum_{j=0}^{n-1} \mu(T^{-j}(A) \cap B) = \int_X \left(\frac{1}{n} \sum_{j=0}^{n-1} \chi_A \circ T^j \right) \chi_B \, d\mu$$

also converges. $\square$

Exercises:

II.1.1. Let $(X, \mathcal{O}, \mu)$ be a probability space, $T : X \to X$ a measure-preserving map and $\{F_n\}_{n \geq 0}$ a dominated sequence of functions in $\mathcal{L}^1(X, \mathcal{O}, \mu)$ which converges almost everywhere to a function $F \in \mathcal{L}^1(X, \mathcal{O}, \mu)$. Prove that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} F_j(T^j(x)) = \tilde{F}(x) \qquad \text{a.e.}$$

II.1.2. Let $(X, \mathcal{O}, \mu)$ be a probability space and $T : X \to X$ a measure-preserving map.

(a) Prove that for every $A \in \mathcal{O}$ with positive measure there exists a subset $A_0 \in \mathcal{O}$ contained in A having positive measure such that, for every $x \in A_0$, we have $\tau(x, A) \geq \mu(A)$. Hint: Put $A_1 := \{x : \tau(x, A) \geq \mu(A)\}$, and prove that $\mu(A_1) > 0$ and $A_1 = \cup_{n \geq 0} T^{-n}(A_1 \cap A)$.

(b) Prove that $\tau(x, A) > 0$ for almost every point $x \in A$.

(c) Let X be a separable metric space and $\mathcal{O}$ the $\sigma-$algebra of Borel sets with a T-invariant probability. Prove that for almost every $x \in X$ we have $\tau_U(x) > 0$ for every neighborhood U of x.

II.1.3. Let $(X, \mathcal{O}, \mu)$ be a probability space and $T : X \to X$ a measure-preserving map.

(a) Prove that if $C : X \to (0, +\infty)$ is measurable, then

$$\liminf_{n\to\infty} \frac{1}{n} C(T^n(x)) = 0 \qquad \text{a.e.}$$

Hint: If the property does not hold, there exist $A \in \mathcal{O}$, $K_1 > 0$ and $K_2 > 0$ such that

$$\liminf_{n\to\infty} \frac{1}{n} C(T^n(x)) \geq K_1$$

$$C(x) \leq K_2 \qquad \text{for all } x \in A$$

and $\mu(A) > 0$. Use the Poincaré recurrence theorem (Theorem I.5.1) to derive a contradiction between the two inequalities.

(b) Prove that if $C : X \to (0, +\infty)$ is measurable and $C \circ T - C$ is integrable, then

$$\lim_{n\to\infty} \frac{1}{n} C(T^n(x)) = 0 \qquad \text{a.e.}$$

Hint: Write

$$\frac{1}{n} C(T^n(x)) = \frac{1}{n} \sum_{j=0}^{n-1} (C \circ T - C)(T^j(x)) + \frac{1}{n} C(x)$$

and apply the Birkhoff-Khinchin theorem.

II.2 Ergodicity

Consider a measure-preserving map T of a probability space $(X, \mathcal{O}, \mu)$. Recall that a set $A \in \mathcal{O}$ is called T-invariant if $T^{-1}(A) = A$.

Definition. T is said to be ergodic if every T-invariant set has measure 0 or 1.

Bernoulli shifts are ergodic. The proof uses the following fact, which should be checked by the reader: If $A_i \subset B(p_1, \ldots, p_n)$, $i = 1, 2$, are cylinders, there exists $m_0 > 0$ such that for all $m > m_0$

$$\mu(\sigma^{-m}(A_1) \cap A_2) = \mu(A_1)\mu(A_2) \qquad \text{(II.2.1)}$$

Now assume that $A \subset B(p_1, \ldots, p_n)$ is $\sigma-$invariant. Since the $\sigma-$algebra of $B(p_1, \ldots, p_n)$ is generated by cylinders, there exists, for any $\varepsilon > 0$, a finite union A_0 of disjoint cylinders such that

$$\mu(A_0 \triangle A) \leq \varepsilon.$$

Property (II.2.1) is easily seen to hold for finite unions of cylinders as well as for single cylinders. Thus, for some $m \geq 0$,

$$\mu(\sigma^{-m}(A_0^c) \cap A_0) = \mu(A_0)\mu(A_0^c)$$

$$\mu(\sigma^{-m}(A_0) \cap A_0^c) = \mu(A_0)\mu(A_0^c)$$

It follows that

$$\begin{aligned} \mu(\sigma^{-m}(A_0) \triangle A_0) &\leq \mu(\sigma^{-m}(A) \triangle \sigma^{-m}(A_0)) + \mu(\sigma^{-m}(A) \triangle A) + \mu(A_0 \triangle A) \\ &= 2\mu(A \triangle A_0) \leq 2\varepsilon \qquad \text{(II.2.2)} \end{aligned}$$

On the other hand,

$$\begin{aligned} \mu(\sigma^{-m}(A_0) \triangle A_0) &= \mu(\sigma^{-m}(A_0) \cap A_0^c) + \mu(\sigma^{-m}(A_0^c) \cap A_0) \\ &= 2\mu(A_0)\mu(A_0^c) \\ &= 2\mu(A_0)(1 - \mu(A_0)) \qquad \text{(II.2.3)} \end{aligned}$$

Form (II.2.2) and (II.2.3) we get

$$2\mu(A_0)(1 - \mu(A_0)) \leq 2\varepsilon$$

and, since ε is arbitrary,

$$\mu(A)(1 - \mu(A)) = 0$$

or, again, $\mu(A) = 0$ or 1.

Proposition II.2.1 *The following properties are equivalent:*

(1) *T is ergodic;*

(2) *If $f \in \mathcal{L}^1(X)$ is T-invariant, then f is constant almost everywhere;*

(3) *If $f \in \mathcal{L}^p(X)$ is T-invariant, then f is constant almost everywhere;*

(4) *For every $A, B \in \mathcal{O}$ we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} \mu(T^{-m}(A) \cap B) = \mu(A)\mu(B)$$

(5) *For every $f \in \mathcal{L}^1(X)$ we have $\tilde{f}(x) = \int_X f\, d\mu$ almost everywhere.*

Proof. (3) $\Rightarrow$ (1). If $A \in \mathcal{O}$ is T-invariant, its characteristic function χ_A is T-invariant and belongs in $\mathcal{L}^p(X)$. Thus χ_A is constant almost everywhere, i.e. $\mu(A) = 0$ or 1.

(1) $\Rightarrow$ (2). If $f \in \mathcal{L}^1(X)$ is T-invariant, the set $A_c := \{x : f(x) \le c\}$ is invariant for each c. Since T is ergodic, this means $\mu(A_c) = 0$ or 1 for each c. We leave it to the reader to show that this implies that f is constant almost everywhere.

(2) $\Rightarrow$ (5) . Since $\tilde{f}$ belongs in $\mathcal{L}^1(X)$ and is T-invariant, it must be a constant. From

$$\int_X \tilde{f}\, d\mu = \int_X f\, d\mu$$

the assertion follows.

(5) $\Rightarrow$ (4). By the Birkhoff-Khinchin theorem,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} \chi_A(T^j(x)) = \tilde{\chi}_A = \int_X \chi_A\, d\mu = \mu(A)$$

almost everywhere. By Dominated Convergence Theorem I.6.3,

$$\begin{aligned}
\mu(A)\mu(B) &= \int_X (\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \chi_A(T^j(x))\chi_B\, d\mu \\
&= \lim_{n\to\infty} \frac{1}{n} \int_X \sum_{j=0}^{n-1} \chi_A(T^j(x))\chi_B\, d\mu \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \mu(T^{-j}(A) \cap B)
\end{aligned}$$

(4) $\Rightarrow$ (1). If A is T-invariant, we apply (4) to the sets A and A^c. Then

$$\mu(A)\mu(A^c) = \lim_{n\to\infty} \sum_{j=0}^{n-1} \mu(T^{-j}(A) \cap A^c) = 0$$

so that $\mu(A) = 0$ or 1. $\square$

In fact, condition (5) only needs to hold in a dense set of $\mathcal{L}^1(X)$ to imply the other four:

Proposition II.2.2 *If there exists a dense set* $F \subset \mathcal{L}^1(X)$ *such that*

$$\tilde{f}(x) = \int_X f\,d\mu \qquad\qquad \text{a.e.}$$

for every $f \in F$, *then* T *is ergodic.*

Proof. Since the sequence

$$\frac{1}{n}\sum_{j=0}^{n-1} f \circ T^j$$

converges to $\tilde{f}$ in $\mathcal{L}^1(X)$, it is enough to check that for every $f \in \mathcal{L}^1(X)$

$$\lim_{n\to\infty} \left\| \frac{1}{n}\sum_{j=0}^{n-1} f \circ T^j - \int_X f\,d\mu \right\|_1 = 0$$

Take $\varepsilon > 0$, and choose $g \in F$ such that $\|g - f\|_1 \le \varepsilon/3$. Let n_0 be such that $n \ge n_0$ implies

$$\left\| \frac{1}{n}\sum_{j=0}^{n-1} g \circ T^j - \int_X g\,d\mu \right\|_1 \le \varepsilon/3$$

Then, for all $n \ge n_0$,

$$\begin{aligned}
\left\| \frac{1}{n}\sum_{j=0}^{n-1} f \circ T^j - \int_X f\,d\mu \right\|_1 \quad &\le \quad \frac{1}{n}\left\| \sum_{j=0}^{n-1} f \circ T^j - \sum_{j=0}^{n} g \circ T^j \right\|_1 \\
&\quad + \left\| \frac{1}{n}\sum_{j=0}^{n-1} g \circ T^j - \int_X g\,d\mu \right\|_1 \\
&\quad + \left| \int_X g\,d\mu - \int_X f\,d\mu \right|
\end{aligned}$$

Since

$$\|f \circ T^j - g \circ T^j\|_1 = \|(f-g) \circ T^j\|_1 = \|f-g\|_1$$

and since

$$\left| \int_X g \, d\mu - \int_X f \, d\mu \right| \leq \|g-f\|_1$$

we conclude that

$$\left\| \frac{1}{n} \sum_{j=0}^{n-1} f \circ T^j - \int_X f \, d\mu \right\|_1 \leq \|g-f\|_1 + \varepsilon/3 + \|g-f\|_1 \leq \varepsilon$$

The proposition is proved. □

Another characterization of ergodicity can be given in terms of the average time $\tau(x,A)$ spent by a point x in a set A, which exists by the Birkhoff-Khinchin theorem.

Proposition II.2.3 *T is ergodic if and only if $\tau(x,A) = \mu(A)$ a.e. for every $A \in \mathcal{O}$.*

Proof. If T is ergodic then

$$\tau(x,A) = \tilde{\chi}_A(x) = \int_X \chi_A \, d\mu = \mu(A) \qquad \text{a.e.}$$

Conversely, let $A \in \mathcal{O}$ be T-invariant. Assume $\mu(A) > 0$. Since $\tau(x,A) = 1$ for $x \in A$, it follows that $\mu(A) = 1$. □

The following "uniqueness theorem" holds for ergodic maps:

Proposition II.2.4 *If T is ergodic and $\mu_1 : \mathcal{O} \to [0,1]$ is another T-invariant probability measure, the following conditions are equivalent:*

(a) $\mu_1 = \mu$;

(b) $\mu_1 \ll \mu$;

(c) *There exists no T-invariant set $A \in \mathcal{O}$ such that $\mu(A) = 0$ and $\mu_1(A) \neq 0$.*

Proof. (b) $\Rightarrow$ (a). If $\mu_1 \ll \mu$ and both μ_1 and μ are invariant, the Radon-Nikodym derivative $d\mu_1/d\mu$ is an invariant function. Since T is ergodic, $d\mu_1/d\mu$ is a.e. constant, and so $\mu_1 = \mu$.

(c) $\Rightarrow$ (b). Assume (b) does not hold, and take $A_0 \in \mathcal{O}$ such that $\mu(A_0) = 0$ and $\mu_1(A_0) \neq 0$. The set $A := \cup_{n\leq 0} T^n(A_0)$ contradicts (c).

(a) $\Rightarrow$ (c). Trivial. $\Box$

If X is a set and $\mathcal{O}$ is a $\sigma-$algebra on X, the set of all signed measures on $\mathcal{O}$ has an obvious vector space structure. If $T : X \to X$ is a measurable map, the set $m_T(X,\mathcal{O})$ of T-invariant probability measures on $\mathcal{O}$ is a convex subset of this vector space. The next proposition characterizes ergodic maps with respect to T. Note: The expression "μ is ergodic with respect to T" evidently means that $\mu \in m_T(X,\mathcal{O})$ and T is an ergodic map of $(X,\mathcal{O},\mu)$.

Proposition II.2.5 *The measure $\mu \in m_T(X,\mathcal{O})$ is ergodic if and only if μ is an extremal point*[1] *of $m_T(X,\mathcal{O})$.*

Proof. Assume that $\mu \in m_T(X,\mathcal{O})$ is ergodic and $\mu = \lambda\mu_1 + (1-\lambda)\mu_2$ with some $\mu_1, \mu_2 \in m_T(X,\mathcal{O})$ and $0 < \lambda < 1$. We have $\mu_1 \ll \mu$ since $\lambda \neq 0$, so Proposition II.2.4 implies that $\mu_1 = \mu$. Similarly, we have $\mu_2 = \mu$, which shows that $\mu_1 = \mu_2$, hence μ is extremal.

Now if $\mu \in m_T(X,\mathcal{O})$ is not ergodic, there exists a T-invariant set $A_0 \in \mathcal{O}$ satisfying $0 < \mu(A_0) < 1$. We define measures $\mu_i \in m_T(X,\mathcal{O})$, $i = 1,2$, by

$$\mu_1(A) = \frac{1}{\mu(A_0)}\mu(A \cap A_0) \qquad \text{and} \qquad \mu_2(A) = \frac{1}{\mu(A_0^c)}\mu(A \cap A_0^c)$$

for all $A \in \mathcal{O}$. Then we can write

$$\mu = \mu(A_0)\,\mu_1 + \mu(A_0^c)\,\mu_2$$

showing that μ is not extremal. $\Box$

Exercises:

[1] An extremal point x of a convex set C is one which cannot be presented as $x = \lambda y + (1-\lambda)z$ for two distinct $y, z \in C$ and some $0 < \lambda < 1$.

II.2.1. Let X be a set, $\mathcal{O}$ a $\sigma-$algebra on X and $T : X \to X$ a measurable map. If $\mu_i \in m_T(X, \mathcal{O})$, $i = 1, \dots, n$ are ergodic measures and $\mu_i \not\ll \mu_j$ for $i \neq j$, prove that there exist disjoint sets $A_i \in \mathcal{O}$, $i = 1, \dots, n$, such that $\cup_{i=1}^n A_i = X$ and $\mu_j(A_i) = \delta_{ij}$.

II.2.2. Let T be a continuous map of a compact metric space X and $\mu \in m_T(X)$ an ergodic measure.

(a) Prove that there exists a full measure subset $A \subset X$ such that for every $x \in A$ we have:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j(x)) = \int_X f \, d\mu$$

for *every* continuous function $f : X \to \mathbb{C}$.

(b) Prove that for almost every $x \in X$ we have

$$\tau(x, A) = \mu(A)$$

for every $A \in \mathcal{O}$ such that $\mu(\partial A) = 0$.

(c) Prove that for a.e. $a \in X$ there exists a countable set $S \subset (0, +\infty)$, such that for $r \notin S$ we have

$$\tau(a, B_r(a)) = \mu(B_r(a))$$

II.2.3. Let $(X, \mathcal{O}, \mu)$ be a probability space and $T : X \to X$ a measure-preserving map. Prove that if $f \in \mathcal{L}^2(X)$ satisfies

$$\sum_{n=0}^{\infty} \left| \langle U_T^n f, f \rangle - \left(\int_X f \, d\mu \right)^2 \right| < +\infty$$

then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j(x)) = \int_X f \, d\mu$$

for almost every $x \in X$. Hint: Use Chebyshev's inequality:

$$\mu(S_\varepsilon) \leq \frac{\sigma^2(f)}{\varepsilon^2}$$

where $\sigma^2(f) = \langle f, f \rangle - (\int f d\mu)^2$ and $S_\varepsilon = \{x \in X : |f(x) - \int f \, d\mu| > \varepsilon\}$.

II.3 Ergodicity of Translations and Linear Maps of the Torus

In Chapter I (section 4) we introduced the translations and linear maps of the torus $\mathbb{T}^n = S^1 \times \cdots \times S^1$. A translation $L_k : \mathbb{T}^n \to \mathbb{T}^n$, where $k = (k_1, \ldots, k_n)$, was defined by $L_k(x) = (k_1x_1, \ldots, k_nx_n)$. Linear maps of $\mathbb{T}^n$ were defined as maps $f : \mathbb{T}^n \to \mathbb{T}^n$ for which there exists a surjective linear map $\tilde{f} : \mathbb{R}^n \to \mathbb{R}^n$, called the linear lifting of f, such that $\tilde{f}(\mathbb{Z}^n) \subset \mathbb{Z}^n$ (i.e. the entries of the matrix of $\tilde{f}$ are integers and $\det \tilde{f} \neq 0$) and $f\pi(x) = \pi\tilde{f}(x)$, where $\pi : \mathbb{R}^n \to \mathbb{Z}^n$ is the projection map defined by $\pi(x_1, \ldots, x_n) = (\exp(2\pi x_1 i), \ldots, \exp(2\pi x_n i))$. The eigenvalues of f are the eigenvalues of $\tilde{f}$.

Let λ be the Lebesgue probability on $\mathbb{T}^n$ (which, as we proved, is invariant under both translations and linear maps) and denote by $\mathcal{L}^2(\mathbb{T}^n)$ the space of complex valued functions on $\mathbb{T}^n$ that are $\mathcal{L}^2$ with respect to λ, endowed with its usual structure of Hilbert space.

Theorem II.3.1 *A linear map $f : \mathbb{T}^n \to \mathbb{T}^n$ is ergodic if and only if none of its eigenvalues is a root of unity.*

Theorem II.3.2 *If $x \in \mathbb{R}^n$, then the translation $L_{\pi(x)} : \mathbb{T}^n \to \mathbb{T}^n$ is ergodic if and only if $\langle k, x\rangle \notin \mathbb{Z}$ for any $k \in \mathbb{Z}^n$, $k \neq 0$.*

The proofs utilize the orthonormal basis of $\mathcal{L}^2(\mathbb{T}^n)$ obtained from the Fourier basis $\{e^{2\pi i\langle k,x\rangle} | \, k \in \mathbb{Z}^n\}$ of $\mathcal{L}^2([0,1] \times \cdots \times [0,1])$. The formal description of this basis and its basic properties are the subject of the following lemma.

Lemma II.3.3 *There exists an orthonormal basis $\{\phi_k | \, k \in \mathbb{Z}^n\}$ of $\mathcal{L}^2(\mathbb{T}^n)$ such that*

(a) $\phi_0 = 1$;

(b) *For every linear map $f : \mathbb{T}^n \to \mathbb{T}^n$ we have*

$$\phi_k \circ f = \phi_{\tilde{f}^*(k)}$$

where $\tilde{f}^*$ is the adjoint of $\tilde{f}$.

(c) *For every $x \in \mathbb{R}^n$ the translation $L_{\pi(x)}$ satisfies*

$$\phi_k \circ L_{\pi(x)} = e^{i\langle k, 2\pi x\rangle} \phi_k$$

Proof. For each $\psi : \mathbb{R}^n \to \mathbb{C}$ satisfying $\psi(x) = \psi(x+m)$ for all $x \in \mathbb{R}^n$, $m \in \mathbb{Z}^n$, there exists a unique $\check{\psi} : \mathbb{T}^n \to \mathbb{C}$ such that $\check{\psi} \circ \pi = \psi$. Let $\psi_k := \mathbb{R}^n \to \mathbb{C}$ be defined by $\psi_k(x) = e^{i\langle k, 2\pi x\rangle}$ and $\phi_k : \mathbb{T}^n \to \mathbb{C}$ by $\phi_k = \check{\psi}_k$. Then

$$\phi_k \circ f \circ \pi = \check{\psi}_k \circ f \circ \pi = \check{\psi}_k \circ \pi \circ \tilde{f} = \psi_k \circ \tilde{f} = \psi_{\tilde{f}^*(k)}$$

proving part (b). Now, for $x \in \mathbb{R}^n$,

$$\phi_k \circ L_{\pi(x)} \circ \pi = \phi_k \circ \pi \circ L_x = \psi_k \circ L_x = e^{i\langle k, 2\pi x\rangle} \psi_k$$

giving part (c). It remains to prove that $\{\phi_k | \, k \in \mathbb{Z}^n\}$ is an orthonormal basis of $\mathcal{L}^2(\mathbb{T}^n)$. Since $\pi : [0,1] \times \cdots \times [0,1] \to \mathbb{T}^n$ is a measure-preserving map, we have

$$\langle \psi_1, \psi_2 \rangle = \langle \psi_1 \circ \pi, \psi_2 \circ \pi \rangle$$

for every $\psi_1, \psi_2 \in \mathcal{L}^2(\mathbb{T}^n)$; here the inner products are taken in $\mathcal{L}^2(\mathbb{T}^n)$ and $\mathcal{L}^2([0,1] \times \cdots \times [0,1])$, respectively. Thus, for any $\psi \in \mathcal{L}^2(\mathbb{T}^n)$ satisfying

$$\langle \psi, \phi_k \rangle = 0$$

for all $k \in \mathbb{Z}^n$, it follows that

$$\langle \psi \circ \pi, \phi_k \circ \pi \rangle = \langle \psi \circ \pi, \psi_k \rangle = 0$$

for all $k \in \mathbb{Z}^n$. This shows that $\psi \circ \pi = 0$, because $\{\psi_k | \, k \in \mathbb{Z}^n\}$ is the the orthonormal Fourier basis of $\mathcal{L}^2([0,1] \times \cdots \times [0,1])$. Since π is surjective, we get $\psi = 0$. $\square$

Proof of Theorem II.3.1. Since f is measure-preserving, we have

$$\langle \psi_1 \circ f, \psi_2 \circ f \rangle = \langle \psi_1, \psi_2 \rangle$$

for every $\psi_1, \psi_2 \in \mathcal{L}^2(\mathbb{T}^n)$. Thus, for $\phi \in \mathcal{L}^2(\mathbb{T}^n)$ such that $\phi \circ f = \phi$, we get

$$\langle \phi \circ f, \phi_{\tilde{f}^*(k)} \rangle = \langle \phi \circ f, \psi_k \circ f \rangle = \langle \phi, \phi_k \rangle$$

Thus, for every $k \in \mathbb{Z}^n$ and $i \geq 0$

$$\langle \phi \circ f, \ \phi_{\tilde{f}^{*i}(k)} \rangle = \langle \phi, \phi_k \rangle \tag{II.3.1}$$

There are two possibilities for the sequence $\{\tilde{f}^{*i}(k)\}$ – either all of its elements are distinct, or, for some i, we have $\tilde{f}^{*i}(k) = k$. If $k \neq 0$ the second case means that 1 is an eigenvalues of $\tilde{f}^{*i}$, hence of $\tilde{f}^i$, and so f has a root of unity as an eigenvalue. Thus, if no eigenvalue of f is a root of unity, we can only have the first case, and

$$\|\phi\|^2 = \sum_{k \in \mathbb{Z}^n} |\langle \phi, \phi_k \rangle|^2 \geq \sum_{i \in \mathbb{Z}} |\langle \phi, \phi_{\tilde{f}^{*i}(k)} \rangle|^2$$

In the latter sum all terms are equal, and thus zero since the sum is finite. Thus, for all $k \neq 0$ we have $\langle \phi, \phi_k \rangle = 0$, showing that ϕ is a.e. equal to $\langle \phi, \phi_0 \rangle \phi_0$, a constant, so f is ergodic.

Conversely, assume that f has a root of unity as an eigenvalue. Then, for some $i \geq 0$, we have $\det(\tilde{f}^{*i} - I) = 0$. Since the components of the matrix $\tilde{f}^{*i} - I$ are integers, there exists some $0 \neq k \in \mathbb{Z}^n$ such that $\tilde{f}^{*i}(k) = k$. Setting

$$\phi := \sum_{j=0}^{i-1} \phi_{\tilde{f}}^{*j}(k)$$

we conclude that

$$\phi \circ f = \sum_{j=0}^{i-1} \phi_{\tilde{f}^{*j}(k)} \circ f = \sum_{j=0}^{i-1} \phi_{\tilde{f}^{*j+1}(k)} = \phi$$

showing that f is not ergodic. □

Proof of Theorem II.3.2. Let $\phi \in \mathcal{L}^2(\mathbb{T}^n)$ and $\phi \circ L_{\pi(x)} = \phi$. Then

$$\langle \phi, \phi_k \rangle = \langle \phi \circ L_{\pi(x)}, \phi_k \rangle = \langle \phi, \phi_k \circ L_{\pi(-x)} \rangle = e^{-i\langle k, 2\pi x \rangle} \langle \phi, \phi_k \rangle$$

Thus if $\langle k, x \rangle \notin \mathbb{Z}$ for every $0 \neq k \in \mathbb{Z}^n$, we have $\langle \phi, \phi_k \rangle = 0$ for all $0 \neq k \in \mathbb{Z}^n$, i.e. ϕ is a constant. On the other hand, if there exists $0 \neq k_0 \in \mathbb{Z}^n$ such that $\langle k_0, x \rangle \in \mathbb{Z}$, then the function ϕ_{k_0} satisfies $\phi_{k_0} \circ L_{\pi(x)} = \phi_{k_0}$. □

II.4 Ergodic Hierarchy

In Section I.2 we proved that a measure-preserving map T of a probability space $(X, \mathcal{O}, \mu)$ is ergodic if and only if, for every $A, B \in \mathcal{O}$

$$\lim_{m \to \infty} \frac{1}{m} \sum_{j=0}^{m-1} \mu(T^{-j}(A) \cap B) = \mu(A)\mu(B)$$

In this case, if the limit of $\mu(T^{-j}(A) \cap B)$ as $j \to \infty$ exists, its value must be equal to $\mu(A)\mu(B)$.

Definition. A measure preserving endomorphism T of a probability space $(X, \mathcal{O}, \mu)$ is said to be *mixing* if for any pair $A, B \in \mathcal{O}$

$$\lim_{n\to\infty} \mu(T^{-n}(A) \cap B) = \mu(A)\mu(B)$$

A pictorial example in a classical book by Arnold and Avez [AA] explains what a mixing map does. Suppose a cocktail shaker M, $\mu(M) = 1$ is filled by 85% of pisco and 15% of lemon juice. Let A be the part of the cocktail shaker originally occupied by the juice and B any part of the shaker. Let $T\colon M \to M$ be the transformation of the content of the shaker made during one move by the bartender (who is shaking the cocktail repeatedly). Then after n moves the fraction of juice in the part B will be $\mu(T^n(A) \cap B)/\mu(B)$. As the bartender keeps shaking the cocktail $(n \to \infty)$, the fraction of juice in any part B approaches $\mu(A) = 15\%$, i.e. the lemon juice will spread uniformly in the mixture.

We note that the definition we gave for a mixing map is good for both invertible and noninvertible maps (endomorphisms).

Proposition II.4.1 *Any mixing map is ergodic.*

Proof. Let A be any T-invariant measurable set, then $T^{-n}(A) = A$ and $\mu(A \cap B) = \lim_{n\to\infty} \mu(T^{-n}(A) \cap B) = \mu(A)\mu(B)$. In particular, for $A = B$ we have $\mu(A) = \mu^2(A)$. This means $\mu(A) = 0$ or 1, hence T is ergodic. □

We note that not all ergodic maps are mixing, see Exercise II.4.1. Therefore, mixing is a stronger property than ergodicity.

Definition. If X is a topological space, a transformation $T : X \to X$ is *topologically mixing* if for any pair of open sets $U, V \subset X$ there exist $N \in \mathbb{N}$ such that $T^{-n}(U) \cap V \neq \emptyset$ for all $n \geq N$.

Proposition II.4.2 *If X is a topological space, $\mathcal{O}$ is the Borel σ-algebra and μ a probability measure positive on open sets, then whenever $T : X \to X$ is mixing, it is also topologically mixing.*

Proof. Since, for any open sets U, V we have $\lim_{n\to\infty} \mu(T^{-n}(U) \cap V) = \mu(U)\mu(V) \neq 0$ it results that, for all $n \geq N$, $\mu(T^{-n}(U) \cap V) \neq 0$, hence $T^{-n}(U) \cap V \neq \emptyset$. □

Definition. A measure preserving automorphism T of a probability space $(X, \mathcal{O}, \mu)$ is Bernoulli if it is equivalent to a Bernoulli shift. See Example 3 in Section I.4 and Section I.8 (the shifts can be defined on a probability space).

It can be proved that every Bernoulli automorphism is mixing, but not vice versa. All these results are discussed in the book by Mañe [Mn], Sections II.8 and II.11. It proves that there is the so called *ergodic hierarchy*:

$$\text{Bernoulli} \Rightarrow \text{Kolmogorov (or K} - \text{mixing)} \Rightarrow \text{Mixing} \Rightarrow \text{Ergodic} \quad \text{(II.4.1)}$$

Each word in this row represents the set of measure preserving maps of a probability space $(X, \mathcal{O}, \mu)$ that satisfy the corresponding definition. We do not provide the definition of K-mixing maps, because it is quite complicated.

We note that all the implications in (II.4.1) are one-way only, none of them can be reversed. This means that ergodicity does not imply mixing (as noted above), mixing does not imply K-mixing, etc.

Exercises:

II.4.1. Let λ be the Lebesgue probability measure on the unit circle S^1. Prove:

(a) Irrational rotations of S^1 are ergodic.

(b) Rotations of S^1 are never mixing.
Hint: let, for example R_ω be the rotation of S^1 through an angle $0 < \omega < 1/2$. Take two arcs A, B of length $\pi\omega$. Then note that if $R_\omega^{-n} A \cap B \neq \emptyset$, then $R_\omega^{-n-1} A \cap B = \emptyset$.

II.5 Statistical Properties of Dynamical Systems

For those familiar with probability theory we present an additional discussion here. This section can be safely ignored, since it is not essential for further reading.

In the language of probability theory, a Bernoulli shift represents a sequence of independent random variables. This immediately follows from the product formula (I.4.1) in Chapter I. For this reason the Bernoulli property is regarded as a *statistical* property of a dynamical system. It establishes an equivalence between a dynamical system and a purely random sequence of independent trials - a canonical model in probability theory.

This a a very interesting and important observation. A dynamical system $T : X \to X$ is, by nature, completely deterministic. This means that if you have a point $x \in X$, its entire future $\{T^n x\}$, $n \geq 1$, is uniquely determined and can be computed precisely. When the map T is invertible, the past $\{T^n x\}$, $n \leq -1$, is uniquely determined and computable, too. One can look at it this way: knowing the *present state* of a dynamical system (given by $x \in X$), one can determine its future and, often, its past. This is the precise meaning we give to the word "deterministic".

On the other hand, in a sequence of independent trials the outcome of any trial gives no clue of what the outcomes of other trials would be (or have been). So, knowing the present state tells just nothing about the future or the past, the outcome of every trial being completely random and unpredictable[2]. A paradox? In a sense, it is, and there are relevant discussions in physics where the theory of dynamical systems finds most applications. We do not elaborate on this topic here. See [Lb1, Lb2].

We do make a few extra comments, though. While the Bernoulli property is a manifestation of an utter randomness or chaoticity, strangely, it has little relevance to direct physical applications. Why? Because the equivalence between a dynamical system and a Bernoulli shift is, usually, given by just a measurable map with a very complicated structure, not at all smooth or even continuous. In physics, on the other hand, the laws of motion are usually specified by differential equations (like Newton or Hamiltonian equations), and all interesting functions (such as temperature, energy, pressure) are smooth as well. Hence, only the properties of dynamical systems expressed by smooth maps and smooth functions are relevant in physics.

For these reasons, assuming that X is a manifold, $T : X \to X$ a smooth map preserving a probability μ and $f : X \to \mathbb{R}$ a smooth function, one can characterize the system in a physically meaningful way as follows. Consider

$$S_n = f + f \circ T + f \circ T^2 + \cdots + f \circ T^{n-1} \qquad \text{(II.5.1)}$$

The quotient S_n/n is called the *time average* of the function f. Adopting physical notation, we denote by $\langle \cdot \rangle$ the expected value of a function with

[2]We note that despite this fact, the long time behavior of a sequence of independent trials can be quite accurately described by the laws of probability theory.

respect to μ, e.g. $\langle f \rangle = \int_X f \, d\mu$. The integral $\langle f \rangle$ is also called the *space average* of f.

Now the Birkhoff Ergodic Theorem asserts that if T is ergodic, then S_n/n converges almost everywhere to $\langle f \rangle$ as $n \to \infty$. In physical language, it means that **time averages converge to space averages**. In probability theory, this fact is also called the *strong law of large numbers*.

An important characteristic of a dynamical system is the *time correlation function*

$$C_f(n) = \langle f \cdot (f \circ T^n) \rangle - \langle f \rangle^2. \tag{II.5.2}$$

If the map T is mixing, one can show that $C_f(n) \to 0$ as $n \to \infty$ (see Ex. II.5.1), i.e. the correlations *decay*, as physicists call it. The asymptotic speed of convergence $C_f(n) \to 0$ characterizes the "speed of mixing" in the system. See [Vi, Ba, CY]

Next, we say that f satisfies the *central limit theorem* if

$$\lim_{n\to\infty} \mu \left\{ x \colon \frac{S_n(x) - n\langle f \rangle}{\sqrt{n}} < z \right\} = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{z} e^{-\frac{s^2}{2\sigma^2}} \, ds \tag{II.5.3}$$

for all $-\infty < z < \infty$. Here $\sigma = \sigma_f \geq 0$ is a constant related to the correlation function:

$$\sigma_f^2 = C_f(0) + 2\sum_{n=1}^{\infty} C_f(n) \tag{II.5.4}$$

Equation (II.5.3) is equivalent to the convergence of $(S_n - n\langle f \rangle)/\sqrt{n}$ in distribution to the normal random variable $N(0, \sigma_f^2)$. We remark that the central limit theorem is considerably more refined than the Birkhoff Ergodic Theorem; it tells us that the distribution of the deviations of the time average S_n/n from its limit value $\langle f \rangle$, when scaled by $1/\sqrt{n}$, is asymptotically Gaussian.

It is clear from (II.5.4) that the central limit theorem only holds if $\sum_n |C_f(n)| < \infty$. Actually, the proof of this theorem requires even a more rapid convergence of $C_f(n)$ to zero. For these (and other) reasons the speed of that convergence is regarded as an important statistical characteristic of the system. Two main types of convergence are *exponential*, when $|C_f(n)| <$ const $\cdot\, e^{-an}$, $a > 0$, and *polynomial*, when $|C_f(n)| <$ const $\cdot\, n^{-b}$, $b > 0$. Systems with exponential decay of correlations are the "most chaotic", they possess many features necessary for applications in statistical physics. Systems with polynomial decay are regarded as being intermediate ("intermittent") between "regular" and "chaotic", and their behavior is very sensitive to the exact value of the power $b > 0$ and other factors which may limit their applications in statistical physics.

It is also interesting to note that if one relaxes the requirement that the function f in (II.5.2) be smooth, then one totally loses control over the decay of correlations. In all known mixing dynamical systems, the convergence $C_f(n) \to 0$ is indeed arbitrarily slow for generic integrable functions, even for generic continuous functions. So, the smoothness of f is essential.

On the other hand, quite surprisingly, for many interesting dynamical systems one can actually prove the above central limit theorem, and obtain good estimates on the decay of correlations for smooth functions f. This opens the door to close interaction between the theory of dynamical systems and probability theory and statistical mechanics, which is currently a very active area of research.

Exercises:

II.5.1. Let $T : X \to X$ be a mixing map preserving a probability μ. Prove that for any $f, g \in L^2(X)$

$$\int_X f(T^n x) g(x)\, d\mu \to \int_X f(x)\, d\mu \cdot \int_X g(x)\, d\mu$$

Hint: first consider two simple functions f, g, and then approximate L^2 functions by simple ones.

II.5.2. Under the conditions of Poincaré recurrence theorem, assume that $f : X \to \mathbb{C}$ is a measurable function such that $f(x) \neq 0$ almost everywhere. Show that the sequence $S_n(x)$ defined by (II.5.1) diverges almost everywhere. Hint: consider the sets $A_n = \{x \in X \colon |f(x)| > 1/n\}$.

Chapter III

Lyapunov exponents. Pesin Theory

III.1 Lyapunov Exponents

Let p be a fixed point of a diffeomorphism $f : A \to A$ of an open set $A \subset \mathbb{R}^d$. We want to study the behaviour of f^n $(n \in \mathbb{Z})$ in a neighborhood of p. As a first approximation to f we consider its linear part (i.e., the derivative) $f'_p : \mathbb{R}^d \to \mathbb{R}^d$.

Let

$$\alpha_1, \alpha_2, \ldots, \alpha_r, \alpha_{r+1}, \bar{\alpha}_{r+1}, \ldots, \alpha_s, \bar{\alpha}_s$$

be all the distinct roots of the characteristic polynomial of f'_p. In this sequence, we denoted by $\alpha_1, \ldots, \alpha_r$ all the real roots and by $\alpha_j, \bar{\alpha}_j$, $r+1 \le j \le s$, all the conjugate pairs of complex roots. Let also $\tilde{m}_j$, $1 \le j \le s$, be the respective multiplicities. The theorem of Jordan (real canonical form) says that the roots (eigenvalues of f'_p) are associated to f'_p-invariant generalized eigenspaces $\tilde{E}_j$, $1 \le j \le s$, whose respective dimensions equal $\tilde{m}_j$ for $1 \le j \le r$ and $2\tilde{m}_j$ for $r < j \le s$ (in the latter case the space $\tilde{E}_j$ is associated to the pair $\alpha_j, \bar{\alpha}_j$). Moreover, $\mathbb{R}^d = \oplus_{j=1}^s \tilde{E}_j$. Note that since f is a diffeomorphism, then $\det f'_p \neq 0$, hence we have $\alpha_j \neq 0$ for all j.

If v_i is an eigenvector of $\tilde{E}_i$, then $(f^n)'_p v_i = \alpha_i^n v_i$, hence

$$\log \|(f^n)'_p v_i\| = n \log |\alpha_i| + \log \|v_i\|$$

for all $n \in \mathbb{Z}$. While this is not true for *any* vector $v_i \in \tilde{E}_i$, it is true that

$$\lim_{n\to\pm\infty} \frac{1}{n} \log \|(f^n)'_p\, v_i\| = \log|\alpha_i| \tag{III.1.1}$$

for every $\vec{0} \neq v_i \in \tilde{E}_i$, see Exercise III.1.1. This shows that when $|\alpha_i| > 1$, the vector $(f^n)'_p\, v_i$ grows exponentially fast as $n \to \infty$ and shrinks exponentially as $n \to -\infty$. If $|\alpha_i| < 1$, then it is vice versa. If $|\alpha_i| = 1$, then there is no exponential growth or contraction, but there might be a slow (e.g., linear in n) growth or contraction of the vector $(f^n)'_p\, v_i$.

The equation (III.1.1) suggests that we study not the eigenvalues α_i, but the logarithms of their moduli,

$$\lambda_i = \log|\alpha_i|$$

which are called the *Lyapunov exponents* of the map f at the fixed point p. We note that some distinct eigenvalues $\alpha_i \neq \alpha_j$ correspond to the same Lyapunov exponent if $|\alpha_i| = |\alpha_j|$. In this case each nonzero vector of the direct sum of the corresponding subspaces $\tilde{E}_i \oplus \tilde{E}_j$ satisfies (III.1.1). So, we have a decomposition of $\mathbb{R}^d$ as a direct sum of subspaces $E_1 \oplus \cdots \oplus E_{m(p)}$ such that if $\vec{0} \neq v_i \in E_i$, then

$$\lim_{n\to\pm\infty} \frac{1}{n} \log \|(f^n)'_p\, v_i\| = \lambda_i(p) \tag{III.1.2}$$

where $\lambda_i(p)$ is the Lyapunov exponent associated to E_i.

Moreover, if we suppose that there are no zero Lyapunov exponents at the fixed point p (we will say that such a point p is *hyperbolic*), we can sum all the subspaces with negative Lyapunov exponents and all the subspaces with positive Lyapunov exponents to obtain, respectively, subspaces E^s and E^u, such that

(E1) $\mathbb{R}^d = E^s \oplus E^u$,

(E2) $f'_p(E^s) = E^s$ and $f'_p(E^u) = E^u$

(E3) there exist $\lambda > 0$ and $n_0 \geq 1$ such that for all $|n| \geq n_0$

$$\frac{1}{n} \log \|(f^n)'_p\, v\| \leq -\lambda \qquad v \in E^s, \quad \|v\| = 1$$

and

$$\frac{1}{n} \log \|(f^n)'_p\, v\| \geq \lambda \qquad v \in E^u, \quad \|v\| = 1$$

This means that the vectors of E^s are contracted by forward iterations of f'_p and expanded by backward iterations of f'_p. It is vice versa for the vectors of E^u. We note that it might happen that $E^u = \{\vec{0}\}$ or $E^s = \{\vec{0}\}$. We will be primarily interested in the study of hyperbolic points.

The theorem of Grobman-Hartman assures that at a hyperbolic fixed point p the behaviour of f is similar to its linear part. Precisely, to (f'_p)-invariant subspaces E^s, E^u there correspond f-invariant submanifolds $W^s(p)$, $W^u(p) \subset A$ (differentiably immersed in A) such that
(W1) $\mathcal{T}_p W^s(p) = E^s$ and $\mathcal{T}_p W^u(p) = E^u$,
(W2) there is a neighborhood $U(p) \subset A$ such that $f(W^s(p) \cap U(p)) \subset W^s(p)$ and $f(W^u(p) \cap U(p)) \subset W^u(p)$ and
(W3) we have

$$\lim_{n\to\infty} f^n(y) = p \qquad y \in W^s(p)$$

and

$$\lim_{n\to\infty} f^{-n}(y) = p \qquad y \in W^u(p)$$

We refer to [PM, KH] for a proof, and only note that the idea is representing f in a local coordinate system at p, associated with E^s and E^u, and then constructing $W^s(p)$ as a graph of a function in those coordinates by successful iterations of f^{-1}. The result for W^u follows by substitution of f for f^{-1}.

The above results easily extend to any diffeomorphism $f : A \to A$ of an open subset $A \subset M$ of a Riemannian manifold M, rather than $A \subset \mathbb{R}^d$. A Riemannian structure in M is necessary for the norm $\|\cdot\|$ to be well defined. The proofs are essentially the same as in the case of $\mathbb{R}^d$. Henceforth we assume that f is defined on an open subset of a Riemannian manifold.

It is necessary for our subsequent study to specify the speed of convergence in (W3). We denote by $\mathrm{dist}(\cdot,\cdot)$ the distance on the Riemannian manifold M. Let $\lambda > 0$ be such that no Lyapunov exponent lies in the interval $(-\lambda, \lambda)$. Then we can specify (W3) as
(W3') we have

$$\mathrm{dist}(f^n(y), p) \le Ce^{-\lambda n} \qquad y \in W^s(p)$$

and

$$\mathrm{dist}(f^{-n}(y), p) \le Ce^{-\lambda n} \qquad y \in W^u(p)$$

for all $n \ge 0$ and some constant $C > 0$.

If p is not a fixed point, but a periodic one, with period k, all these results apply to the map $f^k : A \to A$.

If p is not a periodic point, we still can define Lyapunov exponents in a similar way:

Definition. Let the map f^n be differentiable at a point $p \in M$ for all $n \in \mathbb{Z}$. Assume that the tangent space $\mathcal{T}_p M$ is a direct sum of subspaces $E_1 \oplus \cdots \oplus E_{m(p)}$ such that if $\vec{0} \neq v_i \in E_i$, then

$$\lim_{n\to\pm\infty} \frac{1}{n} \log \|(f^n)'_p\, v_i\| = \lambda_i(p) \tag{III.1.3}$$

Then the values $\lambda_i(p)$ are called *Lyapunov exponents* at the point p, whose multiplicities are $\dim E_i$.

We note that the existence of the limit (III.1.3) is not guaranteed for any point $p \in A$, we will return to this issue in the next section. For now, we will say that p has all Lyapunov exponents if the above limits exist.

If a point p has all Lyapunov exponents and none of them is zero, we call p a *hyperbolic point*. For a hyperbolic point $p \in M$, we have $\mathcal{T}_p M = E_p^s \oplus E_p^u$, where

$$E_p^s = \oplus_{\lambda_i(p)<0} E_i \quad \text{ and } \quad E_p^u = \oplus_{\lambda_i(p)>0} E_i \tag{III.1.4}$$

All the above results apply to hyperbolic (nonperiodic) points, including the existence of submanifolds $W^s(p)$ and $W^u(p)$, but the property (W3') above has to be modified accordingly:
(W3") we have

$$\mathrm{dist}(f^n(y), f^n(p)) \leq Ce^{-\lambda n} \qquad y \in W^s(p)$$

and

$$\mathrm{dist}(f^{-n}(y), f^{-n}(p)) \leq Ce^{-\lambda n} \qquad y \in W^u(p)$$

for all $n \geq 0$ and some constant $C > 0$.

Let us look at the above properties closely. They mean that the forward orbits of the points of $W^s(p)$ are getting close to each other (converge) exponentially fast. For this reason $W^s(p)$ is called the *stable manifolds* (the term comes from differential equations, where the convergence of solutions is interpreted as stability). The forward orbits of the points of $W^u(p)$ get separated (diverge) exponentially fast, and for this reason $W^u(p)$ is called the *unstable manifold*. Note, though, that the backward orbits of $W^u(p)$ converge, and the backward orbits of $W^s(p)$ diverge.

We see that the orbits of all the points near a hyperbolic point p are very unstable: they diverge (get separated) exponentially fast either in the future or in the past, or both. Indeed, if $\dim W^s(p) \neq 0$ and $\dim W^u(p) \neq 0$, then for any point y close to p and not exactly lying on $W^u(p)$ or $W^s(p)$, the trajectory of y separates from that of p both in the future and in the past! That exponential separation of trajectories is the main source of instability, turbulence, mixing – all that we call *chaos*.

The studies of dynamical systems with hyperbolic points have began long ago. Around 1900 J. Hadamard proved the hyperbolicity for geodesic flows on manifolds of constant negative curvature. In the 1930s J. G. Hedlund and E. Hopf studied the ergodic properties of these flows.

Based on the studies of geodesic flows, in the 1960s D. Anosov (and S. Smale, in a different form) introduced general classes of diffeomorphisms with hyperbolic points. We provide the definition of what is now called an Anosov diffeomorphism (which Anosov himself originally called "diffeomorphisms with a condition", or C-diffeomorphisms, C stands for "condition").

Definition. A diffeomorphism $f : M \to M$ of a compact Riemannian manifold is said to be *Anosov* if there are constants $K > 0$, $\lambda > 0$ such that at each point $p \in M$ a decomposition $\mathcal{T}_pM = E_p^s \oplus E_p^u$ exists with the properties

(A1) E_p^s and E_p^u are f'-invariant, i.e. $f'_p(E_p^s) = E_{f(p)}^s$ and $f'_p(E_p^u) = E_{f(p)}^u$,

(A2) for all $n \geq 0$

$$\|(f^n)'_p v\| \leq K e^{-n\lambda} \|v\| \qquad v \in E_p^s$$

$$\|(f^{-n})'_p v\| \leq K e^{-n\lambda} \|v\| \qquad v \in E_p^u$$

(A3) the spaces E_p^s and E_p^u depend on p continuously.

The condition (A3) is included by tradition, it actually follows from the other two.

Note that Anosov's conditions (A1) and (A2) are actually weaker than the requirements in the definition of Lyapunov exponents. Indeed, Lyapunov exponents need not exist at all points for an Anosov diffeomorphism. Nonetheless, the conditions (A1) and (A2) capture the essence of hyperbolicity, in particular at all $p \in M$ there exist stable and unstable manifolds $W^s(p)$ and $W^u(p)$ satisfying (W3"), and their tangent planes are E_p^s and E_p^u, respectively. For these reasons, points that satisfy the conditions (A1) and (A2) are frequently called *hyperbolic* (even when Lyapunov exponents technically do not exist there). This may seem a confusing terminology, but it will be essentially clarified in the next section.

Remark. The constant K in the condition (A2) obviously plays an auxiliary role. It is remarkable that one can always change the Riemannian metric on M so that the condition (A2) will hold with $K = 1$, i.e. the constant K can be dropped. In this case the contraction of stable vectors $v \in E_p^s$ and the expansion of unstable vectors $v \in E_p^u$ under $(f^n)'_p$ will be monotonic in n. Such a metric, in which $K = 1$, is called the *adapted metric* or *Lyapunov metric*.

Exercises:

III.1.1. Verify the formula (III.1.1). Assume, for simplicity, that $\dim \tilde{E}_i = 2$. There are two cases here. If α_i is a real root of multiplicity 2, then f' restricted to $\tilde{E}_i$ is given by a Jordan matrix $J = \begin{pmatrix} \alpha_i & 1 \\ 0 & \alpha_i \end{pmatrix}$. Verify that $J^n = \begin{pmatrix} \alpha_i^n & n\alpha_i^{n-1} \\ 0 & \alpha_i^n \end{pmatrix}$ for all $n \in \mathbb{Z}$ and then derive (III.1.1). If $\alpha_i = a+bi$ is a complex root, $b \neq 0$, then the corresponding Jordan canonical form is $J = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$. Verify that $J^n = |\alpha_i|^n \begin{pmatrix} \cos n\varphi & \sin n\varphi \\ -\sin n\varphi & \cos n\varphi \end{pmatrix}$ for some $\varphi \in [0, 2\pi)$ and all $n \in \mathbb{Z}$, and then derive (III.1.1).

III.2 Oseledec's Theorem

Here we are concerned with the existence of Lyapunov exponents, in other words the existence of the limit (III.1.3). The following example shows that there may be plenty of points where Lyapunov exponents fail to exist.

Example. Let $f : S^1 \to S^1$ be a circle diffeomorphism given by $f(x) = x + \frac{1}{3\pi} \sin 2\pi x$, where $0 \leq x < 1$ is the cyclic coordinate on S^1. We have two fixed points here, $x_0 = 0$ and $x_1 = 1/2$, both with Lyapunov exponents: $\lambda(x_0) = \log f'(x_0) = \log(5/3) > 0$ and $\lambda(x_1) = \log f'(x_1) = \log(1/3) < 0$. Since $\lambda(x_0) > 0$, the point x_0 is unstable (a repeller), likewise, x_1 is a stable point (an attractor). For any point $p \in (0, 1/2)$ we have $f^n(p) \to 1/2$ and $f^{-n}(p) \to 0$ as $n \to \infty$. Hence, by the chain rule, for any nonzero vector $v \in \mathcal{T}_p(S^1)$ we have

$$\log(5/3) = \lim_{n \to -\infty} \frac{1}{n} \log \|(f^n)'_p v\| \neq \lim_{n \to \infty} \frac{1}{n} \log \|(f^n)'_p v\| = \log(1/3)$$

This shows that the limit in (III.1.3) does not exist. The same conclusion holds for any $p \in (1/2, 1)$.

We now see that Lyapunov exponents only exist at two fixed points, x_0 and x_1, and nowhere else on S^1. Hence, it seems like Lyapunov exponents are "very rare" in any reasonable sense: by simple count, topologically, with respect to the Lebesgue measure... But stop! Why should we care about the Lebesgue measure, if it is not invariant under f? The measure-theoretic point of view developed in Chapter II prescribes to study f with the help of invariant measures. It is not hard to find all the invariant measures of the map f here (see Ex. III.2.1). Any invariant measure is supported by the two point set $\{x_0, x_1\}$. So, with respect to any invariant measure, Lyapunov exponents exist almost everywhere!

It is remarkable that the above fact is very general, and this is the content of Oseledec[1] multiplicative ergodic theorem, which we call shortly Oseledec's theorem. The version of that theorem which we state below is sufficient for all piecewise smooth maps studied in this book, including billiards in Chapter IV, but there is a more general version involving *cocycles* that we will not discuss.

First, we introduce a general class of maps that are called smooth maps with singularities.

Definition. Let M be a finite union of compact Riemannian manifolds $M_1, M_2, \dots, M_s$ (perhaps with boundaries and with corners), all of the same dimension $d \geq 2$, glued along a finite number of C^1 submanifolds of positive codimensions. These submanifolds are contained in G, the union of a finite number of C^1 compact submanifolds of positive codimension in $M_1, \dots, M_s$. Then $V = M \setminus G$ is an open dense subset of M. Lastly, let $N \subset V$ be an open set and $f : N \to V$ a C^r diffeomorphism ($r \geq 1$) between N and its image, i.e. a diffeomorphic embedding of N into V. We call f a *smooth map with singularities.*

Note that the inverse map f^{-1} is well defined on the open set $f(N)$. Hence, f^{-1} is also a smooth map with singularities. We denote by

$$H = \cap_{n=-\infty}^{\infty} f^n N$$

the set of points where all the iterates of f are defined.

[1] The name is pronounced *Oseledets*.

Theorem III.2.1 (Oseledec, [Os]) *Assume that the map $f : N \to V$ preserves a Borel probability measure μ on M and $\mu(H) = 1$. If*

$$\int_M \log^+ \|(f)'_p\| \, d\mu(p) < \infty \quad \text{and} \quad \int_M \log^+ \|(f^{-1})'_p\| \, d\mu(p) < \infty$$

where $\log^+ s = \max\{\log s, 0\}$, then there exists an f-invariant set $E \subset H$, $\mu(E) = 1$, such that for every point $p \in E$ all the Lyapunov exponents exist.

In other words, Lyapunov exponents exist almost everywhere (with respect to any invariant measure). Since it is standard in ergodic theory to ignore sets of zero measure, we can simplify matters by thinking that "Lyapunov exponents always exist". This explains why the points satisfying Anosov's conditions (A1) and (A2) in the previous section are said to be *hyperbolic* - the mere existence of Lyapunov exponents (almost everywhere) is guaranteed by Oseledec's theorem.

For $p \in \Gamma$ we denote by $\lambda_1(p) > \lambda_2(p) > \cdots > \lambda_{m(p)}(p)$ all distinct Lyapunov exponents and by $E_1(p), \ldots, E_{m(p)}(p)$ the corresponding subspaces in the tangent space $\mathcal{T}_pM$. For any real number $\kappa \in \mathbb{R}$ and $p \in \Gamma$ denote

$$E^-_{p,\kappa} = \oplus_{\lambda_i(p) \le \kappa} E_i(p) \qquad E^+_{p,\kappa} = \oplus_{\lambda_i(p) > \kappa} E_i(p)$$

(this is a generalization of E^s_p and E^u_p from the previous section).

Remark. Oseledec's theorem also includes the following fact, which we state separately. Let $\gamma_\kappa(p)$ denote the angle between the spaces $E^-_{p,\kappa}$ and $E^+_{p,\kappa}$. Then

$$\lim_{n \to \pm\infty} \frac{1}{n} \log \gamma_\kappa(f^n(p)) = 0 \tag{III.2.1}$$

i.e. the angle $\gamma_\kappa(f^n(p))$ slowly changes with n (more slowly than any exponential function).

Let $p \in N$ be a periodic point of period k and μ the atomic invariant measure supported by the finite set $\{p, f(p), \ldots, f^{k-1}(p)\}$, i.e. let $\mu(f^i(p)) = 1/k$ for $0 \le i < k$. Then the hypotheses of Osedelec's Theorem are satisfied, hence all the Lyapunov exponents exist at the points $p, f(p), \ldots, f^{k-1}(p)$. So, this theorem includes the results obtained in the previous section for periodic points as a particular case.

The proof of Oseledec's theorem is very difficult, and we omit it.

The existence of Lyapunov exponents leads the following question: can we construct stable and unstable manifolds as we did for fixed points in the previous section? In the next section we give an affirmative answer in a somewhat different context.

Exercises:

III.2.1. In the example $f : S^1 \to S^1$ discussed in the beginning of this section, find all f-invariant probability measures on S^1. Hint: consider a small interval $I = (a, b) \subset (0, 1/2)$ such that $f(I) \cap I = \emptyset$ and show that $\mu(I) = 0$ for any invariant probability μ.

III.2.2. In the context of Oseledec's Theorem, prove that $m(p)$ and $\lambda_1(p)$, $\ldots$,$\lambda_{m(p)}(p)$ are f-invariant functions and $f'_p(E_i(p)) = E_i(f(p))$ for all $1 \leq i \leq m(p)$. This has an important implication: if the measure μ is ergodic, then all the Lyapunov exponents are constant almost everywhere.

III.3 Pesin Theory. Nonuniform Hyperbolicity

We now turn back to the notion of hyperbolicity introduced in Section III.1. Recall that a point p is hyperbolic if it has all Lyapunov exponents and none of them equals zero. In addition, we agreed to call a point p hyperbolic if it satisfied Anosov's conditions (A1) and (A2) given in the end of Section III.1. We said that while those conditions technically did not imply the existence of Lyapunov exponents, they guaranteed that those exponents are different from zero (and are at least λ in absolute value!) whenever they existed.

It is now important to notice that the conditions (A1) and (A2) are actually much stronger than mere hyperbolicity, because the constant K is the same for all $p \in M$. Indeed, if we just assume that a point p is hyperbolic and none of its Lyapunov exponents lie in an open interval $(-\lambda, \lambda)$, then we can only deduce (see Exercise III.3.1) that for any $\varepsilon > 0$ there is a $K(p, \varepsilon) > 0$ such that for all $n \geq 1$

$$\|(f^n)'_p v\| \leq K(p, \varepsilon)\, e^{-n(\lambda-\varepsilon)} \|v\| \qquad v \in E^s_p \tag{III.3.1}$$

$$\|(f^{-n})'_p v\| \leq K(p, \varepsilon)\, e^{-n(\lambda-\varepsilon)} \|v\| \qquad v \in E^u_p \tag{III.3.2}$$

where E_p^s and E_p^u are defined by (III.1.4).

It is important that the constant $K(p,\varepsilon)$ above depends, generally, on the point p. In other words, it is the *uniformity* of the factor K in the conditions (A1)-(A2) that distinguishes Anosov diffeomorphisms from more general hyperbolic maps. Accordingly, maps that satisfy the conditions (A1)-(A2) with a uniform constant K for all points, are called *uniformly hyperbolic*. More general hyperbolic maps, which only satisfy the conditions (III.3.1)-(III.3.2) above, are called *nonuniformly hyperbolic*.

We note that when $K(p,\varepsilon)$ is large, then the effect of contraction of stable vectors and that of expansion of unstable vectors can only be seen for large enough n. So, the contraction and expansion are only asymptotic, and for arbitrarily long periods of time neither may occur. Moreover, stable vectors may temporarily expand under $(f^n)'_p$. This is all very characteristic to nonuniformly hyperbolic systems.

A systematic study of nonuniformly hyperbolic diffeomorphisms with absolutely continuous invariant measures was done by Ya. Pesin in mid-seventies [Pe1, Pe2]. In the eighties, Pesin's theory was extended to more general nonuniformly hyperbolic maps by F. Ledrappier, A. Katok, D. Ruelle, and L.-S. Young, among others. We follow the exposition by A. Katok and J.-M. Strelcyn [KS] that covers a large class of nonuniformly hyperbolic maps with singularities. That class includes physically important billiard models, which we will study separately in Chapter IV.

Let f be a smooth map with singularities as defined in the previous section. We will use the same notation M, G, V, N, and H. Let μ be an f-invariant measure satisfying the conditions of Oseledec's theorem. We additionally require that the class of smoothness of f be at least $r \geq 2$. The set $S := M \setminus N$ where the map f is not defined will be called the *singularity set*.

Let $d(x,S) = \inf\{\,\mathrm{dist}(x,y) : \ y \in S\}$ be the distance from $x \in N$ to S. For $x \in N$, we denote by $\exp_x\colon \mathcal{T}_x N \to N$ the exponential map (it is defined by $\exp_x(v) = \gamma(x,v,1)$, where $\gamma(x,v,t)$ is the geodesic in N defined by $\gamma(x,v,0) = x$ and $\dot\gamma(x,v,0) = v$). Let $R(x,N)$ be the radius of injectivity of the map $\exp_x\colon \mathcal{T}_x N \to N$, i.e. $R(x,N) = \sup\{r\colon \exp_x$ is one-to-one on the ball $B_r(0) \subset \mathcal{T}_x N\}$. The map $\exp_x$ is defined and injective on the ball $B_{r(x,N)}(0) \subset \mathcal{T}_x N$ where $r(x,N) = \min\{R(x,N), d(x,S)\}$.

We define $f_{ox} = \exp_{f(x)}^{-1} \circ f \circ \exp_x$. This is a proper way to represent f in a linear coordinate system near the point x. This map is well defined in a neighborhood of $0 \in \mathcal{T}_x N$.

We now make two additional assumptions made by Katok and Strelcyn [KS]:
(KS1) There are constants $C_1 > 0$ and $a > 0$ such that for all $\varepsilon > 0$ the μ-measure of the ε-neighborhood of S satisfies

$$\mu(U_\varepsilon(S)) \le C_1\, \varepsilon^a$$

i.e. the measure μ does not build up too much near the singularity set S.
(KS2) There are constants $C_2 > 0$ and $b > 0$ such that for every $x \in N$ and $v \in \mathcal{T}_x N$, $\|v\| \le r(x, N)$ we have

$$\|f''_{ox}(v)\| \le C_2\, d(\exp_x(v), S)^{-b}$$

i.e. the second derivative f''_{ox} does not grow too fast near the singularity set S.

Oseledec's theorem implies that μ-almost every point $x \in H$ has all Lyapunov exponents $\lambda_1(x) < \cdots < \lambda_{m(x)}(x)$. The set of all hyperbolic points in N is often called the *Pesin region* of f:

$$\Sigma(f) = \{x \in H : \lambda_i(x) \neq 0, \text{ for every } i = 1, \ldots, m(x)\}$$

We note that the Pesin region $\Sigma(f)$ is invariant under f. At every point $x \in \Sigma(f)$ we have the usual subspaces E^s_x and E^u_x defined by (III.1.4), and let

$$\lambda^+(x) = \min\{\lambda_i(x) > 0\} \qquad \text{and} \qquad \lambda^-(x) = \max\{\lambda_i(x) < 0\}$$

be the smallest (in absolute value) positive and negative exponents, respectively.

The following theorem is, in fact, a generalization of our early conditions (III.3.1), (III.3.2), cf. also Exercise III.3.1.

Theorem III.3.1 *Given $\varepsilon > 0$, there is a measurable function $C(x, \varepsilon)$ on $\Sigma(f)$ such that for all $x \in \Sigma(f)$, $n \ge 1$ and $m \in \mathbb{Z}$*

$$\|(f^n)'_{f^m x} v\| \le C(x,\varepsilon)\, e^{\lambda^-(x)n + \varepsilon n + \varepsilon|m|}\|v\| \qquad v \in E^s_x$$

$$\|(f^n)'_{f^m x} v\| \ge C^{-1}(x,\varepsilon)\, e^{\lambda^+(x)n - \varepsilon n - \varepsilon|m|}\|v\| \qquad v \in E^u_x$$

and the angle $\gamma(f^m x)$ between $E^s_{f^m x}$ and $E^u_{f^m x}$ satisfies

$$\gamma(f^m x) \ge C^{-1}(x,\varepsilon) e^{-\varepsilon|m|}$$

This theorem allows us to use the same argument as in Grobman-Hartman theorem and construct stable and unstable manifolds:

Theorem III.3.2 *For μ-almost every $x \in \Sigma(f)$ there is a $\delta(x) \in (0, r(x,N))$ such that for any small $\varepsilon > 0$ the set*

$$W^s(x) = \{y \in \exp_x B_{\delta(x)}(0) \colon \limsup_{n\to\infty} \frac{1}{n} \log \operatorname{dist}(f^n(x), f^n(y)) \le \lambda^-(x) + \varepsilon\}$$

is a C^r differentiable (stable) manifold. Similarly,

$$W^u(x) = \{y \in \exp_x B_{\delta(x)}(0) \colon \liminf_{n\to -\infty} \frac{1}{n} \log \operatorname{dist}(f^n(x), f^n(y)) \ge \lambda^+(x) - \varepsilon\}$$

is a C^r differentiable (unstable) manifold. We also have

$$\mathcal{T}_x W^s(x) = E^s_x \qquad \text{and} \qquad \mathcal{T}_x W^u(x) = E^u_x$$

We note that the stable and unstable manifolds at x are transversal, i.e. they intersect in the point x alone and the angle between them is positive. We also note that $W^s(x)$ and $W^u(x)$ exist almost everywhere, but not necessarily everywhere, on $\Sigma(f)$. Their existence is an important issue, so we provide some more details. These manifolds fail to exist at points $x \in \Sigma(f)$ whose trajectories approach the singularity set S too fast. More specifically, let

$$d(f^{n_k}x, S) < \beta^{n_k} \qquad \text{for some} \qquad \beta < e^{-\lambda_{m(x)}(x)}$$

and for some infinite sequence $n_k \to \infty$. Then $W^s(x)$ cannot exist. Indeed, if it did exist, it would have contracted too slowly under the iterates of f and so would hit the set S sooner or later, i.e. for some $n_k > 0$ we would have $f^{n_k}(W^s(x)) \cap S \neq \emptyset$, which is impossible by the definition of $W^s(x)$ given in the above theorem.

Fortunately, our assumption (KS1) guarantees that for almost every point $x \in N$ there are $c(x) > 0$ and $\kappa > 0$ such that

$$d(f^n(x), S) > c(x)\, |n|^{-\kappa}$$

for all $0 \neq n \in \mathbb{Z}$ (see Exercise III.3.2). This fact, in turn, can be used to derive the existence of $W^s(x)$ and $W^u(x)$ and even estimate their size, we omit further details.

Stable and unstable manifolds can be efficiently used in the study of ergodic properties of the map f. We explain how it goes. Let $B \subset H$ be an f-invariant set such that $\mu(B) > 0$. Then the map $f_B := f|_B$ (the restriction of f to B) preserves the probability μ_B, which is obtained by conditioning the measure μ on B. We recall that the map $f_B\colon B \to B$ is ergodic iff every function $g \in \mathcal{L}^1(B)$ that is f_B-invariant is constant almost everywhere on B, cf. Section II.2.

A classical method to construct a set B on which the map f_B is ergodic (with respect to the measure μ_B) uses stable and unstable manifolds and goes back to E. Hopf. Take any point $x \in \Sigma(f)$ and its stable and unstable manifolds $W^s(x)$, $W^u(x)$. Let g be an f-invariant function, and assume, for simplicity, that it is continuous on M (integrable functions then can be approximated by continuous ones, but this step is purely technical, and we omit it). Then one can show that g is constant on $W^u(x) \cup W^s(x)$, see Exercise III.3.3. Since this fact applies to any other point of the set $\Sigma(f)$ as well, we can proceed as follows. Start by fixing $x \in \Sigma(f)$ and construct a set B_1 (a first approximation to B) as the union of all unstable manifolds $W^u(y)$, $y \in W^s(x)$, and all stable manifolds $W^s(z)$, $z \in W^u(x)$. The function g then must be constant on B_1.

We can defining sets B_n, $n \geq 2$, recursively by

$$B_n = \cup\{W^u(y) \cup W^s(y) : \, y \in B_{n-1}\}$$

One can easily show, by induction on n, that the function g is constant on B_n for each n. We now put $B_\infty = \cup_n B_n$. Clearly, the function g is constant on the entire set B_∞. Then it is enough to put

$$B = \cup_{n=-\infty}^{\infty} f^n(B_\infty) \tag{III.3.3}$$

and the map $f_B : B \to B$ will be ergodic.

We now describe the set B_∞. One can easily verify that B_∞ consists of all the points $y \in \Sigma(f)$ for which there exists a finite sequence $x = z_0, z_1, \ldots, z_{k-1}, z_k = y$ with the property that for all $0 \leq i \leq k-1$ either $W^s(z_i) \cap W^u(z_{i+1}) \neq \emptyset$ or $W^u(z_i) \cap W^s(z_{i+1}) \neq \emptyset$ (clearly, all $z_i \in B_\infty$). In other words, for any $y \in B_\infty$ there is a chain of stable and unstable manifolds that joins the point y with the original point x. Such a chain is usually called *Hopf chain* or a *zig-zag*. Now one can say that the set B_∞ is the union of all Hopf chains (or zig-zags) starting at x.

Now let us look how big the set B_∞ is. The stable and unstable manifolds, $W^s(y)$ and $W^u(y)$, are transversal to each other at any $y \in \Sigma(f)$, i.e.

the angle between them is positive (and, in fact, it is bounded away from zero in a neighborhood of x). Note also that $\dim W^s(y) + \dim W^u(y) = \dim M$ by hyperbolicity. Assume for simplicity that all the stable and unstable manifolds $W^s(y)$, $W^u(y)$ in a neighborhood of x are large enough, say, let $\mathrm{dist}(y, \partial W^s(y)) \geq c$ and $\mathrm{dist}(y, \partial W^u(y)) \geq c$ for some constant $c > 0$ and all y close to x. Then one can easily prove that the set B_∞ contains an open neighborhood of x in the Pesin region $\Sigma(f)$. This fact is called *local ergodicity* (or sometimes local ergodic theorem). We will return to local ergodicity in the next section.

In systems with singularities, though, stable and unstable manifolds can be arbitrarily short. This happens for the same reason as why they sometimes fail to exist at all, as we have shown above. Hence, B_∞ may not cover any open neighborhood of x in $\Sigma(f)$, i.e. there might be some tiny islands left out arbitrarily close to x. But it is still possible to show that the set B_∞ has a positive μ measure. The set B defined by (III.3.3) is called an *ergodic component* of the map f. The following theorem given without proof summarizes our discussion:

Theorem III.3.3 (Pesin [Pe2]) *Let $\mu(\Sigma(f)) > 0$. There exist sets $\Sigma_i \subset \Sigma(f)$, $i = 0, 1, 2, \ldots, J$ $(J \leq +\infty)$ such that*
(i) $\Sigma_i \cap \Sigma_j = \emptyset$ *for* $i \neq j$ *and* $\cup_i \Sigma_i = \Sigma(f)$*;*
(ii) $\mu(\Sigma_0) = 0$ *and* $\mu(\Sigma_i) > 0$ *for* $i > 0$*;*
(iii) $f(\Sigma_i) = \Sigma_i$ *for* $i \geq 0$
(iv) $f|_{\Sigma_i}$ *is ergodic with respect to* μ_{Σ_i} *for* $i > 0$.
Furthermore, for every $i > 0$ we have $\Sigma_i = \Sigma_{i,1} \cup \cdots \cup \Sigma_{i,J(i)}$ with some $1 \leq J(i) < \infty$ such that
(v) $\Sigma_{i,j} \cap \Sigma_{i,k} = \emptyset$ *for* $j \neq k$*;*
(vi) $f(\Sigma_{i,j}) = \Sigma_{i,j+1}$ *for* $1 \leq i < J(i)$ *and* $f(\Sigma_{i,J(i)}) = \Sigma_{i,1}$*;*
(vii) *the map $f^{J(i)}$ restricted to $\Sigma_{i,j}$ is mixing for every $1 \leq j \leq J(i)$.*

According to a tradition, the partition of $\Sigma(f)$ into the sets $\Sigma_{i,j}$ is called *spectral decomposition*. The above statement is referred to as the *spectral decomposition theorem*. This name comes from the fact that the numbers J and $J(i)$, $i \geq 1$, determine the leading eigenvalues (those, whose absolute values equal one), and their multiplicities, of the unitary operator associated with the map f on $\Sigma(f)$, as defined in Section I.8.

Definition. If $\mu(\Sigma(f)) = 1$, that is, if the Pesin region has full measure in N, we will say the map f is *nonuniformly hyperbolic*, or has *chaotic behaviour*.

Next, we state two other important theorems in the Pesin theory. They involve the entropy and periodic points of the map f. Their proofs, even on a heuristic level, are beyond the scope of this book, so we omit them completely.

We learned in Section I.9 that the entropy is a numerical characteristic of a chaotic dynamical system. It represents the exponential rate of growth of the complexity of the map f^n as $n \to \infty$. Now we also know that positive Lyapunov exponents characterize the exponential rate of separation of trajectories under f^n as $n \to \infty$. It should then come as no surprise that the entropy $h_\mu(f)$ is closely related to the positive Lyapunov exponents of f, as stated below:

Theorem III.3.4 (Pesin Formula) *Assume that the measure μ is absolutely continuous with respect to the Lebesgue measure on N, and assume an extra technical condition (KS3) stated below. Then the entropy of f with respect to the measure μ is*

$$h_\mu(f) = \int_H \sum_{\lambda_i(x)>0} \lambda_i(x) \cdot \dim E_i(x)\, d\mu \tag{III.3.4}$$

i.e. the entropy equals the average sum of all positive Lyapunov exponents, counting multiplicities.

The extra technical condition in the theorem is this:
(KS3) there exist $C_3 > 0$ and $c > 0$ such that for every $x \in N$ we have

$$\|f'_x\| < C_3\, d(x, S)^{-c}$$

It somewhat strengthens the assumptions of Oseledec's theorem, but not very much.

Remark. Actually, Theorem III.3.4, as stated, is the original version of Pesin's formula. Later (III.3.4) has been extended to measures μ that are not absolutely continuous on N but whose conditional measures on unstable manifolds are absolutely continuous. The latter means the following: if we take a measurable partition ξ of N such that each atom (element) $A \in \xi$ is a domain on an unstable manifold, i.e. $A \subset W^u(x)$ for some $x \in N$, then the conditional measure $\mu|_\xi$ is absolutely continuous with respect to the Lebesgue measure on unstable manifolds. Such measures are now usually called *Sinai-Ruelle-Bowen measures*[2] (or *SRB measures*).

[2] These measures were first studied by Sinai [Si3], Ruelle [Ru1] and Bowen [Bo2] in the seventies.

Pesin formula (III.3.4) implies that if $\mu(\Sigma(f)) > 0$, then $h_\mu(f) > 0$ for absolutely continuous (and SRB) measures. Katok showed that this fact is actually more general:

Proposition III.3.5 (Katok [Ka]) *Assume that the measure μ has no atoms, i.e. $\mu(\{x\}) = 0$ for every point $x \in N$. If $\mu(\Sigma(f)) > 0$ and the restriction of f to the Pesin region $\Sigma(f)$ is ergodic, then $h_\mu(f) > 0$.*

Another theorem in the Pesin theory involves periodic points of the map f. Let

$$P_n(f) = \#\{x \in N : f^n(x) = x\}$$

be the number of periodic points of period n (more precisely, this is the number of fixed points for f^n) We also denote by $\operatorname{supp}\mu$ the support of the measure μ (i.e. the minimal closed set of full measure), defined by $\operatorname{supp}\mu = \cap\{E\colon \mu(E) = 1 \text{ and } E \text{ is closed}\}$.

Theorem III.3.6 (Katok [Ka]) *If $\mu(\Sigma(f)) = 1$, then*
(a) *Periodic points are dense in the set* $\operatorname{supp}\mu$*;*
(b) $\limsup_{n\to\infty} \frac{\log P_n(f)}{n} \geq h_\mu(f)$

Katok's theorem III.3.6 implies that the number of periodic points $P_n(f)$ grows at least exponentially in n, and the exponent is at least the entropy of f. We note that for Anosov diffeomorphisms defined in Section III.1 a more precise result holds:

Theorem III.3.7 (Bowen [Bo1]) *If $f : M \to M$ is an Anosov diffeomorphism of a compact manifold M, then*

$$\lim_{n\to\infty} \frac{\log P_n(f)}{n} = h(f) := \sup_\mu h_\mu(f)$$

where the supremum is taken over all f-invariant measures on M. The quantity $h(f)$ is independent of a measure and called the topological entropy *of f.*

One might wonder what makes periodic points so interesting in the theory of smooth dynamical systems, because in most cases there are only countably many of periodic points in M and their measure is zero. So, from the measure-theoretic point of view, they seem to mean nothing. However, periodic points play an important role in physical and experimental works, and there are several reasons why they are so relevant.

First, periodic orbits are very simple – they are completely specified by finite iterations of the map – which makes numerical calculations of their Lyapunov exponents and other characteristics feasible with high accuracy. Second, they are relatively easy to find in physically interesting examples – given n, a computer program can find all periodic points of period n. Third, by part (a) of Katok's theorem III.3.6, any finite orbit $\{f^i(x)\}$, $1 \leq i \leq n$, can be arbitrarily well approximated by a periodic orbit, i.e. for any $\varepsilon > 0$ there is a periodic point y such that $\text{dist}(f^i(x), f^i(y)) < \varepsilon$ for all $1 \leq i \leq n$. The last property actually allows to approximate any invariant measure μ by measures supported on periodic orbits. Such approximations were found by Bowen for Anosov diffeomorphisms and later have been used quite efficiently in physical and experimental studies of other chaotic systems.

To summarize it, periodic points allow us to study many delicate properties of smooth chaotic systems. Said H. Poincaré in 1892: "What makes these periodic solutions so precious to us is that they are, so to speak, the only breach through which we may attempt to penetrate an area hitherto deemed inaccessible", see [Po], Vol 1 ,#36.

We finally discuss relations between three most important properties in the theory of smooth chaotic dynamical systems:

1. Hyperbolicity (i.e., nonvanishing of all Lyapunov exponents, when $\mu(\Sigma(f)) = 1$);

2. Positivity of entropy: $h_\mu(f) > 0$;

3. Ergodicity (and its stronger version - **mixing**).

Each of these properties represents *chaos* in dynamical systems in some way. Even though there is no commonly adopted definition of the term *chaos*, each of the above properties could be (and at times, was) regarded as a possible formal definition of chaos.

However, these properties are not equivalent, and in fact, none of them logically implies any of the other two (even though, under some conditions, the hyperbolicity does imply the positivity of entropy, as in Pesin's Theorem III.3.4 and Katok's Proposition III.3.5). But it is more important for us to realize that these three properties characterize different aspects of chaotic behavior of dynamical systems. The positivity of entropy is the most *local* characteristic of all – even a tiny f-invariant subset $E \subset M$ of a small but positive measure, $\mu(E) > 0$, can be used to make $h_\mu(f) > 0$, while on the rest of the space $M \setminus E$ the map f may not be chaotic in any sense. The hyperbolicity is also a *local* condition (it characterizes the separation of *nearby* trajectories), but at least it must hold almost everywhere, since it requires $\mu(\Sigma(f)) = 1$. Still, there is no guarantee that a hyperbolic map f is ergodic or mixing – according to the spectral decomposition theorem III.3.3, the Pesin region $\Sigma(f)$ can be decomposed into up to a countably many noninteracting subregions!

The ergodicity is the most *global* property of the three. However, ergodicity alone does not imply hyperbolicity, or chaos for that matter. For example, consider a circle rotation $T : S^1 \to S^1$ through an irrational angle $\theta > 0$. It is ergodic, see Exercise II.4.1a, but it is not truly chaotic – its only Lyapunov exponent is obviously zero, and so is its entropy (see Exercise I.9.3).

One final remark. The failure of ergodicity in systems with positive entropy and large Pesin regions is a fairly common phenomenon in Hamiltonian systems. It was extensively studied by Kolmogorov, Arnold and Moser, whose theory is now called *KAM theory*. They discovered that for typical Hamiltonian maps $f : M \to M$, there are some f-invariant regions $D \subset M$ of positive measure where the dynamics is very stable. This means that all Lyapunov exponents in D are zero and each region D contains a subset of positive measure, which is the union of f-invariant curves or tori (those are called *KAM tori*) – a feature that makes the dynamics inside D almost integrable. Such stable regions occur around periodic points $p \in M$, $f^n(p) = p$, when the spectrum of the derivative $(f^n)'_p \colon \mathcal{T}_pM \to \mathcal{T}_pM$ lies on the unit circle (there are some other technical conditions that we omit). Such periodic points are called *elliptic* (as opposed to hyperbolic), and the stable regions around them are called *elliptic islands* or *stability islands*. In a typical Hamiltonian system, stability islands coexist with a large Pesin region $\Sigma(f)$ of positive measure. The Pesin region is sometimes called the *chaotic sea*. In that sea, one can find small stable islands separated from $\Sigma(f)$ by some f-invariant curves. Such a picture is typical as various numerical experiments show. The phenomenon of coexistence of chaotic and stable behavior in dynamical systems is not very well understood yet, and it is one of the most intriguing issues in modern mathematical physics.

Exercises:

III.3.1. Derive the conditions (III.3.1)-(III.3.2) from the definition of Lyapunov exponents (III.1.3). Hint: first show that for any given nonzero vector $v \in E_p^s$ (resp., $v \in E_p^u$), there is a $K(p,\varepsilon)$ that satisfies (III.3.1)-(III.3.2), but depends on v. Then pick an orthonormal basis $e_1, \dots, e_k$ in E_p^s (resp., E_p^u), ensure (III.3.1)-(III.3.2) with the same constant $K(p,\varepsilon)$ for all vectors $e_1, \dots, e_k$ and use the triangle inequality to derive (III.3.1)-(III.3.2) for an arbitrary vector v.

III.3.2. Let $H_\kappa = \{x \in H\colon d(f^n(x), S) > c(x)\,|n|^{-\kappa}$ for some $c(x) > 0$ and all $0 \neq n \in \mathbb{Z}\}$. Show that $\mu(H_\kappa) = 1$ for some $\kappa > 0$. Hint: note that the set $\tilde{H}_{\kappa,n} := \{d(f^n(x), S) < |n|^{-\kappa}\}$ coincides with $f^{-n}(U_{|n|^{-\kappa}}(S))$, hence $\mu(\tilde{H}_{\kappa,n}) = \mu(U_{|n|^{-\kappa}}(S))$. Now put $\kappa = 2/a$ and use (KS1) to prove that $\sum_n \mu(\tilde{H}_{\kappa,n}) < +\infty$. Notice that H_κ consists of points that belong to only finitely many sets $\tilde{H}_{\kappa,n}$.

III.3.3. Let g be a continuous f-invariant function on M. Show that g is constant on every stable and unstable manifold. Hint: Since M is compact, the function g is uniformly continuous. Then use the fact that $f^n(W^s(x))$ and $f^{-n}(W^u(x))$ shrink as $n \to \infty$.

III.4 Sufficient Conditions for Hyperbolicity and Ergodicity in Smooth Maps with Singularities

In this section we continue studying smooth maps with singularities and using the same notation M, G, V, N, H, μ, f as in the previous sections.

Here we describe sufficient conditions under which two main chaotic properties – hyperbolicity and ergodicity – hold. These conditions are simple enough to verify, so that they have been successfully used in the studies of many physically interesting models. The next chapter presents our main class of examples – billiards.

Quadratic forms. We recall that a quadratic form B in $\mathbb{R}^d$ is a function $B : \mathbb{R}^d \to \mathbb{R}$ such that $B(u) = Q(u,u)$, where Q is a bilinear symmetric function on $\mathbb{R}^d \times \mathbb{R}^d$. Equivalently, $B : \mathbb{R}^d \to \mathbb{R}$ is a quadratic form if there exists a symmetric matrix A such that $B(u) = u^t A u$ for $u \in \mathbb{R}^d$ (here u^t means transposition of a column-vector u). A *quadratic form* B on M is a function $B : \mathcal{T}M \to \mathbb{R}$ such that its restriction B_x to $\mathcal{T}_x M$ at every point $x \in M$ is a quadratic form in the usual sense.

We say that a quadratic form B is *nondegenerate* at x if $B_x(u) \neq 0$ for every nonzero vector $u \in \mathcal{T}_x M$, $u \neq 0$ (equivalently, $\det A \neq 0$ for the corresponding symmetric matrix A). We say that B is *positive* (nonegative) if at every point x the form B_x is positive definite (positive semidefinite), i.e. $B_x(u) > 0$ (respectively, $B_x(u) \geq 0$) for all $0 \neq u \in \mathcal{T}_x M$. For a form B and $x \in M$, let $K_B(x)$ be the number of positive eigenvalues of the matrix that defines the form B_x (i.e., the maximal dimension of a subspace of $\mathcal{T}_x M$ on which the form is positive).

If $f : M \to M$ is a diffeomorphism, we denote by $f^\# B$ (the pull back of B by f) the quadratic form $(f^\# B)_x u = B_{f(x)}(f'_x u)$. One can easily verify that $f^\# B$ is also a quadratic form, and that $f^\# B$ is nondegenerate at x iff B is nondegenerate at $f(x)$. We note that

$$P = f^\# B - B$$

is a quadratic form, too.

For a quadratic form B defined on the orbit of a point x, we put

$$S_x := \{u \in \mathcal{T}_x M : B((f^n)'_x u) < 0, \ \ n \geq 0\}$$

$$U_x := \{u \in \mathcal{T}_x M : B((f^n)'_x u) > 0, \ \ n \leq 0\}$$

Theorem III.4.1 (Markarian [Ma6]) *Let* $B : \mathcal{T}M \to \mathbb{R}$ *be a nondegenerate quadratic form such that*
(i) B_x *depends measurably on* x,
(ii) *For every* $x \in H$ *we have* $0 < K_B(x) < \dim M$ *and* $K_B(f^n(x)) = K_B(x)$ *for all* $n \in \mathbb{Z}$.
(iii) $P_x = (f^\# B - B)_x$ *is positive for every* $x \in H$;
Then $\mu(\Sigma(f)) = 1$, *i.e. the map* f *is nonuniformly hyperbolic. Moreover, for every* $x \in \Sigma(f)$ *we have* $S_x = E^s_x$ *and* $U_x = E^u_x$.

We outline the proof for the case $\dim M = 2$ in Exercise III.4.1.

Remark. The condition (iii) can be slightly relaxed: it is enough to require that $P \geq 0$ and P is positive eventually, i.e. for almost every $x \in M$ there exists $k = k(x) \in \mathbb{N}$ such that $B((f^{k+1})'u) - B((f^k)'u) > 0$ for every non zero vector $u \in \mathcal{T}_x M$.

We note that the requirement $K_B(f^n(x)) = K_B(x)$ in hypothesis (ii) is automatically satisfied in two important cases:
(a) M has dimension 2. In that case $K_B(x) = 1$ for all $x \in H$;

(b) B is continuous on M, and M is connected. In this case $K_B(x)$ is a continuous integral-valued function on M, hence it is constant.

If the quadratic form is continuous on M, the spaces E^u_x and E^s_x depend on x continuously. In this case the theorem was proved in [Ma1].

Note that the converse also holds:

Theorem III.4.2 (Markarian [Ma6]) *If* $\mu(\Sigma(f)) = 1$, *there exist a quadratic form* $B : \mathcal{T}M \to \mathbb{R}$ *such that the conditions* (i) - (iii) *of Theorem III.4.1 are satisfied.*

Cones. Let $L \subset \mathbb{R}^d$ be a subspace, and $\alpha > 0$. Denote by $L^\perp$ the orthogonal complement to L. The set

$$C = \{u + v : \, u \in L, \, v \in L^\perp, \, \|u\| \geq \alpha \|v\|\}$$

is called a *cone* and L its *axis*. In the case $d = 3$ and $\dim L = 1$, it is the union of two symmetric circular cones with the common axis L. In the case $d = 2$ and $\dim L = 1$ the cone C is the union of two symmetric sectors, hence C is also called a *sector* in this case.

Let C be a cone in $\mathbb{R}^d$. Note that the complement $\mathbb{R}^d \setminus C$ is a set whose closure is also a cone (called the *complementary cone*, C^-), whose axis is $L^\perp$. The set

$$\partial C = \{u + v : \, u \in L, \, v \in L^\perp, \, \|u\| = \alpha \|v\|\}$$

is the common boundary of C and C^-.

Now let a cone $C(y)$ be defined in $\mathcal{T}_y M$ for almost every point $y \in M$. We call the collection $\{C(y)\}$ a *cone field* on M. The cone field is *invariant* under f if

$$f'_y(C(y)) \subset C(f(y)) \tag{III.4.1}$$

for almost every $y \in M$. It is *eventually strictly invariant* under f if for almost every $y \in M$ there exist $n = n(y)$ such that

$$(f^n)'_y(C(y)) \subset \operatorname{int} C(f^n(y)) \tag{III.4.2}$$

In a similar way, the invariance and eventual strict invariant under f^{-1} is defined, by replacing f' with $(f^{-1})'$ in (III.4.1) and $(f^n)'$ with $(f^{-n})'$ in (III.4.2).

We say that a cone field $\{C^u(x)\}$ is *unstable* if it is invariant and eventually strictly invariant under f and for almost every $x \in M$

$$\|f'_x(u)\| > \|u\| \qquad \text{for all} \;\; u \in \operatorname{int} C^u(x) \tag{III.4.3}$$

We say that a cone field $\{C^s(x)\}$ is *stable* if it is invariant and eventually strictly invariant under f^{-1} and for almost every $x \in M$

$$\|(f^{-1})'_x(u)\| > \|u\| \qquad \text{for all} \quad u \in \operatorname{int} C^s(x) \tag{III.4.4}$$

For a stable cone field $\{C^s\}$ defined on the orbit of a point x, we put

$$S_x := \cap_{n\geq 0}(f^{-n})'_x[C^s(f^n x)]$$

For an unstable cone field $\{C^u\}$ defined on the orbit of a point x, we put

$$U_x := \cap_{n\geq 0}(f^n)'_x[C^u(f^{-n} x)]$$

Theorem III.4.3 *Let $\{C^u(x)\}$ be an unstable cone field and $\{C^s(x)\}$ a stable cone field, both depending measurably on x. Let $\dim L^u(x) + \dim L^s(x) = \dim M$ for almost every x, where $L^u(x)$ and $L^s(x)$ are the axes of $C^u(x)$ and $C^s(x)$, respectively. Then $\mu(\Sigma(f)) = 1$. Moreover, for almost every $x \in \Sigma(f)$ we have $S_x = E^s_x$ and $U_x = E^u_x$.*

The proof of this theorem is outlined in Exercises III.4.1 and III.4.2.

We can relax the requirement of eventual strict invariance of the cone fields if, instead, assume the expansion (III.4.3) for *all* vectors $u \in C^u(x)$ and the backward expansion (III.4.4) for *all* vectors $u \in C^s(x)$.

Theorem III.4.3 and its various versions have been used in the studies of physical models since early seventies. A nice touch to this technique was done by M. Wojtkowski in [W1, W2]:

Theorem III.4.4 (Wojtkowski [W1]) *Let $\dim M = 2$ and the invariant measure μ absolutely continuous with respect to the Lebesgue measure on M. If there exists an invariant and eventually strictly invariant cone field $C(x)$ on M depending measurably on x, then $\mu(\Sigma(f)) = 1$.*

Note: in this theorem no requirement is made on the expansion (or contraction) of vectors in the cone $C(x)$! The expansion of the vectors $u \in C(x)$ under $(f^n)'_x$ for large $n > 0$ follows from the eventual strict invariance of the cone field and the absolute continuity of the invariant measure. Wojtkowski and Liverani [LW] also generalized Theorem III.4.4 to multidimensional maps, $\dim M > 2$, but then a special additional structure – an f-invariant symplectic form – is necessary. We omit this case.

Ergodicity of piecewise smooth hyperbolic maps. We now assume that the map f is hyperbolic, i.e. $\mu(\Sigma(f)) = 1$, and describe sufficient conditions under which f is ergodic. As we have seen in Theorem III.3.3, this is generally false – the Pesin region $\Sigma(f)$ only decomposes into finite or countable number of subregions Σ_i on which the restriction of f is ergodic. The sets Σ_i are called the ergodic components of f. Generally, they can be rather complicated and disconnected.

Under certain additional conditions, one can prove that f is ergodic, i.e. there is exactly one ergodic component $\Sigma_1 \subset \Sigma(f)$ and $\mu(\Sigma_1) = 1$. We assume that μ is absolutely continuous with respect to the Lebesgue measure on M. We also assume that the diffeomorphism $f : N \to f(N) \subset V$ is regular in the following sense:

(R1) the sets N and $f(N)$ are dense in M;

(R2) the boundaries ∂N and $\partial f(N)$ consist of of a finite number of C^1 submanifolds of codimension one;

(R3) on each connected component of N the map f can be extended by continuity to the boundary ∂N, and on each connected component of $f(N)$ the inverse map f^{-1} can be extended by continuity to the boundary $\partial f(N)$.

For technical reasons,we restrict ourselves to 2-dimensional maps, i.e. assume $\dim M = 2$. Hence, the sets ∂N and $\partial f(N)$ are one-dimensional, i.e. consist of smooth compact curves.

To visualize such a map f, let us imagine a unit square $M = [0,1]\times[0,1]$. We set $V = \operatorname{int} M$ and $G = \partial M$. Let us partition M into finitely many open domains in two ways:

$$M = B_1^+ \cup \cdots \cup B_r^+ = B_1^- \cup \cdots \cup B_r^-$$

so that the boundaries $\partial B_i^\pm$ are made by finitely many compact C^1 curves. The map f is defined separately on each domain B_i^+, $1 \le i \le r$, so that f is a C^r ($r \ge 2$) diffeomorphism of the interior of B_i^+ onto the interior of B_i^- and a homeomorphisms of B_i^+ onto B_i^-. Then we have $N = \cup_i \operatorname{int} B_i^+$ and $f(N) = \cup_i \operatorname{int} B_i^-$.

Such maps as above are called *piecewise smooth*. We will see more examples of this sort in the next chapter.

We have assumed that f is hyperbolic, i.e. $\mu(\Sigma(f)) = 1$. Hence, at almost every point $x \in M$ the stable and unstable spaces E_x^s and E_x^u exist (both are one-dimensional, since $\dim M = 2$). For simplicity, we assume that stable and unstable vectors expand monotonically, i.e.

$$\|f'_x(u)\| \ge \|u\|, \quad u \in E_x^u \qquad \text{and} \qquad \|(f^{-1})'_x(u)\| \ge \|u\|, \quad u \in E_x^s \qquad \text{(III.4.5)}$$

It is often possible to define the Riemannian metric on M so that (III.4.5) holds (such a metric is called adapted or Lyapunov metric, cf. Remark at the end of Section III.1). In some cases, a pseudo-metric satisfying (III.4.5) suffices, we will show this in Chapter IV. For now, we assume that an adaptive metric exists.

Definition. A point $x \in V$ is said to be *u-essential* if for any $A > 1$ there exist an $n \geq 1$ and a neighborhood U of the point x such that for all $y \in U$

$$\|(f^n)'_y(u)\| > A\|u\|, \ u \in E^u_y$$

(whenever the space E^u_y exists). Similarly, s-essential points are defined (by replacing f^n with f^{-n} and E^u_y with E^s_y).

Definition. A point $x \in V$ is said to be *sufficient* if there exist $A > 1$ and two integers $n < m$, such that $f^n(x)$ is defined, and there is a neighborhood U of the point $f^n(x)$ such that for all $y \in U$ we have

$$\|(f^{m-n})'_y(u)\| > A\|u\|, \ u \in E^u_y \quad \text{and} \quad \|(f^{m-n})'_y(v)\| < A^{-1}\|v\|, \ v \in E^s_y$$

(whenever the spaces E^u_y, E^s_y exist).

Let us also denote by $S^+ = \partial N$ the singularity set for the map f and by $S^- = \partial f(N)$ the singularity set for the map f^{-1}. For $n \geq 1$ we put $S^+_n = f^{-n+1}(S^+)$, this is the singularity set for f^n, and similarly $S^-_n = f^{n-1}(S^-)$, the singularity set for f^{-n}. Note that all these sets consist of compact C^1 curves.

Now we state five technical conditions for proving ergodicity:

Property 1 ("double singularities"). For any $n \geq 1$ the intersection $S^+_n \cap S^-_n$ is a finite or countable set.

Property 2 ("thickness of neighborhoods of singularities"). For any $\delta > 0$ let U_δ be the δ-neighborhood of the set $S^+ \cup S^-$. Then $\mu(U_\delta) \leq \text{const} \cdot \delta$.

Property 3. The families of stable and unstable subspaces E^s_x and E^u_x are continuous on their domains[3]. Furthermore, the limit spaces $\lim_{y \to x} E^u_y$ and $\lim_{y \to x} E^s_y$ are always transversal to each other at every sufficient point x, even if E^u_x or E^s_x does not exist.

Property 4 ("ansatz"). Almost every point of S^+ (with respect to the Lebesgue length on it) is u-essential, and almost every point of S^- is s-essential.

[3]This condition is automatically satisfied if the dependence of the quadratic form or the cone field on $x \in H$ is continuous.

Property 5. At almost every point $x \in S^+$ the subspace $E^u(x)$ is defined and transversal to S^+, and at almost every point $x \in S^-$ the subspace $E^s(x)$ is defined and transversal to S^-.

Theorem III.4.5 (Local ergodic theorem) *Under the above conditions, every sufficient point x has an open neighborhood U that belongs to one ergodic component, i.e. there is a component $\Sigma_i \subset \Sigma(f)$ such that $\mu(U \setminus \Sigma_i) = 0$.*

The first version of this theorem was obtained by Sinai [Si2] in 1970 in the studies of mathematical billiards. Then it was improved by Sinai and Chernov [SC2] in 1987. Later it was generalized by Kramli, Simanyi and Szasz [KSS2], Liverani and Wojtkowski [LW], and Markarian [Ma4] to other classes of hyperbolic systems of physical origin. In the most general form it was stated by Chernov [C1].

The idea of the proof of this theorem is essentially described in Section III.3: given a sufficient point x, one must connect it with almost all points $y \in U$ by Hopf chains. In other words, the union of all stable and unstable manifolds in U must be an arcwise connected subset of U of full measure.

One should remember, though, that stable and unstable manifolds (and, hence, Hopf chains) can be arbitrary short in M due to the singularities of f. In fact, the union of all singularity sets $\cup_n S_n^{\pm}$ is dense in M, hence stable and unstable manifolds can break or terminate abruptly anywhere. This is a serious problem in the proof of the local ergodic theorem III.4.5, which was first solved by Sinai in the context of billiards. He called his version of Theorem III.4.5 *the fundamental theorem in the theory of billiards.* Now the local ergodic theorem still remains one of the most advanced and difficult results in the theory of hyperbolic maps with singularities.

We now proceed to the proof of ergodicity of f. The local ergodic theorem essentially shows that the ergodic components Σ_i consist of open balls around sufficient points. Now we make the last assumption:

Property 6 ("abundance of sufficient points"). The set $R \subset M$ of sufficient points is arcwise connected and has full measure.

Theorem III.4.6 (Global ergodic theorem) *Under all the above conditions, the map f is ergodic, i.e. $\mu(\Sigma_1(f)) = 1$.*

Proof. Let $x, y \in R$. By the property 6, there is a continuous curve $\gamma \subset R$ that joins x and y. By the local ergodic theorem, for every point $z \in \gamma$ there is a neighborhood $U(z)$ that belongs in one component Σ_p (which yet may depend on z). Since γ is compact, it can be covered by a finite number of such open neighborhoods, i.e. there are $z_1, \dots, z_k$ such that $\gamma \subset \cup_{i=1}^k U(z_i)$. The neighborhoods $U(z_i)$ obviously overlap, hence they must belong to one ergodic component Σ_p, which then contains both x and y. Since R has full measure, Σ_p is the only ergodic component, and $\mu(\Sigma_p) = 1$. □

In practice, one has to verify the properties 1-6 to ensure the ergodicity of a hyperbolic map. This may not be so easy, but it was done successfully for many physically interesting systems.

Exercises:

III.4.1. Here we outline the proof of Theorem III.4.1 in dimension $d = 2$. The proof consists of several steps whose verification is left as an exercise:
(a) Let $x \in H$ and put $x_n = f^n(x)$ for $n \in \mathbb{Z}$. For each $n > 0$, pick a vector $w_n \in \mathcal{T}_{x_n} M$ such that $B_{x_n}(w_n) < 0$. Put $v_n = (f^{-n})'_{x_n} w_n$. As $P > 0$, it results that $B_x(v_n) < 0$. Then the sequence of normalized vectors $v_n / \|v_n\|$ has a subsequence that converges to a unit vector $v \in \mathcal{T}_x M$. One can now verify that $v \in S_x$;
(b) In a similar way we construct a unit vector $u \in U_x$. Hence both subspaces U_x and S_x exist, and so they must be one-dimensional;
(c) Given $a > 0$, denote by $D_a \subset M$ the set of points $y \in M$ such that $a\, B_y(w) \le \min\{\|w\|^2, P_y(w)\}$ for all $w \in \mathcal{T}_y M$. If a is small enough, then $\mu(D_a) > 0$, and in fact $\lim_{a \to 0} \mu(D_a) = 1$;
(d) By Birkhoff-Khinchin ergodic theorem applied to the indicator function I_{D_a} (see Exercise II.1.2b), for almost every point $x \in D_a$ we have $\lim_n \tau(x, D_a, n)/n = \tau(x, D_a) > 0$, where $\tau(x, D_a, n) = \#\{0 \le j < n : f^j(x) \in D_a\}$;
(e) From the positivity of P it now follows that $B((f^n)'_x u) \ge (1+a)^{\tau_n} B(u)$ for $u \in U_x$, where $\tau_n = \tau(x, D_a, n)$. Hence, $\liminf_n n^{-1} \log \|(f^n)'_x u\| \ge \log(1 + a)\tau(x, D_a)/2 > 0$. Therefore, x has a positive Lyapunov exponent and $E^u_x \supset U_x$ almost everywhere in D_a, hence in M;
(f) Similarly, $E^s_x \supset S_x$ almost everywhere.

III.4.2. By following the lines of the previous exercise, prove Theorem III.4.3. Hint: due to the eventual strict invariance of the cone field $\{C^u(x)\}$, for almost every $x \in M$ there is a $k = k(x)$ such that $(f^k)'_{f^{-k}x}(C^u(f^{-k}x)) \subset \operatorname{int} C^u(x)$; therefore, there is an $a = a(x) > 0$ such that for every $u \in (f^k)'_{f^{-k}x}(C^u(f^{-k}x))$ we have $\|f'_x(u)\| \geq (1+a)\|u\|$.

Chapter IV

Billiards

Here we get to the main objective of this book. We describe a class of dynamical systems called billiards. It illustrates all the concepts and phenomena discussed in the previous chapters: invariant measures, ergodicity and mixing, Lyapunov exponents and singularities. The study of billiards is especially important because they have plentiful applications in physics.

The popularity of billiards among mathematicians and physicists alike is also due to relative simplicity of the rules defining the billiard dynamics. As one mathematician, A. Katok, put it, billiards are a playground for researchers. Here one can play all kinds of games and everyone can have fun...

IV.1 Planar Billiards

Imagine a particle (a point mass) moving freely (without friction) on a table and bouncing off the edges of the table elastically. The table is just a bounded region of the plane.

To some extend, this model resembles a popular game of billiards (or pool) where a few balls are pushed by a cue in order to drive some of them into netted pockets in the corners of the table. But our model looks much simpler - we only have one ball, the ball is just a dimensionless pointlike particle, it moves without friction, and the table has no pockets, just edges. On the other hand, the shape of our table is not necessarily rectangular. It can be rather arbitrary, and that is what actually makes this model interesting – the shape of the table determines the character of the dynamics and all its properties.

Billiard table. We denote by $Q \subset \mathbb{R}^2$ an open bounded connected domain, which we call a billiard table. Its boundary $\Gamma = \partial Q$ is supposed to be a finite union of smooth (C^k, $k \geq 3$) compact curves:

$$\partial Q = \Gamma = \Gamma_1 \cup \cdots \cup \Gamma_s \tag{IV.1.1}$$

which are disjoint but may have common endpoints. For example, is Q is a polygon, then Γ_i are its sides. The set

$$\Gamma^* = \partial\Gamma_1 \cup \cdots \cup \partial\Gamma_s \tag{IV.1.2}$$

of the endpoints of our curves will play an important role. Normally, the set Γ^* consists of the corner points of the table Q.

The moving particle has position $q \in Q$ and velocity vector $v \in \mathbb{R}^2$, which are functions of time. If $q \in Q$, then the particle moves freely, i.e.

$$\dot{q} = v \qquad \text{and} \qquad \dot{v} = 0 \tag{IV.1.3}$$

where dots indicate the derivative with respect to time. We can assume that the particle has mass one, then its momentum is $p = v$, and the above equations are Hamiltonian ones, with Hamiltonian function $H(p, q) = \|p\|^2/2$. When $q \in \partial Q$, the velocity v of the particle changes discontinuously, according to the classical rule *the angle of incidence is equal to the angle of reflection*. So, the new (outgoing) vector v_+ is related to the old (incoming) vector v_- by

$$v_+ = v_- - 2\langle v_-, n(q)\rangle n(q) \tag{IV.1.4}$$

Here $\langle \cdot, \cdot \rangle$ stands for the scalar product and $n(q)$ is the inward unit normal vector to the boundary ∂Q of the table Q at the point q.

To use the reflection rule (IV.1.4), we need the normal vector $n(q)$, hence the rule cannot be applied at points $q \in \Gamma^*$, where such a vector fails to exist. To be precise, one might define $n(q)$ by continuity at points of Γ^*, but this would give more than one normal vector $n(q)$ at every corner point of the table, where two different curves $\Gamma_i, \Gamma_j \subset \Gamma$ meet, hence the dynamics would be multiply defined. This is not good. We adopt a standard convention that the reflection is not defined at any $q \in \Gamma^*$. Hence, whenever the moving particle hits a point of Γ^* (called the singular set of the boundary), it stops and ceases to exist.

Example. Define a domain Q in polar coordinates ρ, θ by $Q = \{0 < \rho < 1,\ \theta \neq 0\}$. This is a unit circle with one radial segment removed. In this case $\partial Q = \{\rho = 1\} \cup \{\theta = 0\}$ consists of two smooth curves, and Γ^* consists of two points: the origin $\rho = 0$ and the point $(\rho = 1, \theta = 0)$. Note that the line $\theta = 0$ is an "inner boundary" – billiard trajectories bounce off it on both sides.

There is one more reason why the dynamics may not be defined. This is quite peculiar. If the particle experiences infinitely many reflections on a finite interval of time, then its motion cannot be defined beyond the point where the reflections accumulate. Weird as it sounds, this phenomenon can indeed occur, either near some corner points of the table or even when ∂Q is entirely smooth (Halpern, 1977). We will discuss the effect of this strange fact later.

Phase space. We now construct the phase space of the system. The dynamics (IV.1.3)-(IV.1.4) preserves the norm $\|v\|$, hence we can set $\|v\| = 1$ throughout. Then the phase space of the system will be $\mathcal{M} = \bar{Q} \times S^1$. Here $\bar{Q}$ stands for the closure of the billiard table Q, and S^1 is the unit circle of all velocity vectors[1]. For each $q \in \partial Q$ the points (q, v_-) and (q, v_+) related by (IV.1.4) are identified[2]. The resulting one-parameter group of transformations (flow) on $\mathcal{M}$ is denoted by Φ^t, where $t \in \mathbb{R}$ is time.

It is standard in ergodic theory to reduce the study of flows to maps by constructing a cross-section. The latter is a hypersurface transversal to the flow. For the flow Φ^t, a hypersurface in $\mathcal{M}$ can be very naturally constructed with the help of the boundary of the table Q. Let

$$M = \{x = (q, v) \in \mathcal{M}\colon\ q \in \partial Q,\ \langle v, n(q)\rangle \geq 0\} \tag{IV.1.5}$$

This is a two-dimensional submanifold in $\mathcal{M}$. It consists of all possible outgoing velocity vectors resulting from reflections at ∂Q. Clearly, any trajectory of the flow Φ^t crosses the surface M every time it reflects at ∂Q. This defines the *return map* $T\colon M \to M$ by $Tx = \Phi^{\tau(x)+0}x$ where

$$\tau(x) = \min\{t > 0 : \Phi^{t+0}x \in M\} \tag{IV.1.6}$$

The map T is often called the *billiard map* or *billiard ball map*.

---

[1] In the literature, a billiard table Q is sometimes defined as the closure of an open domain. There is slight problem with that definition, since it would not allow ∂Q to have "inner" lines like one in the above example.

[2] On "inner boundary" lines the identification must be defined separately on each side of the line.

We introduce coordinates r, φ on M, where r stands for the arclength parameter on ∂Q and $\varphi \in [-\pi/2, \pi/2]$ for the angle between v and $n(q)$. These coordinates are oriented as shown on Fig. IV.1, and the reference point on ∂Q (where $r = 0$) can be chosen arbitrarily.

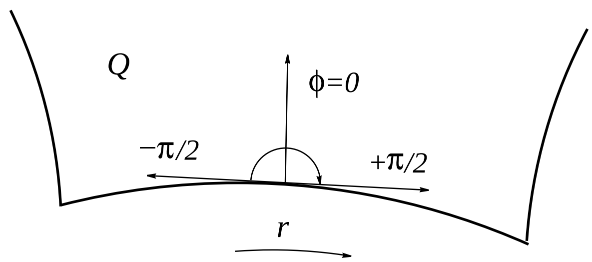

Figure IV.1: The orientation of the coordinates r and φ.

If the boundary ∂Q is one closed curve (this is the case when Q is as a polygon or a disk), then M is, topologically, a cylinder $M = \partial Q \times [-\pi/2, \pi/2]$, and r is a cyclic coordinate. If ∂Q consists of several closed curves (this happens when Q is not simply connected), then M is a union of several cylinders. In all cases $\partial M = \{|\varphi| = \pi/2\}$. Let

$$S_0 = \{|\varphi| = \pi/2\} \cup \{(q, v)\colon q \in \Gamma^*\}$$

This is the set where the return map T has singularities:

Lemma IV.1.1 *Let the boundary ∂Q consists of C^k smooth curves. Then the map T is C^{k-1} at every point $x \in M \setminus S_0$ such that $Tx \in M \setminus S_0$.*

Proof. Let $x = (r, \varphi) \in M$ and $Tx = (r_1, \varphi_1)$. Denote, as in (IV.1.6), by $\tau = \tau(x)$ the return time (travel time) between x and Tx, and by $K = K(r)$ the curvature of the boundary ∂Q at r (so that the angle between normal vectors $n(r)$ and $n(r + dr)$ equals $K\,dr + o(dr)$). To fix the sign of K, we assume that $K(r) > 0$ if ∂Q is concave (convex inward) at r (as in Fig. IV.2) and $K(r) < 0$ if ∂Q otherwise. Similarly, we put $K_1 = K(r_1)$.

A detailed (but elementary) geometric analysis illustrated on Fig. IV.2 gives the derivative of T:

$$DT(x) = -\frac{1}{\cos\varphi_1}\begin{pmatrix} \tau K + \cos\varphi & \tau \\ \tau K K_1 + K\cos\varphi_1 + K_1\cos\varphi & \tau K_1 + \cos\varphi_1 \end{pmatrix} \tag{IV.1.7}$$

Note that, since $\cos\varphi_1 \neq 0$ and $\cos\varphi \neq 0$, this matrix is defined and is nonsingular. Also, since the first derivative of T involves the curvature K of ∂Q (related to the second derivative of Γ_i's), then the smoothness of T is only C^{k-1}. $\square$

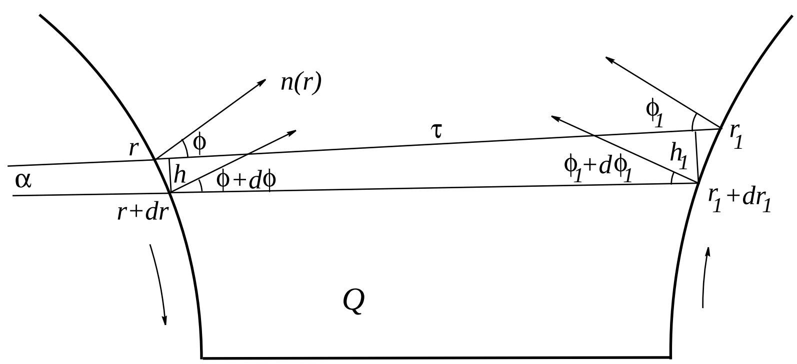

Figure IV.2: The calculation of DT: note that, infinitesimally, $h = dr\cos\varphi$, $h_1 = -dr_1\cos\varphi_1$, $\alpha = d\varphi + K\,dr$ (where α is the angle between the lines from r to r_1 and $r + dr$ to $r_1 + dr_1$), $h_1 = h + \tau\alpha$ and $d\varphi_1 = K_1\,dr_1 - \alpha$.

Remark. It is easy to compute that $|\det DT(x)| = \cos\varphi/\cos\varphi_1$.

Invariant measures. The Hamiltonian equations (IV.1.3) preserve the Liouville measure $dq\,dv$ on the phase space $\mathcal{M}$, where dq and dv are uniform measures on Q and S^1, respectively. One can check by direct inspection (which we omit) that this measure is also invariant under reflections (IV.1.4). Since the table Q is compact, the measure $dq\,dv$ on $\mathcal{M}$ is finite and can be normalized, so we get a probability measure

$$d\mu = c_\mu\,dq\,dv \tag{IV.1.8}$$

where

$$c_\mu = (2\pi|Q|)^{-1} \tag{IV.1.9}$$

and $|Q|$ stands for the area of the domain Q.

The invariant measure μ induces a T-invariant measure on the surface M. To find it, we introduce a special coordinate system on the space $\mathcal{M}$.

For every point $x = (q, v) \in \mathcal{M}$ let $s(x) = -\max\{t < 0 \colon \Phi^{t+0}x \in M\}$ and $y = (q', v') = \Phi^{-s(x)+0}x$. Plainly, $s(x)$ is the time elapsed since the last reflection, $q' \in \partial Q$ is the point of that reflection, and v' is the outgoing velocity vector. We note that $v' = v$, and the point q' and the value of s can be determined uniquely by solving the equation $q - sv \in \partial Q$. Let r, φ be the coordinates of the point $(q', v') \in M$. Now the manifold $\mathcal{M}$ can be parametrized by three new coordinates r, φ, s.

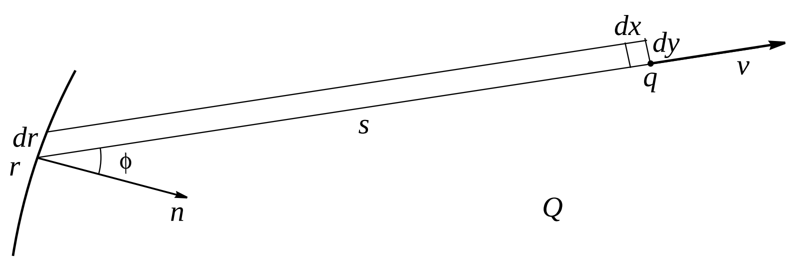

Figure IV.3: The coordinates r, φ, s.

It takes a little geometric reasoning (illustrated on Fig. IV.3) to verify that the infinitesimal volume in $\mathcal{M}$ can be represented as $dq\, dv = \cos\varphi\, dr\, d\varphi\, ds$. Hence, the measure μ on $\mathcal{M}$ satisfies

$$d\mu = c_\mu \cos\varphi\, dr\, d\varphi\, ds \tag{IV.1.10}$$

It is now easy to conclude that since the flow Φ^t has constant speed (it equals one), the measure $\cos\varphi\, dr\, d\varphi$ is invariant for the map $T : M \to M$. We normalize it and get a probability

$$d\nu = c_\nu \cos\varphi\, dr\, d\varphi \tag{IV.1.11}$$

where

$$c_\nu = (2|\partial Q|)^{-1} \tag{IV.1.12}$$

and $|\partial Q|$ stands for the total length of the boundary ∂Q (perimeter of Q). We note that the invariance of the measure (IV.1.11) under T also follows from the remark after Lemma IV.1.1, as one can see by making change of variables in direct integration over any Borel set $U \subset M$:

$$\nu(T(U)) = c_\nu \int_{T(U)} \cos\varphi_1\, dr_1\, d\varphi_1 = c_\nu \int_U \cos\varphi_1 \frac{\cos\varphi}{\cos\varphi_1}\, dr\, d\varphi = \nu(U)$$

Lemma IV.1.2 *There is a subset $M' \subset M$ of full measure, i.e. $\nu(M') = 1$, on which the map T^n is defined for all $n \in \mathbb{Z}$. Likewise, there is a subset $\mathcal{M}' \subset \mathcal{M}$ of full measure, i.e. $\mu(\mathcal{M}') = 1$, on which the flow Φ^t is defined at all times $t \in \mathbb{R}$.*

Proof. The map T is not defined at $x \in M$ only if the next reflection occurs at some $q \in \Gamma^*$. Such points x belong to a finite or countable union of smooth curves in M. Hence, the points $x \in M$ whose forward or backward images under T ever hit the singular set Γ^* lie on a countable union of smooth curves in M, hence their total measure is zero. This proves the first part of the lemma.

The flow Φ^t is not defined at $x \in \mathcal{M}$ if the trajectory of x either (i) hits a singularity point $q \in \Gamma^*$ or (ii) experiences infinitely many reflections on a finite interval of time. The set of points of type (i) has zero measure by the first part of the lemma. It turns out, that the set of points of type (ii) also has zero measure, see Exercise IV.1.1. □

Billiards on a torus. An interesting (and very popular in physics) modification of billiard model is obtained by taking a table Q on a unit 2-D torus $\mathbb{T}^2$, instead of the plane $\mathbb{R}^2$. For example, let $D \subset \mathbb{T}^2$ be a small disk and $Q = \mathbb{T}^2 \setminus D$. Such a system can be thought of as a billiard inside a unit square with periodic boundary conditions, as shown on Fig IV.4. In this case the return time $\tau(x)$, as a function on M, is unbounded. However, the map $T : M \to M$ is still well defined, i.e. if $x \in M$ then the point $Tx \in M$ always exists, see Exercise IV.1.2.

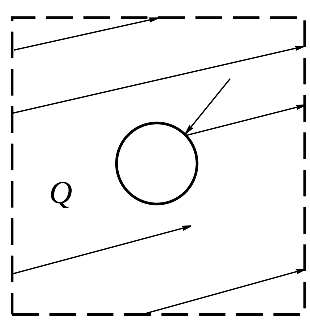

Figure IV.4: A billiard on a torus.

Involution. The dynamical system $(\mathcal{M}', \mu, \Phi^t)$ has an interesting property called *involution*: for any $x = (q, v) \in \mathcal{M}'$ the point $I(x) = (q, -v) \in M'$ satisfies

$$\Phi^t(I(x)) = I(\Phi^{-t}x)$$

for all $t \in \mathbb{R}$. Hence, the involution $I\colon \mathcal{M}' \to \mathcal{M}'$ anticommutes with the dynamics Φ^t, which we can write as $\Phi^t \circ I = I \circ \Phi^{-t}$. Note that the map I also preserves the measure μ.

The map $T : M' \to M'$ also admits an involution, I_1, defined by $(r, \varphi) \mapsto (r, -\varphi)$. It anticommutes with T, i.e. $T^k \circ I_1 = I_1 \circ T^{-k}$ for all $k \in \mathbb{Z}$. We also note that the map $I_1 : M \to M$ preserves the measure ν.

Lyapunov exponents. We now prove that Oseledec's Theorem III.2.1 applies to the billiard map T.

Theorem IV.1.3 *If the absolute value of the curvature of ∂Q is uniformly bounded, then Osedelec's theorem applies to the billiard map T.*

Proof. We need to verify that the functions $\log^+ \|DT(x)\|$ and $\log^+ \|DT^{-1}(x)\|$ have finite integrals over M. By the involution property, it is enough to do this only for $DT(x)$. It follows from (IV.1.7) that $\|DT(x)\| \le C/\cos\varphi_1$, where $C > 0$ is a constant. Then

$$\begin{aligned} c_\nu \int_M \log^+ \|DT(x)\|\, d\nu &\le c_\nu \int_M |\log C + \log\cos\varphi_1| \cos\varphi\, d\varphi\, dr \\ &\le |\log C| + c_\nu \int_M |\log\cos\varphi| \cos\varphi\, d\varphi\, dr \end{aligned}$$

where we have used the invariance of the measure ν. Finally,

$$\begin{aligned} \int_M |\log\cos\varphi| \cos\varphi\, d\varphi\, dr &= |\partial Q| \int_{-\pi/2}^{\pi/2} |\log\cos\varphi| \cos\varphi\, d\varphi \\ &= |\partial Q|\,(2 - \log 4) < \infty \end{aligned}$$

The theorem is proved. □

Remark. The condition on the curvature of ∂Q is necessary. Katok and Strelcyn [KS] give an example of a simply connected billiard table whose boundary is C^∞ everywhere, except for one point, at which the curvature turns infinite, for which $\int_M \log^+ \|DT(x)\|\, d\nu = \infty$.

The above theorem ensures the existence of two Lyapunov exponents, $\lambda_1(x)$ and $\lambda_2(x)$, and the corresponding invariant subspaces, $E_1(x)$ and $E_2(x)$, at almost every point $x \in M'$. We note also that if $\lambda_1(x) \neq \lambda_2(x)$, then the angle $\lambda(x)$ between the lines $E_1(x)$ and $E_2(x)$ is defined. By a remark in Section III.3,

$$\lim_{n\to\pm\infty} \frac{1}{n} \log \gamma(T^n x) = 0 \tag{IV.1.13}$$

Corollary IV.1.4 *We have $\lambda_1(x) + \lambda_2(x) = 0$ almost everywhere on M.*

Proof. Let Π be a parallelogram in $\mathcal{T}_x M$ with sides s_1, s_2 parallel to $E_1(x)$ and $E_2(x)$, respectively. Then its area is $|\Pi| = s_1 s_2 \sin\gamma(x)$. Its image under DT^n is a similar parallelogram in $\mathcal{T}_{T^n x} M$ with sides that we call $s_1(n), s_2(n)$. Therefore,

$$\|DT^n(x)\| = \frac{\sin\gamma(T^n x)}{\sin\gamma(x)} \, \frac{s_1(n)}{s_1} \, \frac{s_2(n)}{s_2} \tag{IV.1.14}$$

We also note that

$$\lim_{n\to\pm\infty} \frac{1}{n} \log \frac{s_i(n)}{s_i} = \lambda_i(x) \tag{IV.1.15}$$

for $i = 1, 2$. Consider the point $T^n x = (r_n, \varphi_n)$. It follows from the remark after Lemma IV.1.1 that $\|DT^n(x)\| = \cos\varphi / \cos\varphi_n$. Hence, $\|DT^n(x)\|$ stays bounded away from 0 and infinity most of the time. Now taking logarithm in (IV.1.14), dividing by n, taking a limit as $n \to \pm\infty$, and using (IV.1.13) and (IV.1.15) proves the corollary. □

We see that almost every point $x \in M$ either has both zero Lyapunov exponents or else is completely hyperbolic.

A mechanical model. Consider a system of two point particles of masses m_1 and m_2 on a unit interval $0 \le x \le 1$. The particles move freely and collide elastically with each other and with the "walls" at $x = 0$ and $x = 1$. Let x_1, x_2 be the positions of the particles and u_1, u_2 their velocities. Assume that $x_1 \le x_2$. When a particle collides with a wall, it simply reverses its velocity. When the two particles collide with each other, they preserve the total momentum $m_1 u_1 + m_2 u_2$ and the total kinetic energy $(m_1 u_1^2 + m_2 u_2^2)/2$.

Now we introduce new variables:

$$q_i = x_i \sqrt{m_i} \quad \text{and} \quad v_i = \dot{q}_i = u_i \sqrt{m_i} \tag{IV.1.16}$$

for $i = 1, 2$. Then the state of the system is described by a point $q = (q_1, q_2) \in \mathbb{R}^2$ (called the *configuration point*) and its velocity vector $v = (v_1, v_2)$. The set of all configuration points (called the *configuration space*) is the right triangle

$$Q = \{q = (q_1, q_2) : 0 \leq q_1/\sqrt{m_1} \leq q_2/\sqrt{m_2} \leq 1\} \qquad \text{(IV.1.17)}$$

One can check by direct inspection (we leave it to the reader) that the trajectory of the system in Q is governed by the billiard rules (IV.1.3) and (IV.1.4). The only nontrivial part here is to show that collisions between the particles correspond to specular reflections at the hypotenuse of the triangle Q. This was the reason why the multipliers $\sqrt{m_i}$, $i = 1, 2$, are introduced in (IV.1.16).

Therefore, the study of this mechanical model reduces to the study of billiard dynamics in the triangle Q. We will see more examples of such reduction in the next section.

Exercises:

IV.1.1. Show that for μ-almost every point $x \in \mathcal{M}$ and every $T > 0$ the segment $\Phi^t x$, $0 < t < T$, contains finitely many reflections at ∂Q. Hint: it is enough to show the same property for ν-almost every $x \in M$, which can be done with the help of Poincaré recurrence theorem applied to the map $T : M \to M$.

IV.1.2. Let $Q \subset \mathbb{T}^2$ be any billiard table on a torus. Prove that for every point $x \in M$ its future semiorbit $\Phi^t x$, $t > 0$, necessarily crosses ∂Q, hence the point Tx exists (even though it may belong to the singular set S_0). Hint: if the semiorbit did not cross ∂Q, it would be either periodic or dense on the torus, see [PM] or [KH].

IV.2 Billiards in Higher Dimensions

This section is more advanced than the previous one. For a novice, a slow pace is recommended. It is also possible to skip this section and move to the following one, but later come back, if necessary, and cover selected topics from this section.

Let Q be an open bounded domain in a d-dimensional space $\mathbb{R}^d$ or in a d-dimensional torus $\mathbb{T}^d$. Assume that the boundary $\partial Q = \Gamma$ either is a C^k

($k \geq 3$) smooth $(d-1)$-dimensional surface or consists of C^k a finite number of smooth $(d-1)$-dimensional surfaces $\Gamma_1, \dots, \Gamma_s$ with boundary, so that (IV.1.1) again applies. For example, if Q is a d-dimensional polyhedron, then ∂Q consists of $(d-1)$-dimensional faces, whose boundaries are $(d-2)$-dimensional edges. The set Γ^* given by (IV.1.2) is again called the singular set of ∂Q.

The billiard particle moves freely inside Q according to the equations (IV.1.3) and reflects elastically off the boundary ∂Q by the equation (IV.1.4), unless $q \in \Gamma^*$, in which case the reflection is not defined. These equations preserve the norm $\|v\|$, and we set $\|v\| = 1$. The domain Q is called now the *configuration space* of the billiard system.

Many results of the previous section extend to multidimensional billiards with obvious modifications. We repeat those results briefly, emphasizing the differences between planar and spatial billiard.

The phase space is $\mathcal{M} = \bar{Q} \times S^{d-1}$, where $\bar{Q}$ is again the closure of the billiard domain Q, and S^{d-1} is now the $(d-1)$-dimensional unit sphere of all velocity vectors. At each $q \in \partial Q$ the points (q, v_-) and (q, v_+) related by (IV.1.4) are identified.

The billiard dynamics induces a flow Φ^t on the space $\mathcal{M}$ that has invariant measure $d\mu = c_\mu dq\, dv$, where dq and dv are the Lebesgue measures on Q and S^{d-1}, respectively, and c_μ is the normalizing factor. It is easy to see that

$$c_\mu = (|Q| \cdot |S^{d-1}|)^{-1} \tag{IV.2.1}$$

where $|Q|$ is the d-dimensional volume of the domain Q and $|S^{d-1}|$ is the $(d-1)$-dimensional volume of the sphere S^{d-1}.

The measure $d\mu$ can be represented in a new coordinate system in a way similar to (IV.1.10):

$$d\mu = c_\mu \langle v, n(q) \rangle \, dr\, dv\, ds \tag{IV.2.2}$$

where r and s are shown on Figure IV.3 in Section IV.1. Note that the coordinate φ in (IV.1.10) is replaced by $v \in S^{d-1}$, and $\cos\varphi$ by $\langle v, n(q)\rangle$.

The cross-section M of the phase space $\mathcal{M}$ is still defined by (IV.1.5). Note that $\dim M = 2d-2$. The billiard map $T : M \to M$ and the return time function $\tau(x)$ on M are defined naturally, in the same way as in Section IV.1. If the smooth components of the boundary ∂Q are of class C^k, then the map T is C^{k-1} smooth at all points $x \in M \setminus S_0$ such that $Tx \in M \setminus S_0$. Here

$$S_0 = \{(q, v) \in M : \langle v, n(q)\rangle = 0\} \cup \{(q, v) \in M : q \in \Gamma^*\}$$

The billiard map $T : M \to M$ preserves the measure

$$d\nu = c_\nu \langle v, n(r) \rangle \, dr \, dv$$

where dr and dv are the Lebesgue measures on ∂Q and S^{d-1}, respectively, and c_ν is the normalizing factor. By a direct integration over M one can compute c_ν:

$$c_\nu = (|\partial Q| \cdot |B^{d-1}|)^{-1} \tag{IV.2.3}$$

where $|\partial Q|$ is the $(d-1)$-dimensional volume of ∂Q and $|B^{d-1}|$ is the $(d-1)$-dimensional volume of the unit ball $B^{d-1} \subset \mathbb{R}^{d-1}$. We leave the computation of c_ν as an exercise.

There is a subset $\mathcal{M}' \subset \mathcal{M}$ of full μ-measure where the flow Φ^t is defined at all times $t \in \mathbb{R}$. Likewise, there is a subset $M' \subset M$ of full ν-measure where the map T^k is defined for all $k \in \mathbb{Z}$.

The involution $I : \mathcal{M}' \to \mathcal{M}'$ is defined by $I(q, v) = (q, -v)$ as before. It anticommutes with the flow ($\Phi^t \circ I = I \circ \Phi^{-t}$) and preserves the measure μ.

The map $T : M' \to M'$ also admits an involution, I_1, defined by $(r, v_-) \mapsto (r, v_+)$ in the notation of (IV.1.4). The involution I_1 anticommutes with T, i.e. $T^k \circ I_1 = I_1 \circ T^{-k}$ for all $k \in \mathbb{Z}$, and preserves the measure ν.

If the absolute value of all sectional curvatures of ∂Q is uniformly bounded, then Osedelec's theorem applies to the billiard map T and ensures the existence of $2d-2$ Lyapunov exponents $\{\lambda_1(x), \ldots, \lambda_{2d-2}(x)\}$ at almost every point $x \in M$. As in Corollary IV.1.4, their sum vanishes:

$$\lambda_1(x) + \cdots + \lambda_{2d-2}(x) = 0 \tag{IV.2.4}$$

Remark. A much stronger fact than (IV.2.4) can be derived by means of symplectic geometry. Every billiard is a Hamiltonian system and preserves a naturally defined symplectic form, we refer the reader to [LW] for details. It then follows that if we order the set of Lyapunov exponents, $\lambda_1(x) \geq \cdots \geq \lambda_{2d-2}(x)$, then

$$\lambda_i(x) + \lambda_{2d-1-i}(x) = 0$$

for all $1 \leq i \leq 2d-2$ and almost every $x \in M$. This fact is usually referred to as the *symmetry of Lyapunov exponents*.

A mechanical model. This is a generalization of the 2-particle system from Section IV.1. Consider n point particles of masses $m_1, \ldots, m_n$ on a unit interval $0 \le x \le 1$. The particles move freely and collide elastically with each other and with the "walls" at $x = 0$ and $x = 1$. Let $x_1, \dots, x_n$ be the positions of the particles and $u_1, \dots, u_n$ their velocities. Assume that the particles are ordered so that $x_1 \le \cdots \le x_n$.

We introduce new variables:

$$q_i = x_i \sqrt{m_i} \quad \text{ and } \quad v_i = \dot{q}_i = u_i \sqrt{m_i}$$

for $i = 1, \dots, n$. Then the state of the system is described by a point $q = (q_1, \dots, q_n) \in \mathbb{R}^n$ (called the *configuration point*) and its velocity vector $v = (v_1, \dots, v_n)$. The set of all configuration points (called the *configuration space*) is a right pyramid in $\mathbb{R}^n$:

$$Q = \{q : 0 \le q_1/\sqrt{m_1} \le \cdots \le q_n/\sqrt{m_n} \le 1\}$$

As in Section IV.1, one can check by direct inspection that the trajectory of the system in Q is governed by the billiard rules (IV.1.3) and (IV.1.4). Therefore, the study of the mechanical model on n particles on an interval can be reduced to the study of billiard dynamics in the pyramid Q.

This reduction has many interesting implications. For example, recall that the invariant measure $d\mu$ is a product of the Lebesgue measures on both the pyramid Q and the $(n-1)$-dimensional sphere S_v^{n-1} of velocity vectors. Hence, for every particle i the distribution of its velocity v_i is a marginal measure of the uniform distribution on the sphere S_v^{n-1}. Let us set the norm of the velocity vector v so that

$$v_1^2 + \cdots + v_n^2 = Kn, \qquad K = \text{const} \tag{IV.2.5}$$

Then the distribution of each v_i has a limit as $n \to \infty$. This limit is a Gaussian measure with zero mean and variance K, i.e. $v_i \sim N(0, K)$ as $n \to \infty$. Normal distribution of velocity vectors is known as Maxwellian distribution in statistical mechanics, and it is characteristic for molecules in gases and fluids.

Note that (IV.2.5) can be translated into the original variables q_i, u_i, and turns to be a condition on the total kinetic energy:

$$\frac{m_1 u_1^2}{2} + \cdots + \frac{m_n u_n^2}{2} = \frac{Kn}{2}$$

Note that $K/2$ is the mean kinetic energy per particle. Now, since v_i^2 has the same probability distribution for all i, so does the kinetic energy $m_i u_i^2/2$

of every particle. This is an important fact: each particle (independently of its mass) has the same distribution (and, of course, the same mean value, equal to $K/2$) for its kinetic energy. This fact is known in statistical mechanics as *equipartition of energy* – the total energy in a multiparticle system is equally divided between the particles.

Gases of hard disks and balls. The system of particles moving on an interval discussed above is a very simplified model of a gas. Consider a more realistic model of n disks moving on a plane, or n balls moving in space. These two models are very similar, so that we can discuss them in parallel. For simplicity, let all the disks (balls) have the same radius r and the same mass m. Each ball (disk) moves freely, i.e. with constant velocity, until it hits another moving ball. When two balls (disks) collide, they change their velocities according to the laws of elastic collision.

This law means the following. Let two balls collide. Denote by q_1 and q_2 their centers and by v_1 and v_2 their velocity vectors at the moment of collision. Let L be the line through the centers q_1 and q_2. We decompose

$$v_i = v_i^0 + v_i^\perp$$

for $i = 1, 2$, where v_i^0 is the component of the vector v_i parallel to L and $v_i^\perp$ is the one perpendicular to L. Then the new, outgoing, velocities of the balls are

$$v_1^{\text{new}} = v_1^\perp + v_2^0 \quad \text{and} \quad v_2^{\text{new}} = v_2^\perp + v_1^0$$

In other words, the balls exchange the velocity components parallel to the center line L and retain the orthogonal components. We note that the laws of elastic collision imply preservation of the total kinetic energy $\sum m\|v_i\|^2/2$ and the total momentum $\sum mv_i$ of the system of n balls (disks). We also note that a collision of two hard balls with centers q_1 and q_2 can only occur if $\text{dist}(q_1, q_2) = 2r$, i.e. $\|q_1 - q_2\|^2 = (2r)^2$.

The system of n balls (disks) moving in the open space (or on the plane) without walls is dynamically not very interesting. As it is intuitively clear (and proven mathematically, see below), the total number of collisions between balls is always finite, and after the last collision the balls will fly freely forever. Furthermore, the number of collisions between n balls in the open space is uniformly bounded by a constant that only depends on n. This last fact was proved very recently – in 1998 – by Burago, Ferleger and Kononenko [BFK2], see Corollary IV.2.4 below.

Let us consider n balls or disks enclosed in a bounded domain R, called a container (or reservoir). The balls (disks) collide elastically with each other and with the walls of the container. Precisely, if a ball with center q hits a wall at a point $w \in \partial R$, then we decompose its velocity vector as $v = v^0 + v^\perp$, where v^0 is the component parallel to the line passing through q and w, and $v^\perp$ is perpendicular to that line. The new, outgoing, velocity of the ball is $v^{\text{new}} = v^\perp - v^0$. Note that this rule preserves the total kinetic energy of the system, but *not* its total momentum.

Now we reduce the system of n hard balls in a container R to a billiard. We denote by $q_i = (q_i^1, q_i^2, q_i^3)$ the center of the ith ball and by $v_i = (v_i^1, v_i^2, v_i^3)$ its velocity vector, $1 \le i \le n$. For disks on a plane, we have two coordinates instead of three, of course. Now the entire system can be described by a configuration point

$$q = (q_1^1, q_1^2, q_1^3, q_2^1, \dots, q_n^2, q_n^3) \in \mathbb{R}^{3n}$$

and its velocity vector

$$v = (v_1^1, v_1^2, v_1^3, v_2^1, \dots, v_n^2, v_n^3) \in \mathbb{R}^{3n} \tag{IV.2.6}$$

(for systems of disks, we need to replace $3n$ by $2n$).

We note that $q \in R^n = R \times \cdots \times R$. It is also important to observe that not the entire region R^n is available for the configuration point q. By the rules of elastic collisions, the balls cannot overlap – the moment they bump into each other, they collide. This rule requires exclusion of configurations that satisfy

$$(q_i^1 - q_j^1)^2 + (q_i^2 - q_j^2)^2 + (q_i^3 - q_j^3)^2 < (2r)^2 \tag{IV.2.7}$$

for some $1 \le i < j \le n$ (here r is the radius of the balls). The inequality (IV.2.7) specifies a spherical cylinder in $\mathbb{R}^{3n}$, which we denote by C_{ij}. For the model of hard disks on a plane, we get a circular cylinder C_{ij} in $\mathbb{R}^{2n}$. The cylinders C_{ij}, $1 \le i < j \le n$, contain all forbidden configurations of the balls (disks), hence they must be removed from the available space. As a result, we get a smaller domain

$$Q = R^n \setminus \cup_{i \ne j} C_{ij}$$

This domain Q is called the *configuration space* of the system.

Now one can check by direct inspection (we leave it to the reader as an exercise, rather tedious one, though) that the trajectory of the configuration point q in Q is governed by the billiard rules (IV.1.3) and (IV.1.4). Specular reflections at the surface of a cylinder C_{ij} correspond to collisions between the balls i and j. Thus, the study of the mechanical model of n balls or disks is reduced to the study of billiard dynamics in the domain Q. We note that the conservation of the total kinetic energy $\sum_i m\|v_i\|^2/2$ is equivalent to the preservation of the norm $\|v\|$ of the velocity vector (IV.2.6).

The singularity set Γ^* contains all intersection of the cylindrical surfaces ∂C_{ij} with each other. Such intersections correspond to simultaneous collisions of three or more balls. The outcome of such multiple collisions is not defined. It is our general rule, though, to ignore billiard trajectories that hit Γ^*.

The gas of hard balls is a classical model in statistical physics. Its study goes back to L. Boltzmann in the XIX century. Many physical laws have been first established for gases of hard balls, and then experimentally verified for other gases. Boltzmann was first to state the celebrated *ergodic hypothesis*. He assumed that gases of hard balls are, in general, ergodic and used this assumption to justify the laws of statistical mechanics (on a "heuristic" level). Since then, it remains a major challenge for physicists and mathematicians to prove this hypothesis, as well as to make use of the ergodicity to build the mathematical foundation of statistical mechanics.

In early sixties, Ya. Sinai studied a specific version of Boltzmann's model – the gas of n hard balls (or disks) on a torus $\mathbb{T}^d$, $d \geq 2$. In that case the container R is a torus, so there are no walls (i.e., $\partial R = \emptyset$). Hence, the balls only collide with each other. Therefore, in addition to the total kinetic energy, the total momentum is conserved. Sinai conjectured that if one sets the total momentum to zero and fixes the center of mass, then the resulting reduced system would be ergodic.

Attempts to prove the Boltzmann-Sinai conjecture spanned almost 40 years, and they had a colorful and sometimes dramatic history described in [Sz1]. See also [Sz2]. It appears that the problem is almost solved by now due to very recent works of N. Simanyi and D. Szasz [SS1, Sm]. But that solution is beyond the scope of this book.

Lorentz gas. Here is another popular physical model. Imagine a point particle moving between fixed rigid balls in space $\mathbb{R}^3$. The balls are of the same radius r. They can be positioned either randomly or make a regular crystalline-like structure (for example, their centers may be located at cites of the lattice $\mathbb{Z}^3$). The balls do not move, only one point particle moves in between and collides with the balls elastically. The balls act like obstacles (like bumps in a pinball machine).

This model was introduced by H. Lorentz in 1905 [Lo] in the study of electrons in metals. The moving pointlike particle represented an electron, and the balls played the role of molecules in metal. When the balls are located at cites of a regular lattice, the Lorentz gas is said to be periodic (as opposed to random). We only discuss periodic Lorentz gases here.

Instead of balls, we can place other identical bodies at the lattice cites, and require the moving particle bounce off the boundaries of the obstacles elastically. The bodies may have cubical or more general polyhedral shape. Various modifications of this model are studied in physics [Hg]. One can also define a Lorentz gas on a plane, in this case 2-D obstacles are placed at cites of a 2-d lattice, e.g., $\mathbb{Z}^2$.

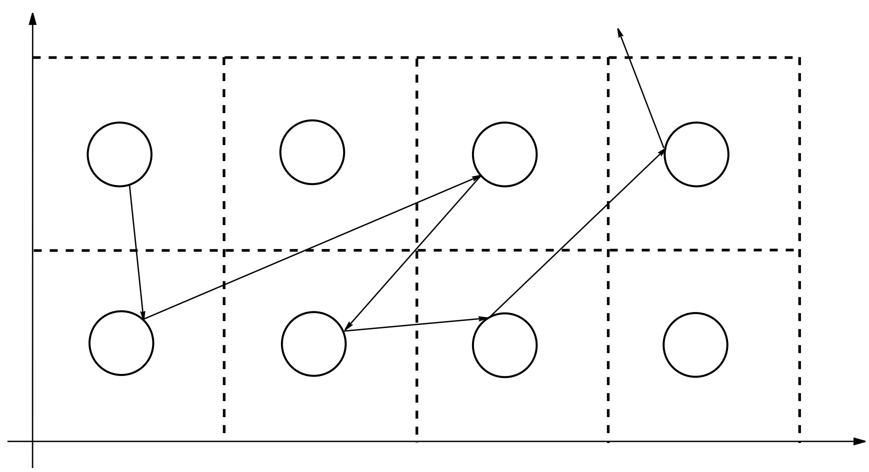

Figure IV.5: A periodic Lorentz gas.

We now reduce the Lorentz gas to a billiard. It is already "almost" a billiard, except that the available domain for the moving particle is unbounded and has infinite volume. This can be fixed easily. Since the obstacles are located at the cites of a lattice, their structure is periodic, and one can find a *fundamental domain* D, i.e. the domain whose parallel translations will cover the whole space. The domain D contains just a few (sometimes just one) obstacle, and all the other obstacles are obtained by parallel translations of those in D. Fig. IV.5 shows a periodic Lorentz gas on a plane, where fixed disks make a periodic array, and the fundamental domain is just a square containing one disk.

Now the motion of the particle can be projected down to the fundamental domain D, and we get a billiard system in D with periodic boundary conditions, i.e a billiard on a torus $\mathbb{T}^d$, $d = 2, 3$. The billiard domain Q is obtained by removing one or several obstacles from the torus. Such an example was already shown on Fig. IV.4 in Section IV.1.

We conclude this section by two properties of billiards that have geometric, rather than dynamical, nature.

Mean free path. Recall that the function $\tau(x)$ on M is the return time for the map $T : M \to M$, or the time elapsed between reflections at the points x and Tx. Since the speed of the billiard particle is set to one, $\tau(x)$ also equals the distance in Q between these points of reflection, which is called the *free path* of the billiard trajectory. We now want to estimate the *mean free path*, i.e. the asymptotic of value

$$\bar{\tau}(x) = \lim_{n\to\infty} \frac{\tau(x) + \tau(Tx) + \cdots + \tau(T^{n-1}x)}{n} \tag{IV.2.8}$$

By the Birkhoff ergodic theorem, the value $\bar{\tau}(x)$ exists almost everywhere in M and its average value is

$$\bar{\tau} := \int_M \bar{\tau}(x)\, d\nu(x) = \int_M \tau(x)\, d\nu(x) \tag{IV.2.9}$$

If the billiard map T is ergodic, then, furthermore, the function $\bar{\tau}(x)$ is constant almost everywhere, and it equals $\bar{\tau}$.

The calculation of $\bar{\tau}$, in terms of geometric characteristics of the domain Q, is remarkably simple. We write

$$\tau(x) = \int_0^{\tau(x)} ds$$

where s is the parameter introduced in (IV.2.2), and then

$$\bar{\tau} = \int_M \tau(x)\, d\nu(x) = c_\nu \int_M \int_0^{\tau(x)} \langle v, n(r)\rangle\, ds\, dr\, d\varphi$$

Then we use the identity (IV.2.2) and get

$$\bar{\tau} = \frac{c_\nu}{c_\mu} \int_{\mathcal{M}} d\mu = \frac{c_\nu}{c_\mu} \tag{IV.2.10}$$

Now using (IV.2.1) and (IV.2.3) completes the calculation:

$$\bar{\tau} = \frac{|Q| \cdot |S^{d-1}|}{|\partial Q| \cdot |B^{d-1}|} \tag{IV.2.11}$$

We note that the mean free path between collisions only depends on the volume of the domain Q and the surface area of its boundary, but not on its shape.

In particular, for planar billiards $d = 2$, and we have $|S^1| = 2\pi$, $|B^1| = 2$, hence the formula (IV.2.11) turns very simple:

$$\bar{\tau} = \frac{\pi\,|Q|}{|\partial Q|} \tag{IV.2.12}$$

The formula (IV.2.11) is well known in geometric probability and integral geometry. Its planar version (IV.2.12) is often referred to as Santalo formula, since it was first given in Santalo's book [Sa].

For example, consider again a billiard table Q on a unit torus $\mathbb{T}^2$ where a small disk D of radius r is removed, as shown on Fig. IV.4. Clearly, for small r the billiard particle can move freely for a long time between collisions with the disk D, and the function $\tau(x)$ can take arbitrarily large values. The Santalo formula (IV.2.12) gives its mean value:

$$\bar{\tau} = \frac{\pi(1 - \pi r^2)}{2\pi r} = \frac{1 - \pi r^2}{2r}$$

i.e. the mean free path is asymptotically equal $\frac{1}{2r}$ as $r \to 0$. We will see later that the map T is ergodic in this example, so that $\bar{\tau}(x) = \bar{\tau}$ almost everywhere.

As a far more complicated example, consider a system of N hard balls on a unit (d-dimensional) torus. It reduces to a billiard in multidimensional domain Q whose boundary consists of cylinders (note: $\dim Q = Nd$). The mean free time between collisions can now be estimated by (IV.2.11). This requires computing the volume of Q and surface area of its boundary. This is a difficult but feasible job, which was done in [C5]. Quite remarkably, the final expression coincided with the classical Boltzmann's formula for the mean intercollision time used in statistical physics for decades. We refer the interested reader to [C5] and references therein.

Estimates for the number of reflections. Here we consider the following problem: given a piece of a billiard trajectory of length L, how many reflections at ∂Q can be there on that piece? In particular, can the number of reflections, n, be infinite? We showed in Sect. IV.1 that $n < \infty$ with probability one. But, from the geometric point of view, one would like to know if n can ever be infinite, and how large n can be. These questions also arise in the studies of ergodic properties of billiards.

We start with a simple case - a billiard trajectory moving between two lines, l_1 and l_2, which intersect at a point A at angle $\alpha > 0$, see Fig. IV.6. We call Q the (infinite) domain bounded by these lines. When the trajectory hits either line, say l_1, it gets reflected, but its mirror image across l_1 will continue moving straight on the other side of l_1. We will follow that mirror image, rather than the trajectory itself. It will continue moving in the domain Q_1 that is the mirror images of Q across the line l_1. When our trajectory hits the other line l_2, its mirror image also hits the other side of the domain Q_1, etc. We will keep reflecting the domains Q_i across their sides and following the straight line made by the mirror images of our trajectory. This will look like a mirror room in an amusement park, with multiple reflections in different mirrors. This method of reflecting the billiard table Q across its sides and following the images of the trajectory is called *unfolding.* The result is shown on Fig. IV.6.

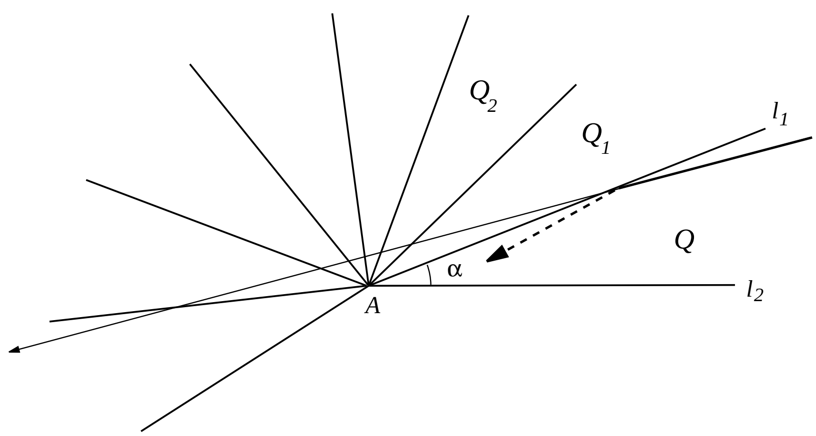

Figure IV.6: The unfolding of a billiard trajectory.

It is clear from Fig. IV.6 that the total angle made by the unfolding images $Q_1, Q_2, \ldots$, between the first reflection and the last one, cannot exceed π, hence

$$n < \frac{\pi}{\alpha} + 1$$

This simple estimate gives an upper bound on the number of reflections in Q. We note that this bound is uniform, i.e. the same for all billiard trajectories in Q.

The above estimate has a multidimensional version. Suppose that several hyperplanes in $\mathbb{R}^d$, $d \geq 2$, intersect at one point A, so that they make a "polyhedral angle" with vertex at A. Ya. Sinai proved in 1978 [Si4] that the number of reflections experienced by any billiard trajectory inside such an angle is uniformly bounded, the bound only depends on the configuration of the hyperplanes.

It is more difficult to estimate the number of reflections in billiard domains with curvilinear boundary. Here we have two distinct cases. One is a billiard domain Q with a convex boundary, such as a disk on a plane or a ball in $\mathbb{R}^d$. Near a convex boundary ∂Q, a short piece of trajectory clearly can experience arbitrary many reflections. This happens when the velocity vector v at a point of reflection $q \in \partial Q$ is almost tangent to the boundary ∂Q. Such a trajectory would simply "slide" along ∂Q experiencing many "grazing" collisions with ∂Q in rapid succession.

It is even possible to construct convex billiard tables $Q \subset \mathbb{R}^2$ where a short piece of trajectory experiences infinitely many reflections accumulating at a point of ∂Q where the curvature vanishes. Such "anomalous" examples were found by Halpern [Hn].

On the contrary, when the boundary of Q is concave (i.e., convex inward), the number of reflections can be well bounded. To picture a billiard table with concave boundary, take a polygon and bow each side inward a bit. Or recall the table on a torus where a disk is removed, Fig. IV.4.

It is clear that near one smooth concave piece of ∂Q any short billiard trajectory can only have one reflection. Consider now a corner point, i.e. a vertex A where two (or more) concave pieces of ∂Q meet. Estimates on the number of reflections near such corners have been extensively studied by Galperin [Ga], Vasserstein [Va] and others. Here we present the most general estimate obtained by Burago, Ferleger & Kononenko in 1998 [BFK2].

To define a corner point with concave walls in $\mathbb{R}^d$, one can consider finitely many closed convex subsets $B_i \subset \mathbb{R}^d$, $i = 1, \ldots, n$, whose boundaries are C^1 hypersurfaces and define a billiard domain by

$$Q = \mathbb{R}^d \setminus (\cup_{i=1}^n B_i) \tag{IV.2.13}$$

It is clearly enough to assume that

$$B := \bar{Q} \cap (\cap_{i=1}^n B_i) \neq \emptyset$$

and consider short billiard trajectories near B (the set B plays the role of a vertex).

For any two points $X, Y \in Q$ we denote by $T(X,Y)$ the piece of billiard trajectory starting at X and ending at Y (if one exists), and by $|T(X,Y)|$ its length. The following lemma compares $|T(x,y)|$ to the distance from X and Y to the "bottom of the corner" – the set B.

Lemma IV.2.1 (Comparison Lemma) *For every $X, Y \in Q$ and every $A \in B$*

$$|XA| + |AY| \geq |T(X,Y)|$$

The inequality is strict if one of the reflections occurs at a strictly concave part of the boundary of Q.

Proof. Denote by $X_i \in \partial Q$, $i = 1, \ldots, k$, the reflection points of the trajectory $T(X,Y)$ and put $X_0 = X$, $X_{k+1} = Y$. Consider the triangulated surface made by the triangles AX_iX_{i+1}, $i = 0, \ldots, k$. Let $A'X_0'X_1' \cdots X_{k+1}' \subset \mathbb{R}^2$ be the *unfolding* (or *development*) of this surface made by putting those triangles on a plane with adjacent sides, see Fig. IV.7. In Exercise IV.2.1, it is proved that the curve $\gamma = X_0 \cdots X_{k+1}$ is convex: for any $i = 0, \ldots, k$ it lies on one side of the line X_iX_{i+1} (opposite to the point A'). The convexity of γ implies

$$\begin{aligned} |XA| + |AY| &= |X_0'A'| + |A'X_{k+1}'| \\ &\geq \sum_{i=0}^{k} |X_i'X_{i+1}'| = \sum_{i=0}^{k} |X_iX_{i+1}| = |T(X,Y)| \qquad \square \end{aligned}$$

Theorem IV.2.2 ([Ga], [Va]) *For any billiard trajectory of finite length in Q the number of reflections is finite.*

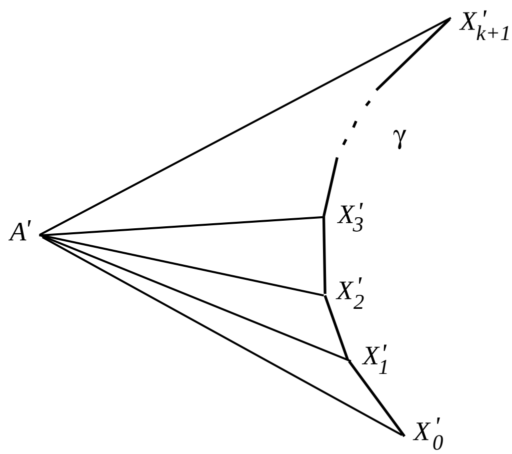

Figure IV.7: The development of a triangulated surface on a plane.

Proof. Assume the opposite – a trajectory T starting at $X \in Q$ has infinitely many reflection points that accumulate at a point $A \in B$ (if $A \notin B$, we can remove some B_i's from our construction). Let $X_1, X_2, \ldots$ be the points of reflection, and $X_i \to A$ as $i \to \infty$. Clearly the length of the straight segment $X_1 A$ is smaller than the length $|T(X_1, A)|$ of the entire trajectory between X_1 and A. Therefore we can find X_k sufficiently close to A so that $|X_1 A| + |AX_k| < |T(X_1, X_k)|$, which contradicts to the comparison Lemma. □

A uniform bound on the number of reflections requires some extra conditions on the billiard domain. Indeed, if a corner point A of a planar billiard table $Q \subset \mathbb{R}^2$ with concave boundary is a cusp, i.e. made by two concave curves tangent to each other at A, then a short billiard trajectory can experience arbitrary many reflections near A, see Exercise IV.2.2. Therefore, some sort of transversality of B_i's at their intersection B is necessary. Such a condition was found in [BFK2]:

Definition. A billiard domain Q given by (IV.2.13) is *nondegenerate* in a subset $U \subset \mathbb{R}^d$ with constant $C > 0$ if for any $I \subset \{1, \cdots, n\}$ and for any

$y \in (U \cap Q) \setminus (\cap_{i\in I} B_i)$

$$\max_{k\in I} \frac{\mathrm{dist}(y, B_k)}{\mathrm{dist}(y, \cap_{i\in I} B_i)} \geq C$$

whenever $\cap_{i\in I} B_i$ is nonempty.

Roughly speaking, this means that if a point is d-close to all the walls from I, then it is d/C-close to their intersection.

Theorem IV.2.3 ([BFK2]) *Let a semidispersing billiard Q be nondegenerate in an open domain $U \subset \mathbb{R}^d$. Then for any point $x \in U$ there exist a number $M_x < \infty$ and a smaller neighborhood U_x of x such that every billiard trajectory entering U_x leaves it after making no more than M_x collisions with the boundary ∂Q.*

Corollary IV.2.4 ([BFK2]) *The system of N hard balls of arbitrary masses and arbitrary radii in an open space $\mathbb{R}^d$ can experience no more than a certain number, $M < \infty$, collisions. Here*

$$M = \left(32 \sqrt{\frac{m_{\max}}{m_{\min}}} \, \frac{r_{\max}}{r_{\min}} \, N^{\frac{3}{2}} \right)^{N^2}$$

where $m_{\max}$ and $m_{\min}$ are the maximum and minimum masses and $r_{\max}$ and $r_{\min}$ are the maximum and minimum radii of the balls, respectively.

It is remarkable that all the above facts admit rather elementary and geometrically explicit proofs, see [BFK2]. Those are, however, quite involved to be included in this book.

Exercises:

IV.2.1. Prove that the curve $\gamma = X_0' \cdots X_{k+1}'$ defined in the proof of Comparison Lemma is indeed convex. Hint: observe the relation between the "development" of the tangent line at X_i and the unfolding line $A'X_i'$ and use the fact that the angle of reflection is equal to the angle of incidence.

IV.2.2. Let $Q \subset \mathbb{R}^2$ be a billiard table with a corner point A at which two concave components $\gamma_1, \gamma_2 \subset \partial Q$ of the boundary meet and make a cusp (i.e., intersect at zero angle). Show that a short billiard trajectory can experience arbitrary many reflections at ∂Q near the point A. Hint:

consider a common tangent line L to γ_1 and γ_2 at A and start the trajectory on $L \cap Q$ with an initial velocity vector pointing almost (but not exactly) into A. Show that at each reflection the velocity vector turns by an arbitrary small amount, hence it may take arbitrary many reflections to turn the trajectory around and drive it away from A.

IV.3 Simple Examples

In the study of billiards, even simple examples may exhibit surprisingly rich dynamical properties.

Circles. We start with a circular billiard table: let Q be a disk of radius one. The surface $M = \partial Q \times [-\pi/2, \pi/2]$ (with the coordinates r, φ) is a cylinder whose base is a circle of length 2π and whose height is π.

To see how the map $T : M \to M$ acts, let $x = (r, \varphi) \in M$. The next reflection occurs at the point $Tx = (r + \pi - 2\varphi, \varphi)$, and the n-th reflection at

$$T^n x = (r + n(\pi - 2\varphi), \varphi)$$

In this formula r is assumed to be a cyclic coordinate, i.e. $r + n(\pi - 2\varphi)$ is taken modulo 2π. Two conclusions immediately follow.

(i) The angle of reflection φ is preserved by the map T, i.e. every curve $\varphi =$ const in M is invariant under T. Each such curve is a circle parallel to the base of the cylinder M.

(ii) The map T restricted to any curve $\varphi =$ const rotates it through a constant angle, $\pi - 2\varphi$. Hence, T acts as a circle rotation on each invariant curve.

We also note that the return time function $\tau(x) = 2\cos\varphi$ is constant on each invariant circle $\varphi =$ const. All the links of a billiard trajectory starting with a given angle φ are tangent to one circle of radius $R = \sin\varphi$ concentric to the billiard table, see Fig. IV.8.

If the angle of rotation $\pi - 2\varphi$ is a *rational multiple* of π, i.e. if

$$\frac{\pi - 2\varphi}{2\pi} = \frac{m}{n}$$

with $m, n \in \mathbb{Z}$, then the map T on the circle $\varphi =$ const is periodic with period n. Moreover, each point on this circle is periodic with the same period.

If the angle of rotation $\pi - 2\varphi$ is *irrational*, i.e. if φ/π is an irrational number, then the circle rotation is ergodic and every trajectory is dense and uniformly distributed on the circle. In this case the links of any billiard trajectory densely fill a ring on the billiard table with the inner radius $R = \sin\varphi$ (see Fig. IV.8).

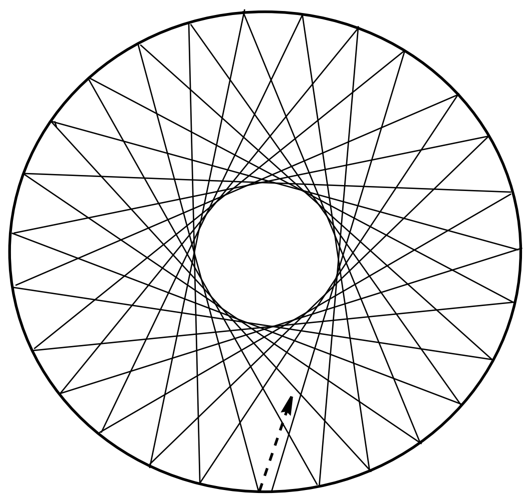

Figure IV.8: Links of a nonperiodic trajectory.

One can clearly see on Fig. IV.8 that the links look denser near the inner circle of the ring. In other words, the links focus on the inner circle. If our billiard trajectory were the path of a laser ray and the border of the billiard table Q were a perfect mirror, then it would feel "very hot" there, on the inner circle. For this reason, the inner circle is called a *caustic* (which means "burning"). In general, a caustic for a billiard is a curve such that if one link of a billiard trajectory is tangent to it, then every other link (or the line that contains that link) of the same trajectory is tangent to the caustic.

Since the surface M is foliated by invariant curves $\varphi = \text{const}$, the map T is not ergodic. Any measurable set that is a union of invariant curves will be T-invariant. Another way to see nonergodicity is to find a nonconstant invariant function. Here the function $F : M \to \mathbb{R}$ defined by $F(r, \varphi) = \varphi$ is a smooth nonconstant function invariant under T, i.e. it satisfies $F(Tx) = F(x)$ for all $x \in M$.

Definition. If a smooth dynamical system $T : M \to M$ on a manifold M admits a smooth nonconstant function invariant under T, then F is called a *first integral* and T is said to be *integrable*.

If $T : M \to M$ is integrable, then every level surface $S_c = \{F(x) = c\}$ is T-invariant, i.e. M can be foliated by invariant hypersurfaces. If $\dim M = d$ and $T : M \to M$ admits $d - 1$ independent first integrals $F_1, \ldots, F_{d-1}$, then M can be foliated by one-dimensional T-invariant submanifolds $\{F_1(x) = c_1, \ldots, F_{d-1}(x) = c_{d-1}\}$, where $c_1, \ldots, c_{d-1} \in \mathbb{R}$.

Definition. If M can be foliated by one-dimensional T-invariant submanifolds (curves), then T is said to be *completely integrable*.

The billiard in a circle is then completely integrable.

The derivative of our map T is

$$DT(x) = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} \tag{IV.3.1}$$

(this follows from (IV.1.7), where we need to set $\varphi_1 = \varphi$, $\tau = 2\cos\varphi$ and $K = -1$). Hence,

$$DT^n(x) = \begin{pmatrix} 1 & -2n \\ 0 & 1 \end{pmatrix}$$

for all $n \in \mathbb{Z}$. This easily implies that both Lyapunov exponents equal zero at every point $x \in M$. As a result, the entropy of the map T is zero as well. Hence, by all standards, the map T is not chaotic.

Note, however, that typical tangent vectors do grow under DT^n as $n \to \infty$, but only linearly in n (rather than exponentially). Such maps, where the separation of nearby trajectories is typically linear, are usually said to exhibit *parabolic* behavior (as opposed to hyperbolic one that we learned in Chapter III).

Ellipses. Consider now a billiard system inside an ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{IV.3.2}$$

with some $a > b > 0$. The surface M is then a cylinder with base (IV.3.2) and height π.

The billiard inside ellipse has many nice geometric properties. We refer to [CFS, Ta1, Bi] for proofs, which are all very elementary but sometimes a little involved. One can also use the methods of projective geometry (the so called Poncelet theorem) to derive all these properties.

Denote by F_1 and F_2 the foci of the ellipse (IV.3.2), note that they lie on the x axis. If a billiard trajectory passes through one focus, then it reflects at ∂Q and passes through the other focus, and so on (see Exercise IV.3.2). The trajectories that pass through the foci are special. They make a closed curve on the surface M (the ∞-shaped curve shown on Fig. IV.9).

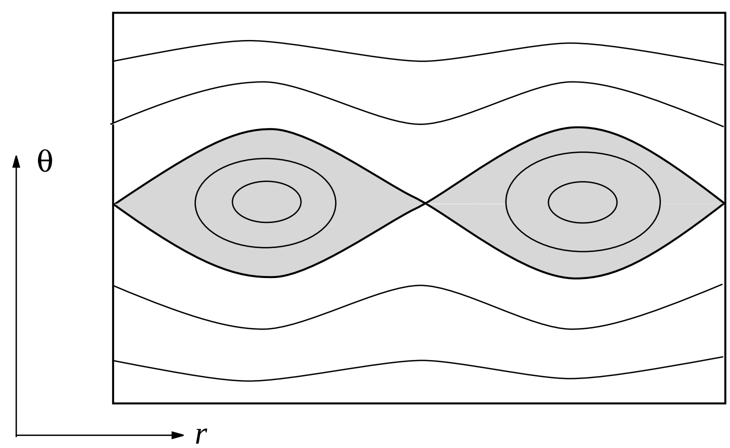

Figure IV.9: The surface M for the elliptic billiard.

If a trajectory crosses the segment F_1F_2 joining the foci, then it reflects at ∂Q and crosses this segment again. Such trajectories fill two domains in M bounded by the ∞-shaped curve (marked grey on Fig. IV.9).

The remaining trajectories never cross the segment F_1F_2. Every time such a trajectory crosses the major axis (the x axis), the intersection point lies away from F_1F_2 (alternatively, to the left and to the right of the segment F_1F_2). Such trajectories fill two domains in M (white area above and below the ∞-shaped curve on Fig. IV.9).

There are two distinguished periodic orbits - one lying on the major axis (it is represented by the points of intersection of the two branches of the ∞-shaped curve) and one lying on the minor axis (represented by the centers of the two parts of the ∞-shaped domain).

Every billiard trajectory in ellipse not passing through the foci and different from the distinguished periodic orbits has a caustic. For trajectories crossing the segment F_1F_2 the caustics are hyperbolas confocal to the ellipse (IV.3.2). For the other trajectories the caustics are ellipses confocal to the given ellipse (IV.3.2) and lying inside that ellipse (those elliptical caustics generalize concentric circles in the circular billiard discussed above).

All the trajectories tangent to one elliptic caustic lie on a curve in M that is invariant under T (horizontal "waves" in the white area on Fig. IV.9). Those invariant curves appear as deformations of the invariant lines $\varphi = \text{const}$ in the circular billiard. All the trajectories tangent to one hyperbolic caustic lie on two closed curves in M, one inside each half of the ∞-shaped domain, each curve being invariant under T^2. Therefore, the surface M is completely foliated by invariant curves. In this sense, the elliptical billiard is similar to the circular one – both are completely integrable.

On each invariant curve the map T (or T^2) is conjugate to a rigid circle rotation through some angle (called the *rotation number*). That number changes continuously and monotonically with the invariant curve. The special ∞-shaped curve that separates the two regions with different types of caustics (hyperbolic and elliptic) is called a *separatrix.* It is also invariant under T, but it does not rotate. Instead, every trajectory starting on the separatrix converges (both in the future and in the past) to a 2-periodic orbit lying on the major axis.

The formula for the derivative of the map T in the r, φ coordinates is complicated and would not tell us much. But one can choose a special curvilinear coordinate system in M (apart from the separatrix) with one axis along an invariant curve, in which the matrix DT will be triangular and have ones on the diagonal. Then one can easily derive that the Lyapunov exponents are zero (since they do not depend on the choice of the coordinate system or even the metric). Hence, the entropy of the map T is zero as well. Again, by all our standards, the map T is not chaotic.

The dynamical behavior of the map T, apart from the separatrix and the distinguished periodic orbits, is parabolic, as in the case of circular billiard. But there is more to the picture now. The periodic trajectory lying on the major axis is hyperbolic, and its stable and unstable manifolds lie on the separatrix, see Exercise IV.3.4. The other distinguished periodic trajectory, lying on the minor axis (the y axis) is neither hyperbolic nor parabolic.

Definition. Let x be a periodic point, i.e. $T^n x = x$ for some $n \geq 1$. If the derivative $DT^n(x)$ has only simple complex (nonreal) eigenvalues lying on the unit circle, then the point x is said to be *elliptic*.

When $\dim M = 2$, then the derivative map DT^n is just a rotation through some angle. In that case a neighborhood of x often contains many T^n-invariant closed curves (ellipse-like ovals, hence the term *elliptic*). This fact is related to the KAM theory mentioned at the end of Section III.3. Our periodic trajectory lying on the minor axis is elliptic, and its neighborhood is completely foliated by T^2-invariant ovals, see Exercise IV.3.3.

Hence, the the billiard inside an ellipse displays a combination of parabolic, hyperbolic and elliptic structures.

Smooth convex tables. One may wonder if there are completely integrable billiards other than the circular and elliptic ones. This question is known as

Birkhoff conjecture. The only completely integrable billiards are those in circles and ellipses.

Surprisingly, this conjecture is still open. Most mathematicians believe that it is correct, though.

On the other hand, billiards in general oval-shaped tables, even if not completely integrable, have many common features with billiards in ellipses. In 1973 Lazutkin [La] proved: if Q is a strictly convex domain (the curvature of the boundary never vanishes) with sufficiently smooth boundary, then there exists a positive measure set $N \subset M$ that is foliated by invariant curves. The set N accumulates near the boundary ∂M. All trajectories starting in the set N have caustics, which are convex closed curves lying inside Q. Of course, the billiard cannot be ergodic since $\nu(N) > 0$. The Lyapunov exponents for points $x \in N$ are zero. However, away from N the dynamics might be quite different.

Originally Lazutkin demanded 553 continuous derivatives in his theorem, but later R. Douady [Dy] proved that 6 are enough (therein it is conjectured that the curve may be C^4). This improvement was a consequence of the improvements in the KAM theorem. There exist examples that show that at least two derivatives are necessary for the existence of caustics[3]. Also, J. Mather [Mz] proved that even a single point on the

[3]Convex billiard tables whose boundaries are C^1 curves were studied by Hubacher [Hu]. Their second derivatives exist and are continuous except for a finite set of points

boundary of a C^2 convex table where the curvature vanishes prevents the existence of caustics near ∂Q.

Stadium. We will now modify a little our circular billiard table. Take two semicircles of radius one, put them together so that they make a whole circle and then move them slightly apart, by a distance 2ε. The resulting domain can be described as

$$Q = \{|y| < 1,\ |x| < \varepsilon + \sqrt{1 - y^2}\}$$

This is a convex domain with C^1 (but not C^2) boundary. It resembles a stadium.

L. Bunimovich first introduced and studied stadia in the seventies [Bu2]. He discovered a striking fact: the billiard in a stadium has nonzero Lyapunov exponents a.e., i.e. it is completely hyperbolic. Its entropy is positive. Moreover, it is ergodic, mixing and isomorphic to a Bernoulli shift. Hence, one can perturb a completely integrable circular billiard by an arbitrary small amount to get a fully chaotic dynamical system.

Similar modifications can be done with ellipses. Let us cut the ellipse along its minor axis, move the two halves of the ellipse apart in the direction of the major axis, and joint both parts by two segments of length equal $2h$. We obtain the *elliptical stadium*. It is also hyperbolic an ergodic for $h > h_{\min} > 0$, where $h_{\min}$ depends on the eccentricity of the ellipse.

There is another way of modifying circular and elliptic billiards to obtain a chaotic system. Take a circle and cut a slice of it along a chord (which can be arbitrarily small). The billiard in such a "truncated" circle is hyperbolic and ergodic, as Bunimovich showed in [Bu1]. Or take the ellipse (IV.3.2) and cut two symmetric slices along the lines $y = \pm(b - h)$. The billiard in such a truncated ellipse is hyperbolic and ergodic, too [De].

We note that all the billiards mentioned above are convex with C^1 (but not C^2) boundary. The map $T : M \to M$ is continuous but not always differentiable. There are points where DT does not exist (singular points). These are points $x \in M$ such that $Tx \in \Gamma^*$. It appears that the singularities of T always exist in chaotic billiards.

---

where one-sided limits exist, but the curvature function is discontinuous at those points. It is proved that caustics do not exist in a vicinity of the boundary, but it is observed that they could exist away from the boundary.

Exercises:

IV.3.1. Let Q be a ring – a domain bounded by two concentric circles. Describe the phase space M of the billiard in Q. Note that the map $T : M \to M$ is not continuous.

IV.3.2. Let Q be an ellipse with foci F_1 and F_2. Let $A \in \partial Q$ and L the tangent line to ∂Q at A. Prove that the segments AF_1 and AF_2 make equal angles with the line L (this fact is known in projective geometry as Poncelet theorem). Hint: use the characteristic property of the ellipse: the sum $|AF_1| + |AF_2|$ does not depend on A.

IV.3.3.
(a) Let x be a 2-periodic point for the billiard map T. Note that in this case $\cos\varphi = \cos\varphi_1 = 1$ in the formula (IV.1.7). Then show that the eigenvalues of the map DT^2 are complex (nonreal) if and only if

$$\tau K K_1 + K + K_1 < 0 \quad \text{and} \quad (\tau K + 1)(\tau K_1 + 1) > 0$$

in the notation of (IV.1.7).
(b) Prove that the periodic orbit along the minor axis of the ellipse is elliptic.

IV.3.4. Prove that the periodic orbit along the major axis of the ellipse is hyperbolic. Describe precisely its stable and unstable manifolds.

IV.4 Dispersing Billiards

From now on we will study chaotic billiards only. They are characterized, first of all, by non-zero Lyapunov exponents, and then by positive entropy, ergodicity and mixing. For simplicity, we will restrict ourselves to planar billiards, i.e. assume $d = 2$.

Wave fronts. Recall that Lyapunov exponents for a map $T : M \to M$ are characteristics of tangent vectors $u \in \mathcal{T}M$. Tangent vectors admit an explicit geometric representation in billiard systems. First, a tangent vector $u \in \mathcal{T}_x M$ at a point $x = (q, v) \in M$ can be represented by an infinitesimal curve $\gamma \subset M$ passing through x in the direction of u. The trajectories of the points $y \in \gamma$, when they just leave the boundary ∂Q and enter the domain Q, make a one-parameter family of oriented lines, which we call a *bundle of rays*.

It is actually more convenient to work with a bundle of rays in the domain Q rather than with the curve γ in M. Let us take an orthogonal cross-section of that bundle, which passes through the point $x = (q, v)$. We call that cross-section Σ, see Fig. IV.10. It is a curve in Q that intersects every ray of our bundle perpendicularly. Velocity vectors of the points on that curve are thus normal vectors to it. Hence, Σ is smooth curve equipped with a family of normal vectors pointing in the direction of motion. We call Σ a *wave front*, the term borrowed from physics.

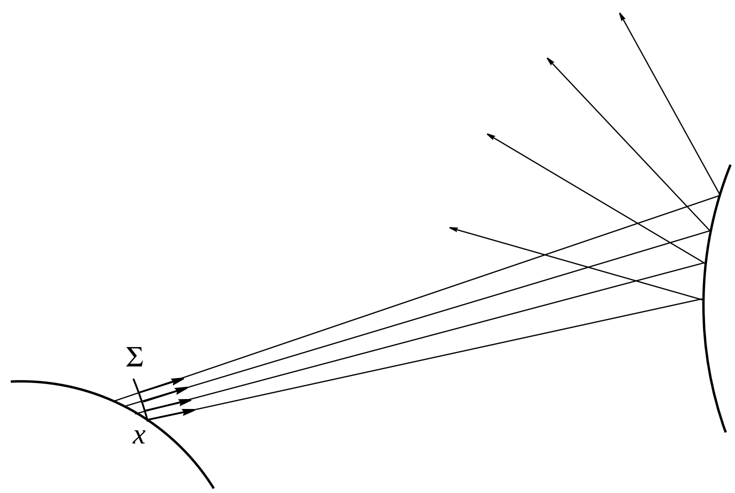

Figure IV.10: A divergent wave front in Sinai billiard.

The curvature of the front Σ plays a crucial role in our analysis. The sign of the curvature is chosen according to the following rule. If the front Σ is *divergent*, as the one shown on Fig. IV.10, then its curvature is positive. If the front is *convergent*, its curvature is negative. If the front is made by parallel rays (Σ is then a perpendicular line), then the curvature is zero, and such fronts are said to be *neutral*.

The curvature of the boundary ∂Q will play an equally important role. Remember our convention in Section IV.1 (made in the proof of

Lemma IV.1.1): the curvature of the boundary is positive if it is concave (convex inward), and negative if the boundary is convex.

Definition. A smooth component of the boundary $\Gamma_i \subset \partial Q$ is said to be *dispersing*, if its curvature is positive, *focusing* if its curvature is negative and *neutral* if its curvature is zero.

If the curvature of Γ_i changes sign, we divide it into smaller components whose curvature has constant sign. So, we assume that the boundary ∂Q consists of the components of the above three types – dispersing, focusing and neutral.

Definition. If all the components of the boundary ∂Q are dispersing, the billiard is said to be *dispersing*. If ∂Q consists of dispersing and neutral components, the billiard is said to be *semidispersing*.

Dispersing billiards are also known as *Sinai billiards*, since Ya. Sinai introduced them in 1970 and thus began mathematical study of chaotic billiards. Dispersing and semidispersing billiards make the main (but not only) classes of chaotic billiards.

It is geometrically evident that when a neutral (parallel) bundle of rays falls upon a dispersing boundary, then the outgoing rays (after the reflection) make a divergent bundle. It is also easy to see that at the subsequent reflections at dispersing boundaries that bundle "opens up" even more, and its rays rapidly diverge from each other, see Fig. IV.10. This is the main reason of exponential growth of wave fronts in dispersing billiards, which will translate into the existence of a positive Lyapunov exponent. We know from Section IV.1 that the sum of the two Lyapunov exponents of the map $T : M \to M$ is zero, hence the other Lyapunov exponent will be automatically negative.

It is interesting that one can explicitly construct the wave front $\Sigma^-(x)$ corresponding to the negative Lyapunov exponent at $x \in M$. It must exponentially converge, i.e. shrink, as time goes on. The existence of such a front is reminiscent the idea of *stability* in the theory of differential equations. Recall that a trajectory (a solution of a differential equation) is said to be stable if all the nearby trajectories do not deviate too far from it in the future. Next we will construct $\Sigma^-(x)$.

Let $x = (q, v) \in \mathcal{M}$. Take a large $t > 0$ and consider the point $(q_t, v_t) = \Phi^t(q, v)$, i.e. a distant image of the point (q, v). Reverse the time by using the involution, i.e. take the point $(q_t, -v_t) = I(q_t, v_t)$. Consider a neutral wave front Σ_t^+ passing through $(q_t, -v_t)$, i.e. a parallel bundle of rays around q_t moving in the direction of $-v_t$. Then, by the main property of the involution, the image $\Phi^t(\Sigma_t^+)$ will be a wave front containing the point $(q, -v) = I(q, v)$. Its involution $\Sigma_t^- = I(\Phi^t(\Sigma_t^+))$ will then contain the original point (q, v).

Now, in Sinai billiards (as we remarked above) parallel wave fronts become divergent and grow exponentially fast in time. Hence, the front $\Phi^t(\Sigma_t^+)$ will be exponentially (in t) longer than the front Σ_t^+ (remember that all our fronts are infinitesimally small anyway). Therefore, the front $\Phi^t(\Sigma_t^-)$ will be *exponentially shorter* than the front Σ_t^-. This gives us a wave front starting at the given point (q, v) that shrinks exponentially fast during the time interval $(0, t)$. It remains to take the limit $t \to \infty$ and get a front $\Sigma^-(x) := \lim_{t\to\infty} \Sigma_t^-$ that shrinks exponentially fast all the time. This is our stable family of trajectories that correspond to the negative Lyapunov exponent at the point (q, v). Note that the fronts Σ_t^- and the limit front $\Sigma^-(x)$ are convergent (have negative curvature).

In addition to the stable families represented by convergent wave fronts, we will need their "reversed" counterparts. Given a point $x = (q, v) \in \mathcal{M}$, let $\Sigma^-(I(x))$ be the stable family constructed for the point $I(q, v) = (q, -v)$. Then the wave front $\Sigma^+(x) := I(\Sigma^-(I(x)))$ will be divergent (have positive curvature). It will grow exponentially fast in the future and contract exponentially fast in the past. Thus, its Lyapunov exponent will be positive.

We now turn to exact equations describing the dynamics of wave fronts in billiards.

Let Σ_0 be a wave front, i.e. a C^1 arc equipped with normal vectors pointing in the direction of motion. Let $(q, v) \in \Sigma_0$ be one of its elements. Denote by χ_0 the curvature of the front at the point q whose sign is set by the above rules. Consider the front $\Sigma_t = \Phi^t(\Sigma_0)$, $t > 0$, it contains the point $(q_t, v_t) = \Phi^t(q, v)$. We will compute the curvature χ_t of the front Σ_t at the point q_t.

Assume first that t is small enough, so that during the time interval $(0, t)$ no reflections at ∂Q occurred. Then Φ^t is just a free motion. In that case the equation is very simple:

$$\chi_t = \frac{1}{t + 1/\chi_0} \tag{IV.4.1}$$

It is well known in geometric optics. One can verify it directly, paying attention to the sign rules for the curvature. (Note: it is enough to approximate Σ_0 by a circular arc, then $|\chi_0|$ is the reciprocal of its radius.)

When the wave front bounces off the boundary ∂Q, its curvature instantaneously jumps. One can say that the curvature of the boundary is then "combined" with the curvature of the front itself. The following is one of the basic laws of geometric optics known as *mirror equation*. Denote by χ_- and χ_+ the curvature of the front before and after reflection, respectively. Also, let K_1 be the curvature of the boundary at the point of reflection (whose sign is set by the above mentioned rules), and φ_1 the angle of reflection, see Section IV.1. Then the mirror equation reads

$$\chi_+ = \chi_- + \frac{2K_1}{\cos\varphi_1} \tag{IV.4.2}$$

The proof is left as an exercise (see Exercise IV.4.1) . It is quite involved, but we advise the reader to attempt it anyway – it can help to "see" better what is going on.

By combining Eqs. (IV.4.1) and (IV.4.2), one can compute the curvature of the wave front at any time t.

Continued fractions. For the purposes of describing stable and unstable wave fronts (as outlined above), it is more sensible to express χ_0 in terms of χ_t, $t > 0$. That is what we do next. Denote by χ_1 is the curvature of the wave front right after its first reflection at ∂Q. Then

$$\chi_0 = -\cfrac{1}{\tau + \cfrac{1}{\frac{2K_1}{\cos\varphi_1} - \chi_1}}, \tag{IV.4.3}$$

where τ is the time of the first reflection. Furthermore, the value χ_1 can be in turn expressed through the curvature after the second reflection, etc. Proceeding in this way to the following reflections and repeating the same trick we will append more and more alike fractions to our main fraction in Eq. (IV.4.3) extending it downward. The limiting fraction will be infinite. Such expressions are called *continuous fractions*. Their value is defined as the limit (if one exists) of a sequence of finite, truncated, fractions. If $x = (q, v) \in M$, then our continuous fraction is

$$\kappa^s(x) = -\cfrac{1}{b_1(x) + \cfrac{1}{b_2(x) + \cfrac{1}{b_3(x) + \cfrac{1}{b_4(x) + \cdots}}}} \tag{IV.4.4}$$

$$b_{2k}(x) = \frac{2K(T^k x)}{\cos\varphi(T^k x)}, \quad b_{2k+1}(x) = \tau(T^k x), \quad k \in \mathbb{N}. \tag{IV.4.5}$$

Here $K(T^k x)$ is the curvature of the boundary at the point $T^k x$ of the k-th reflection, and $\varphi(T^k x)$ is the angle of that reflection. Also, $\tau(T^k x)$ is the free path between the kth and $(k+1)$st reflections.

The fraction in Eq. (IV.4.4) has an alternating structure – the odd components are τ's and the even ones are fractions $2K/\cos\varphi$. This reflects the natural alteration of free paths and reflections along the trajectory.

We remark that if one truncates the infinite fraction (IV.4.4) at an even term, adding zero, the value of the truncated fraction is the curvature of the front Σ_t^- constructed above, with t being the time of the corresponding reflection.

For Sinai billiards, the components of the boundary are dispersing, hence $K(T^k x) > 0$ by the sign rules for the curvature. Since $\tau(T^k x) > 0$ and $\cos\varphi(T^k x) > 0$ in any case, then all the terms of our fraction $\kappa^s(x)$ are positive. This simplifies the verification of its convergence. This fact was first observed in [Si2]; see Exercise IV.4.2.

The limit value of the fraction in (IV.4.4) is then positive, thus $\kappa^s(x)$ is negative. We see, once again, that a stable wave front in Sinai billiards is convergent.

Let

$$\kappa^u(x) = a_0(x) + \cfrac{1}{a_1(x) + \cfrac{1}{a_2(x) + \cfrac{1}{a_3(x) + \cfrac{1}{a_4(x) + \cdots}}}} \tag{IV.4.6}$$

for $(x, v) \in M$, where

$$a_{2k}(x) = \frac{2K(T^{-k}(x))}{\cos\varphi(T^{-k}(x))}, \quad a_{2k+1}(x) = \tau(T^{-k}(x)), \; k \geq 0 \tag{IV.4.7}$$

The continuous fraction (IV.4.6) is the curvature of the unstable wave front at x. It can be deduced in a similar way as we deduced (IV.4.4) working towards the past (with negative iterations of T). The extra term $a_0(x)$ appears due to the mirror equation (IV.4.2). The convergence of the fraction (IV.4.6) depends on the same factors. For dispersing billiards it is always convergent and $\kappa^u(x)$ is now positive, i.e. the unstable front is divergent.

Theorem IV.4.1 *In dispersing billiards, the fractions (IV.4.4) and (IV.4.6) converge at every point $x \in \mathcal{M}'$. The value $\kappa^s(x) < 0$ gives the curvature* *of the stable wave front. The value $\kappa^u(x) > 0$ gives the curvature of the unstable wave front.*

In semidispersing billiards, the curvature of the neutral components of the boundary is zero, hence some even terms in (IV.4.4) and (IV.4.6) may vanish, see also Exercise IV.4.5. This does not affect convergence, hence $\kappa^s(x)$ and $\kappa^u(x)$ still exist at every point $x \in M'$. However, they may be equal to zero. If $\kappa^s(x) = 0$ or $\kappa^u(x) = 0$, they no longer represent stable or unstable wave fronts. In this case the Lyapunov exponents at x either vanish or do not exist. In particular, when ∂Q consists of neutral components only (i.e., Q is a polygon), then $\kappa^s(x) = \kappa^u(x) = 0$ at every $x \in M'$, and both Lyapunov exponents vanish everywhere.

In fact, the convergence of the continuous fractions (IV.4.4) and (IV.4.6) does not imply hyperbolicity or even existence of the Lyapunov exponents. For example, those fractions converge for the circular billiards (see Exercise IV.4.4) and elliptical ones, their values are nonzero, but the Lyapunov exponents vanish a.e.. L. Bunimovich proved that those fractions converge for very large classes of billiards, chaotic and nonchaotic. In fact, there are no examples of a billiard table where those fractions fail to converge on a set of positive measure.

On the other hand, for all chaotic billiards the above fractions (IV.4.4) and (IV.4.6) do represent the curvature of stable and unstable fronts, related to the negative and positive Lyapunov exponent, respectively.

Equations for wave fronts and their transversality. We now derive equations for the tangent subspaces corresponding to the positive and negative Lyapunov exponents in $\mathcal{T}M$. We will call them *unstable* and *stable* subspaces, respectively, and denote by E^u and E^s.

Let $x = (q, v) \in M$ be a point having coordinates (r, φ) and $u = (dr, d\varphi) \in \mathcal{T}_x M$ a tangent vector at x. Note that dr and $d\varphi$ are infinitesimal quantities. As we explained above, every tangent vector $u \neq 0$ can be represented by a curve in M, for example $(r+s\,dr, \varphi+s\,d\varphi)$, where $0 < s < \varepsilon$ is a small parameter. This curve gives a bundle of outgoing trajectories, whose cross-section is a wave front Σ passing through the point q. The curvature of this front, call it χ, at the point q can be now computed as

$$\chi = \frac{1}{\cos\varphi}\left(\frac{d\varphi}{dr} + K(r)\right) \tag{IV.4.8}$$

This follows from infinitesimal analysis like we used in the calculation of the derivative DT of the billiard map – see Lemma IV.1.1, we leave it as an exercise (see Exercise IV.4.3). Note that the front Σ is not uniquely defined by the vector u, but the curvature χ is completely determined by u. Also, χ only depends on the direction of u, and not on its length.

Solving the equation (IV.4.8) for $d\varphi/dr$ gives

$$\frac{d\varphi}{dr} = -K(r) + \chi \cos\varphi$$

Therefore, the unstable subspace E^u is spanned by a vector $(dr^u, d\varphi^u)$ that satisfies

$$\frac{d\varphi^u}{dr^u} = -K(r) + \kappa^u \cos\varphi$$

and the stable subspace E^s is spanned by a vector $(dr^s, d\varphi^s)$ that satisfies

$$\frac{d\varphi^s}{dr^s} = -K(r) + \kappa^s \cos\varphi$$

where κ^u is given by (IV.4.6) and κ^s by (IV.4.4).

For simplicity, we assume that the curvature of dispersing and focusing components of ∂Q is bounded above and below, i.e.

$$K_{\max} = \max_r |K(r)| < \infty$$

and

$$K_{\min} = \min_r |K(r)| > 0$$

where the minimum is taken over dispersing and focusing components.

We note that in Sinai billiards $\kappa^s < 0$ and $K(r) > 0$, hence $d\varphi^s/dr^s < 0$, and moreover

$$\frac{d\varphi^s}{dr^s} < -K(r) \le K_{\min} < 0 \tag{IV.4.9}$$

Also, according to (IV.4.6), we have $\kappa^u \ge 2K(r)/\cos\varphi$, hence

$$\frac{d\varphi^u}{dr^u} > K(r) \ge K_{\min} > 0 \tag{IV.4.10}$$

In other words, unstable directions are always positive, or increasing (in the r, φ coordinates), and stable directions are negative, or decreasing. Moreover, both are bounded away from the horizontal direction $d\varphi = 0$.

They may not be bounded from the vertical direction, though, i.e. the derivatives (IV.4.10) and (IV.4.9) may be arbitrarily large. But in typical cases they are bounded: one can easily see that $|\kappa^s(x)| < 1/\tau(x)$ and $|\kappa^u(x) - a_0(x)| < 1/\tau(x)$. Therefore,

$$\left|\frac{d\varphi^a}{dr^a}\right| \le \max_r |K(r)| + \frac{1}{\min_x \tau(x)}$$

for $a = u, s$.

Our analysis shows that the lower and upper bounds on the function $\tau(x)$ are involved in the studies of Sinai billiards. The following are simple facts:

(i) if a Sinai billiard table has *corner points* (intersections of smooth components of ∂Q where they make angles $< \pi$), then $\min_x \tau(x) = 0$; otherwise $\tau_{\min} = \min_x \tau(x) > 0$;

(ii) for any Sinai billiard table $Q \subset \mathbb{R}^2$ we have $\tau_{\max} = \max_x \tau(x) < \infty$;

(iii) for Sinai billiards on a torus, $Q \subset \mathbb{T}^2$, the value $\tau_{\max}$ may be either finite or infinite. In the first case we say that Q has *finite horizon*.

We now see that the stable and unstable directions can approach the vertical direction $dr = 0$ when $\tau(x) \approx 0$, which only occurs near the corner points in Sinai billiard tables. If there are no corner points, $\tau_{\min} > 0$ and the derivatives (IV.4.10) and (IV.4.9) are bounded above. In that case the stable and unstable subspaces, E^s and E^u are uniformly transversal – the angle between them is bounded away from zero. In this respect, Sinai billiards are similar to Anosov diffeomorphisms.

We say that a Sinai billiard table with corner points is a *proper* table if all the corners have positive angles, i.e. the sides of ∂Q intersect each other transversally. Originally, Sinai and his school only studied proper billiard tables. Improper tables, where some corners are *cusps* (making zero angle) have somewhat different properties.

It is interesting that in proper Sinai billiard tables with corner points the stable and unstable spaces E^s and E^u cannot approach the vertical direction $dr = 0$ simultaneously. There is a geometric argument described in [BSC2] that prevents this anomaly. Therefore, the angle between E^s and E^u remains bounded away from zero for all proper Sinai billiards! The transversality of E^u and E^s plays an important role in the study of ergodic properties of chaotic billiards.

Expansion and contraction rates of wave fronts. As we explained above, in Sinai billiards divergent fronts constantly grow as they move forward because of obvious geometric reasons. On the other hand, the derivative DT of the map T, cf. Lemma IV.1.1, does not always expand unstable vectors $u = (dr^u, d\varphi^u)$, and we will see that shortly. This fact suggests that the standard Euclidean norm $|u| = [(dr)^2 + (d\varphi)^2]^{1/2}$ of tangent vectors may not be so convenient for the study of divergent fronts in Sinai billiards. Indeed, it seems more natural to measure the length of a tangent vector $u = (dr, d\varphi)$ by the width of the orthogonal cross-section Σ associated to the wave front. This width is denoted by $|u|_p$ and called the p-norm. One can easily see that

$$|u|_p = \cos\varphi\,|dr|$$

Strictly speaking, this is not a norm, since $|u| = 0$ for all vertical vectors $(0, d\varphi)$. It should be then called a *pseudonorm.* But we have seen above that stable and unstable vectors are never vertical, so they will always have a positive p-norm. That justifies our interest in using the p-norm instead of the Euclidean norm.

The p-norm of a tangent vector $u \in \mathcal{T}_x M$ changes under DT by a simple formula

$$\begin{aligned}\frac{|DT(u)|_p}{|u|_p} &= |1 + \tau(x)\chi| \\ &= \left|1 + \frac{\tau(x)}{\cos\varphi}\left(\frac{d\varphi}{dr} + K(r)\right)\right| \qquad \text{(IV.4.11)}\end{aligned}$$

where χ is the curvature of the corresponding wave front at x.

In Sinai billiards, unstable vectors u satisfy (IV.4.10), hence the quantity (IV.4.11) is always greater than one. In other words, unstable vectors always expand under T in the p-norm, just as we expected. Moreover, if there are no corner points on the table, then $\tau(x) \geq \tau_{\min} > 0$, and the quantity (IV.4.11) has a lower bound $\Lambda > 1$, i.e.

$$|DT(u)|_p \geq \Lambda |u|_p, \qquad \Lambda > 1 \qquad \text{(IV.4.12)}$$

for all unstable vectors. This demonstrates uniform hyperbolicity of T in the sense of Chapter III, another common feature between Sinai billiards and Anosov diffeomorphisms.

It is interesting that proper Sinai billiard tables with corner points are also uniformly hyperbolic, at least after some minor modifications. Indeed, the number of rapid reflections near any corner point is uniformly bounded, see Section IV.2. Hence, there is a constant $m \geq 1$ such that no trajectory can experience more than m reflections near any corner point. In that case the map T^m is uniformly hyperbolic, i.e. (IV.4.12) holds with T replaced by T^m.

On the other hand, the expansion factor in the p-metric is unbounded. If x lies near S_0, where $\cos\varphi \approx 0$, then the quantity (IV.4.11) can be arbitrarily large. This constitutes an important difference between Sinai billiards and Anosov maps. It is also easy to see that in Sinai billiard tables with finite horizon, $\cos\varphi \approx 0$ is the only reason why the quantity (IV.4.11) may approach infinity. In other words, we have

$$c_1/\cos\varphi \le |DT(u)|_p/|u|_p \le c_2/\cos\varphi \qquad \text{(IV.4.13)}$$

with some constants $c_1, c_2 > 0$.

We note that in the study of statistical properties of Sinai billiards (their decay of correlations and limit theorems, see Chapter II), a very precise control on the expansion factor $|DT(u)|_p/|u|_p$ is a necessity. The above estimate shows that such a control is difficult in the vicinity of the set $\partial M = \{\cos\varphi = 0\} = \{|\varphi| = \pi/2\}$. In order to improve the control, one can partition the surface M into countably many strips by the lines $|\varphi| = \pi/2 - a_k$, with some $a_k \to 0$ as $k \to \infty$. For example, it is customary to set $a_k = 1/k^2$. Then the critical area near the lines $|\varphi| = \pi/2$ is divided into narrow strips where the expansion factor $|DT(u)|_p/|u|_p$ can be controlled easier. Those narrow regions are called *homogeneity strips* in the literature, they were first introduced in [BSC2].

We now consider the evolution of tangent vectors in the standard, Euclidean norm $|u|$. The relation between our two norms is

$$J(u) = \frac{|u|_p}{|u|} = \cos\varphi \left[1 + \left(\frac{d\varphi}{dr}\right)^2\right]^{-1/2} \qquad \text{(IV.4.14)}$$

Then

$$\frac{|DT(u)|}{|u|} = \frac{|DT(u)|_p \, J(u)}{|u|_p \, J(DT(u))} \qquad \text{(IV.4.15)}$$

It is not hard to see that unstable vectors u may not be expanded by DT in the Euclidean norm, if $J(u)$ is small. Indeed, when $\cos\varphi \approx 0$, and $\tau(x)$ is small, then we may have $|DT(u)|/|u| \approx 0$, i.e. the unstable vector u may shrink by an arbitrarily large factor! This is definitely an undesirable effect, again justifying our interest in the p-norm.

Invariant cones. In the above, we described the stable and unstable subspaces, E^s and E^u, at every point $x \in M'$ for Sinai billiards. This was done along the lines of original Sinai's studies of dispersing billiards. Recently, however, the invariant cone techniques became very popular in the studies of hyperbolic dynamics. We will construct invariant cones for Sinai billiards now.

Let a point $x = (q, v) \in M$ have coordinates (r, φ). Every tangent vector $u = (dr, d\varphi)$ at x can be represented by a wave front Σ passing through x and having curvature χ.

Definition. We define the *stable cone* at x as consisting of vectors such that $\chi < 0$, i.e. whose wave front Σ is convergent. The *unstable cone* consists of vectors u for which the wave front Σ *was* divergent *before* the reflection at the point x.

Hence, the unstable cone consists of vectors u such that $\chi_- > 0$, i.e. $\chi - 2K(r)/\cos\varphi > 0$, according to the mirror equation (IV.4.2).

Due to our equation (IV.4.8), the stable cone can be described by

$$\mathcal{C}^s(x) = \{u : K(r) \le d\varphi/dr < \infty\}$$

and the unstable cone is similarly described by

$$\mathcal{C}^u(x) = \{u : -\infty < d\varphi/dr \le -K(r)\}$$

Hence, in the r, φ coordinates, the unstable cone lies in the first and third quadrants (increasing directions) and the stable cone lies in the second and fourth quadrants (decreasing directions). For this reason, sometimes unstable directions are called increasing and stable – decreasing, in the studies of dispersing billiards [BSC2].

It follows immediately from our previous observations that the cone families $\mathcal{C}^u(x)$ and $\mathcal{C}^s(x)$ are invariant:

$$DT(\mathcal{C}^u(x)) \subset \mathcal{C}^u(Tx) \qquad DT^{-1}(\mathcal{C}^s(x)) \subset \mathcal{C}^s(T^{-1}x)$$

In fact, they are strictly invariant in Sinai billiards.

We note that instead of $\mathcal{C}^u(x)$ and $\mathcal{C}^s(x)$ one can fix an $m \ge 1$ and consider the cone families $DT^m(\mathcal{C}^u(x))$ and $(DT^{-m}(\mathcal{C}^s(x))$. Those families will be invariant, too, and for large m the cones can be made arbitrarily narrow. And, of course, the angle between these cones will be positive for any $m \ge 1$.

Stable and unstable manifolds. In dispersing billiards, the map $T : M \to M$ is uniformly hyperbolic in the sense of Chapter III, but not smooth. This last feature makes a sharp distinction between billiards and Anosov diffeomorphisms. The map T is discontinuous at points $x \in M$ such that either $x \in S_0$ or $Tx \in S_0$, see Lemma IV.1.1. It is customary to denote the discontinuity set by $S_{-1} = S_0 \cup T^{-1}S_0$.

Recall that $M = \partial Q \times [-\pi/2, \pi/2]$ is a two-dimensional manifold consisting of smooth pieces $M_k := \Gamma_k \times [-\pi/2, \pi/2]$ for $1 \le k \le s$, see (IV.1.1). It is easy to see that $\partial M = \cup_{k=1}^{s} \partial M_k = S_0$, i.e. the set S_0 makes the narural boundary of the manifold M.

The set $T^{-1}S_0$ has a more complicated structure. Without going into detail, we claim that $T^{-1}S_0$ is a union of smooth compact curves in M. (For billiards on the plane the number of curves is finite, but for billiards on a torus, like the Lorentz gas, the number of curves may be infinite.)

Remark. It is important that all the curves in $T^{-1}S_0$ are decreasing, i.e. are given by equations $\varphi = f(r)$ with some $f'(r) < 0$. They correspond to convergent wave fronts. To see this, draw a dispersing billiard table Q whose boundary contains a corner point and a family of rays starting out on ∂Q and converging at the corner point.

Similarly, the map T^n, $n > 0$, has singularities on the set

$$S_{-n} := S_0 \cup T^{-1}S_0 \cup \cdots \cup T^{-n}S_0$$

which also is a union of decreasing curves in M. The map T^{-n} for $n > 0$ has singularities on the set S_n defined similarly and consisting of increasing curves.

In Section III.3 we introduced the general class of smooth maps with singularities. The billiard map $T : M \to M$ always belongs to this class, in particular satisfies the technical requirements by Katok and Strelcyn (KS1), (KS2) and (KS3). In fact, the entire class was introduced by these authors in [KS] with the main goal to cover billiard maps. The verification of all technical conditions is not an easy task, though. It is done for general billiards by Katok and Strelcyn in their fundamental book.

It now follows by Theorem III.3.2 that stable and unstable manifolds for the map T exist at ν-almost every point $x \in \Sigma(T)$. For dispersing billiards, $\nu(\Sigma(T)) = 1$, hence stable and unstable manifolds exist almost everywhere:

Theorem IV.4.2 *In dispersing billiards, for ν-almost every point $x \in M$ there is a stable manifold $W^s(x)$ and an unstable manifold $W^u(x)$ through x.*

Those manifolds are one-dimensional curves in M. The curve $W^s(x)$ corresponds to a convergent wave front (which, moreover, stays convergent at all times). The curve $W^u(x)$ corresponds to a divergent wave front (which, moreover, stays divergent at all times).

Each stable manifold $W^s(x)$ and each unstable manifold $W^u(x)$ is a smooth compact curve of finite length. Its length is determined by the iterations of x and how closely the images of x come to the singularity set S_0, see the discussion after Theorem III.3.2 . It is also true that the images of the endpoints of $W^s(x)$ and $W^u(x)$ hit the singular set S_0.

Entropy. Other theorems stated in Section III.3 apply to billiards as well. In particular, Pesin formula (Theorem III.3.4) expresses the entropy of the billiard map T:

$$h_\nu(T) = \int_M \lambda_+(x)\, d\nu(x) \tag{IV.4.16}$$

where $\lambda_+(x)$ is the (only) positive Lyapunov exponent of the map T at x.

This formula has a definite theoretical value but practically is not always convenient, since the Lyapunov exponent is defined by (III.1.3) in Chapter III as a limit of an expression involving all iterates of the map T. There is a remarkable simplification of the equation (IV.4.16).

For each point $x \in M$ let

$$D^u(x) = ||(DT)_x(v)||/||v||, \qquad v \in E^u_x$$

be the expansion factor of unstable vectors $v \in E^u_x$. Then, by the chain rule, the definition (III.1.3) in Chapter III can be rewritten as

$$\lambda_+(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \log D^u(T^i x) \tag{IV.4.17}$$

Now combining (IV.4.16), (IV.4.17) and Birkhoff Ergodic Theorem gives

$$h_\nu(T) = \int_M \log D^u(x)\, d\nu(x) \tag{IV.4.18}$$

The integral here does not depend on the metric used (see Exercise IV.4.6), hence we can define $D^u(x)$ in the p-norm by (IV.4.11). This leads to the following important

Theorem IV.4.3 *In dispersing billiards, the entropy is given by*

$$h_\nu(T) = \int_M \log|1 + \tau(x)\kappa^u(x)|\, d\nu(x) \tag{IV.4.19}$$

This theorem can be generalized to any billiard system in any dimension, see [C5].

In dynamical systems, one can also define the entropy of a flow Φ^t. It is defined to be the entropy of the map Φ^1 (which is obtained by setting $t = 1$; the map Φ^1 is called "the time one map"), so

$$h_\mu(\{\Phi^t\}) := h_\mu(\Phi^1)$$

This definition is based on the standard fact that for any real $t \in \mathbb{R}$ we have $h_\mu(\Phi^t) = |t| h_\mu(\Phi^1)$, so $h(\Phi^t)$ is sufficient to determine $h(\Phi^t)$ for all t.

There is a relatively simple formula by Abramov [Ab] that relates the entropy of the flow Φ^t to that of the map T:

$$h_\mu(\{\Phi^t\}) = h_\nu(T)/\bar{\tau}$$

where $\bar{\tau}$ is the mean free path, see (IV.2.9) and (IV.2.10). Therefore, we obtain

Theorem IV.4.4 *In dispersing billiards,*

$$h_\mu(\{\Phi^t\}) = h_\nu(T) c_\mu c_\nu^{-1} = h_\nu(T)\, |\partial Q|\, (\pi|Q|)^{-1} \tag{IV.4.20}$$

The above formulas for the entropy were essentially obtained by Sinai in 1970 [Si2]. They were generalized to any dimensions by himself in 1979 [Si5]. For a more recent exposition and extensions to other classes of billiards see [C5].

One can employ Eqs (IV.4.19) and (IV.4.20) to compute or estimate the entropy of many physically interesting models, such as Lorentz gases and hard ball systems. Some results were obtained in [C5, CM]. In particular, for the d-dimensional Lorentz gas ($d \geq 2$) with a single spherical obstacle of a small radius r in the unit torus, the entropies are given by

$$h_\nu(T) = -d(d-1)\log r + O(1)$$

$$h_\mu(\{\Phi^t\}) = -d(d-1)\, |B^{d-1}|\, r^{d-1} \log r + O(r^{d-1})$$

as $r \to 0$. We note that the entropy of T grows to infinity, while that of Φ^t vanishes. The same holds for the Lyapunov exponents of T and Φ^t, respectively. We encourage the reader to find a qualitative explanation why these Lyapunov exponents behave so differently as $r \to 0$.

Exercises:

IV.4.1. Prove the equation (IV.4.2). Hint: fix the sign of K and χ_-; for example assume that $K > 0$ and $\chi_- > 0$, i.e. the fronts are divergent and the boundary is dispersing.

IV.4.2. Prove the convergence of the continued fraction (IV.4.4) for dispersing billiards. Hint: Consider the truncated fractions at even and odd components, call them A_{2n} and A_{2n+1}, respectively. Prove that $A_{2n+2} < A_{2n+4} < A_{2n+3} < A_{2n+1}$ for every $n \geq 0$. Also observe that the sum of all the odd terms is infinite. (The convergence of continued fractions with positive elements is known as Seidel-Stern theorem.)

IV.4.3. Prove the equation (IV.4.8). Hint: the curve Σ can be approximated by a an arc of a circle. Then if it is parameterized by arclength s, then difference between normal vectors $n(s)$ and $n(s+ds)$ has length $\kappa(s)\,ds + o(ds)$ where $\kappa(s)$ is the curvature at the point s.

IV.4.4. Prove that for the billiard in a circle of radius R the expression (IV.4.6) converges to $1/\cos\varphi$. (Recall that the Lyapunov exponents vanish!) Hint: $a_{2k+1}(x) = \frac{2}{R\cos\varphi}$, $a_{2k}(x) = 2R\cos\varphi$.

IV.4.5. In semidispersing billiards, some elements of the continued fraction (IV.4.4) are zero. Precisely, $b_{2k} = 0$ whenever the point $T^k x$ lies on a neutral component of the boundary. Show that of $b_{2k} = 0$, then we can rewrite (IV.4.4) by skipping the k-th reflection and adding the free paths $\tau(T^{k-1}x) + \tau(T^k x)$. Therefore, one can just ignore reflections at neutral components in (IV.4.4) altogether.

IV.4.6. Complete the details in the Proof of Theorem IV.4.3. In particular, verify that the integral in (IV.4.18) does not depend on the metric. Hint: use formula (IV.4.15) and the invariance of the measure ν.

IV.5 Other Hyperbolic Billiards

Since Sinai's discovery of dispersing billiards in 1970, they remained, for several years, the only known class of chaotic billiards. Moreover, in 1973 Lazutkin proved that billiards in generic convex domains had caustics, hence could not be chaotic, cf. Sect. IV.3. It seemed therefore that chaoticity in billiards could only be produced by concave (dispersing) boundaries. Or at least, if chaotic billiards with focusing boundary existed, they would have to have a very special boundary, because in simple convex domains – circles and ellipses – billiards are completely integrable (which is almost opposite to chaotic).

It came then as a big surprise when in mid-seventies Sinai's student L. Bunimovich constructed chaotic billiards whose boundaries were focusing and consisted of simple circular arcs!

Bunimovich billiard tables. Bunimovich's construction is based on the phenomenon of defocusing, which we describe below. Recall that in Sinai's billiards a neutral (parallel) wave front becomes divergent after colliding with the boundary, and during free runs between collisions it stretches out (grows in size). The factor of growth is given by

$$\Lambda = |1 + \tau\chi| \tag{IV.5.1}$$

where τ is the time of the free run between collisions and $\chi > 0$ is the curvature of the front at the beginning of the free run, see (IV.4.11).

Now, consider a billiard table with focusing boundary. A parallel wave front colliding with a focusing component of the boundary will become convergent, i.e. $\chi < 0$. Then it will certainly shrink in size as it moves back into Q. But it can focus (converge to a point inside Q) and then emerge from the focusing point as a divergent wave front! This transformation of convergent fronts into divergent ones is called *defocusing*. Then our front will travel further and grow in size. When it reaches the boundary ∂Q and gets reflected again, its size may be even larger than the size of the original front, see Fig. IV.11. The factor of growth is still given by (IV.5.1). It is true that $\chi < 0$, hence Λ need not be greater than one. Nonetheless, we have $\Lambda > 1$ under an additional condition $\tau\chi < -2$, i.e. when

$$1/|\chi| < \tau/2 \tag{IV.5.2}$$

It is clear that an infinitesimal convergent wave front whose curvature is $\chi < 0$ will focus (converge to a point) at time $1/|\chi|$. Hence, the above condition (IV.5.2) means that the wave front focuses before it reaches the midpoint between collisions. And then $\Lambda > 1$, so our wave front will grow before the following collision occurs. This is the key condition for the hyperbolicity of the map T.

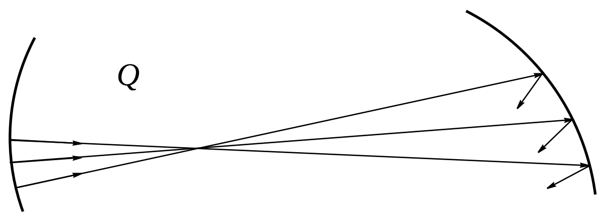

Figure IV.11: Defocusing of a convergent front.

How can we ensure (IV.5.2) when constructing a billiard table with focusing boundary? If $\Gamma_i \subset \partial Q$ is a focusing component of the boundary, then we do not want to position other components of ∂Q too close to Γ_i. Indeed, wave fronts leaving Γ_i need to travel freely and defocus before they hit ∂Q again, and they need enough room to grow after that. This explains the requirement $K_i \subset Q$ in the following definition by Bunimovich.

Definition. Let Q be a billiard table whose boundary consists of dispersing, neutral and focusing components. Each focusing component Γ_i is an arc of a circle, call that circle K_i and its radius R_i. Assume that $K_i \subset Q$. Now, a tangent vector $u \in \mathcal{T}_x M$ at a point $x \in M$ represented by a wave front Σ is called *unstable* if either
(i) x belongs to a dispersing or neutral component of ∂Q and Σ is divergent, or
(ii) x belongs to a focusing component $\Gamma_i \subset \partial Q$ and Σ is convergent, and its curvature satisfies

$$\chi < -(R_i \cos\varphi)^{-1} \qquad \text{(IV.5.3)}$$

One can easily see that (IV.5.3) is equivalent to (IV.5.2). The following is the key theorem by Bunimovich.

Theorem IV.5.1 ([Bu1], [Bu2]) *Let Q be a billiard table define above. If u is an unstable tangent vector, then so is $DT(u)$.*

Proof. See Exercise IV.5.1. □

Therefore, the unstable tangent vectors make a cone field $\mathcal{C}^u(x)$ invariant under DT. This indicates that the methods of Section III.4 apply and give the hyperbolicity. However, the invariant cone field does not imply hyperbolicity yet. Indeed, the above theorem, technically, holds in circular billiards, which are not at all hyperbolic. So, we may have an invariant cone family but no hyperbolicity. The methods of Section III.4 require that the cone field is strictly invariant, at least eventually. In the Bunimovich billiards defined above, the strict invariance of the cones, i.e. the property

$$DT(\overline{\mathcal{C}^u(x)}) \subset \operatorname{int} \mathcal{C}^u(Tx) \tag{IV.5.4}$$

holds in two cases: (a) the point x belongs to a dispersing component of the boundary or (b) the point x belongs to a focusing component Γ_i, but its image Tx belongs to any component Γ_j other than Γ_i. These facts can be verified by direct inspection going slightly beyond the proof of Theorem IV.5.1.

It must be checked now that almost every trajectory on a Bunimovich billiard table does contain points of the above types (a) or (b). If it does, we have hyperbolicity. A simple geometric analysis shows that there might be two kinds of trajectories that fail to contain such points. We describe them below:

A. These are trajectories which only reflect at the same focusing component Γ_i.

B. These are trajectories which only reflect at neutral components of ∂Q.

One can easily show that type A trajectories are periodic and their total measure is zero (one needs to remember that Γ_i is only an arc of a circle, not an entire circle). Concerning type B, in particular examples it might be easy to check that such trajectories make a set of zero measure. Such are billiard tables that have one (or none) neutral component of ∂Q, or two parallel neutral components, etc. However, in general it remains an open (and challenging) problem to prove that the type B trajectories make a null set, see some partial results in [CT2]. The following is a theorem proved by Bunimovich:

Theorem IV.5.2 ([Bu1], [Bu2]) *Let Q be a billiard table define above. Assume that the type B trajectories make a set of zero measure. Then the billiard map T is hyperbolic.*

The stadium mentioned in the end of Section IV.3 is, probably, the most celebrated Bunimovich billiard table. One can construct other nice examples, such as a "flower table" - the union of several circles, of which one is large and placed in the center, and the others are small and placed all around the central one overlapping with it.

Taking a circle, cutting it along a chord, and removing the smaller part gives another simple example of a Bunimovich billiard table. This one does not technically satisfy his requirements, but this can be easily remedied. Indeed, let us reflect the table along the chord that makes a part of its boundary. Now the union of the original table and its mirror image across the chord will make a perfect Bunimovich table. It is easy to check that the hyperbolicity and ergodicity of the new, larger table implies those of the original, smaller one.

Absolutely focusing arcs. Since Bunimovich's discovery of the defocusing mechanism many mathematicians were trying to construct other classes of chaotic billiards with focusing components of the boundary. It was not easy to do, because the circle was too rigid a figure whose properties played essential role in Bunimovich's calculations.

Only in mid-eighties Wojtkowski [W2] and then Markarian [Ma1] constructed new classes of hyperbolic billiards with focusing boundaries other than circular arcs. Later V. Donnay [Do1] and Bunimovich [Bu6] summarized those constructions and presented a unified theory of what they called absolutely focusing arcs.

Our discussion of this theory combines analytic machinery of quadratic forms (Theorem III.4.1) with geometric arguments referring to the evolution of wave fronts (Section IV.4).

Let $\Gamma \subset \mathbb{R}^2$ be a C^r smooth curve with endpoints A_1 and A_2. Suppose its curvature has constant sign and never vanishes. Denote by $\alpha(\Gamma)$ the angle by which the tangent line to Γ turns as one travels from A_1 to A_2 along Γ. We will assume that Γ is a focusing component of the boundary of a billiard table Q, i.e. Γ is convex from inside of Q.

Definition. A curve Γ is an *absolutely focusing arc* if
(AF1) $\alpha(\Gamma) \le \pi$, and
(AF2) any infinitesimal parallel wave front falling upon Γ from Q will focus before the next collision, and if that collision occurs with Γ again, then it will focus before the following collision, and so on, until it hits a component of ∂Q other than Γ.

As we said, this type of curves were introduced and studied carefully by Donnay and Bunimovich, see [Do1], [Bu6]. The importance of the requirement (AF2) was demonstrated above – the wave front grows after passing through a focusing point, which is a key to eventual expansion.

Quadratic forms. We now develop a general machinery that applies to practically any billiard and allows us to verify hyperbolicity. Many recent studies of chaotic billiards start with theorems on hyperbolicity (non-vanishing of Lyapunov exponents). All such theorems are proved either by construction of invariant cones or by using increasing quadratic forms. As we have seen in Chapter III, both methods are essentially equivalent. We use the second method (quadratic form) in a rather general context of planar billiards.

Let $x = (q, v) \in M$ be a point with coordinates (r, φ) and $(dr, d\varphi)$ a tangent vector. We introduce a new coordinate system in the tangent space $\mathcal{T}_x M$, which is directly related to α and h used in Fig. IV.2 in Section IV.1. The new coordinates (U, V) are defined with the help of the orthogonal projection of $(dr, d\varphi)$ on the subspace perpendicular to v:

$$U = \cos\varphi\, dr \qquad V = K dr + d\varphi$$

Now, if $D_x T(U, V) = (U_1, V_1) \in \mathcal{T}_{Tx} M$ is the image of our vector (U, V), with $U_1 = \cos\varphi_1\, dr_1$ and $V_1 = K_1 dr_1 + d\varphi_1$, then a simple computation gives that

$$U = -(\tau K_1 + \cos\varphi_1) dr_1 + \tau d\varphi_1 \qquad V = K_1 dr_1 - d\varphi_1$$

where all the symbols have their usual meaning.

We now define a measurable nondegenerate quadratic form

$$B_x(U, V) = aU^2 + 2bUV + cV^2$$

where a, b, c are measurable functions of x (we will make them piecewise continuous below) and $ac - b^2 \neq 0$. The coefficients a, b, c should satisfy the requirement (iii) of Theorem III.4.1, i.e. the quadratic form must be increasing, $P_x(U, V) \geq 0$, where

$$\begin{aligned} P_x(U,V) &= (T^{\#}B - B)_x(U,V) \\ &= a_1 U_1^2 + 2b_1 U_1 V_1 + c_1 V_1^1 - (aU^2 + 2bUV + cV^2) \end{aligned}$$

Denote $\varphi' = d\varphi_1/dr_1$, then

$$\begin{aligned} P_x(U,V) &= (dr_1)^2[(a_1 - a)\cos^2\varphi_1 + 2(b_1 - b)(K_1 + \varphi')(\cos\varphi_1) \\ &\quad +(c_1 - c)(K_1 + \varphi')^2 + a(2\tau\cos\varphi_1(K_1 + \varphi') - \tau^2(K_1 - \varphi')^2) \\ &\quad +2b(-2K_1\cos\varphi_1 + \tau(K_1 - \varphi')^2) + c4K_1\varphi'] \end{aligned}$$

Then the sign of P depends on the following polynomial of degree two in φ':

$$\begin{aligned}&(c_1 - c - a\tau^2 + 2b\tau)(\varphi')^2 +\\ &+2[(b_1 - b)\cos\varphi_1 + K_1(c_1 - c) + a\tau\cos\varphi + a\tau^2 K_1 - 2b\tau K_1 + 2cK_1]\varphi' +\\ &+(a_1 - a)\cos^2\varphi_1 + 2(b_1 - b)K_1\cos\varphi_1 + (c_1 - c)K_1^2 -\\ &-2a\tau K_1\cos\varphi_1 - a\tau^2 K_1^2 + 4bK_1\cos\varphi_1 + 2b\tau K_1^2 \qquad \text{(IV.5.5)}\end{aligned}$$

We set $b = b_1 = 1$ in order to have an easy control on the nondegeneration of B and to simplify our computations. Then $P > 0$ for every φ' iff

$$c_1 - c - a\tau^2 + 2\tau > 0, \qquad \text{(IV.5.6)}$$

and

$$\begin{aligned}(c - 2\tau \quad + \quad & a\tau^2)[4c_1 - 2E + (a_1 - a)E^2/4]\\ + \quad & c_1 E\,[2 - 2a\tau - (a_1 - a)E/4] + a^2\tau^2 E^2/4 < 0 \quad \text{(IV.5.7)}\end{aligned}$$

if $K_1 \neq 0$, where $E = -2K_1^{-1}\cos\varphi$. If the component where q_1 belongs is focusing $(K_1 < 0)$, then $E = L_1 = 2R_1\cos\varphi_1$ is the time the trajectory arriving at (or departing from) q_1 spends in the circle of curvature of ∂Q at q_1. In the case $K_1 = 0$ the condition (IV.5.7) becomes

$$(a_1 - a)(c - c_1 - 2\tau) + aa_1\tau^2 < 0 \qquad \text{(IV.5.8)}$$

Remark. We note that our formalism gives another proof of the fact that was mentioned already: a semidispersing billiard is hyperbolic if the set of points whose trajectories only hit neutral components of the boundary has measure zero. Indeed, if we take $a \equiv c \equiv 0$, then (IV.5.7) is satisfied for dispersing boundaries $(K > 0)$. On neutral components we consider the same quadratic form. Then $P_x \geq 0$ (it is actually positive on $\mathcal{T}_x M$, with the exception of one direction). Also, P becomes strictly positive whenever the trajectory hits a dispersing component.

We study now local conditions which a focusing arc Γ must satisfy in order to be suitable for the construction of a hyperbolic billiard. The following heuristic observations are very important. We look at the behavior of expressions (IV.5.6) and (IV.5.7) during series of successive reflections at Γ, with $\varphi \cong \pm\pi/2$. We have that $L \cong \tau \cong 0$, and a, c are continuous in a neighborhood of the points $(q_1, \pm\pi/2)$. Then, if $c \gg \tau$, the discriminant

Δ of (IV.5.5) satisfies : $\Delta \cong 4cc_1 \cong 4c_1^2$. As we must have $\Delta < 0$, we deduce that $\lim_{\varphi \to \pm\pi/2} c = 0$. If $c \ll \tau$, we have $\Delta \cong 4\tau E \cong 4\tau^2$. In conclusion, we must have $c \cong \tau$, this justifies the choice of $c = \tau$ (or L).

If $a \equiv 0$, $c = L$, and $c_1 = L_1$, then an immediate computation of (IV.5.7) gives

$$L + L_1 < 2\tau \tag{IV.5.9}$$

Wojtkowski [W2] proved that curves that verify this condition are absolutely focusing. The proof is based on the fact that an incoming wave front falling onto ∂Q at the same angle focuses at distance $L/2$ along the trajectory departing from the point $x = (q, v)$. See Exercise IV.5.2.a. If the component of the boundary is C^4 smooth, the condition (IV.5.9) is equivalent to

$$\frac{d^2 R}{ds^2} < 0 \tag{IV.5.10}$$

where s is the arc length of the focusing arc and $R = -1/K$ is the radius of its curvature. See Exercise IV.5.2.b.

If $a \equiv 0$, $c = \tau$, and $c_1 = \tau_1$, then we obtain the condition

$$L_1(\tau + \tau_1) < 2\tau\tau_1 \tag{IV.5.11}$$

Curves that satisfy (IV.5.11) are also absolutely focusing (see Theorem IV.5.5 below). The above condition is equivalent (see Exercise IV.5.3) to

$$\frac{d^2 R^{1/3}}{ds^2} > 0 \tag{IV.5.12}$$

It is interesting that the conditions (IV.5.10) and (IV.5.12) are seemingly opposite, but both define absolutely focusing arcs.

We will say that a smooth curve can be a part of a hyperbolic billiard if it is possible to define a quadratic form that is increasing for trajectories experiencing repeated collisions with that curve (this definition can also be expressed in terms of invariant cone fields). Now we sketch a proof of the following surprising result. It defines *short focusing arcs*:

Theorem IV.5.3 ([Ma6]) *Any sufficiently small C^4 focusing arc can be a regular component of the boundary of a hyperbolic billiard.*

Proof. Let $q_1 + r(A)e^{iA}q_1'$ be the parametrization of a C^k curve ($k \geq 4$) in polar coordinates, where A is the polar angle and $r(0) = 0$. Let us expand $r(A)$ into a Taylor series $r(A) = \dot{r}A + \frac{\ddot{r}}{2}A^2 + \cdots$ so that $\dot{r} = \dot{r}(0) = -2/K_1 > 0$.

If the coefficients a and c are differentiable with respect to s, B (here B is the angle between the oriented tangent line and the inward direction) then they may be also expanded as

$$\begin{aligned} c(s,B) &= C_1 s + C_2 B + C_3 s^2 + C_4 sB + C_5 B^2 \\ &\quad + C_6 s^3 + C_7 s^2 B + C_8 sB^2 + C_9 B^3 + \cdots \end{aligned}$$

$$a(s,B) = A_0 + A_1 s + A_2 B + A_3 s^2 + \cdots$$

At the point q_1 we have $s = 0$ and $B = A$. At the point q we have $s = s(-A)$ and $c_1 = (-C_1 + 1)\dot{r}A + \left(C_1 \frac{\ddot{r}}{2} + \frac{\ddot{r}}{2} + \dot{r}^2 C_3 - \dot{r}C_4 + C_5\right) A^2$. If we substitute these expressions and the expansion of τ and L in (IV.5.6) and (IV.5.7), we obtain that (IV.5.6) is immediately verified and that the left hand side of (IV.5.7) is $\zeta A^3 + \sigma A^4 + \cdots$ with

$$\zeta = 3\dot{r}\ddot{r} + 2\dot{r}^3 C_3 - 2\dot{r}^2 C_4$$

$$\begin{aligned} \sigma = \ddot{r}^2 - \tfrac{4}{3}\dot{r}(\dot{r} + \ddot{r}) \quad &+ \quad 2\ddot{r}C_5 + 4C_5^2 + 2\dot{r}^2(-\ddot{r}C_3 - \dot{r}^2 C_6 + \dot{r}C_7 - C_8) \\ &+ \quad 2\dot{r}^2 A_0 C_5 - \dot{r}^2\ddot{r}A_0 + \tfrac{\dot{r}^4}{4}(A_0^2 - 2A_1) \end{aligned} \tag{IV.5.13}$$

Given any C^4 small focusing arc (that is, given $\dot{r} > 0$ in our polar coordinates), we can choose C_i, A_i conveniently, so that $\zeta = 0$ and $\sigma < 0$. Then (IV.5.7) will be satisfied. For example, we can take C_6 in such a way that $-2\dot{r}^4 C_6$ dominates all the other terms (all details are in [Ma6]). $\square$

General hyperbolic billiard tables. We have described classes of focusing arcs that *could be* used in the construction of hyperbolic billiards. This does not mean, of course, that any billiard table made of these arcs is hyperbolic. Here we describe how to assemble boundary pieces of various kind in order to make a hyperbolic billiard table. Certain simple rules ("building code") must be followed. We need to verify conditions on interior angles made by adjacent components, distanced between different components, etc. The following theorem shows how to construct plane billiards with hyperbolic behaviour whose boundary components are of any type with the only requirement that focusing components satisfy condition (IV.5.9) or

(IV.5.11). We remark that case (iii) below includes all Bunimovich billiard tables.

Theorem IV.5.4 *Hyperbolic billiards can be constructed in the following ways:*
(i) C^3 *components of the boundary can be of any type, as long as the focusing ones satisfy (IV.5.9) (for* C^4 *curves (IV.5.9) can be replaced by (IV.5.10)). The circles of semicurvature at each point of every focusing component must not contain points of other components or circles of semicurvature of other focusing components. Adjacent focusing components must make interior angles greater than* π*. Focusing and dispersing adjacent components must make interior angles not smaller than* π*. Focusing and neutral adjacent components must make interior angles greater than* $\pi/2$.
(ii) C^3 *components of the boundary can be of any type as long as the focusing components satisfy (IV.5.11) (for* C^4 *curves (IV.5.11) can be replaced by (IV.5.12)). The circles of curvature of focusing components must not contain points of other boundary components. Conditions on adjacent arcs are the same as in* (i).
(iii) C^3 *components of the boundary are as in* (i) *or* (ii)*, but equalities are allowed in (IV.5.9), (IV.5.11). Almost every trajectory must bounce off at least two distinct components of the boundary.*

Proof. We prove (ii). The proof of (i) follows the same ideas. We will define convenient quadratic forms with $a = 0$, $b = 1$ in all cases, and set $c = \tau$ if the reflection occurs at a focusing component and $c = 0$ otherwise.

Several cases arise. If both q and q_1 lie on dispersing or neutral components of the boundary, the argument in the remark made in the previous subsection works. If q and q_1 lie on the same focusing component, condition (IV.5.11) is satisfied and then the quadratic form is increasing. If q and q_1 belong in different focusing components and q_2 is not in the intersection of a neutral component with the circle of curvature at q_1, then $\tau, \tau_1 > L_1$. Let $\tau \geq \tau_1$, then $L_1(\tau + \tau_1) < \tau_1(\tau + \tau) = 2\tau\tau_1$, and the condition (IV.5.11) is satisfied. If q, q_1 lie on focusing components and q_2 belongs in an adjacent neutral component, we observe that $\tau_1 + \tau_2 > L_1$ since q_3 is not in the circle of curvature of q_1. Moreover, if q_2 lies on a flat component, then $D_{x_2}T \cdot D_{x_1}T$ acts as if the point x_2 is skipped and the intercollision times are added, see Exercise IV.4.5. Finally, if q_1 is in a neutral component contained in the circle of curvature at q_2, the analysis is similar to the previous one, taking into account that $\tau + \tau_1 > L_2$.

The part (iii) is proved using the same quadratic forms, which in this case will be increasing eventually (see Remark after Theorem III.4.1.). $\square$

The above theorem remains valid if in both parts (i) and (ii) we also allow short focusing arcs defined in Theorem IV.5.3. We sketch the proof in the case when the focusing curves are either short focusing arcs or satisfy the condition (IV.5.9). By C we denote a short focusing arc. The details are left to the (interested) reader. We define quadratic forms with $a = 0$:

(i) If $q_1, q_2 \in C$ and $q_3 \notin C$ we define $c_2 = L_2$. Then (IV.5.6) is immediately satisfied since the first term, in linear approximation, is equal to $2\dot{r}A$. Now (IV.5.7) is satisfied if (IV.5.13) holds with $C_5 = 0 = A_0 = A_1$, it is sufficient to take C_0 big enough.

(ii) If $q_1 \in C$ and $q_2 \notin C$, we consider initially only the condition $c_2 \geq 0$. In the first term of (IV.5.6) we have $c_2 - L_1 + 2\tau_1 > 2\tau_1 - L_1 > 0$ if $2\tau_1 > L_1$; this means that the component where q_2 belongs must be outside of the circle of semicurvature of any point of C. If $K_2 = 0$, the first term of (IV.5.8) turns zero. So, in order to maintain the form B increasing along the trajectories, they must eventually hit non-neutral components.
If $K_2 \neq 0$, the first term of (IV.5.7) is $(L_1 - 2\tau_1)(4c_2 - 2E) + 2c_2E$.
If $K_2 > 0$, then $E < 0$ and it is enough again to consider $2\tau_1 > L_1$.
If $K_2 < 0$, we define $c_2 = L_2$ and then the previous expression becomes $2L_2(L_1 + L_2 - 2\tau_1)$ which is negative if $L_1 + L_2 < 2\tau_1$. This means that circles of semicurvature of focusing components must not intersect themselves.

(iii) If $q_1, q_3 \notin C$, $q_2 \in C$, then let $c_1 \geq 0$, $c_2 = L_2$. The first term of (IV.5.6) is $= L_2 - c_1 + 2\tau_1 > 2\tau_1 - c_1$; so (IV.5.6) is satisfied if $2\tau_1 \geq c_1$. The first term in (IV.5.7) is $(c_1 - 2\tau_1)2L_2 + 2L_2^2 = 2L_2(c_1 + L_2 - 2\tau_1)$. If the component where q_1 belongs is dispersing or neutral, we define $c_1 = 0$; if it is focusing, we define $c_1 = L_1$. Now (IV.5.7) holds if the arcs satisfy the conditions that appear in part (ii).

(iv) If $q_1 \notin C$ and $q_2, q_3 \in C$, the verification of (IV.5.6) does not need new conditions and, since $c_2 \simeq L_2$, the first term of (IV.5.7) is $(c_1 - 2\tau_1)(4c_2 - 2L_2) + 2c_2L_2 \simeq 2c_2(c_1 + L_2 - 2\tau_1)$, which is negative under the same conditions as those found in (iii).

(v) If a segment of a trajectory runs between two components that are not short focusing arcs, the quadratic form is defined with $c = L$ if q is in a focusing component and with $c = 0$ otherwise.

A similar argument works if the focusing boundaries satisfy (IV.5.11).

See [Ma6].

Invariant cones and continued fractions. It was observed in Section III.4 that a nongenederate quadratic form B automatically defines a field of cones in $\mathcal{T}M$ consisting of vectors whose B-values are nonnegative. If the quadratic form is increasing under the action of DT, these cones are invariant under T. Given an increasing nondegenerate quadratic form B we define

$$C^u(x) = \{v \in \mathcal{T}_x M : \ B_x v \geq 0\} \qquad \text{(IV.5.14)}$$

This field of cones satisfies the invariance property $D_x T(C^u(x)) \subset C^u(Tx)$.

Cone fields can be defined, of course, without quadratic forms. In fact it is sometimes easier to construct an invariant cone field by using direct geometrical or "optical" properties of the system. This was exactly the case with dispersing billiards in Section IV.4.

Such constructions have been successfully used to prove hyperbolicity and ergodicity of billiards. In all the cases some version of Theorems III.4.3 and III.4.4 of Chapter III were used. Those methods are very well explained in papers by Wojtkowski [W2], Donnay [Do1] and Bunimovich [Bu6]. They include detailed studies of defocusing properties of wave fronts bouncing off focusing boundaries of a billiard table.

The hyperbolicity of billiards with focusing boundaries satisfying condition (IV.5.9) was first proved using these geometric cone techniques. Some other convex billiard tables (not satisfying any of the previously studied conditions) have been proved to be hyperbolic using other cone techniques. See examples at the end of this Section.

A natural question arises for hyperbolic billiards with focusing components of the boundary. Is it possible to find expressions for the stable and unstable tangent vectors, similar to continued fractions of Section IV.4? The following theorem gives a positive answer.

Theorem IV.5.5 *Consider billiard tables satisfying the conditions of Theorem IV.5.4. For ν-almost every point the continued fractions (IV.4.4) and (IV.4.6) converge and give the slopes of the lines E^s_x, E^u_x according to (IV.4.8).*

Proof. We begin by recalling an interesting theorem on the convergence of continued fractions whose elements may be both negative and positive

([Bu6], Theorems 10 and 11). A continued fraction

$$a_1(x) + \cfrac{1}{a_2(x) + \cfrac{1}{a_3(x) + \cfrac{1}{a_4(x) + \cfrac{1}{a_5(x) + \cdots}}}} \tag{IV.5.15}$$

is convergent if the following two conditions are satisfied:

(a) all even elements are positive and their sum is infinite,

(b) for any two negative elements $a_{2k'-1}, a_{2k''+1}$, $k' < k''$, for which there are no other negative elements in between, we have

$$a_{2k'} \geq |a_{2k'-1}|^{-1}(2 + \delta_{k'}) + |a_{2k''+1}|^{-1}\left(2 - \frac{\delta_{k''}}{1 + \delta_{k''}}\right) \tag{IV.5.16}$$

for some $\delta_l \geq -1$, $l = 1, 2, \cdots$.

We prove Theorem IV.5.5 in the unstable case. We observe that conditions (IV.5.9) and (IV.5.11) are similar in terms of the components of $k^u(x)$ since they can be rewritten, respectively, as

$$a_{2k} \geq |a_{2k-1}|^{-1} + 2|a_{2k+1}|^{-1} \tag{W}$$

$$|a_{2k+1}| \geq 2a_{2k}^{-1} + 2a_{2k+2}^{-1}, \quad k = 1, 2, \cdots \tag{M}$$

So, if the convergence is proved assuming, for example, (IV.5.9), the convergence assuming (IV.5.11) will automatically follow, because the latter correspond to a shift by one position in the continued fraction. Condition (W) for two consecutive bounces in the same focusing component is exactly (IV.5.16) for $k' = k''$, with $\delta_{k'} = 0$. If $T^{-k'+1}x$ and $T^{-k''}x$ are successive bounces in (non-adjacent) focusing components, then conditions on non-adjacent components in Theorem IV.5.4 directly imply (IV.5.16) with $\delta_{k'} = \delta_{k''} = 0$. Now, the fact that the fractions (IV.4.4) and (IV.4.6) give the stable and unstable tangent vectors follows from our construction of stable and unstable fronts. □

We note that hyperbolic billiards with focusing boundaries have positive entropy, which can be computed by the same formulas as in the case of dispersing billiards. In particular, (IV.4.16), (IV.4.19) and (IV.4.20) remain valid, see [CM] and [C5] for a more detailed account on this issue.

Lastly, we prove that curves that satisfy (M) are absolutely focusing arcs. Recall that the convergence of $k^u(x)$ implies the convergence of

$$k^-(x) = \cfrac{1}{\tau(T^{-1}x) + \cfrac{1}{\cfrac{2K(T^{-1}x)}{\cos\varphi(T^{-1}x)} + \cdots}} \tag{IV.5.17}$$

which is the curvature of the unstable manifold before the reflection at x (cf. the mirror equation (IV.4.2)). As curves that satisfy (W) are absolutely focusing arcs – this was remarked immediately after (IV.5.9) – the sign of $k_W(x)$, i.e. the sign of the first term of $k_W(x)$, is negative, hence the sign of k_W^-, i.e. the sign of the second term in k_W, is positive. Then the sign of k_M^-, i.e. the sign of the first term of k_M^-, is positive and, finally, $k_M(x)$ (considered as a part of $k_M(Tx)$) has the sign of its third term, i.e. it is negative. As these relations are satisfied at each reflection at a focusing component, the focusing point of the unstable wave front ($k_M(x) < 0$) lies before the next reflection ($k_M^-(Tx) > 0$). If this is true for divergent (before the reflection) fronts, it must be true for parallel fronts as well. □

Examples:

1. Curves that satisfy the condition $d^2R/ds^2 < 0$. In this group are the epicycloid, the hypocycloid, the cycloid and, in particular, the cardioid (a closed curve whose equation in polar coordinates is $r(t) = 1 + \cos t$, $-\pi \le t \le \pi$). The arc of ellipse $x = a\cos t,\ y = b\sin t,\ b^2 > a^2,$ with $-\pi/4 < t < \pi/4$ also satisfies this condition. See [W2].

2. Curves that satisfy the condition $L_2(\tau_1 + \tau_2) < 2\tau_1\tau_2$. In the previously mentioned ellipse, such is the arc defined by $\sin^2 t > b^2/(b^2 + a^2)$. Curiously, this arc is disjoint from the one in part 1. See [Ma1].

3. We observe that the conditions in parts (i) and (ii) of Theorem IV.5.4 are C^3 open. Hence, the non-neutral components of the boundaries of such billiards can be perturbed in the C^3-topology and hyperbolicity will be preserved. If equalities are allowed – as in (iii) of the same Theorem – then perturbations may destroy hyperbolic behaviour of the new billiard table. This is the case when arcs of circles are part of the boundary. If circular arcs are perturbed, the conditions (IV.5.9) and (IV.5.11) may fail, and the hyperbolicity may not survive. Anyway, in [Ma6] it is proved that arcs smaller than half a circumference can be C^4 perturbed maintaining hyperbolic behaviour. A similar result for C^6 perturbations is obtained in [Do1].

4. Donnay [Do1] proved that absolutely focusing arcs can be part of hyperbolic billiards if they are joined by sufficiently long segments making a convex region. His proof is based on a subtle study of the properties of focusing wave fronts in integrable billiards. It is proved that in the above mentioned ellipse, its half corresponding to $x \geq 0$ is a focusing arc if and only if $a/b < \sqrt{2}$. See also [MOP].

5. It is important to distinguish hyperbolicity and ergodicity (or mixing property) – they are not equivalent. See Exercise IV.5.5.

Exercises:

IV.5.1. Prove Theorem IV.5.1. There are three types of a vector u to consider. First, let u and $DT(u)$ belong to the same focusing component Γ_i. Then use the mirror equation (IV.4.2) to verify that $DT(u)$ is an unstable vector. Second, let $DT(u)$ belong to a component other than Γ_i. Show that the wave front representing u arrives at that component as a divergent front. Third, if u belongs to a component other than Γ_i show, by the same argument, that the wave front arrives at Γ_i as a divergent front. In all cases use the assumption $K_j \subset Q$ for all focusing components Γ_j's that are involved in your arguments.

IV.5.2.

(a) Prove that focusing components of the boundary that satisfy the condition (IV.5.9) are absolutely focusing arcs. Hint: Consider parallel wave fronts and fronts which fall on such a curve with a constant angle of incidence. Compare the relative positions of their focusing points on the outgoing trajectory.

(b) Prove that conditions (IV.5.9) and (IV.5.10) on a focusing C^4 arc Γ are equivalent. Hint: let q_0, q_1 be distinct points of Γ, choose cartesian coordinates (x, y) on the plane in such a way that the x axis passes through q_0 and q_1. Assume that q_0 is the origin, q_1 has positive x-coordinate, and the arc q_0q_1 of Γ is below the x axis, i.e. has nonpositive y-coordinates. If $\Gamma(s) = (x(s), y(s))$ is parametrized by the arc length, let $\phi(s)$ be the angle which the tangent line to γ makes with the x-axis at point s. Note that the radius of curvature of γ is $R = ds/d\phi$. Let $\tau = |q_0q_1|$. Then

$$L = \int_a^b \frac{dx}{ds}\, ds = R(b)\sin\phi(b) - R(a)\sin\phi(a) - \int_a^b \sin\phi\, \frac{dR}{ds}\, ds$$

and

$$2\tau - L - L_1 = \int_a^b y(s)\, \frac{d^2R}{ds^2}\, ds$$

IV.5.3. Prove that conditions (IV.5.11) and (IV.5.12) on a focusing C^4 arc Γ are equivalent. Hint: see [Ma1].

IV.5.4. Complete the details in the proofs of Theorems IV.5.3 and IV.5.4.

IV.5.5. Consider an ellipse, cut it along the major axis. Move both halves apart along the minor axis. Join the two halves by lines parallel to the minor axis. Call the resulting figure Q. Insert a "wall" in Q, which is a linear segment joining the original positions of the foci. Prove that this billiard is hyperbolic but has at least three distinct ergodic components. Hint: [W2]

IV.5.6. Compute the product of matrices $D_{x_1}T \cdot D_xT$, where $x_1 = Tx$ and x_1 belongs in a neutral component of the boundary.

IV.6 Final Remarks and Outline of Perspectives

This chapter is designed to introduce the reader to the modern theory of chaotic billiard. It gives a technical and thorough introduction, after which one should be able to read main research papers in the area. But still, our chapter is no more than an introduction to the subject. It covers basics, but all major results in the theory of chaotic billiards are beyond the scope of the book.

The real difficulties appear in the study of ergodic properties of chaotic billiards. In Section III.4 we outlined a plan for the proof of ergodicity for general hyperbolic systems with singularities. It applies to dispersing billiards and some other hyperbolic billiards, but the verification of all conditions involves very serious work. There are several expository papers on the issue cited in III.4, dealing with essentially the same method in somewhat different ways.

While the machinery for the proof of ergodicity is quite well understood in two dimensions (when $\dim M = 2$), the picture is much less bright in higher dimensions. The geometry of multidimensional billiards is very complicated. The structure of singularity manifolds and the behavior of the billiard map T are yet to be described in full detail.

One of the most interesting models is the celebrated gas of hard balls, for which Boltzmann conjectured ergodicity (in some obscure words) in late 1800's. Sinai specified Boltzmann's conjecture in 1963 [Si1] by claiming that the system of N hard balls on the unit torus (in the space of any dimension $d \geq 2$) is ergodic. He obtained a proof for $N = d = 2$ (i.e., for 2 hard disks on a 2D torus) in 1970 [Si2], by reducing the system to a Lorentz gas on the plane. The case $N = 3$ turned out to be far more complicated, it was solved in 1991 by Kramli, Simanyi and Szasz [KSS3]. The conjecture for general N remains open.

Despite all difficulties, it can be safely expected that most hyperbolic billiards described in our Sections IV.2, IV.4, IV.5 are, indeed, ergodic. Hence, in this respect they should be similar to Anosov diffeomorphisms.

It turns out that chaotic billiards also have more advanced ergodic properties. As Sinai noted in his seminal 1970 paper [Si2], ergodicity of dispersing billiards almost automatically implies the mixing property and K-mixing property, see Chapter II. Later, Gallavotti and Ornstein [GO] proved the Bernoulli property, and then Chernov and Haskell [CH], Ornstein and Weiss [OW] proved that billiards with K-property are always Bernoulli, i.e. isomorphic to a Bernoulli shift. This was the highest degree of chaoticity in the measure-theoretic sense, putting billiards on the same level with Anosov diffeomorphisms.

For more practical purposes (applications to physics), the measure-theoretic Bernoulli property is of little help, though. As we explained in Section II.5, the statistics of iterations of smooth functions plays the key role. Most importantly, the rate of the decay of correlations and the central limit theorem (CLT) are two major issues. For Anosov diffeomorphisms, the CLT was proved in early seventies, and at the same time it was established that the correlations decay exponentially, which is accepted as the fastest possible rate of decay. This was done by Sinai, Ruelle and Bowen [Si3, Ru1, Ru2, Bo2]; the key element in their works was the construction of Markov partitions.

Since 1980, a major project was to extend the Sinai-Ruelle-Bowen techniques to chaotic billiards. The work has already spanned 20 years and is still in progress. First, Markov partitions were constructed [BS2, BSC1], then the CLT was proved [BS3, BSC2]. On the rate of the decay of correlations, first a weaker result was established [BS3, BSC2] - a subexponential upper bound. In the late nineties, it was shown that correlations decayed exponentially by a different approach [Y1, C6]. It finally become clear that for the purpose of physical applications, chaotic billiards behave just like Anosov diffeomorphisms. This was summarized by Gallavotti and Cohen in their "Axiom C" [GC].

It should be noted that all the above statistical properties are obtained only for dispersing billiards and only in dimension two. For other chaotic billiards, and especially in higher dimensions, the issue is wide open. Moreover, the rate of the decay of correlations for nonuniformly hyperbolic billiards, such as the stadium and the gas of hard balls, cannot be exponential, it is expected to be polynomial, see Section II.5.

Other fine properties of chaotic billiards remain to be investigated. For example, the asymptotics of periodic orbits is partially estimated [St3, C2], but nothing close to Theorem III.3.7 is obtained yet.

Lately, the studies of more "realistic" modifications of billiards became popular. There, one replaces elastic reflections at the rigid boundary ∂Q by a repulsive potential, or adds external forces acting on the moving particle inside Q. A steady stream of new interesting results in this direction indicates that the theory of chaotic billiards is not restricted to billiards alone – it plays an ever growing role in the modern mathematical physics.

Bibliography

[Ab] L. M. Abramov, *On the entropy of a flow* (in Russian), Dokl. Akad. Nauk SSSR **128**, 873 –875 (1959).

[Al] R. Alexander, *Time Evolution of Infinitely Many Hard Spheres*, Commun. Math. Phys. **49** (1976), 217–232

[An] D. V. Anosov, *Geodesic flows on closed Riemannian manifolds with negative curvature*, Proc. Steklov Inst. Math. **90** (1967).

[AS] D.V. Anosov & Ya.G. Sinai, *Some smooth ergodic systems*, Russ. Math. Surveys **22** (1967), 103–167.

[AA] V. I. Arnold & A. Avez, *Problèmes ergodiques de la Mècanique Classique.* Gauthier-Villars, Paris (1967).

[Ar] P. Arnoux, *Erogidicité générique des billiards polygonaux* (Séminaire Bourbaki), Astérisque **161-162** (1988), 203-221.

[ACG] R. Artuso, G. Casati & I. Guarneri, *Numerical Experiments on Billiards*, J. Statist. Phys. **83** (1996), 145–166.

[BC] A. Bäcker & N. Chernov, *Generating partitions for two-dimensional hyperbolic maps*, Nonlinearity **11** (1998), 79–87.

[Ba] V. Baladi, *Positive transfer operators and decay of correlations.* World Scientific, River Edge, N. Y. (2000).

[Be] G. Benettin, *Power law behavior of Lyapunov exponents in some conservative dynamical systems*, Phys. D **13** (1984), 211–220.

[BS] G. Benettin, & J. M. Strelcyn, *Numerical experiments on the free motion of a point mass moving in a plane convex regime: Stochastic transition and entropy*, Physical Review A **17** (1978), 773–785.

[Bi] G. D. Birkhoff, *Dynamical systems*, A.M.S. Colloquium Publications, New York (1927).

[Bo] L. Boltzmann, *Einige allgemenine Sätze über das Wärmegleichgewicht*, Wien. Ber. **63** (1871) 679-711.

[BlD] J.-P. Bouchaud and P. Le Doussal, *Numerical study of a d-dimensional periodic Lorentz gas with universal properties*, J. Statist. Phys. **41** (1985), 225–248.

[Bo1] R. Bowen, *Topological entropy and Axiom A*. In *Global Analysis (Proc. Symp. Pure Math., Vol XIV*, A.M.S., Providence, R.I. (1970), 23–41.

[Bo2] R. Bowen, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Lect. Notes Math. **470**, Springer, Berlin (1975).

[BG] D. S. Broomhead & E. Gutkin, *The dynamics of billiards with no-slip collisions*, Phys. D **67** (1993), 188–197.

[Bu1] L. A. Bunimovich, *On billiards close to dispersing*, Math. USSR. Sb. **23** (1974), 45–67.

[Bu2] L. A. Bunimovich, *On ergodic properties of nowhere dispersing billiards*, Commun. Math. Phys. **65** (1979), 295–312.

[Bu3] L. A. Bunimovich, *Many dimensional nowhere dispersing billiards with chaotic behavior*, Physica D **53**(1988), 58–64.

[Bu4] L. A.Bunimovich, *A Theorem on Ergodicity of Two-Dimensional Hyperbolic Billiards*, Commun. Math. Phys **130** (1990), 599–621.

[Bu5] L. A.Bunimovich, *Conditions of stochasticity of two-dimensional billiards*, Chaos **1** (1991), 187–193.

[Bu6] L. A.Bunimovich, *On absolutely focusing mirrors*, In *Ergodic Theory and related topics, III (Güstrow, 1990)*. Edited by U. Krengel et al., Lecture Notes in Math. **1514**, Springer, Berlin (1992), 62–82.

[Bu7] L. A. Bunimovich, *Continued fractions and geometrical optics*. In *Sinaĭ's Moscow Seminar on Dynamical Systems*, Amer. Math. Soc. Transl. Ser. 2 **171**, Amer. Math. Soc., Providence, RI, (1996) 45–55.

[BLPS] L. A. Bunimovich, C. Liverani, A. Pellegrinotti & Y. Suhov, *Ergodic systems of n balls in a billiard table*, Comm. Math. Phys. **146** (1992), 357–396.

[BR1] L. A. Bunimovich & J. Rehacek, *Nowhere dispersing* 3*D billiards with non-vanishing Lyapunov exponents*, Comm. Math. Phys. **189** (1997), 729–757.

[BR2] L. A. Bunimovich & J. Rehacek, *How high-dimensional stadia look like*, Comm. Math. Phys. **197** (1998), 277–301.

[BR3] L. A. Bunimovich & J. Rehacek, *On the ergodicity of many-dimensional focusing billiards* Ann. Inst. H. Poincaré Phys. Théor. **68** (1998), 421–448.

[BS1] L. A. Bunimovich, & Ya. G. Sinai, *On a fundamental theorem in the theory of dispersing billiards*, Math. USSR Sbornik **19** (1973), 407-423.

[BS2] L. A. Bunimovich & Ya. G. Sinai, *Markov partitions for dispersed billiards*, Commun. Math. Phys. **73** (1980), 247–280.

[BS3] L. A. Bunimovich & Ya. G. Sinai, *Statistical properties of Lorentz gas with periodic configuration of scatterers*, Commun. Math. Phys. **78** (1981), 479–497.

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

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

[BFK1] D. Burago, S. Ferleger & A. Kononenko, *A geometric approach to semi-dispersing billiards*, Ergodic Theory & Dynam. Systems**18** (1998), 303–319.

[BFK2] D. Burago, S. Ferleger & A. Kononenko, *Uniform estimates on the number of collisions in semi-dispersing billiards*, Ann. of Math. **147** (1998), 695–708.

[BFK3] D. Burago, S. Ferleger & A. Kononenko, *Topological entropy of semi-dispersing billiards*, Ergodic Theory & Dynam. Systems **18** (1998), 791–805.

[CMOP] E. Canale, R. Markarian, S. K. Oliffson & S. Pinto, *A lower bound for chaos on the elliptical stadium*, Phys. D **115** (1998), 189–202.

[C1] N. I. Chernov, *A new proof of Sinai's Formula of entropy of hyperbolic billiards. Its application to Lorentz gas ans stadium*, Funct. An. and its Appl. **25** (1991), 204–219.

[C2] N. I. Chernov, *Topological entropy and periodic points of two-dimensional hyperbolic billiards*, Functional Anal. Appl. **25** (1991), 39–45.

[C3] N.I. Chernov, *Statistical properties of the periodic Lorentz gas. Multidimensional case*, J. Stat. Phys. **74** (1994), 11–53.

[C4] N. I. Chernov, *Limit theorems and Markov approximations for chaotic dynamical systems*, Prob. Th. Rel. Fields **101** (1995), 321–362.

[C5] N. Chernov, *Entropy, Lyapunov exponents, and mean free path for billiards*, J. Statist. Phys. **88** (1997), 1–29.

[C6] N.I. Chernov, *Decay of correlations and dispersing billiards*, J. Statist. Phys. **94** (1999), 513–556.

[C7] N.I. Chernov, *Statistical properties of piecewise smooth hyperbolic systems in high dimensions*, Discrete and Continuous Dynamical Systems **5** (1999), 425–448.

[CH] N.I. Chernov & C. Haskell, *Nonuniformly hyperbolic K-systems are Bernoulli*, Ergodic Theory and Dynamical Systems, **16** (1996), 19-44.

[CM] N. I. Chernov & R. Markarian, *Entropy of non-uniformly hyperbolic plane billiards*, Bol. Soc. Bras. Mat. **23** (1992) 121–135.

[CT1] N. I. Chernov & S. Troubetzkoy, *Measures with infinite Lyapunov exponents for the periodic Lorentz gas*, J. Statist. Phys. **83** (1996), 193–202.

[CT2] N. I. Chernov & S. Troubetzkoy, *Ergodicity of billiards in polygons with pockets*, Nonlinearity **11** (1998), 1095–1102.

[CY] N. I. Chernov & L.-S. Young, *Decay of correlations of Lorentz gases and hard balls.* In *Hard ball systems and Lorentz gas*, D. Szász, editor, Springer, Berlin (2000), 89–120.

[CFS] I. P. Cornfeld, S. V. Fomin & Ya. G. Sinai, *Ergodic Theory.* Springer, Berlin (1982).

[DDL1] M. Degli Esposti, G. Del Magno & M. Lenci, *An infinite step billiard*, Nonlinearity **11** (1998), 991–1013.

[DDL2] M. Degli Esposti, G. Del Magno & M. Lenci, *Escape orbits and ergodicity in infinite step billiards*, Nonlinearity **13** (2000), 1275–1292.

[De] G. Del Magno, Manuscript (2001)

[Do1] V. Donnay, *Using Integrability to Produce chaos: Billiards with Positive Entropy*, Commun. Math. Phys. **141** (1991), 225–257.

[Do2] V. Donnay, *Elliptic islands in generalized Sinai billiards*, Ergodic Theory & Dynam. Systems **16** (1996), 975–1010.

[Do3] V. Donnay, *Non-ergodicity of two particles interacting via a smooth potential*, J. Statist. Phys. **96** (1999), 1021–1048.

[DoL] V. Donnay & C. Liverani, *Potentials on the two-torus for which the Hamiltonian flow is ergodic*, Comm. Math. Phys. **135** (1991), 267–302.

[Dy] R . Douady, *Applications du théorème des tores invariants.* These. Université Paris VII (1982).

[DvL] M. M. Dvorin & V. F. Lazutkin, *Existence of an infinite number of elliptic and hyperbolic periodic trajectories for convex billiards*, Functional Anal. Appl. **7** (1973), 103–109.

[EW] J.J. Erpenbeck & W.W. Wood, *Molecular-dynamics calculations of the velocity-autocorrelation function. Methods, hard-disk results*, Phys. Rev. A **26** (1982), 1648–1675.

[F] P. J. Fernandez, *Medida e integracao*. IMPA, Rio de Janeiro (1996).

[FM] B. Friedman & R. F. Martin, *Behavior of the velocity autocorrelation function for the periodic Lorentz gas*, Physica D **30** (1988), 219–227.

[Ga] G. Gallavotti, *Lectures on the billiards*. In *Dynamical systems, theory and applications (Rencontres, Battelle Res. Inst., Seattle., Wash., 1974)*, Lecture Notes in Physics **38**, Springer, Berlin(1975), 236-295.

[GO] G. Gallavotti & D. S. Ornstein, *Billiards and Bernoulli Schemes*, Commun. Math. Phys. **38** (1974), 83-101.

[GC] G. Gallavotti & E. D. G. Cohen, *Dynamical ensembles in stationary states*, J. Stat. Phys. **80** (1995), 931–970.

[Ga] Galperin, G. A. , *On systems of locally interacting and repelling particles moving in space*, Trans. Moscow Math. Soc. **43** (1981), 159-214.

[GG] P. Garrido & G. Gallavotti, *Billiards correlation function*, J. Statist. Phys. **76** (1994), 549–585.

[GD] P. Gaspard & J. R. Dorfman, *Chaotic scattering theory, thermodynamic formalism, and transport coefficients*, Phys. Rev. E **52** (1995), 3525–3552.

[Gu1] E. Gutkin, *Billiards in polygons*, Physica D **19** (1986), 311–333.

[Gu2] E. Gutkin, *Billiards in polygons: survey of recent results*, J. Statist. Phys. **83** (1996), 7–26.

[GK] E. Gutkin & A. Katok, *Caustics for inner and outer billiards*, Comm. Math. Phys. **173** (1995), 101–133.

[GSG] B. Gutkin, U. Smilansky & E. Gutkin, *Hyperbolic billiards on surfaces of constant curvature*, Comm. Math. Phys. **208** (1999), 65–90.

[Ha] P. R. Halmos, *Measure Theory*. D. van Nostrand, New York (1950); Springer, New York (1974) .

[Hn] B. Halpern, *Strange billiard tables*, Trans. Am. Math. Soc. **232** (1977), 297–305.

[Hg] E. H. Hauge, *What can one learn from Lorentz models? Transport phenomena*, (Sitges Internat. School Statist. Mech., Sitges, 1974), Lecture Notes in Phys., Vol. 31, Springer, Berlin (1974) 337–367.

[Hy] A. Hayli & Th. Dumont, *Experiences numeriques sur les billiards* C^1 *formés de quatre arcs de cercles*, Celestial Mechanics **38** (1986), 23–66.

[HDM] A. Hayli, Th. Dumont, J. Moulin-Ollagnier, & J. M. Strelcyn, *Quelques résultats nouveaux sur les billiards de Robnik*, J. Phys. A: Math. Gen. **20** (1987), 3237–3249.

[HW] M. Hénon & J. Wisdom, *The Benettin Strelcyn oval billiard revisited*, Physica D **8** (1983), 157–169.

[He] M. R. Herman, *Sur les courbes invariantes par les difféomorphismes de l'anneau, Vol. I. Completé par un appendice au chapitre I de Albert Fathu*, Asterisque **103-104** (1983), 3–221.

[Hu] A. Hubacher, *Instability of the boundary on the billiard ball problem*, Commun. Math. Phys. **108** (1987), 483–488.

[Ka] A. Katok, *Lyapunov exponents, entropy and periodic orbits for diffeomorphisms*, Publ. Math. I.H.E.S. **51** (1980), 137–173.

[KH] A. Katok & B. Hasselblatt, *Introduction to the modern theory of dynamical systems*. Cambridge Univ. Press (1995).

[KS] A. Katok & J.-M. Strelcyn, with the collaboration of F. Ledrappier and F. Przytycki, *Invariant manifolds, entropy and billiards; smooth maps with singularities*, Lect. Notes Math., **1222**, Springer, New York (1986).

[KMS] S. Kerckhoff, H. Masur & J. Smillie, *Ergodiciy of billiard flows and quadratic differentials*, Ann. of Math. **124** (1986), 293–311.

[Ki1] Jonathan. L. King, *Three problems in search of a measure*, Amer. Math. Monthly **101** (1994), 609–628.

[Ki2] Jonathan. L. King, *Billiards inside a cusp*, Math. Intelligencer **17** (1995), 8–16.

[KMOP1] J. Koiller, R. Markarian, S. Oliffson & S. Pinto, *Time-dependent billiards*, Nonlinearity **8** (1995), 983–1003.

[KMOP2] J. Koiller, R. Markarian, S. Oliffson & S. Pinto, *Static and time-dependent perturbations of the classical elliptical billiard*, J. Statist. Phys. **83** (1996), 127–143.

[Ko] A. N. Kolmogorov, *On the Conservation of Conditionally Periodic Motions under Small Perturbations of the Hamiltonian*, Dokl. Akad. Nauk SSSR, **98** (1954), 527–530

[KT] V. V. Kozlov & D. V. Treshchëv, *Billiards. A genetic introduction to the dynamics of systems with impacts*. Translations of Mathematical Monographs, **89** American Mathematical Society, Providence, RI (1991).

[KSS1] A. Krámli, N. Simányi & D. Szász, *Ergodic Properties of Semi–Dispersing Billiards I. Two Cylindric Scatterers in the 3–D Torus*, Nonlinearity, **2** (1989), 311–326

[KSS2] A. Krámli, N. Simányi & D. Szász, *A "Transversal" Fundamental Theorem for Semi-Dispersing Billiards*, Commun. Math. Phys., **129** (1990), 535–560

[KSS3] A. Krámli, N. Simányi & D. Szász, *The K–Property of Three Billiard Balls*, Annals of Mathematics, **133** (1991), 37–72

[KSS4] A. Krámli, N. Simányi & D. Szász, *The K–Property of Four Billiard Balls*, Commun. Math. Phys. **144** (1992), 107–142.

[KS] A. Krámli & D. Szász, *Convergence to Equilibrium of the Lorentz Gas*, Colloquia Math. Soc. János Bolyai

[KPT] T. Krüger, L. D. Pustyl'nikov & S. Troubetzkoy, *Acceleration of bouncing balls in external fields*, Nonlinearity **8** (1995), 397–410.

[Kr1] N. S. Krylov, *The Processes of Relaxation of Statistical Systems and the Criterion of Mechanical Instability* Thesis (1942). In *Development of Krylov's Ideas*, Princeton University Press, Princeton, N. J. (1977), 193–238

[Kr2] N. S. Krylov, *Works on the foundation of statistical physics.* Princeton U. Press, Princeton, N. J. (1979).

[La] V. F. Lazutkin, *On the existence of caustics for the billiard ball problem in a convex domain*, Math. USSR Izv. **7** (1973), 185–215.

[Lb1] J. Lebowitz, *Microscopic origins of irreversible macroscopic behavior*, Physica A **263** (1999), 516–527.

[Lb2] J. Lebowitz, *A century of statistichal mechanics: a selective review of two central issues*, Reviews of Modern Physics **71** (1999), 346–357.

[Le] M Lenci, *Escape orbits for non-compact flat billiards*, Chaos **6** (1996), 428–431.

[LL] J. Lewowicz & E. Lima De Sa, *Analytic Models of pseudo-Anosov maps*, Ergodic Theory & Dyn. Sys. **6** (1986), 385–392.

[LW] C. Liverani & M. Wojtkowski, *Ergodicity in Hamiltonian Systems* Dynamics reported, Dynam. Report. Expositions Dynam. Systems (N.S.) **4**, Springer, Berlin (1995), 130–202.

[LM1] A. Lopes & R. Markarian, *Billiards and conditionally invariant probabilities*, Resenhas **1** (1994), 421–439.

[LM2] A. Lopes & R. Markarian, *Open billiards: Cantor sets, invariant and conditionally invariant probabilities*, SIAM J. of Applied Math. **56** (1996), 651–680.

[Lo] H. A. Lorentz, *The motion of electrons in metallic bodies*, Proc. Amst. Acad. **7** (1905), 438–453.

[Mc] J. Machta, *Power Law Decay of Correlations in a Billiard Problem*, J. Statist. Phys. **32** (1983), 555–564.

[MR] J. Machta & B. Reinhold, *Decay of correlations in the regular Lorentz gas*, J. Statist. Phys. **42** (1986), 949–959.

[Mn] R. Mañe, *Introducao à Teoria Ergódica*. IMPA, Rio de Janeiro (1983). English Ed. *Ergodic Theory and Differentiable Dynamics*, Springer, Berlin, Heidelberg, New York (1987).

[Ma1] R. Markarian, *Billiards with Pesin region of measure one*, Comm. Math. Phys. **118** (1988), 87–97.

[Ma2] R. Markarian, *Non Uniform Hyperbolicity. Billiards*. In Workshop on Dynamical Systems, ICTP, Trieste, 1988. Edited by Z. Coelho and E. Shiels. Longman (1989), 129–135.

[Ma3] R. Markarian, *New ergodic billiards: exact results*, Nonlinearity **6** (1993), 819–841.

[Ma4] R. Markarian, *The fundamental theorem of Sinai-Chernov for dynamical systems with singuularities*. In *Dynamical Systems. Santiago de Chile 1990*, R. Bamón, R. Labarca, J. Lewowicz, and J. Palis, editors, Longman, Harlow (1993), 131–158.

[Ma5] R. Markarian, *Introduction to the ergodic theory of plane billiards*. In *Dynamical Systems. Santiago de Chile 1990*, R. Bamón, R. Labarca, J. Lewowicz, and J. Palis, editors, Longman, Harlow (1993), 327–439.

[Ma6] R. Markarian, *Non-uniformly hyperbolic billiards*, Ann. Fac. Sci. Toulouse Math. **3** (1994), 223–257.

[Ma7] R. Markarian, *Statistical properties of dynamical systems with singularities*, J. Statist. Phys. **80** (1995), 1207–1239.

[MOP] R. Markarian, S. Oliffson & S. Pinto, *Chaotic properties of the elliptical stadium*, Commun Math. Phys., **174** (1996), 661–679.

[Mz] J.N. Mather, *Glancing billiards*, Erg. Th. & Dynam. Syst. **2** (1982), 397–403.

[Mo1] T. Morita, *The symbolic representation of billiards without boundary condition*, Trans. Amer. Math. Soc. **325** (1991), 819–828.

[Mo2] T. Morita, *Billiards without boundary and their zeta functions*, Zeta functions in geometry (Tokyo, 1990), Adv. Stud. Pure Math., **21** (1992), 173–179.

[Mo3] T. Morita, *On the length spectrum of the bounded scattering billiards table.* In *Algorithms, fractals, and dynamics (Okayama/Kyoto, 1992)*, Plenum, New York, (1995), 169–177.

[OP] S. Oliffson & S. Pinto, *Bounded gain of energy on the breathing circle billiard*, Nonlinearity **12** (1999), 1363–1371.

[Or] D. S. Ornstein, *Bernoulli shifts with the same entropy are isomorphic*, Advances in Math. **4** (1970), 337–352.

[OW] D. S. Ornstein & B. Weiss, *On the Bernoulli nature of systems with some hyperbolic structure*, Erg. Th. & Dynam. Syst. **18** (1998), 441–456.

[Os] V. I. Oseledec, *A multiplicative ergodic theorem*, Trans. Moskow Math. Soc., **19** (1968), 197–231.

[PM] J. Palis & W. de Melo, *Introducao aos Sistemas Dinâmicos.* IMPA, Rio de Janeiro (1978). English Ed. *Geometric Theory of Dynamical Systems*, Springer, New York - Berlin (1982).

[Pr1] R. Peirone, *Reflections can be trapped*, Amer. Math. Monthly, **101** (1994), 259–260.

[Pr2] R. Peirone, *Billiards in tubular neighborhoods of manifolds of codimension 1*, Comm. Math. Phys., **207** (1999), 67–80.

[Pe1] Ya. B. Pesin, *Families of invariant manifolds corresponding to non zero Characteristic exponent* Math. USSR Izvestija **10** (1976), 1261–1305.

[Pe2] Ya. B. Pesin, *Characteristic Lyapunov exponents and smooth ergodic theory*, Russ. Math. Surv. **32** (1977), 55–114.

[PS] V. M. Petkov & L. Stojanov, *On the number of periodic reflecting rays in generic domains*, Ergodic Theory & Dynamical Systems **8** (1988), 81–91.

[Po] H. Poincaré, *Les Méthodes Nouvelles de la Mécanique Céleste* (1892-1899). 3 Volumes. Dover Publications, Inc., New York, N.Y., (1957).

[Pt] Hillel Poritsky, *The billiard ball problem on a table with a convex boundary—an illustrative dynamical problem* Ann. of Math. (2) **51** (1950), 446–470.

[Pu1] L. D. Pustyl'nikov, *On the Fermi-Ulam model*, Soviet Math. Dokl. **35** (1987), 88–92.

[Pu2] L. D. Pustyl'nikov, *Existence of invariant curves for maps close to degenerate maps and a solution of the Fermi-Ulam problem.*, Russian Acad. Sci. Sb. Math **82** (1995), 231–241.

[Re] J. Rehácek, *On the ergodicity of dispersing billiards*, Random Comput. Dynam. **3** (1995), 35–55.

[Ro] M. Robnik, *Classical dynamics of a family of billiards with analytic boundaries*, J. Phys. A; Math. Gen. **16** (1983), 3971–3986.

[Rd] W. Rudin, *Real and Complex Analysis.* Mc Graw-Hill, New York (1966).

[Ru1] D. Ruelle, *A measure associated with Axiom-A attractors*, Amer. J. Math. **98** (1976), 619–654.

[Ru2] D. Ruelle, *Thermodynamic Formalism*, Addison-Wesley, Reading, Mass. (1978).

[Ry] M. R. Rychlik, *Periodic points of the billiard ball map in a convex domain*, J. Differential Geom. **30** (1989), 191–205.

[Sa] L. A. Santaló, *Integral Geometry and geometric probability.* Addison Wesley, Reading, Mass. (1976).

[Sm] N. Simányi, *Ergodicity of hard spheres in a box*, Ergodic Theory & Dynam. Systems **19** (1999), 741–766.

[SS1] N. Simányi & D. Szász, *Hard ball systems are completely hyperbolic*, Ann. of Math. **149** (1999), 35–96.

[SS2] N. Simányi & D. Szász, *Non-integrability of cylindric billiards and transitive Lie group actions*, Ergodic Theory & Dynam. Systems **20** (2000), 593–610.

[SW] N. Simányi & M. Wojtkowski, *Two-particle billiard system with arbitrary mass ratio* Ergodic Theory & Dynamical Systems **9** (1989), 165–171.

[Si1] Ya. G. Sinai, *On the Foundation of the Ergodic Hypothesis for a Dynamical System of Statistical Mechanics*, Dokl. Akad. Nauk SSSR, **153** (1963), 1261–1264.

[Si2] Ya.G. Sinai, *Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards*, Russ. Math. Surv. **25** (1970), 137–189.

[Si3] Ya. G. Sinai, *Gibbs measures in ergodic theory*, Russ. Math. Surveys **27** (1972), 21–69.

[Si4] Ya. G. Sinai, *Billiard trajectories in a polyhedral angle*, Russ. Math. Surveys **33** (1978), 229–230.

[Si5] Ya. G. Sinai, *Entropy per particle for the system of hard spheres.* Harvard Univ. Preprint, 1978.

[Si6] Ya.G. Sinai, *Development of Krylov's ideas*, Afterwards to N. S. Krylov, *Works on the foundations of statistical physics.* Princeton Univ. Press, Princeton, N. J. (1979), 239–281.

[Si7] Ya.G. Sinai, *Hyperbolic billiards.* In *Proceedings of the International Congress of Mathematicians. Kyoto 1990* Math. Soc. Japan, Tokyo (1991), 249–260.

[SC1] Ya.G. Sinai & N.I. Chernov, *Entropy of a gas of hard spheres with respect to the group of space-time translations*, Trudy Sem. Petrovsk. **8**, (1982), 218–238. In Russian. In *Dynamical Systems*, Ed. by Ya. Sinai, Adv. Series in Nonlinear Dynamics, **1** [4]

[SC2] Ya.G. Sinai & N.I. Chernov, *Ergodic properties of some systems of 2-dimensional discs and 3-dimensional spheres*, Russ. Math. Surv. **42** (1987), 181–207.

[St1] L. Stoyanov, *Generic properties of periodic billiard trajectories*, Russian Math. Surveys **42** (1987), 235–236.

[St2] L. Stojanov, *Generic properties of periodic reflecting rays*, Ergodic Theory Dynamical Systems **7** (1987), 597–609.

[St3] L. Stojanov, *An Estimate from above of the number of periodic orbits for semi-dispersed billiards*, Commun. Math. Phys. **124** (1989), 217–227.

[St4] L. Stojanov, *Note on the periodic points of the billiard*, J. Differential Geom. **34** (1991), 835–837.

[St5] L. Stoyanov, *Instability and entropy for a class of dispersing billiards*, C. R. Acad. Sci. Paris Sér. I Math. **326** (1998), 737–742.

[St6] L. Stoyanov, *Exponential instability for a class of dispersing billiards*, Ergodic Theory & Dynam. Systems **19** (1999), 201–226.

[Sz1] D. Szász, *Boltzmann's ergodic hypothesis, a conjecture for centuries?*, Studia Sci. Math. Hung. **31** (1996), 299–322.

[Sz2] D. Szász, *Hard ball systems and the Lorentz gas*, Edited by D. Szász. Springer, Berlin (2000).

[Ta1] S. Tabachnikov, *Billiards*. Panor. Synth. No. 1, SMF, Paris (1995).

[Ta2] S. Tabachnikov, *Exact transverse line fields and projective billiards in a ball*, Geom. Funct. Anal. **7** (1997), 594–608.

[Ta3] S. Tabachnikov, *Introducing projective billiards*, Ergodic Theory & Dynam. Systems **17** (1997), 957–976.

[Va] L . N. Vaserstein, *On Systems of particles with finite - range and/or repulsive interactions*, Commun. Math. Phys. **69** (1979), 31–56.

[Vi] M. Viana, *Stochastic dynamics of deterministic systems*. IMPA, Rio de Janeiro (1997).

[VC] F. Vivaldi, G. Casati & I. Guarneri, *Origin of long-time tails in strongly chaotic systems*, Phys. Rev. Letters **51** (1983), 727–730.

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

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

[W3] M. Wojtkowski, *Measure theoretic entropy of the system of hard spheres*, Ergodic Theory Dynamical Systems **8** (1988), 133–153.

[W4] M. Wojtkowski, *A system of one-dimensional balls with gravity*, Comm. Math. Phys. **126** (1990), 507–533.

[W5] M. Wojtkowski, *The system of one-dimensional balls in an external field. II*, Comm. Math. Phys. **127** (1990), 425–432.

[W6] M. Wojtkowski, *Linearly stable orbits in* 3*-dimensional billiards* Comm. Math. Phys. **129** (1990), 319–327.

[W7] M. Wojtkowski, *Two applications of Jacobi fields to the billiard ball problem*, J. Differential Geom. **40** (1994), 155–164.

[Y1] L.-S. Young, *Statistical properties of dynamical systems with some hyperbolicity*, Annals of Math. **147** (1998), 585–650.

[Y2] L.-S. Young, *Recurrence times and rates of mixing*, Israel J. of Mathematics **110** (1999), 153–188.