University of Crete
Department of Mathematics

Notes on
Ergodic Theory of Dynamical Systems
from a geometric point of view

Konstantin Athanassopoulos

.

These notes grew out of two series of lectures I gave in the years 1994-95. The first was given jointly with D. Gatzouras in the Department of Mathematics of the University of Crete and coverd general Ergodic Theory. Chapters 1, 4 and the first two sections of chapter 3 are based on parts of this series of lectures. The rest of these notes are based on lectures given to the members of the research group of P. Strantzalos.

I want to thank all colleagues who attended the lectures for their comments and suggestions. Especially, I want to thank D. Gatzouras for his help, and E. Menioudaki who red the notes and lectured on a large part of them in the Department of Mathematics of the University of Crete during the academic year 2001-2002.

Contents

Chapter 1

Introduction

1.1 Measurable dynamical systems

Let $(X, \mathcal{A}, \mu)$ be a probability space. An *endomorphism* of X is a map $T : X \to X$ such that $T^{-1}(A) \in \mathcal{A}$ and $\mu(A) = \mu(T^{-1}(A))$ for every $A \in \mathcal{A}$. If T is invertible and T^{-1} is also an endomorphism, then T is called *automorphism*. A *measurable flow* on X is a one parameter group of automorphisms $(\phi_t)_{t\in\mathbb{R}}$, such that the evaluation map $\phi : \mathbb{R} \times X \to X$ is also measurable. Any of the above is called a measurable dynamical system.

Two measurable dynamical systems, say $T_k : X_k \to X_k$ on probability spaces $(X_k, \mathcal{A}_k, \mu_k)$, $k = 1, 2$, are called *measurably isomorphic* if there are T_k-invariant sets $Y_k \in \mathcal{A}_k$ with $\mu_k(Y_k) = 1$, $k = 1, 2$, and an isomorphism $h : (Y_1, \mathcal{A}_1, \mu_1) \to (Y_2, \mathcal{A}_2, \mu_2)$ such that $T_2 \circ h = h \circ T_1$ on Y_1 and $T_1 \circ h^{-1} = h^{-1} \circ T_2$ on Y_2.

We shall give in this introductory section three examples. Further examples will be given in later chapters. Let first G be a compact topological group. The Haar measure μ on G is the unique Borel probability measure invariant under left and right translations of G. For instance, if G is a torus, then the Haar measure is the normalized Lebesgue measure. Let $T : G \to G$ be a continuous group epimorphism. If $\nu(A) = \mu(T^{-1}(A))$ for any Borel set $A \subset G$, then $\nu(T(x)A) = \mu(xT^{-1}(A)) = \nu(A)$. Since T is onto, it follows that $\nu = \mu$. So T preserves the Haar measure. In particular, for the case $G = S^1$ we have that the map $T(z) = z^n$ preserves the normalized Lebesgue measure, for any $n \in \mathbb{Z}^+$.

Let $(X, \mathcal{A}, \mu)$ be a probability space and $(E, \mathcal{F})$ be a measurable space. A *random variable* with values in E is a measurable function from X to E. A *stochastic process* with values in E and parameter space J, which is usually one of $\mathbb{Z}^+$, $\mathbb{Z}$, $\mathbb{R}^+$ or $\mathbb{R}$, is a family of random variables $f = (f_j)_{j\in J}$ with values in E. If on the product E^J we consider the product σ-algebra $\mathcal{F}^J$, which is by definition the smallest σ-algebra that contains $\pi_j^{-1}(\mathcal{F})$, $j \in J$, where $\pi_j : E^J \to E$ is the j-projection, then f is just a random variable with values in the measurable space $(E^J, \mathcal{F}^J)$.

The *distribution* of a random variable $g : X \to E$ is the probability measure $g_*\mu = \mu \circ g^{-1}$ on $(E, \mathcal{F})$. The distribution of a stochastic process $f = (f_j)_{j\in J} : X \to E^J$ is the probability measure $f_*\mu$ on $(E^J, \mathcal{F}^J)$. This is the unique probability

measure such that

$$f_*\mu(\pi_{j_1}^{-1}(A_{j_1}) \cap ... \cap \pi_{j_n}^{-1}(A_{j_n})) = \mu(f_{j_1}^{-1}(A_{j_1}) \cap ... \cap f_{j_n}^{-1}(A_{j_n}))$$

for every finite set $\{j_1, ..., j_n\} \subset J$ and A_{j_1},...$A_{j_n} \in \mathcal{F}$.

If $\{\mu_j : j \in J\}$ is a family of probability measures on $(E, \mathcal{F})$, there exists a unique probability measure μ^J on $(E^J, \mathcal{F}^J)$ such that

$$\mu^J(\pi_{j_1}^{-1}(A_{j_1}) \cap ... \cap \pi_{j_n}^{-1}(A_{j_n})) = \mu_{j_1}(A_{j_1})...\mu_{j_n}(A_{j_n})$$

for every finite set $\{j_1, ..., j_n\} \subset J$ and A_{j_1},...$A_{j_n} \in \mathcal{F}$. The measure μ^J is called the product measure of the family $\{\mu_j : j \in J\}$.

The random variables $f_j : X \to E$, $j \in J$, are called *independent* if

$$\mu(f_{j_1}^{-1}(A_{j_1}) \cap ... \cap f_{j_n}^{-1}(A_{j_n})) = \mu(f_{j_1}^{-1}(A_{j_1}))...\mu(f_{j_n}^{-1}(A_{j_n}))$$

for every finite set $\{j_1, ..., j_n\} \subset J$ and A_{j_1},...$A_{j_n} \in \mathcal{F}$. In other words, they are independent if and only if the distribution $f_*\mu$ of the stochastic process $f = (f_j)_{j \in J}$ coincides with the product measure μ^J, where μ_j is the distribution of the random variable f_j. The random variables are called *identically distributed* if their distributions are equal.

Let now $f = (f_k)_{k \in \mathbb{Z}^+}$ be a sequence of independent and identically distributed random variables and $\tau : E^{\mathbb{Z}^+} \to E^{\mathbb{Z}^+}$ be the *shift*, that is τ is the map defined by $\tau((x_k)_{k \geq 0}) = (x_{k+1})_{k \geq 0}$. Then the distribution $f_*\mu$ is preserved by τ. In general we have the following.

1.1.1. Lemma. *The product measure $\mu^{\mathbb{Z}^+}$ of a sequence of probability measures $(\mu_k)_{k \in \mathbb{Z}^+}$ on $(E, \mathcal{F})$ is preserved by the shift if and only if $\mu_k = \mu_l$ for every $k, l \in \mathbb{Z}^+$.*

Proof. If k_1,...,$k_n \in \mathbb{Z}^+$ and A_{k_1},...,$A_{k_n} \in \mathcal{F}$, then

$$\tau^{-1}(\pi_{k_1}^{-1}(A_{k_1}) \cap ... \cap \pi_{k_n}^{-1}(A_{k_n})) = \pi_{k_1+1}^{-1}(A_{k_1}) \cap ... \cap \pi_{k_n+1}^{-1}(A_{k_n}).$$

So, if $\mu^{\mathbb{Z}^+}$ is τ-invariant, we have

$$\mu_k(A) = \mu^{\mathbb{Z}^+}(\tau^{-1}(\pi_k^{-1}(A)) = \mu^{\mathbb{Z}^+}(\pi_{k+1}^{-1}(A)) = \mu_{k+1}(A)$$

for every $k \in \mathbb{Z}^+$ and $A \in \mathcal{F}$. Conversely, if $\mu_k = \mu_0$ for every $k \in \mathbb{Z}^+$, then

$$\tau_*\mu^{\mathbb{Z}^+}(\pi_{k_1}^{-1}(A_{k_1}) \cap ... \cap \pi_{k_n}^{-1}(A_{k_n})) = \mu^{\mathbb{Z}^+}(\tau^{-1}(\pi_{k_1}^{-1}(A_{k_1}) \cap ... \cap \pi_{k_n}^{-1}(A_{k_n}))) =$$

$$\mu^{\mathbb{Z}^+}(\pi_{k_1+1}^{-1}(A_{k_1}) \cap ... \cap \pi_{k_n+1}^{-1}(A_{k_n})) = \mu_0(A_{k_1})...\mu_0(A_{k_n}) =$$

$$\mu^{\mathbb{Z}^+}(\pi_{k_1}^{-1}(A_{k_1}) \cap ... \cap \pi_{k_n}^{-1}(A_{k_n})).$$

Since $\mu^{\mathbb{Z}^+}$ and $\tau_*\mu^{\mathbb{Z}^+}$ are equal on cylinders, they are everywhere equal. □

A stochastic process $f = (f_k)_{k \in \mathbb{Z}^+}$ is call *stationary* if its distribution is preserved by the shift. This is equivalent to saying that

$$\mu(f_{k_1}^{-1}(A_{k_1}) \cap ... \cap f_{k_n}^{-1}(A_{k_n})) = \mu(f_{k_1+1}^{-1}(A_{k_1}) \cap ... \cap f_{k_n+1}^{-1}(A_{k_n}))$$

for every $k_1,...,k_n \in \mathbb{Z}^+$ and $A_{k_1},...,A_{k_n} \in \mathcal{F}$. So every stochastic process of independent and identically distributed random variables is stationary.

1.1.2. Proposition. *Let $(X, \mathcal{A}, \mu)$ be a probability space, $T : X \to X$ be an endomorphism and $(E, \mathcal{F})$ be a measurable space. For every measurable $f : X \to E$, the sequence of random variables $f_k = f \circ T^k$, $k \in \mathbb{Z}^+$ is a stationary stochastic process.*

Proof. For every $k_1,...,k_n \in \mathbb{Z}^+$ and $A_{k_1},...,A_{k_n} \in \mathcal{F}$ we have

$$\mu(f_{k_1}^{-1}(A_{k_1}) \cap ... \cap f_{k_n}^{-1}(A_{k_n})) = \mu(T^{-k_1}(f^{-1}(A_{k_1})) \cap ... \cap T^{-k_n}(f_{k_n}^{-1}(A_{k_n}))) =$$

$$\mu(T^{-1}(T^{-k_1}(f^{-1}(A_{k_1})) \cap ... \cap T^{-k_n}(f_{k_n}^{-1}(A_{k_n})))) = \mu(f_{k_1+1}^{-1}(A_{k_1}) \cap ... \cap f_{k_n+1}^{-1}(A_{k_n})). \square$$

Let $(X, \mathcal{A}, \mu)$ be a probability space, $k \in \mathbb{N}$ and $f_n : X \to \{0, 1, ..., k-1\}$, $n \in \mathbb{Z}^+$, be a sequence of random variables, where on $\{0, 1, ..., k-1\}$ we consider its Borel algebra as a discrete space. Let the random variables be independent and identically distributed and suppose that $\mu_0(l) = p_l$, $l = 0, 1, ..., k-1$, where μ_0 is their common distribution. The shift τ on the product space $\{0, 1, ..., k-1\}^{\mathbb{Z}^+}$ with the product measure is called the one-sided *Bernulli shift* on the space of sequences on k symbols with probabilities $p_0,...,p_{k-1}$. Similarly, on the space $\{0, 1, ..., k-1\}^{\mathbb{Z}}$ of doubly infinite sequences on k symbols we have the two sided Bernulli shift with probabilities $p_0,...,p_{k-1}$, which is an automorphism.

Note that the product space $\{0, 1, ..., k-1\}^{\mathbb{Z}^+}$ has a totally disconnected, compact, abelian topological group structure and the shift is a continuous epimorphism. The Haar measure is the product measure coming from probabilities $p_0 = p_1 = ... = p_{k-1} = 1/k$. So in this case the Bernulli shift is a particular case of our first example.

Our third example are the volume preserving vector fields on oriented manifolds. Let M be a compact, connected, smooth manifold, oriented by a volume element ω, whose integral on M is equal to 1. It follows from the Riesz representation theorem that there exists a unique Borel probability measure μ_ω on M such that

$$\int_M f d\mu_\omega = \int_M f\omega,$$

for every continuous $f : M \to \mathbb{R}$. If $h : M \to M$ is an orientation preserving diffeomorphism, then from the change of variables formula we have

$$\int_M f d\mu_\omega = \int_M f\omega = \int_M h^*(f\omega) = \int_M (f \circ h) \cdot h^*\omega = \int_M (f \circ h) d\mu_{h^*\omega}.$$

It follows from this that

$$\mu_\omega(h(A)) = \int_M \chi_{h(A)} d\mu_\omega = \int_M (\chi_{h(A)} \circ h) d\mu_{h^*\omega} = \int_M \chi_A d\mu_{h^*\omega} = \mu_{h^*\omega}(A)$$

for every Borel set $A \subset M$. Since h is orientation preserving, there exists a unique smooth function $\det_\omega h_* : M \to (0, +\infty)$ such that $(h^*\omega)_x = (\det_\omega h_*(x)) \cdot \omega_x$ for every $x \in M$. From the chain rule we have $\det_\omega(g \circ h)_* = ((\det_\omega g_*) \circ h) \cdot (\det_\omega h_*)$.

Let now ξ be a smooth vector field on M. There exists a unique smooth function $\mathrm{div}_\omega \xi : M \to \mathbb{R}$, called the *divergence* of ξ with respect to ω, such that $d(i_\xi \omega) = (\mathrm{div}_\omega \xi)\omega$. If $(U, x^1, ..., x^n)$ is a system of local coodrinates on M, then $\omega | U = f dx^1 \wedge ... \wedge dx^n$ for some smooth function $f : U \to \mathbb{R}$. If $\xi = (\xi^1, ..., \xi^n)$ in the local coordinates of U, then

$$\mathrm{div}_\omega \xi | U = \frac{1}{f} \cdot \sum_{k=1}^{n} \frac{\partial (f \xi^k)}{\partial x^k}.$$

Since the Lie derivative $L_\xi \omega = d(i_\xi \omega) + i_\xi (d\omega) = (\mathrm{div}_\omega \xi)\omega$, we have

$$\mathrm{div}_\omega \xi = \lim_{t \to 0} \frac{\det_\omega \phi_{t*} - 1}{t}$$

where $(\phi_t)_{t \in \mathbb{R}}$ is the flow of ξ. If $x \in M$ and $\psi_x(t) = \det_\omega \phi_{t*}(x)$, $t \in \mathbb{R}$, then $\mathrm{div}_\omega \xi(x) = \psi_x'(0)$ and

$$\psi_x'(t) = \lim_{s \to 0} \frac{\det_\omega \phi_{(t+s)*}(x) - \det_\omega \phi_{t*}(x)}{s} =$$

$$\lim_{s \to 0} \frac{(\det_\omega \phi_{s*}(\phi_t(x))) \cdot (\det_\omega \phi_{t*}(x)) - \det_\omega \phi_{t*}(x)}{s} = \psi'_{\phi_t(x)}(0) \cdot \psi_x(t).$$

We conclude that $\mathrm{div}_\omega \xi = 0$ identically on M if and only if $\det_\omega \phi_{t*} = 1$ for every $t \in \mathbb{R}$ if and only if $\phi_t^* \omega = \omega$ for every $t \in \mathbb{R}$. In other words the divergenceless smooth vector fields are preciesely the volume preserving ones.

1.2 Poincaré recurrence

It is clear from the definitions we gave that the notion of measurable dynamical system is too general and in order to conclude useful properties we shall need a minimum of additional information on the nature of a system. There is however a general remarkable theorem due to H. Poincaré, which is qualitative in nature. We prove it first in the measure theoretical setting.

1.2.1. Theorem (Poincaré-Gibbs). *Let T be an endomorphism of a probability space $(X, \mathcal{A}, \mu)$. Let $A \in \mathcal{A}$ and*

$$A_0 = \{x \in A : T^n(x) \in A \text{ for infinitely many } n \geq 0\}.$$

Then, $A_0 \in \mathcal{A}$ and $\mu(A_0) = \mu(A)$.

Proof. Let $C_n = \{x \in A : T^m(x) \notin A \text{ for every } m \geq n\}$. Then, $A_0 = A \setminus \bigcup_{n=1}^{\infty} C_n$. It suffices to prove that $C_n \in \mathcal{A}$ and $\mu(C_n) = 0$ for every $n \in \mathbb{N}$. We observe first that

$$C_n = A \setminus \bigcup_{m \geq n} T^{-m}(A) = A \setminus T^{-n}(\bigcup_{m \geq 0} T^{-m}(A)).$$

Since T is an endomorphism, $T^{-m}(A) \in \mathcal{A}$ and thus $C_n \in \mathcal{A}$. Moreover,

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

and hence

$$\mu(C_n) \leq \mu(\bigcup_{m\geq 0} T^{-m}(A)) - \mu(\bigcup_{m\geq n} T^{-m}(A)) =$$

$$\mu(\bigcup_{m\geq 0} T^{-m}(A)) - \mu(T^{-n}(\bigcup_{m\geq 0} T^{-m}(A))) = 0.\square$$

We shall give now the topological version of Poincaré's recurrence theorem in the case of continuous time. A *continuous flow* on a metric space X is a continuous one parameter group of homeomorphisms $(\phi_t)_{t\in\mathbb{R}}$ of X, that is the evaluation map $\phi : \mathbb{R} \times X \to X$ is continuous. The set

$$L^+(x) = \{y \in X : \phi_{t_n}(x) \to y \text{ for some } t_n \to +\infty\}$$

is called *the positive limit set* of x and is closed and invariant under the flow. The *negative limit set* $L^-(x)$ is defined in the obvious way and has similar properties. Let $P^{\pm} = \{x \in X : x \in L^{\pm}(x)\}$ and $P = P^+ \cap P^-$. The closure of P is called the *Birkhoff center* of the flow. The points of P^+ are called *positively recurrent* and of P^- *negatively recurrent*.

1.2.2. Lemma. *A point $x \in X$ is positively recurrent if and only if for every neighbourhood V of x there exists $t \geq 1$ such that $\phi_t(x) \in V$.*

Proof. Only the converse requires proof. Let $\{V_n : n \in \mathbb{N}\}$ be a neighbourhood base of x. According to the hypothesis, there exist $t_n \geq 1$ such that $\phi_{t_n}(x) \in V_n$, $n \in \mathbb{N}$. Then, $\phi_{t_n}(x) \to x$ and either $t_n \to +\infty$ or the sequence $(t_n)_{n\in\mathbb{N}}$ has a convergent subsequence. In the second case there exists some $t \geq 1$ such that $\phi_t(x) = x$, because of the continuity of the flow, and therefore $\phi_{nt}(x) = x$ for every $n \in \mathbb{N}$. Hence in any case $x \in P^+$. $\square$

1.2.3. Theorem. *Let $(\phi_t)_{t\in\mathbb{R}}$ be a continuous flow on a separable metric space X, which preserves a Borel probability measure μ. Then P contains a Borel set of full measure.*

Proof. For any Borel set $A \subset X$, the sets

$$A^+ = A \setminus \bigcup_{n=1}^{\infty} A \cap \phi_n(A) \text{ and } A^- = A \setminus \bigcup_{n=1}^{\infty} A \cap \phi_{-n}(A)$$

are Borel. Obviously, $(\phi_t(A))^{\pm} = \phi_t(A^{\pm})$ for every $t \in \mathbb{R}$. For every $k > l \geq 0$ we have $\phi_k(A^+) \cap \phi_l(A^+) = \phi_{k-l}(A^+) \cap (A^+) = \varnothing$. It follows that

$$\sum_{k=0}^{\infty} \mu(A^+) = \sum_{k=0}^{\infty} \mu(\phi_k(A^+)) = \mu(\bigcup_{k=0}^{\infty} \phi_k(A^+)) \leq 1.$$

This can happen only if $\mu(A^+) = 0$. Similarly we have $\mu(A^-) = 0$. Let now $\{A_n : n \in \mathbb{N}\}$ be a countable base of the topology of X. Set $B^{\pm} = \bigcup_{n\in\mathbb{N}} A_n^{\pm}$ and $B = B^+ \cup B^-$. According to the above remarks, $\mu(B^+) = \mu(B^-) = 0$ and thus $\mu(X \setminus B) = 1$. So it suffices to prove that $X \setminus B \subset P$. Let $x \in X \setminus B$ and A_l be a basic open set containing x. Then, $x \in (X \setminus A_l^+) \cap (X \setminus A_l^-)$. Thus, there exist m, $n > 0$ such that $x \in A_l \cap \phi_m(A_l) \cap \phi_{-n}(A_l)$. It follows from Lemma 1.2.2 that $x \in P$. □

1.2.4. Corollary. *Let $(\phi_t)_{t\in\mathbb{R}}$ be continuous flow on a separable metric space X, which preserves a Borel probability measure μ. Then the support of μ is contained in the Birkhoff center of the flow.*

Chapter 2

Classical mechanical systems

2.1 Hamiltonian systems

A *symplectic vector space* is a finite dimensional real vector space equiped with a non-degenerate, antisymmetric, bilinear form ω. Every symplectic vector space (V, ω) has even dimension, say $2n$ for some $n \in \mathbb{N}$, and a basis $\{e_1, ..., e_n, e_1^*, ..., e_n^*\}$ such that $\omega(e_i, e_j^*) = \delta_{ij}$ and $\omega(e_i, e_j) = \omega(e_i^*, e_j^*) = 0$, for $1 \leq i, j \leq n$. A linear map $f : V \to V$ is called symplectic if $\omega(f(u), f(v)) = \omega(u, v)$ for every u, $v \in V$. Every symplectic f is an isomorphism, $(\det f)^2 = 1$ and is conjugate to $(f^{-1})^t$. Thus, if $\lambda \in \mathbb{C}$ is an eigenvalue of f, then $\bar{\lambda}$, $1/\lambda$ and $1/\bar{\lambda}$ are also eigenvalues.

A *symplectic manifold* is a smooth manifold P equipped with a smooth, closed, non-degenerate, 2-form ω. Thus, the pair (T_xP, ω_x) is a symplectic vector space for every $x \in P$. It follows that every symplectic manifold is even dimensional. The simplest and perhaps most important example is the cotangent bundle of a smooth manifold. Let M be a smooth manifold of any finite dimension and $q : T^*M \to M$ be the cotangent bundle map. Let θ be the 1-form on T^*M defined by $\theta_a = a \circ q_{*a}$ for $a \in T^*M$ and $\omega = -d\theta$. We shall describe θ and ω locally. To a system of local coordinates $(U, q^1, ..., q^n)$ on M corresponds a local trivialization of q, which gives local coordinates $(q^{-1}(U), q^1, ..., q^n, p_1, ..., p_n)$ on T^*M, such that if the local coordinates of $x \in U$ are $(q^1, ..., q^n)$, then the local coordinates of $a \in q^{-1}(U)$ are $(q^1, ..., q^n, p_1, ..., p_n)$, where

$$p_i = a(\frac{\partial}{\partial q^i}), \quad 1 \leq i \leq n.$$

From the definition of θ we have

$$\theta_a(\frac{\partial}{\partial q^i}) = a(q_{*a}(\frac{\partial}{\partial q^i})) = a(\frac{\partial}{\partial q^i}) = p_i$$

and

$$\theta_a(\frac{\partial}{\partial p_i}) = a(q_{*a}(\frac{\partial}{\partial p_i})) = a(0) = 0.$$

This shows that

$$\theta = \sum_{i=1}^{n} p_i dq^i \text{ and } \omega = \sum_{i=1}^{n} dq^i \wedge dp_i.$$

In particular, ω is non-degenerate and therefore (T^*M, ω) is a symplectic manifold.

By the theorem of Darboux, every symplectic $2n$-manifold (P, ω) can be coverd by local coordinates $(W, q^1, ..., q^n, p_1, ..., p_n)$ such that

$$\omega | W = \sum_{i=1}^{n} dq^i \wedge dp_i.$$

In these Darboux local coordinates we have

$$\omega \wedge ... \wedge \omega = (-1)^{[n/2]} \cdot n! \cdot dq^1 \wedge ... \wedge dq^n \wedge dp_1 \wedge .... \wedge dp_n,$$

where the wedge product on the left hand side is taken n times.

A smooth map $f : P \to P$ is called symplectic if $f^*\omega = \omega$. It is evidently a local diffeomorphism and preserves the volume element $\Omega = ((-1)^{[n/2]}/n!)\omega \wedge ... \wedge \omega$.

2..1.1. Definition. Let (P, ω) be a symplectic manifold. A smooth vector field ξ on P is call *Hamiltonian* if there exists a smooth function $H : P \to \mathbb{R}$, called the hamiltonian, such that such $dH = i_\xi \omega$.

Since ω is non-degenerate, every smooth function is the hamiltonian of a Hamiltonian vector field. The integral curves of a Hamiltonian vector field are locally solutions of Hamilton's differential equations. Let $(W, q^1, ..., q^n, p_1, ..., p_n)$ be Darboux local coordinates. On the one hand on W we have

$$dH = \sum_{i=1}^{n} \frac{\partial H}{\partial q^i} \cdot dq^i + \sum_{i=1}^{n} \frac{\partial H}{\partial p_i} \cdot dp_i$$

and on the other hand

$$i_\xi \omega(\frac{\partial}{\partial q^i}) = (\sum_{k=1}^{n} dq^k \wedge dp_k)(\xi, \frac{\partial}{\partial q^i}) = -dq^i(\frac{\partial}{\partial q^i}) \cdot dp_i(\xi) = -\dot{p}_i$$

and

$$i_\xi \omega(\frac{\partial}{\partial p_i}) = (\sum_{k=1}^{n} dq^k \wedge dp_k)(\xi, \frac{\partial}{\partial p_i}) = dq^i(\xi) \cdot dp_i(\frac{\partial}{\partial p_i}) = \dot{q}^i.$$

Thus, the equation $dH = i_\xi \omega$ in the local coordinates of W is equivalent to

$$\sum_{i=1}^{n} \frac{\partial H}{\partial q^i} \cdot dq^i + \sum_{i=1}^{n} \frac{\partial H}{\partial p_i} \cdot dp_i = \sum_{i=1}^{n} (-\dot{p}_i) dq^i + \sum_{i=1}^{n} \dot{q}^i dp_i$$

or equivalently

$$\dot{q}^i = \frac{\partial H}{\partial p_i} \text{ and } \dot{p}_i = -\frac{\partial H}{\partial q^i}, \quad 1 \leq i \leq n,$$

which are Hamilton's equations.

It is obvious that the hamiltonian H of a Hamiltonian vector field ξ is a first integral, since $dH(\xi) = \omega(\xi, \xi) = 0$. Thus, the level sets $H^{-1}(c)$, $c \in \mathbb{R}$, are invariant under the flow of ξ and the qualitative study of its flow falls into the study of the restrictions on these level sets, the topology of the level sets themselves and the way

they fill in P. If c is a regular value of H, then $H^{-1}(c)$ is a submanifold of P. The volume element Ω induces a natural volume element $\tilde{\Omega}$ on $H^{-1}(c)$ defined by

$$\tilde{\Omega}_x(u_1, ..., u_{2n-1}) = \Omega_x(u, u_1, ..., u_{2n-1})$$

where $x \in H^{-1}(c)$, $u_1, ..., u_{2n-1} \in T_x H^{-1}(c) = \ker dHx$, and $u \in T_x P$ is such that $dH(x)u = 1$. The definition is clearly independent of u.

The local flow of the Hamiltonian vector field ξ consists of symplectic diffeomorphisms of open subsets of P, which therefore preserve the volume element Ω. If ϕ_t is a diffeomorphism of the local flow of ξ for some $t \in \mathbb{R}$, then

$$(\phi_t^* \tilde{\Omega})_x(u_1, ..., u_{2n-1}) = \Omega_{\phi_t(x)}(u, \phi_{t*}(x)u_1, ..., \phi_{t*}(x)u_{2n-1})$$

where $x \in H^{-1}(c)$, $u_1, ..., u_{2n-1} \in T_x H^{-1}(c)$, and $u \in T_{\phi_t(x)} P$ is such that $dH(\phi_t(x))u = 1$. Since $\phi_{t*}(x) : T_x P \to T_{\phi_t(x)} P$ is a linear isomorphism, there is a unique $u_0 \in T_x P$ such that $\phi_{t*}(x)u_0 = u$. Differentiating the equation $H \circ \phi_t = H$ we get $dH(\phi_t(x)) \circ \phi_{t*}(x) = dH(x)$. Therefore, $dH(x)u_0 = 1$ and

$$(\phi_t^* \tilde{\Omega})_x(u_1, ..., u_{2n-1}) = (\phi_t^* \Omega)_x(u_0, u_1, ..., u_{2n-1}) = \tilde{\Omega}_x(u_1, ..., u_{2n-1})$$

which shows that $\phi_t^* \tilde{\Omega} = \tilde{\Omega}$.

2.2 Mechanical systems on Riemannian manifolds

Let M be a n-dimensional Riemannian manifold with metric g. There is a natural bundle isomorphism $\mathcal{L} : TM \to T^*M$, such that if $v \in T_x M$ then $\mathcal{L}(v)$ is the linear form on $T_x M$ defined by $\mathcal{L}(v)(w) = g_x(v, w)$. The inner product g_x on $T_x M$ is thus transfered to an inner product g_x^* on $T_x^* M$. If in local coordinates the matrix of g is $G = (g_{ij})$, then in the dual local coordinates the matrix of g^* is $G^{-1} = (g^{ij})$. If $\omega = -d\theta$ is the standard symplectic 2-form on T^*M, then $\mathcal{L}^*\omega = -d(\mathcal{L}^*\theta)$ is a symplectic 2-form on TM.

2.2.1. Definition. A *mechanical system* on the Riemannian manifold M is a Hamiltonian vector field ξ on TM with hamiltonian function of the form

$$E(v) = \frac{1}{2}\|v\|^2 + V(\pi(v))$$

where $V : M \to \mathbb{R}$ is a smooth function, called the *potential energy*, $\pi : TM \to M$ is the tangent bundle projection and $\| \cdot \|$ is the norm on the fibers of the tangent bundle defined by the Riemannian metric.

We shall find Hamilton's equations of motion for a mechanical system on a Riemannian manifold. First we must find local expressions for $\mathcal{L}^*\theta$ and $\mathcal{L}^*\omega$. Let $(U, q^1, ..., q^n)$ be a system of local coordinates on M. Since $\mathcal{L}(x, v) = (x, g_x(v, \cdot))$, its Jacobian is

$$D\mathcal{L}(x, v) = \begin{pmatrix} I_n & 0 \\ \dfrac{\partial}{\partial x} g_x(v, \cdot) & g_x(\cdot, \cdot) \end{pmatrix}$$

or explicitly

$$D\mathcal{L}(x,v)\begin{pmatrix}u\\w\end{pmatrix}=\begin{pmatrix}u\\(\frac{\partial}{\partial x}g_x(v,\cdot))u+g_x(\cdot,w)\end{pmatrix}.$$

It follows that

$$(\mathcal{L}^*\theta)_{(x,v)}(u,w)=\theta_{\mathcal{L}(x,v)}(u,\frac{\partial}{\partial x}g_x(v,\cdot))u+g_x(\cdot,w))=g_x(v,u).$$

This means that if $(q^1,...,q^n,v^1,...,v^n)$ are the corresponding local coordinates of $\pi^{-1}(U)$, then on $\pi^{-1}(U)$we have

$$\mathcal{L}^*\theta=\sum_{i,j=1}^n g_{ij}v^j dq^i$$

and therefore

$$\mathcal{L}^*\omega=\sum_{i,j=1}^n g_{ij}dq^i\wedge dv^j+\sum_{i,j,k=1}^n\frac{\partial g_{ij}}{\partial q^k}\cdot v^j dq^i\wedge dq^k.$$

Note that the local coordinates on $\pi^{-1}(U)$ are not Darboux. Next we have

$$dE=\frac{1}{2}\sum_{i,j,k=1}^n\frac{\partial g_{ij}}{\partial q^k}v^iv^jdq^k+\sum_{i,k=1}^n g_{ik}v^idv^k+\sum_{k=1}^n\frac{\partial V}{\partial q^k}dq^k,$$

and

$$i_\xi\mathcal{L}^*\omega(\frac{\partial}{\partial q^k})=-\sum_{j=1}^n g_{kj}dv^j(\xi)+\sum_{i,j=1}^n\frac{\partial g_{ij}}{\partial q^k}v^jdq^i(\xi)-\sum_{j,l=1}^n\frac{\partial g_{kj}}{\partial q^l}v^jdq^l(\xi),$$

$$i_\xi\mathcal{L}^*\omega(\frac{\partial}{\partial v^k})=\sum_{i=1}^n g_{ik}dq^i(\xi),\quad 1\le k\le n.$$

If I is an open interval, then $(q^1(t),...,q^n(t),v^1(t),...,v^n(t))$, $t\in I$, is an integral curve of ξ if and only if it is a solution of the system of differential equations

$$\sum_{i=1}^n g_{ik}\dot{q}^i=\sum_{i=1}^n g_{ik}v^i$$

$$-\sum_{j=1}^n g_{kj}\dot{v}^j+\sum_{i,j=1}^n\frac{\partial g_{ij}}{\partial q^k}v^j\dot{q}^i-\sum_{i,j=1}^n\frac{\partial g_{kj}}{\partial q^i}v^j\dot{q}^i=\frac{1}{2}\sum_{i,j=1}^k\frac{\partial g_{ij}}{\partial q^k}v^iv^j+\frac{\partial V}{\partial q^k},\quad 1\le k\le n.$$

It is obvious that the first n equations are equivalent to $\dot{q}^i=v^i$, $1\le i\le n$. The rest of them can be written

$$\sum_{j=1}^n g_{kj}\dot{v}^j=-\frac{1}{2}\sum_{i,j=1}^k\frac{\partial g_{ij}}{\partial q^k}v^iv^j+\sum_{i,j=1}^n\frac{\partial g_{ij}}{\partial q^k}v^jv^i-\sum_{i,j=1}^n\frac{\partial g_{kj}}{\partial q^i}v^jv^i-\frac{\partial V}{\partial q^k},\quad 1\le k\le n.$$

or equivalently, since G is symmetric,

$$\dot{v}^k = \sum_{l=1}^{n} g^{kl}\left[\frac{1}{2}\sum_{i,j=1}^{k}\frac{\partial g_{ij}}{\partial q^l}v^i v^j - \sum_{i,j=1}^{n}\frac{\partial g_{lj}}{\partial q^i}v^j v^i - \frac{\partial V}{\partial q^l}\right] = -\sum_{i,j=1}^{n}\Gamma_{ij}^k v^i v^j - \sum_{l=1}^{n} g^{kl}\frac{\partial V}{\partial q^l},$$

where Γ_{ij}^k are the Christofell symbols, because

$$\Gamma_{ij}^k = \frac{1}{2}\sum_{l=1}^{n} g^{kl}\left(\frac{\partial g_{jl}}{\partial q^i} + \frac{\partial g_{li}}{\partial q^j} - \frac{\partial g_{ij}}{\partial q^l}\right)$$

and thus

$$\sum_{i,j=1}^{n}\Gamma_{ij}^k v^i v^j = \sum_{i,j=1}^{n}\sum_{l=1}^{n} g^{kl}\left(\frac{\partial g_{jl}}{\partial q^i} - \frac{1}{2}\frac{\partial g_{ij}}{\partial q^l}\right)v^i v^j.$$

So Hamilton's differential equations can be written locally

$$\dot{q}^k = v^k,$$

$$\dot{v}^k = -\sum_{i,j=1}^{n}\Gamma_{ij}^k v^i v^j - \sum_{i=1}^{n} g^{ki}\frac{\partial V}{\partial q^i}, \quad 1 \le k \le n,$$

which are equivalent to the system of second order differential equations

$$\ddot{q}^k + \sum_{i,j=1}^{n}\Gamma_{ij}^k \dot{q}^i \dot{q}^j = -\sum_{i=1}^{n} g^{ki}\frac{\partial V}{\partial q^i}, \quad 1 \le k \le n.$$

These calculations prove the following.

2.2.2. Proposition. *A smooth curve $\gamma : I \to M$ in a Riemannian manifold M is the projection of an integral curve in TM of the mechanical system with potential energy $V : M \to \mathbb{R}$ if and only if*

$$\nabla_{\dot{\gamma}}\dot{\gamma} = -\mathrm{grad}V.$$

The mechanical system with potential energy $V = 0$ of a Riemannian manifold M is called the *geodesic vector field* of M. The metric on M is by definition *complete* if the geodesic vector field is complete on TM and so defines a flow, called the *geodesic flow* of M. The projected curves on M of the integral curves of the geodesic vector field are the *geodesics*.

2.3 Jacobi's theorem

Let M be a Riemannian manifold with metric g and let $V : M \to \mathbb{R}$ be a smooth function bounded from above. Let $e \in \mathbb{R}$ be such that $V(x) < e$ for every $x \in M$. On M we consider the new Riemannian metric $g_e = (e - V)g$, called the *Jacobi metric*. Let g^* be the induced by $\mathcal{L}$ metric on the fibers of the cotangent bundle

$q : T^*M \to M$. If G is the matrix of g in some local coordinates, then the matrix of g^* is G^{-1}. The induced Jacobi metric is thus

$$g_e^* = \frac{1}{e - V} \cdot g.$$

The mechanical system with potential energy V is equivalent to the Hamiltonian vector field on T^*M with hamiltonian

$$H(a) = \frac{1}{2} g^*(a, a) + V(q(a))$$

and the geodesic vector field of the Jacobi metric is equivalent to the Hamiltonian vector field on T^*M with hamiltonian

$$H_e(a) = \frac{1}{2} \cdot \frac{1}{e - V(q(a))} \cdot g^*(a, a).$$

Observe that $H^{-1}(e) = H_e^{-1}(1)$.

2.3.1. Lemma. *Let (U, ω) be a symplectic vector space and W be a vector subspace of codimension 1. Then, the subspace $K = \{v \in W : \omega(v, w) = 0 \text{ for every } w \in W\}$ is at most 1-dimensional.*

Proof. There exists $u \in U$ such that $U = W \oplus \langle u \rangle$. Since ω is non-degenerate, $\omega(v, u) \neq 0$ for every non-zero $v \in K$. Hence the linear map $\omega(\cdot, u) : K \to \mathbb{R}$ is one-to-one. □

2.3.2. Proposition. *Let (P, ω) be a symplectic manifold and H_1, $H_2 : P \to \mathbb{R}$ be hamiltonians with correspomding Hamiltonian vector fields ξ_1 and ξ_2. If $c_i \in \mathbb{R}$ is a regular value of H_i, for $i = 1, 2$ and $S = H_1^{-1}(c_1) = H_2^{-1}(c_2)$, then there exixts a smooth function $f : S \to \mathbb{R} \setminus \{0\}$ such that $\xi_2 | S = f \cdot (\xi_1 | S)$.*

Proof. For every $x \in S$ and $w \in T_x S$ we have $0 = dH_i(x)w = \omega_x(\xi_i(x), w)$ and so $\xi_1(x)$ and $\xi_2(x)$ are colinear by Lemma 2.3.1, and non-zero because c_i is a regular value of H_i, $i = 1, 2$. Therefore, there exists a function $f : S \to \mathbb{R} \setminus \{0\}$ such that $\xi_2 | S = f \cdot (\xi_1 | S)$, which is easily seen to be smooth. □

2.3.3. Lemma. *Let M be a Riemannian manifold and ξ be the geodesic vector field, which has mechanical energy $E(v) = \frac{1}{2}\|v\|^2$.*

(a) Every $c > 0$ is a regular value of E.

(b) For every c_1, $c_2 > 0$ there exists a diffeomorphism $h : E^{-1}(c_1) \to E^{-1}(c_2)$ such that

$$h_*(\xi | E^{-1}(c_1)) = \sqrt{\frac{c_2}{c_1}} \cdot (\xi | E^{-1}(c_2)).$$

Proof. (a) In local coordinates $(q^1, ..., q., v^1, ..., v^n)$ we have

$$dE = \frac{1}{2} \sum_{i,j,k=1}^{n} \frac{\partial g_{ij}}{\partial q^k} v^i v^j dq^k + \sum_{i,k=1}^{n} g_{ik} v^i dv^k.$$

So for $x \in M$ and $v \in T_xM$ with $E(v) = c$ we have $dE(v)(0, v) = g_x(v, v) = 2c > 0$.

(b) If $h : TM \to TM$ is the diffeomorphism with $h(v) = \big(\sqrt{\frac{c_2}{c_1}}\big)v$, then clearly h satisfies our requirements.□

Summurizing now we have the following.

2.3.4. Theorem (Jacobi). *Let M be a Riemannian manifold and $V : M \to \mathbb{R}$ be a smooth function for which there exists $e \in \mathbb{R}$ such that $V(x) < e$ for every $x \in M$. Let ξ be the mechanical system with potential function V. If e is a regular value of the mechanical energy $E(v) = \frac{1}{2}\|v\|^2 + V(\pi(v))$, then the restricted system on the level of mechanical energy e is a reparametrization of the restricted geodesic flow on the unit tangent bundle of M with respect to the Jacobi metric.*

Recall that by Sard's theorem the set of critical values of a smooth function has Lebesgue measure zero. Thus, if the potential energy is bounded from above, we can always find an upper bound which is a regular value of the mechanical energy. The theorem of Jacobi reduces the theoretical study of mechanical systems to the study of geodesic flows. In the sequel with the term geodesic flow we shall always mean the dynamical system defined on the unit tangent bundle of a complete Riemannian manifold by restriction of the geodesic vector field.

2.4 The Liouville measure

Let M be a complete Riemannian manifold with metric g and ξ be the geodesic vector field on TM. Let $(U, q^1, ..., q^n)$ be a system of local coordinates on M and $(\pi^{-1}(U), q^1, ..., q^n, v^1, ..., v^n)$ be the corresponding system of local coordinates on TM. As we saw in section 2 of this chapter, the symplectic 2-form in these local coordinates of $\pi^{-1}(U)$ is

$$\mathcal{L}^*\omega = \sum_{i,j=1}^{n} g_{ij} dq^i \wedge dv^j + \sum_{i,j,k=1}^{n} \frac{\partial g_{ij}}{\partial q^k} \cdot v^j dq^i \wedge dq^k.$$

So ξ preserves

$$\mathcal{L}^*\omega \wedge ... \wedge \mathcal{L}^*\omega = c(\det G) dq^1 \wedge ... \wedge dq^n \wedge dv^1 \wedge ... \wedge dv^n,$$

where c is a constant depending only on the dimension of M. Let

$$\Omega = \frac{1}{c}\mathcal{L}^*\omega \wedge ... \wedge \mathcal{L}^*\omega.$$

The induced volume element $\tilde{\Omega}$ on the unit tangent bundle T^1M, defines a Borel probability measure preserved by the geodesic flow, called the *Liouville measure* on T^1M. We shall describe locally the volume element $\tilde{\Omega}$ on T^1M. The part of T^1M in the system of local coordinates we consider is described by the equation

$$\sum_{i,j=1}^{n} g_{ij}(q^1, ..., q^n) v^i v^j = 1.$$

For every $(q^1, ..., q^n)$ this equation describes an ellipsoid in $\{(q^1, ..., q^n)\} \times \mathbb{R}^n$. We shall construct a new system of coordinates on $\pi^{-1}(U)$, which converts this ellipsoid to a $(n-1)$-sphere. Since the matrix $G(q^1, ..., q^n)$ is symmetric and positively definite, it has positive eigenvalues and can be diagonalized. More precisely, there exists an orthogonal matrix $A(q^1, ..., q^n) = (a_{ij}(q^1, ..., q^n))$ such that $A^t G A = I_n$. We consider now coordinates $(q^1, ..., q^n, u^1, ..., u^n)$ on $\pi^{-1}(U)$ such that

$$v^i = \sum_{j=1}^{n} a_{ij}(q^1, ..., q^n) u^j,$$

and we have

$$1 = \sum_{i,j=1}^{n} g_{ij} v^i v^j = \sum_{i,j=1}^{n} g_{ij} \left(\sum_{k=1}^{n} a_{ik} u^k \right) \left(\sum_{l=1}^{n} a_{jl} u^l \right) =$$

$$\sum_{k,l=1}^{n} \left(\sum_{i,j=1}^{n} a_{ik} g_{ij} a_{jl} \right) u^k u^l = \sum_{k=1}^{n} (u^k)^2,$$

It follows that in $\pi^{-1}(U)$ we have

$$\Omega = (\det G) dq^1 \wedge ... \wedge dq^n \wedge dv^1 \wedge ... \wedge dv^n =$$

$$(\det G)(\det A) dq^1 \wedge ... \wedge dq^n \wedge du^1 \wedge ... \wedge du^n =$$

$$\sqrt{\det G} dq^1 \wedge ... \wedge dq^n \wedge du^1 \wedge ... \wedge du^n.$$

This means that Ω is locally the product of the Riemannian volume element on M with the euclidean volume element on the fibers of the tangent bundle. Hence the Liouville measure is locally the product of the Riemannian measure with the $(n-1)$-spherical Lebesgue measure on the fibers of the unit tangent bundle.

Chapter 3

Dynamical systems on compact metric spaces

3.1 Invariant measures on compact metric spaces

Let X be a compact metrizable space. Every Borel measure μ on X is regular, that is for every Borel set $B \subset X$ and $\epsilon > 0$ there exist an open set $V \subset X$ and a closed set $C \subset X$ such that $C \subset B \subset V$ and $\mu(V \setminus C) < \epsilon$. Consequently,

$$\mu(B) = \sup\{\mu(C) : C \subset B,\ C \text{ closed in } X\} = \inf\{\mu(V) : B \subset V,\ V \text{ open in } X\}.$$

From the Riesz representation theorem follows that the set $\mathcal{M}(X)$ of all Borel probability measures on X is in a one-to-one, onto correspondence with the set of all positive linear forms $J : C(X) \to \mathbb{R}$ with $J(1) = 1$. The correspondence is defined by the formula

$$J(f) = \int_X f d\mu, \quad f \in C(X), \mu \in \mathcal{M}(X).$$

Since $C(X)$ is separable, $\mathcal{M}(X)$ endowed with the weak topology becomes a compact metrizable space.

3.1.1. Lemma. *Let $T : X \to X$ be a continuous, onto map. A Borel measure μ on X is T-invariant if and only if*

$$\int_X (f \circ T) d\mu = \int_X f d\mu$$

for every $f \in C(X)$.

Proof. If μ is T-invariant, then the conclusion is a direct consequence of the definition of the integral. For the converse, since the measure is regular and T is continuous and onto, it suffices to prove that $\mu(T^{-1}(A)) = \mu(A)$ for every closed set $A \subset X$. Indeed, if A is closed, there exists a decreasing sequence of continuous functions $f_n : X \to [0, 1]$ such that $f_n^{-1}(1) = A$, for every $n \in \mathbb{N}$, which converges pointwise to χ_A. Thus, we have

$$\mu(T^{-1}(A)) = \int_X \chi_{T^{-1}(A)} d\mu = \int_X (\chi_A \circ T) d\mu = \lim_{n \to +\infty} \int_X (f_n \circ T) d\mu =$$

$$\lim_{n\to+\infty}\int_X f_n d\mu = \int_X \chi_A d\mu = \mu(A).\square$$

Every continuous, onto map $T : X \to X$ induces a continuous map $T_* : \mathcal{M}(X) \to \mathcal{M}(X)$ defined by

$$\int_X f dT_*\mu = \int_X (f \circ T)d\mu, \quad f \in C(X),$$

and μ is T-invariant if and only if $T_*\mu = \mu$.

3.1.2. Theorem. *Every continuous, onto map $T : X \to X$ of a compact metrizable space X has a T-invariant Borel probability measure.*

Proof. Let $\mu_0 \in \mathcal{M}(X)$. The sequence

$$\mu_n = \frac{1}{n}\sum_{k=0}^{n-1} T_*^k\mu_0, \quad n \in \mathbb{N},$$

has a weakly convergent subsequence $(\mu_{n_k})_{k\in\mathbb{N}}$ to some $\mu \in \mathcal{M}(X)$, since $\mathcal{M}(X)$ is a compact metrizable space with respect to the weak topology. Then, we have

$$T_*\mu_{n_k} - \mu_{n_k} = \frac{1}{n_k}\big(T_*^{n_k}\mu_0 - \mu_0\big)$$

and for every $f \in C(X)$ we get

$$\Big|\int_X f dT_*^{n_k}\mu_0 - \int_X f d\mu_0\Big| = \Big|\int_X (f \circ T^{n_k})d\mu_0 - \int_X f d\mu_0\Big| \le 2\|f\|.$$

It follows that

$$\Big|\int_X f dT_*\mu - \int_X f d\mu\Big| \le \frac{2\|f\|}{n_k} \to 0,$$

that is $T_*\mu = \mu$. $\square$

In the same way one can prove that every continuous flow $(\phi_t)_{t\in\mathbb{R}}$ on a compact metrizable space X has an invariant Borel probability measure. In this case we get the continuous flow $((\phi_t)_*)_{t\in\mathbb{R}}$ on $\mathcal{M}(X)$ and the measure μ is invariant under the flow on X if and only if it is a fixed point of this flow on $\mathcal{M}(X)$. The proof runs along the lines of the proof of 3.1.2, considering the sequence

$$\mu_n = \frac{1}{n}\int_0^{n-1} (\phi_t)_*\mu_0 dt, \quad n \in \mathbb{N}.$$

It is obvious that the support of every flow invariant measure is a closed invariant subset of X and is contained in the Birkhoff center, by 1.2.4.

In the sequel we shall denote by $\mathcal{M}_T(X)$ the set of T-invariant Borel probability measures of a continuous, onto map T and by $\mathcal{M}_\phi(X)$ the set of invariant Borel probability measures of a flow $(\phi_t)_{t\in\mathbb{R}}$. In both cases it is evident that we have a weakly compact convex set.

3.2 Uniquely ergodic dynamical systems

Let X be a compact metrizable space. A continuous, onto map $T : X \to X$ is called *uniquely ergodic* if there a unique T-invariant Borel probability measure, that is $\mathcal{M}_T(X)$ is a singleton. Similarly a continuous flow is called uniquely ergodic if it has a unique invariant Borel probability measure.

3.2.1. Theorem. *For a continuous, onto map $T : X \to X$ of a compact metrizable space X, the following assertions are equivalent :*

(i) T is uniquely ergodic.

(ii) For every $f \in C(X)$ the sequence of continuous functions

$$\frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k, \quad n \in \mathbb{N},$$

converges uniformly to a constant (the integral of f with respect to the unique invariant measure).

(iii) For every $f \in C(X)$ the sequence of continuous functions

$$\frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k, \quad n \in \mathbb{N},$$

converges pointwise to a constant.

Proof. Suppose first that T is uniquely ergodic and μ is the unique T-invariant Borel probability measure. We shall prove (ii) by contradiction. If (ii) is not true, there exists some $f \in C(X)$ for which there are $\epsilon > 0$, $n_k \to +\infty$ and points $x_k \in X$, $k \in \mathbb{N}$ such that

$$\Big|\frac{1}{n_k}\sum_{i=0}^{n_k-1} f \circ T^i(x_k) - \int_X f d\mu\Big| \geq \epsilon, \quad k \in \mathbb{N}.$$

For every $k \in \mathbb{N}$, the combination of Dirac point measures

$$\mu_k = \frac{1}{n_k}\sum_{i=0}^{n_k-1} \delta_{T^i(x_k)}$$

is an element of $\mathcal{M}(X)$. So we may assume that the sequence $(\mu_k)_{k\in\mathbb{N}}$ converges to some $\nu \in \mathcal{M}(X)$. Obviously,

$$\Big|\int_X f d\nu - \int_X f d\mu\Big| \geq \epsilon,$$

and therefore $\nu \neq \mu$. However, for every $g \in C(X)$ we have

$$\int_X (g \circ T) d\nu = \lim_{k\to+\infty} \frac{1}{n_k}\sum_{i=1}^{n_k} (g \circ T^i)(x_k)$$

and

$$\frac{1}{n_k}\sum_{i=1}^{n_k}(g\circ T^i)(x_k)=\int_X g d\mu_k+\frac{1}{n_k}\big[(g\circ T^{n_k}(x_k)-g(x_k)\big],$$

while

$$\frac{1}{n_k}\big|(g\circ T^{n_k}(x_k)-g(x_k)\big|\leq\frac{2\|g\|}{n_k}\to 0.$$

Hence

$$\int_X(g\circ T)d\nu=\lim_{k\to+\infty}\int_X g d\mu_k=\int_X g d\nu,$$

which means that ν is another T-invariant Borel probability measure. It remains to prove that if (iii) is true, then T is uniquely ergodic. If (iii) is true, then by the Riesz representation theorem there exists a Borel probability measure μ such that

$$\int_X f d\mu=\lim_{n\to+\infty}\frac{1}{n}\sum_{k=0}^{n-1}f\circ T^k$$

for every $f\in C(X)$. Obviously, μ is T-invariant. Let now $\nu\in\mathcal{M}_T(X)$. For every $f\in C(X)$ and $n\in\mathbb{N}$ we have

$$\int_X f d\nu=\int_X\left(\frac{1}{n}\sum_{k=0}^{n-1}f\circ T^k\right)d\nu.$$

Since

$$\big|\frac{1}{n}\sum_{k=0}^{n-1}(f\circ T^k)(x)\big|\leq\|f\|,$$

for every $x\in X$, by dominated convergence we have

$$\int_X f d\nu=\int_X\left(\lim_{n\to+\infty}\frac{1}{n}\sum_{k=0}^{n-1}f\circ T^k\right)d\nu=\int_X\left(\int_X f d\mu\right)d\nu=\int_X f d\mu.\square$$

In the same way, replacing the sums with Riemann integrals and using Fubini's theorem one can prove the following.

3.2.2. Theorem. *For a continuous flow $(\phi_t)_{t\in\mathbb{R}}$ on a compact metrizable space X the following are equivalent :*

(i) The flow is uniquely ergodic.

(ii) For every $f\in C(X)$ we have

$$\lim_{t\to\pm\infty}\frac{1}{t}\int_0^t(f\circ\phi_s)ds=\int_X f d\mu$$

uniformly on X.

(iii) For every $f\in C(X)$ there is a constant $c\in\mathbb{R}$ such that for every $x\in X$ we have

$$\lim_{t\to\pm\infty}\frac{1}{t}\int_0^t f(\phi_s(x))ds=c.$$

The topological behavior of uniquely ergodic flows can be described as follows.

3.2.3. Proposition. *Let $(\phi_t)_{t\in\mathbb{R}}$ be a uniquely ergodic flow on a compact metrizable space X, with invariant Borel probability measure μ with support* $\mathrm{supp}\mu$. *Then,*

(i) $\mathrm{supp}\mu \subset L^+(x) \cap L^-(x)$ *for every* $x \in X$,

(ii) $\mathrm{supp}\mu$ *is a minimal set, that is it is non-empty, closed invariant and has no proper subset with these properties, and*

(iii) the restricted flow on $\mathrm{supp}\mu$ *is uniquely ergodic.*

Proof. Let $z \in \mathrm{supp}\mu$ and W be an open neighbourhood of z in X. Since $\mu(W) > 0$ and the measure is regular, there is some $C \subset W$ closed in X with $\mu(C) > 0$. There exists a continuous function $f : X \to [0,1]$ such that $f^{-1}(1) = C$ and $f^{-1}(0) = X \setminus W$. For every $x \in X$ we have from 3.2.2,

$$\lim_{t\to\pm\infty} \frac{1}{t}\int_0^t f(\phi_s(x))ds = \int_X f d\mu \geq \int_C f d\mu = \mu(C) > 0.$$

So, for every $t_0 > 0$ there are $s < -t_0 < 0 < t_0 < t$ such that $\phi_s(x)$, $\phi_t(x) \in W$. This proves (i), while (ii) and (iii) are immediate consequences. □

A similar proposition is true for uniquely ergodic homeomorphisms. Recall that a closed invariant set is minimal if and only if every orbit in it is dense. In view of 3.2.3, uniquely ergodic dynamical systems with dense orbits are of particular interest.

3.2.4. Theorem. *Let $T : X \to X$ be a continuous, onto map of the compact metrizable space X. If*

(i) the sequence $\{T^k : k \in \mathbb{Z}^+\}$ is equicontinuous, and

(ii) there exists some $x_0 \in X$ such that $\overline{\{T^k(x_0) : k \in \mathbb{Z}^+\}} = X$,

then T is uniquely ergodic.

Proof. For every $f \in C(X)$ the sequence

$$f_n = \frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k, \quad n \in \mathbb{N},$$

is equicontinuous and uniformly bounded by $\|f\|$. Thus, from Ascoli's theorem, there exists a subsequence $(f_{n_k})_{k\in\mathbb{N}}$ which converges uniformly to some $g \in C(X)$. Then,

$$g(T(x)) = \lim_{k\to+\infty} \frac{1}{n_k}\sum_{i=1}^{n_k} (f \circ T^i)(x) = g(x),$$

for every $x \in X$. Our assumption (ii) implies now that g must be constant on X. For every $\mu \in \mathcal{M}_T(X)$ we have

$$\int_X f d\mu = \lim_{k\to+\infty} \frac{1}{n_k}\sum_{i=0}^{n_k-1} \int_X (f \circ T^i) d\mu = \lim_{k\to+\infty} \int_X f_{n_k} d\mu = g(x_0).$$

This proves that $\mathcal{M}_T(X)$ is a singleton. $\square$

Note that if T is a homeomorphism, the sequence $\{T^k : k \in \mathbb{Z}\}$ is equicontinuous if and only if T is an isometry with respect to some metric compatible with the topology of X. Of course, a similar statement like 3.2.4 is true in the case of continuous flows.

Let now G be a 1st countable, compact (Hausdorff) topological group. Then G is metrizable and there is a compatible metric which is invariant under left and right translations. For instance, in the case of the n-torus $T^n = S^1 \times ... \times S^1$ the invariant metric is

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|,$$

where $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$. So, for any $g \in G$, if $T_g : G \to G$ is the left translation $T_g(x) = gx$, then the sequence $\{T_g^k : k \in \mathbb{Z}\}$ is equicontinuous and leaves the Haar measure invariant.

3.2.5. Corollary. *A left translation of a compact, mertizable topological group G is uniquely ergodic if and only if it has a dense orbit in G.*

Proof. If a left translation is uniquely ergodic, then the support of the Haar measure is a minimal set, by 3.2.3. Therefore, every orbit is dense in G. The converse follows immediately form 3.2.4. $\square$

Note that if a left translation of a compact metrizable topological group has a dense orbit, then every orbit is dense and the group must be abelian. In the case of the torus we have the following.

3.2.6. Theorem (Kronecker). *If the real numbers 1, a_1,...,a_k are linearly independent over $\mathbb{Q}$, then every orbit of the translation*

$$T(e^{2\pi i x_1}, ..., e^{2\pi i x_k}) = (e^{2\pi i (x_1 + a_1)}, ..., e^{2\pi i (x_n + a_k)})$$

is dense in the k-torus and so T is uniquely ergodic.

We shall give an elementary proof which makes use of a couple of lemmas.

3.2.7. Lemma. *If a_1,...,a_k are irrational numbers, then for every $\epsilon > 0$ there exist $s \in \mathbb{N}$ and b_1,...,$b_k \in \mathbb{Z}$ such that $|sa_i - b_i| < \epsilon$, $1 \leq i \leq k$.*

Proof. Let $n \in \mathbb{N}$ be such that $1/n < \epsilon$. We consider the partition of the cube $[0, 1]^k$ into n^k subcubes with edges of length $1/n$. The points $(ma_1 - [ma_1], ..., ma_k - [ma_k])$, $0 \leq m \leq n^k$, are $n^k + 1$, and thus at least two of them belong to the same cube of the partition, that is there are $0 \leq m_2 < m_1 \leq n^k$ such that

$$|(m_1 - m_2)a_i - ([m_1 a_i] - [m_2 a_i])| < \frac{1}{n} < \epsilon, \quad 1 \leq i \leq k.$$

It suffices to take now $s = m_1 - m_2$ and $b_i = [m_1 a_i] - [m_2 a_i]$, $1 \leq i \leq k$. $\square$

3.2.8. Lemma. *If a is an irrational number, then for every $\epsilon > 0$ and $x \in \mathbb{R}$ there exist n, $m \in \mathbb{Z}$ such that $|na - m - x| < \epsilon$.*

Proof. By 3.2.7, for every $k \in \mathbb{N}$ there are $s_k \in \mathbb{N}$ and $b_k \in \mathbb{Z}$ such that $\lim_{k\to+\infty}(s_k a + b_k) = 0$. If $t_k = s_k a + b_k$, then $t_k \neq 0$ for every $k \in \mathbb{N}$, because a is irrational. Dividing x with t_k, we find some $l_k \in \mathbb{Z}$ such that $|x - l_k t_k| < |t_k|$ and therefore $\lim_{k\to+\infty} |l_k t_k - x| = 0$. Since $l_k t_k = (l_k s_k)a + (l_k b_k)$, we have the conclusion. □

Proof of 3.2.6. It suffices to prove that the orbit of the point $(1, ..., 1)$ is dense in T^k. We perform induction on k. The case $k = 1$ is precisely 3.2.8. Suppose that we have proved the theorem in dimension $k-1$. Our assumption says in particular that a_1,...,a_k are irrational. By 3.2.7, for every $\epsilon > 0$ there exist $s \in \mathbb{N}$ and b_1,...,$b_k \in \mathbb{Z}$ such that $|sa_i - b_i| < \epsilon/2$, $1 \leq i \leq k$. If

$$a'_i = \frac{sa_i - b_i}{sa_k - b_k}, \quad 1 \leq i \leq k,$$

then $a'_k = 1$ and a'_1,...,$a'_{k-1}, 1$ are linearly independent over $\mathbb{Q}$. Thus, by the induction hypothesis, for every x_1,...,$x_k \in \mathbb{R}$ there are c_1,...$c_k \in \mathbb{Z}$ such that

$$|c_k a'_i - c_i - (x_i - x_k a'_i)| < \frac{\epsilon}{2}, \quad 1 \leq i \leq k-1.$$

Substituting a'_i we find

$$\left| \left(\frac{c_k + x_k}{sa_k - b_k} \right)(sa_i - b_i) - c_i - x_i \right| < \frac{\epsilon}{2}, \quad 1 \leq i \leq k,$$

because and for $i = k$ we have $c_k a'_k - c_k - (x_k - x_k a'_k) = 0$, since $a'_k = 1$. Let now $N \in \mathbb{Z}$ be such that

$$\left| N - \frac{c_k + x_k}{sa_k - b_k} \right| < 1,$$

and set $n = sN$ and $m_i = Nb_i + c_i$, $1 \leq i \leq k$. Then,

$$|na_i - m_i - x_i| = |N(sa_i - b_i) - c_i - x_i|$$

and

$$|N(sa_i - b_i) - \frac{c_k + x_k}{sa_k - b_k} \cdot (sa_i - b_i)| < |sa_i - b_i|.$$

It follows that

$$|N(sa_i - b_i) - c_i - x_i| \leq |sa_i - b_i| + \left| \frac{c_k + x_k}{sa_k - b_k} \cdot (sa_i - b_i) - c_i - x_i \right| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

for every $1 \leq i \leq k$. This proves the theorem. □

3.2.9. Corollary. *If a is irrational, then the rotation $r_a(z) = e^{2\pi i a} z$ of the unit circle $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ is uniquely ergodic.*

A rational rotation of the circle is never uniquely ergodic, because every point is then periodic, and so the normalized sum of the Dirac point measures of the

points of a periodic orbit is another invariant Borel probability measure, apart from the Haar measure. In the next sections of this chapter we shall find more general homeomorphisms of the circle that are uniquely ergodic, not necessarily preserving the Haar measure.

We end this section remarking that the property of unique ergodicity is invariant under topological conjugancies. Recall that two continuous onto maps $T_1, T_2 : X \to X$ of the compact metrizable space X are *topologically conjugate* if there exists a homeomorphism $h : X \to X$ such that $T_1 \circ h = h \circ T_2$. The homeomorphism h is called *topological conjugation*. If h is only continuous and onto, then the maps are called *semi-conjugate* and h is caled *semi-conjugation*.

3.3 Homeomorphisms of the circle

Let $f : S^1 \to S^1$ be a homeomorphism. There is then a homeomorphism $F : \mathbb{R} \to \mathbb{R}$ such that $f(e^{2\pi it}) = e^{2\pi iF(t)}$ for every $t \in \mathbb{R}$. Such an F is called a *lift* of f. Clearly, any two lifts of f differ by an integer. The original homeomorphism f is orientation preserving if and only if F is increasing, and orientation reversing if F is decreasing. It is easy to see that in the later case $F(t + k) = F(t) - k$ for every $k \in \mathbb{Z}$, and f has exactly two fixed points. We shall be concerned exclusively with orientation preserving homeomorphisms f of S^1. Then $F(t + k) = F(t) + k$ for every $k \in \mathbb{Z}$ or equivalently $F - id$ is periodic with period 1. So we have a well defined continuous function $\psi : S^1 \to \mathbb{R}$ with $\psi(e^{2\pi it}) = F(t) - t$, the displacement function.

3.3.1. Lemma. *If $a = \min\{\psi(z) : z \in S^1\}$ and $b = \max\{\psi(z) : z \in S^1\}$, then $b - a < 1$.*

Proof. If $s, t \in \mathbb{R}$ and $s \leq t < s + 1$, then

$$\psi(e^{2\pi is}) - \psi(e^{2\pi it}) = F(s) - s - F(t) + t \leq t - s < 1,$$

because F is increasing. Therefore, $\psi(e^{2\pi is}) < 1 + \psi(e^{2\pi it})$ for every $t \in [s, s + 1)$. Consequently, $\psi(e^{2\pi is}) < 1 + a$ for every $s \in \mathbb{R}$, and so $b < 1 + a$. □

3.3.2. Proposition (Poincaré). *There exists a constant $\rho(F) \in \mathbb{R}$ such that*

$$\lim_{n\to+\infty} \frac{1}{n}(F^n - id) = \rho(F)$$

uniformly on $\mathbb{R}$.

Proof. Let $\mu \in \mathcal{M}_f(S^1)$ and $\psi_n : S^1 \to \mathbb{R}$ be the induced continuous function

$$\psi_n(e^{2\pi it}) = \frac{1}{n}(F^n(t) - t).$$

Then, $\psi = \psi_1$ and

$$\frac{1}{n}\sum_{k=0}^{n-1} (\psi \circ f^k)(e^{2\pi it}) = \frac{1}{n}\sum_{k=0}^{n-1} \psi(e^{2\pi iF^k(t)}) =$$

$$\frac{1}{n}\sum_{k=0}^{n-1}(F-id)(F^k(t)) = \frac{1}{n}\sum_{k=0}^{n-1}F^{k+1}(t)-F^k(t) = \frac{1}{n}(F^n(t)-t) = \psi_n(e^{2\pi it}).$$

Thus, the integral of ψ_n is equal to the integral of ψ and

$$\int_{S^1}\left(n\psi_n - n\int_{S^1}\psi d\mu\right)d\mu = 0.$$

Applying now 3.3.1 for f^n, which lifts to F^n with displacement function $n\psi_n$, we get

$$\min\{n\psi_n(z) - n\int_{S^1}\psi d\mu : z\in S^1\} > \|n\psi_n - n\int_{S^1}\psi d\mu\| - 1,$$

and therefore

$$\|\psi_n - \int_{S^1}\psi d\mu\| < \frac{1}{n},$$

for every $n\in\mathbb{N}$. Hence

$$\lim_{n\to+\infty}\psi_n = \int_{S^1}\psi d\mu$$

uniformly on S^1. □

3.3.3. Remarks. (a) As the proof of 3.3.2 shows, for every $\mu\in\mathcal{M}_f(S^1)$ we have

$$\rho(F) = \int_{S^1}\psi d\mu.$$

(b) $\|F^n - id - n\rho(F)\| < 1$ for every $n\in\mathbb{N}$.
(c) If $a=\rho(F)$, there exists some $t_0\in\mathbb{R}$ such that

$$F^n(t_0) - t_0 - na = n\psi_n(e^{2\pi it_0}) - n\int_{S^1}\psi d\mu = 0.$$

So $F^n(t_0) = R_{na}(t_0)$, or in other words $R_{-na}\circ F^n$ has a fixed point t_0, where $R_{na}:\mathbb{R}\to\mathbb{R}$ is the translation $R_{na}(t) = t+na$.
(d) For every $a\in\mathbb{R}$ we have $\rho(R_a) = a$.
(e) Since $R_k\circ F = F\circ R_k$ for every $k\in\mathbb{Z}$, we have

$$\frac{(R_k\circ F)^n - id}{n} = \frac{R_{nk}\circ F^n - id}{n} = \frac{F^n - id + nk}{n} \to \rho(F) + k.$$

It follows from 3.3.3(e) that the number $\rho(f) = e^{2\pi i\rho(F)}\in S^1$ does not depend on the particular lift F of f. It is called the *Poincaré rotation number* of f.

3.3.4. Proposition. *An orientation preserving homeomorphism $f: S^1\to S^1$ has a periodic orbit if and only if $\rho(f)\in\mathbb{Q}/\mathbb{Z}$.*

Proof. Let F be a lift of f. If $z_0 = e^{2\pi it_0}$ is a periodic point of f of period q, then $z_0 = f^q(e^{2\pi it_0}) = e^{2\pi iF^q(t_0)}$, and therefore $p = F^q(t_0) - t_0\in\mathbb{Z}$. So we have

$$\rho(F) = \lim_{n\to+\infty}\frac{F^{nq}(t_0)-t_0}{nq} = \lim_{n\to+\infty}\frac{np}{nq} = \frac{p}{q}.$$

Conversely, if $\rho(F) = p/q \in \mathbb{Q}$, then from 3.3.3(c), $R_{-p} \circ F^q$ has a fixed point $t_0 \in \mathbb{R}$ or equivalently $F^q(t_0) = t_0 + p$. $\square$

As in the case of flows, if $f : X \to X$ is a homeomorphism of a metric space X, the set

$$L^+(x) = \{y \in X : f^{n_k}(x) \to y \text{ for some } n_k \to +\infty\}$$

is called the *positive limit set* of the point $x \in X$, and is a closed invariant set. Similarly, the *negative limit set* $L^-(x)$ is defined and has the same properties.

3.3.5. Proposition. *If the orientation preserving homeomorphism $f : S^1 \to S^1$ has irrational rotation number, then there exists a compact f-invariant set $K \subset S^1$ with the following properties.*

(i) $L^+(x) = L^-(x) = K$ for every $x \in S^1$, and in particular K is minimal.

(ii) Either $K = S^1$ or K is a Cantor set.

(iii) $\mathrm{supp}\mu = K$ *for every f-invariant Borel probability measure.*

Proof. Let $x \in S^1$ and $K = L^+(x)$. Since K is closed and invariant, we have $L^+(y) \cup L^-(y) \subset K$ for every $y \in K$. The connected components I_n, $n \in \mathbb{Z}$, of $S^1 \setminus K$ are permuted by f. Let now $y \in S^1 \setminus K$. If $L^+(y) \cap (S^1 \setminus K) \neq \varnothing$, there are some n, k, $l \in \mathbb{Z}$ with $k > l$ such that $f^k(y)$, $f^l(y) \in I_n$. This means that $y \in f^{-k}(I_n) \cap f^{-l}(I_n)$ and therefore $f^{k-l}(I_n) \cap I_n \neq \varnothing$. Then, $f^{k-l}(\bar{I}_n) = \bar{I}_n$, and from the intermediate value theorem f^{k-l} must have a fixed point in $\bar{I}_n$. This contradicts 3.3.4, since f is supposed to have irrational rotation number. Hence $L^+(y) \subset K$ and similarly $L^-(y) \subset K$ for every $y \in S^1$. In other words, we have shown that $L^+(y) \cup L^-(y) \subset L^+(x)$ for every x, $y \in S^1$ and similarly $L^+(y) \cup L^-(y) \subset L^-(x)$. Thus $L^+(y) \cup L^-(y) \subset L^+(x) \cap L^-(x)$ for every x, $y \in S^1$, and symmetrically we get

$$L^+(x) \cup L^-(x) \subset L^+(y) \cap L^-(y) \subset L^+(y) \cup L^-(y) \subset L^+(x) \cap L^-(x)$$

for every x, $y \in S^1$. Hence $K = L^+(y) = L^-(y) = L^+(x) = L^-(x)$ for every x, $y \in S^1$. It is clear now that K is a perfect set. If K is not totally disconnected, it contains an open interval $J \subset S^1$. Then, for every $x \in S^1$ there exists $n \in \mathbb{Z}$ such that $f^n(x) \in J$, that is $x \in f^{-n}(J) \subset K$. This shows that $K = S^1$, if it is not a Cantor set. Obviously, $K = \{x \in S^1 : x \in L^+(x)\}$, and so from Poincaré's recurrence theorem we have $\mathrm{supp}\mu \subset K$ for every $\mu \in \mathcal{M}_f(S^1)$. Since K is minimal, we must have equality. $\square$

3.3.6. Lemma. *Let $f : S^1 \to S^1$ be an orientation preserving homeomorphism with irrational rotation number, F be a lift of f and $a = \rho(F)$. If $t \in \mathbb{R}$ and $C(t) = \{F^n(t) + m : n, m \in \mathbb{Z}\}$, then the function $F_t : C(t) \to \mathbb{Z} + a\mathbb{Z}$ with*

$$F_t(F^n(t) + m) = m + an$$

is strictly increasing, onto.

Proof. If $F^n(t) + m < F^k(t) + l$, then $F^{n-k}(t) < t + l - m$ and therefore

$$F^{2(n-k)}(t) < F^{n-k}(t + l - m) = F^{n-k}(t) + (l - m) < t + 2(l - m).$$

Inductively now we have

$$F^{q(n-k)}(t) < F^{n-k} + (q-1)(l-m) < t + q(l-m)$$

for every $q \in \mathbb{N}$. Dividing by q and taking the limit for $q \to +\infty$, we find $(n-k)a \leq l-m$. Since a is irrational, we must have $(n-k)a < l-m$. □

3.3.7. Theorem (Poincaré). *If $f : S^1 \to S^1$ is an orientation preserving homeomorphism with irrational rotation number $\rho(f) = e^{2\pi i a}$, there exists an orientation preserving, continuous, onto map $h : S^1 \to S^1$ such that $h \circ f = r_a \circ h$. If f has a dense orbit in S^1, then h is a homeomorphism.*

Proof. Let K be the unique minimal set of f given by 3.3.5, and let $z_0 = e^{2\pi i t_0} \in K$. Let F be a lift of f. If C is the orbit of z_0, the function $H : p^{-1}(C) \to \mathbb{Z} + a\mathbb{Z}$ with $H(F^n(t_0)+m) = m+an$ is a bijection, by 3.3.6, where $p : \mathbb{R} \to S^1$ is the exponential map $p(t) = e^{2\pi i t}$. Moreover, $H(F^n(t_0)+m+1) = m+1+an = H(F^n(t_0)+m)+1$, and

$$H(F(F^n(t_0)+m)) = H(F^{n+1}(t_0)+m) = m+(n+1)a = R_a(H(F^n(t_0)+m)),$$

or in other words $H \circ F = R_a \circ H$. We extend H to $\overline{p^{-1}(C)}$ setting

$$H(s) = \lim_{t \in p^{-1}(C), t \to s} H(t).$$

The right and left limits exist due to the monotonicity of H, and they are equal because $\mathbb{Z} + a\mathbb{Z}$ is dense in $\mathbb{R}$, since a is irrational. Th function H is continuous and increasing, but not necessarily strictly. Indeed, if I is a connected component of $\mathbb{R} \setminus \overline{p^{-1}(C)}$, then H takes the same value at the endpoints of I. We can extend now H continuously on $\mathbb{R}$ requiring H to take on a connected component I of $\mathbb{R} \setminus \overline{p^{-1}(C)}$ the value it takes at its endpoints. Thus, we get a continuous, onto, increasing map $H : \mathbb{R} \to \mathbb{R}$ such that $H(t+1) = H(t)+1$ for every $t \in \mathbb{R}$ and $H \circ F = R_a \circ H$. Therefore, $h : S^1 \to S^1$ defined by $h(e^{2\pi i t}) = e^{2\pi i H(t)}$ is continuous, onto, preserves the orientation of S^1 and $h \circ f = r_a \circ h$. Moreover, $h(K) = h(f(K)) = r_a(h(K))$, from which follows that $h(K) = S^1$. It is evident from the construction of H that if $K = S^1$, then H is strictly increasing and hence h is a homeomorphism. □

3.3.8. Theorem. *An orientation preserving homeomorphism $f : S^1 \to S^1$ is uniquely ergodic if it has irrational rotation number.*

Proof. Let $\mu \in \mathcal{M}_f(S^1)$. According to 3.3.7, there exists an orientation preserving, onto, continuous map $h : S^1 \to S^1$ such that $h \circ f = r_a \circ h$, where $\rho(f) = e^{2\pi i a}$. For every $g \in C(S^1)$ we have

$$\int_{S^1} (g \circ r_a) dh_* \mu = \int_{S^1} (g \circ r_a \circ h) d\mu = \int_{S^1} (g \circ h \circ f) d\mu =$$

$$\int_{S^1} (g \circ h) df_* \mu = \int_{S^1} (g \circ h) d\mu = \int_{S^1} g dh_* \mu,$$

which means that $h_*\mu$ is invariant under r_a. Hence $h_*\mu$ is the normalized Lebesgue measure, by 3.2.9. If $K = S^1$, then h is a homeomorphism and so μ is unique. Suppose that $K \neq S^1$. and $\mu_1,\ \mu_2 \in \mathcal{M}_f(S^1)$. According to the above, $h_*\mu_1 = h_*\mu_2 = h_*\mu$, where $\mu = \frac{1}{2}(\mu_1 + \mu_2)$. It suffices to prove that $\mu_1(I) = \mu_2(I)$ for every open interval $I = (t, s)$ in S^1. First we observe that for every Borel set $B \subset S^1$ we have

$$\mu_1(h^{-1}(B)) = h_*\mu_1(B) = h_*\mu_2(B) = \mu_2(h^{-1}(B)) = \mu(h^{-1}(B)).$$

The set $J = h(I)$ is an interval or a singleton. If $I' = h^{-1}(J)$, then I' is an interval with endpoints $t' < s'$ containing I. Since $h(I) = h(I')$ and h is monotonous, we have $h(t) = h(t')$ and $h(s) = h(s')$, which implies that $(t', t) \cup (s, s') \subset S^1 \setminus K$. Consequently,

$$\mu(I' \setminus I) = \mu((t', t) \cup (s, s')) = 0,$$

and therefore $\mu_1(I' \setminus I) = \mu_2(I' \setminus I) = 0$. It follows that

$$\mu_1(I) = \mu_1(I') = \mu_1(h^{-1}(J)) = \mu_2(h^{-1}(J)) = \mu_2(I') = \mu_2(I). \quad \square$$

3.4 Denjoy's theorem

In this section and the next we shall study the bevavior of sufficiently smooth diffeomorphisms of the circle. Let $f : S^1 \to S^1$ be an orientation preserving C^2 diffeomorphism and F be a lift of f. Then, $DF(t) > 0$ and $DF(t - [t]) = DF(t)$ for every $t \in \mathbb{R}$. Let

$$c = \sup\{\frac{|D^2F(t)|}{DF(t)} : t \in [0, 1]\}.$$

3.4.1. Lemma. *If $t,\ s \in \mathbb{R}$ and $t < s$, then*

$$\left|\log \frac{DF^n(t)}{DF^n(s)}\right| \leq c \cdot \sum_{k=0}^{n-1} |F^k(t) - F^k(s)|$$

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

Proof. From the chain rule we have

$$DF^n(t) = \prod_{k=0}^{n-1} DF(F^k(t)),$$

and by the mean value theorem

$$\left|\log \frac{DF^n(t)}{DF^n(s)}\right| \leq \sum_{k=0}^{n-1} |\log DF(F^k(t)) - \log DF(F^k(s))| \leq c \cdot \sum_{k=0}^{n-1} |F^k(t) - F^k(s)|. \square$$

The goal of this section is to prove the following.

3.4.2. Theorem (Denjoy). *An orientation preserving C^2 diffeomorphism $f : S^1 \to S^1$ with irrational rotation number $\rho(f) = e^{2\pi i a}$ is topologically conjugate to the rotation r_a.*

Proof. In view of the results of the preceding section, it suffices to prove that the unique minimal set K of f is equal to S^1. We proceed to prove this by contradiction. So suppose that $K \neq S^1$ and let $I = p((t_0, s_0))$ be a connected component of $S^1 \setminus K$, where p is the exponential map. Let l_n be the length of the interval $p^{-1}(f^n(I)) \cap [0,1]$. For every t, $s \in [t_0, s_0]$ we have

$$\left|\log \frac{DF^n(t)}{DF^n(s)}\right| \leq c \sum_{k=0}^{n-1} l_k \leq c$$

from 3.4.1. It follows that $DF^n(t) \leq e^c DF^n(s)$ for every t, $s \in [t_0, s_0]$. By the mean value theorem we get

$$DF^n(t) \leq e^c \cdot \frac{l_n}{l_0}$$

for every $t \in [t_0, s_0]$. Since f has no periodic point, the intervals $p^{-1}(f^n(I)) \cap [0,1]$, $n \in \mathbb{Z}$, are disjoint and therefore

$$\sum_{n \in \mathbb{Z}} l_n \leq 1.$$

Hence $\lim_{n \to \pm\infty} l_n = 0$ and $\lim_{n \to \pm\infty} DF^n = 0$ uniformly on $[t_0, s_0]$. Let $d = l_0 / c e^{c+1}$. Then for every $n \geq 0$ and every $t_0 - d < t < t_0$ we have

$$DF^n(t) \leq e DF^n(t_0).$$

Indeed, this is trivial for $n = 0$. By induction, suppose that we have proven it for all $0 \leq k < n$. By 3.4.1, for all $t_0 - d \leq t < t_0$ we have

$$\left|\log \frac{DF^n(t)}{DF^n(t_0)}\right| \leq c \cdot \sum_{k=0}^{n-1} |F^k(t) - F^k(t_0)|.$$

From the mean value theorem, there exist $u_k \in (t, t_0)$ such that

$$\left|\log \frac{DF^n(t)}{DF^n(t_0)}\right| \leq cd \sum_{k=0}^{n-1} DF^k(u_k),$$

and by the induction hypothesis,

$$\left|\log \frac{DF^n(t)}{DF^n(t_0)}\right| \leq cde \sum_{k=0}^{n-1} DF^k(t_0) \leq cde^{c+1} \sum_{k=0}^{n-1} \frac{l_k}{l_0} \leq \frac{1}{l_0} cde^{c+1} = 1,$$

from which the inequality follows. The above imply now that

$$DF^n(t) \leq e^{c+1} \frac{l_n}{l_0}$$

for every $n \in \mathbb{N}$ and $t \in [t_0 - d, s_0]$, and thus $DF^n \to 0$ uniformly on $[t_0 - d, s_0]$. Let $z_0 = e^{2\pi i t_0}$ and $n_k \to +\infty$ be such that $f^{n_k}(z_0) \to z_0$. There exists some $k \in \mathbb{N}$ such that $DF^{n_k}(t) < 1/2$ for every $t \in [t_0 - d, s]$ and $f^{n_k}(z_0) \in p((t_0 - \frac{d}{2}, s_0))$. Then,

$$|F^{n_k}(t) - F^{n_k}(t_0)| < \frac{d}{2}$$

for every $t \in [t_0 - d, t_0]$, by the mean value theorem, and there exists $m \in \mathbb{Z}$ such that

$$|F^{n_k}(t_0) + m - t_0| < \frac{d}{2}.$$

Thus, $|F^{n_k}(t) + m - t_0| < d$ for every $t \in [t_0 - d, t_0]$. This means that if $J = p([t_0 - d, t_0])$, then $f^{n_k}(J) \subset J$. Since J is an interval, f^{n_k} must have a fixed point in J. This contradiction proves the theorem. □

3.5 C^1 diffeomorphisms of Denjoy

In this section we shall show that 3.4.2 is not true for C^1 diffeomorphisms by constructing an orientation preserving C^1 diffeomorphism $f : S^1 \to S^1$ with irrational rotation number which is not topologically conjugate to a rotation.

Let $a \in \mathbb{R} \setminus \mathbb{Q}$ and $t_0 \in \mathbb{R} \setminus (\mathbb{Z} + a\mathbb{Z})$. Since $\mathbb{Z} + a\mathbb{Z}$ is dense in $\mathbb{R}$, the same is true for $t_0 + \mathbb{Z} + a\mathbb{Z}$. Let $l_n > 0$, $n \in \mathbb{Z}$, be such that $\sum_{n\in\mathbb{Z}} l_n = \rho$, where $0 < \rho \leq 1$. For instance, $l_n = \frac{\rho}{\pi}(\arctan(n+1) - \arctan n)$. We consider the functions $q : \mathbb{R} \to \mathbb{R}^+$ with

$$q(t) = \begin{cases} 0, & \text{if } t \notin t_0 + \mathbb{Z} + a\mathbb{Z} \\ l_n, & \text{if } t = t_0 + m + an \text{ for some } m, n \in \mathbb{Z}. \end{cases}$$

and $J : \mathbb{R} \to \mathbb{R}$ defined by

$$J(t) = \begin{cases} (1-\rho)t + \sum_{0\leq s\leq t} q(s), & \text{if } t \geq 0 \\ (1-\rho)t - \sum_{t<s\leq 0} q(s), & \text{if } t < 0\ . \end{cases}$$

3.5.1. Lemma *The function J is strictly increasing, continuous except at the points of the set $t_0 + \mathbb{Z} + a\mathbb{Z}$, where it is only right continuous and from the left has jump l_n at the point $t_0 + m + an$. Moreover, it has the following properties.*

(i) $J(0) = 0$, $J(t+1) = J(t) + 1$ for every $t \in \mathbb{R}$, and so $J(k) = k$ for $k \in \mathbb{Z}$.

(ii) The set $C = \overline{J(\mathbb{R})}$ is closed, perfect, totally disconnected and $R_1(C) = C$, where $R_1 : \mathbb{R} \to \mathbb{R}$ is the translation $R_1(t) = t + 1$.

(iii) $\mu(C \cap [0, 1]) = 1 - \rho$, where μ is the Lebesgue measure.

Proof. Firstly J is strictly increasing, because if $t_1 > t_2 \geq 0$, then

$$J(t_1) = (1-\rho)t_1 + \sum_{0\leq s\leq t_1} q(s) > (1-\rho)t_2 + \sum_{0\leq s\leq t_2} q(s) = J(t_2),$$

since $t_0+\mathbb{Z}+a\mathbb{Z}$ is dense in $\mathbb{R}$ and so there exists $s \in t_0+\mathbb{Z}+a\mathbb{Z}$, with $t_1 < s < t_2$. Similarly, if $t_1 < t_2 < 0$, then

$$J(t_1) = (1-\rho)t_1 - \sum_{t_1<s\leq 0} q(s) < (1-\rho)t_2 - \sum_{t_2<s\leq 0} q(s) = J(t_2).$$

Finally, if $t_1 < 0 < t_2$, then

$$J(t_1) = (1-\rho)t_1 - \sum_{t_1<s\leq 0} q(s) < (1-\rho)t_1 \leq (1-\rho)t_2 < (1-\rho)t_2 + \sum_{0\leq s\leq t_2} q(s) = J(t_2).$$

For the continuity of J, let $t \geq 0$. If $t_k \searrow t$, then

$$J(t_k) = (1-\rho)t_k + \sum_{0\leq s\leq t_k} q(s) \searrow (1-\rho)t + \sum_{0\leq s\leq t} q(s) = J(t),$$

which shows that J is right continuous at t. If $t_k \nearrow t$, then

$$J(t_k) = (1-\rho)t_k + \sum_{0\leq s\leq t_k} q(s) \nearrow (1-\rho)t + \sum_{0\leq s< t} q(s) = J(t) - q(t).$$

Thus, if $t \notin t_0+\mathbb{Z}+a\mathbb{Z}$, then $q(t) = 0$ and $J(t_k) \to J(t)$, while if $t = t_0 + m + an$, then $q(t) = l_n$ and therefore $J(t_k) \to J(t) - l_n$. This shows that J is left continuous at every $t \geq 0$ with $t \notin t_0+\mathbb{Z}+a\mathbb{Z}$, but at $t = t_0 + m + an$ has jump l_n from the left. Similarly for $t < 0$.
(i) Obviously, $0 \notin t_0+\mathbb{Z}+a\mathbb{Z}$, and so $q(0) = 0$. Hence $J(0) = 0$. Observe that for every $t \in \mathbb{R}$ and for every $n \in \mathbb{Z}$, in the interval $(t, t+1]$ there exists a unique $m \in \mathbb{Z}$ such that $t_0 + m + an \in (t, t+1]$, because $t_0+\mathbb{Z}+a\mathbb{Z}$ is dense in $\mathbb{R}$ and $(t, t+1]$ has unit length. Consequently, for every $t \in \mathbb{R}$ we have

$$\sum_{t<s\leq t+1} q(s) = \sum_{n\in\mathbb{Z}} l_n = \rho.$$

If $t \geq 0$, then

$$J(t+1) = (1-\rho)(t+1) + \sum_{0\leq s\leq t+1} q(s) =$$

$$(1-\rho)t + \sum_{0\leq s\leq t} q(s) + (1-\rho) + \sum_{t<s\leq t+1} q(s) = J(t) + 1 - \rho + \rho = J(t) + 1.$$

If $-1 \leq t < 0$ then

$$J(t+1) = (1-\rho)(t+1) + \sum_{0\leq s\leq t+1} q(s) =$$

$$(1-\rho)t - \sum_{t<s\leq 0} q(s) + (1-\rho) + \sum_{t<s\leq t+1} q(s) = J(t) + 1 - \rho + \rho = J(t) + 1.$$

Finally, if $t < -1$, then $t + 1 < 0$ and

$$J(t+1) = (1-\rho)(t+1) - \sum_{t+1<s\leq 0} q(s) =$$

$$(1-\rho)t - \sum_{t<s\leq 0} q(s) + (1-\rho) + \sum_{t<s\leq t+1} q(s) = J(t) + 1 - \rho + \rho = J(t) + 1.$$

(ii) To show that C is perfect let $t \in \mathbb{R}$. Since $t_0 + \mathbb{Z} + a\mathbb{Z}$ is dense in $\mathbb{R}$, there are $t_k \in t_0 + \mathbb{Z} + a\mathbb{Z}$ such that $t_k \searrow t$ and $t_k \neq t$. By the right continuity of J we have $J(t_k) \searrow J(t)$ and $J(t_k) \neq J(t)$, since J is strictly increasing. Thus every point of C is an accumulation point. If C contains an open interval I, then $I \cap J(\mathbb{R}) \neq \varnothing$. Let $t \in \mathbb{R}$ be such that $J(t) \in I$ and $t_k \in t_0 + \mathbb{Z} + a\mathbb{Z}$ such that $t_k \searrow t$. Then $J(t_k) \searrow J(t) \in I$, and so there is some $k_0 \in \mathbb{N}$ such that $J(t_k) \in I$ for every $k \geq k_0$. Let $k \geq k_0$. Since $t_k \in t_0 + \mathbb{Z} + a\mathbb{Z}$, there exist $m, n \in \mathbb{Z}$ such that $t_k = t_0 + m + an$. Then $\varnothing \neq (J(t_k) - l_n, J(t_k)) \cap I \subset (\mathbb{R} \setminus C) \cap I$, contradiction. Finally, from (i) we have $J(t+1) = J(t) + 1$, that is $J \circ R_1 = R_1 \circ J$ and hence $J(\mathbb{R}) = J(R_1(\mathbb{R})) = R_1(J(\mathbb{R}))$. Thus,

$$C = \overline{J(\mathbb{R})} = \overline{R_1(J(\mathbb{R}))} = R_1(\overline{J(\mathbb{R})}) = R_1(C).$$

(iii) We have

$$\mu(C \cap [0,1]) = \mu([0,1]) - \mu([0,1] \setminus C) = 1 - \sum_{0\leq s\leq 1} q(s) =$$

$$1 - \sum_{0<s\leq 1} q(s) = 1 - \sum_{n\in\mathbb{Z}} l_n = 1 - \rho. \quad \square$$

Recall that

$$J(\mathbb{R}) = \mathbb{R} \setminus \bigcup_{n,m\in\mathbb{Z}} [J(t_0 + m + an) - l_n, J(t_0 + m + an))$$

and

$$C = \overline{J(\mathbb{R})} = \mathbb{R} \setminus \bigcup_{n,m\in\mathbb{Z}} (J(t_0 + m + an) - l_n, J(t_0 + m + an)).$$

Let $I_{n,m} = [J(t_0 + m + an) - l_n, J(t_0 + m + an)]$, $n, m \in \mathbb{Z}$. Then,

$$\mathbb{R} \setminus C = \bigcup_{n,m\in\mathbb{Z}} int(I_{n,m}).$$

Let $H : \mathbb{R} \to \mathbb{R}$ be the function defined by

$$H(x) = \begin{cases} t, & \text{if } x = J(t) \text{ for some } t \in \mathbb{R} \\ t_0 + m + an, & \text{if } x \in I_{n,m}. \end{cases}$$

3.5.2. Lemma. *The function H continuous, increasing, $H \circ J = id$ and $H(C) = \mathbb{R}$. Moreover, the following hold.*

(i) $H(0) = 0$, $H(x+1) = H(x) + 1$ for every $x \in \mathbb{R}$, and so $H(k) = k$ for $k \in \mathbb{Z}$.

(ii) For every $x \in J(\mathbb{R}^+)$ we have $\mu(C \cap [0, x]) = (1 - \rho)H(x)$.

Proof. From the definition of H it is clear that $H \circ J = id$. Therefore, $H(C) = H(\overline{J(\mathbb{R})}) \supset H(J(\mathbb{R})) = \mathbb{R}$. The continuity of H follows easily from the definitions and the fact that $t_0 + \mathbb{Z} + a\mathbb{Z}$ is dense in $\mathbb{R}$.

(i) From the definition of H we have $0 = J(0) \in J(\mathbb{R})$ and so $H(0) = 0$. To show that $H(x+1) = H(x) + 1$ for every $x \in \mathbb{R}$, we consider cases. If $x \in J(\mathbb{R})$, and $x = J(t)$ for some $t \in \mathbb{R}$, then $H(x+1) = H(J(t)+1) = H(J(t+1)) = t+1 = H(x)+1$. Consequently, $H(x+1) = H(x)+1$ for every $x \in C$, by the continuity of $o'W$ffl. If $x \in \mathbb{R} \setminus C = \bigcup_{n,m\in\mathbb{Z}} int(I_{n,m})$, there exist $n, m \in \mathbb{Z}$ such that $x \in int(I_{n,m})$, and so $H(x) = t_0 + m + an$. The open interval $R_1(int(I_{n,m}))$ has right endpoint $J(t_0 + m + an) + 1 = J(t_0 + (m+1) + an)$, and is the connected component of $\mathbb{R} \setminus C$ which contains $x+1$. Since $x + 1 \in I_{n,m+1}$, from the definition of H we have $H(x+1) = t_0 + (m+1) + an = t_0 + m + an + 1 = H(x) + 1$.
(ii) If $x \in J(\mathbb{Z}^+)$, then from 3.5.1 we have $\mu(C \cap [0,1]) = 1 - \rho$ and $R_1(C) = C$ and so $\mu(C \cap [0,x]) = (1-\rho)x = (1-\rho)H(x)$, because $H(x) = x$. If $x \in J(\mathbb{R}^+)$ with $0 \le x < 1$ and $x = J(t)$ for some $t > 0$, we observe that

$$\sum_{\{n,m\in\mathbb{Z}:I_{n,m}\subset[0,x]\}} l_n = \sum_{\{n,m\in\mathbb{Z}:I_{n,m}\subset[0,J(t)]\}} l_n = \sum_{0\le s\le t} q(s),$$

because $I_{n,m} \subset [0,x]$ if and only if $[J(t_0+m+an)-l_n, J(t_0+m+an)] \subset [J(0), J(t)]$ or equivalently $0 < t_0 + m + an \le t$. Therefore,

$$\mu(C \cap [0,x]) = \mu([0,x]) - \mu([0,x] \setminus C) =$$

$$x - \sum_{\{n,m\in\mathbb{Z}:I_{n,m}\subset[0,x]\}} l_n = J(t) - \sum_{0\le s\le t} q(s) =$$

$$(1-\rho)t + \sum_{0\le s\le t} q(s) - \sum_{0\le s\le t} q(s) = (1-\rho)t = (1-\rho)H(x).$$

if $x \in J(\mathbb{R}^+)$, then

$$\mu(C \cap [0,x]) = \mu(C \cap [0,[x]]) + \mu(C \cap [[x],x]) =$$

$$(1-\rho)[x] + \mu(C \cap [0, x-[x]]) = (1-\rho)[x] + (1-\rho)H(x-[x]) =$$

$$(1-\rho)H([x] + x - [x]) = (1-\rho)H(x). \quad \square$$

3.5.3. Proposition. *If $F : \mathbb{R} \to \mathbb{R}$ is an increasing homeomorphism such that $F(x+1) = F(x)+1$ for every $x \in \mathbb{R}$, the following are equivalent.*
(i) $H \circ F = R_a \circ H$, where $R_a : \mathbb{R} \to \mathbb{R}$ is the translation $R_a(x) = x + a$.
(ii) $F(x) = J(H(x) + a)$ for every $x \in J(\mathbb{R})$.

Proof. Suppose that $H \circ F = R_a \circ H$, that is $H(F(x)) = H(x) + a$ for every $x \in \mathbb{R}$. Let $x_0 \in J(\mathbb{R})$ be such that $H(x_0) \notin t_0 + \mathbb{Z} + a\mathbb{Z}$, and so $x_0 \notin I_{n,m}$, and in particular $x_0 \ne J(t_0 + m + an)$ for every $n, m \in \mathbb{Z}$. Then $H(F(x_0)) = H(x_0) + a \notin t_0 + \mathbb{Z} + a\mathbb{Z}$. So $F(x_0) \in J(\mathbb{R})$ and $F(x_0) = J(H(x_0) + a)$. Since $H \circ F = R_a \circ H$, inductively we have $H \circ F^k = (R_a)^k \circ H = R_{ka} \circ H$, that is $H(F^k(x)) = H(x) + ka$ for every $x \in \mathbb{R}$ and $k \in \mathbb{Z}$. For every $k, \lambda \in \mathbb{Z}$ we have $H(F^k(x_0)) + \lambda = H(x_0) + ka + \lambda \notin t_0 + \mathbb{Z} + a\mathbb{Z}$

and therefore $F^k(x_0)+\lambda \in J(\mathbb{R})$. Moreover, $F^k(x_0)+\lambda = J(H(x_0)+ak+\lambda)$ for every $k,\lambda \in \mathbb{Z}$. Let now $x = J(t)$ for some $t \in \mathbb{R}$. The set

$$H(x_0)+\mathbb{Z}+a\mathbb{Z} = \{H(x_0)+\lambda+ak : k,\lambda \in \mathbb{Z}\} = \{H(F^k(x_0)+\lambda) : k,\lambda \in \mathbb{Z}\}$$

is dense in $\mathbb{R}$. Since J is everywhere right continuous, the set $\{F^k(x_0)+\lambda : k,\lambda \in \mathbb{Z}\}$ is dense in $J(\mathbb{R})$, and so in C also. Thus, there are k_n, $\lambda_n \in \mathbb{Z}$, $n \in \mathbb{N}$, such that $H(x_0)+\lambda_n+ak_n \searrow t$, and then

$$F^{k_n}(x_0)+\lambda_n = J(H(x_0)+\lambda_n+ak_n) \searrow x.$$

By the continuity and the monotonicity of H, the right continuity of J and since $H \circ J = id$ we have

$$J(H(x)+a) = J(H(\lim_{n\to+\infty}(F^{k_n}(x_0)+\lambda_n))+a) =$$

$$J(\lim_{n\to+\infty} H(F^{k_n}(x_0)+\lambda_n)+a) = \lim_{n\to+\infty} J(H(F^{k_n}(x_0)+\lambda_n)+a) =$$

$$\lim_{n\to+\infty} J(H(x_0)+\lambda_n+ak_n+a) = \lim_{n\to+\infty} J(H(x_0)+\lambda_n+(k_n+1)a) =$$

$$\lim_{n\to+\infty}(F^{k_n+1}(x_0)+\lambda_n) = \lim_{n\to+\infty} F((F^{k_n}(x_0)+\lambda_n)) =$$

$$F(\lim_{n\to+\infty}(F^{k_n}(x_0)+\lambda_n)) = F(x).$$

Conversely, suppose that $F(x) = J(H(x)+a)$ for every $x \in J(\mathbb{R})$. If $x \in J(\mathbb{R})$, then $H(F(x)) = H(J(H(x)+a)) = H(x)+a$, and the same is true for every $x \in C$ by continuity. If $x \in \mathbb{R}\setminus C$, then $x \in I_{n,m}$ for some $n,m \in \mathbb{Z}$. We observe that $F(J(t)) = J(t+a)$ for every $t \in \mathbb{R}$ and therefore $F(J(\mathbb{R})) = J(\mathbb{R})$. Since F is a homeomorphism we have $F(C) = C$. We also have $F(I_{n,m}) = I_{n+1,m}$. Indeed, if $x = J(t_0+m+an)$, then $H(x) = t_0+m+an$ and so $F(x) = J(H(x)+a) = J(t_0+m+(n+1)a)$, which is the right endpoint of $I_{n+1,m}$. Hence $H(F(x)) = t_0+m+an+a = H(x)+a = R_a(H(x))$ for every $x \in I_{n,m}$. □

3.5.4. Corollary. *If $F : \mathbb{R} \to \mathbb{R}$ is an increasing homeomorphism with $F(x+1) = F(x)+1$ and $H \circ F = R_a \circ H$, then*

(i) $F(C) = C$, and

(ii) the set $\{F^k(x)+\lambda : k,\lambda \in \mathbb{Z}\}$ is dense in C for every $x \in C$.

Proof. The first claim was proved in 3.5.3, where we also proved that if $x \in J(\mathbb{R})$ and $H(x) \notin t_0+\mathbb{Z}+a\mathbb{Z}$, then the set $\{F^k(x)+\lambda : k,\lambda \in \mathbb{Z}\}$ is dense in C. It remains to examine the following two cases. First, if $x = J(t_0+m+an)$ for some $n,m \in \mathbb{Z}$, then form 3.5.3 we have

$$F^k(x)+\lambda = J(H(x)+ka+\lambda) = J(H(J(t_0+m+an))+ka+\lambda) =$$

$$J(t_0+m+an+ka+\lambda) = J(t_0+(m+\lambda)+(n+k)a) \in J(t_0+\mathbb{Z}+a\mathbb{Z}).$$

But since the set $t_0+\mathbb{Z}+a\mathbb{Z}$ is dense in $\mathbb{R}$ and J is right continuous, it follows that $\{F^k(x)+\lambda : k,\lambda \in \mathbb{Z}\}$ is dense in C. Let now $x \in C\setminus J(\mathbb{R}) = \{J(t_0+m+an)-l_n :$

$n, m \in \mathbb{Z}\}$. Since $\lim_{n\to\pm\infty} l_n = 0$, the set $\{J(t_0+m+an)-l_n : n, m \in \mathbb{Z}\}$ is dense in C and since F is a homeomorphism with $F(C) = C$, the set $\{F^k(x) + \lambda : k, \lambda \in \mathbb{Z}\}$ is dense in C. $\square$

For the rest of the section we make the additional assumption that

$$\lim_{n\to\pm\infty} \frac{l_{n+1}}{l_n} = 1 \text{ and } \frac{l_{n+1}}{l_n} > \frac{1}{3} \text{ for every } n \in \mathbb{Z}.$$

This is true for $l_n = \frac{\rho}{\pi}(\arctan(n+1) - \arctan n)$.

3.5.5. Lemma. *There exist C^1 diffeomorphisms $F_{n,0} : I_{n,0} \to I_{n+1,0}$, $n \in \mathbb{Z}$, with the following properties.*

(i) $F'_{n,0}(J(t_0 + an) - l_n) = F'_{n,0}(J(t_0 + an)) = 1$.

(ii) $0 < F'_{n,0}(x) \le 1 + 6|\frac{l_{n+1}}{l_n} - 1|$ *for every* $x \in I_{n,0}$, $n \in \mathbb{Z}$, *and*

$$\lim_{n\to\pm\infty}(\sup\{|F'_{n,0}(x) - 1| : x \in I_{n,0}\}) = 0.$$

Proof. For simplicity in notation we set $a_n = J(t_0 + an) - l_n$, $b_n = J(t_0 + an)$ and $c_n = 6(\frac{l_{n+1}}{l_n} - 1) > -4$. Let $F_{n,0} : I_{n,0} \to I_{n+1,0}$ be defined by

$$F_{n,0}(x) = a_{n+1} + \int_{a_n}^{x} [1 + \frac{c_n}{l_n^2}(y - a_n)(b_n - y)]dy.$$

Then $F_{n,0}$ is obviously C^1, $F_{n,0}(a_n) = a_{n+1}$ and

$$F_{n,0}(b_n) = a_{n+1} + \int_{a_n}^{b_n} dy + \frac{c_n}{l_n^2}\int_{a_n}^{b_n} (y - a_n)(b_n - y)dy =$$

$$a_{n+1} + (b_n - a_n) + \frac{c_n}{l_n^2}\frac{(b_n - a_n)^3}{6} =$$

$$a_{n+1} + l_n + (\frac{l_{n+1}}{l_n} - 1)l_n =$$

$$a_{n+1} + l_n + l_{n+1} - l_n = a_{n+1} + l_{n+1} = b_{n+1}.$$

Also, $F_{n,0}$ is strictly increasing, because for every $x \in I_{n,0}$ we have

$$F'_{n,0}(x) = 1 + \frac{c_n}{l_n^2}(x - a_n)(b_n - x) > 1 - \frac{4}{l_n^2}(x - a_n)(b_n - x) \ge 1 - \frac{4}{l_n^2} \cdot \frac{l_n^2}{4} = 0.$$

Since $F_{n,0}$ is continuous and strictly increasing,

$$F_{n,0}(I_{n,0}) = F_{n,0}([a_n, b_n]) = [a_{n+1}, b_{n+1}] = I_{n+1,0},$$

that is $F_{n,0}$ is a C^1 diffeomorphism onto $I_{n+1,0}$. We also have

$$F'_{n,0}(J(t_0 + an) - l_n) = F'_{n,0}(b_n - l_n) = F'_{n,0}(a_n) = 1 + \frac{c_n}{l_n^2}(a_n - a_n)(b_n - a_n) = 1,$$

and similarly, $F'_{n,0}(J(t_0+an)) = F'_{n,0}(b_n) = 1$. Finally, for every $x \in I_{n,0}$ we have

$$|F'_{n,0}(x) - 1| = \frac{|c_n|}{l_n^2}(x-a_n)(b_n-x) \leq \frac{|c_n|}{l_n^2}(b_n-a_n)^2 = \frac{|c_n|}{l_n^2}\cdot l_n^2 = |c_n|,$$

and hence $\sup_{x\in I_{n,0}} |F'_{n,0}(x)-1| \to 0$, for $n \to \pm\infty$. □

3.5.6. Theorem. *There exists an increasing C^1 diffeomorphism $F : \mathbb{R} \to \mathbb{R}$ such that $H \circ F = R_a \circ H$ and $F(x+1) = F(x)+1$ for every $x \in \mathbb{R}$.*

Proof. For every $n \in \mathbb{Z}$ let $F_{n,0} : I_{n,0} \to I_{n+1,0}$ be the C^1 diffeomorphism of 3.5.5 defined by

$$F_{n,0}(x) = a_{n+1} + \int_{a_n}^{x} [1 + \frac{c_n}{l_n^2}(y-a_n)(b_n-y)]dy.$$

For every $m \in \mathbb{Z}$, define $F_{n,m} : I_{n,m} \to I_{n+1,m}$ by

$$F_{n,m} = R_m \circ F_{n,0} \circ R_{-m}.$$

Then $F_{n,m}$ is an increasing C^1 diffeomorphism and $F'_{n,m}(x) = F'_{n,0}(x-m)$ for every $x \in I_{n,m}$. Let $G : \mathbb{R} \to (0,+\infty)$, be defined by

$$G(x) = \begin{cases} 1, & \text{if } x \in C \\ F'_{n,m}(x), & \text{if } x \in I_{n,m}, n, m \in \mathbb{Z}. \end{cases}$$

Since $\sup_{x\in I_{n,0}} |F'_{n,0}(x)-1| \to 0$ for $n \to \pm\infty$, by 3.5.5, G is continuous and bounded. Let $M > 1$ be such that $0 \leq G(x) \leq M$ for every $x \in \mathbb{R}$. Let now $F : \mathbb{R} \to \mathbb{R}$ be defined by

$$F(x) = J(a) + \int_0^x G(s)ds.$$

Then F is an increasing C^1 diffeomorphism onto $\mathbb{R}$, since $G > 0$. Moreover,

$$F(x) = J(H(x)+a)$$

for every $x \in J(\mathbb{R})$. Indeed, let $x \in J(\mathbb{R})$ and $x \geq 0$. If $I_{n,m} \subset [0,x]$, then

$$\int_{I_{n,m}} G(s)ds = \int_{I_{n,m}} F'_{n,m}(s)ds = \int_{I_{n,0}} F'_{n,0}(s) =$$

$$\int_{a_n}^{b_n} [1 + \frac{c_n}{l_n^2}(s-a_n)(b_n-s)]ds = l_n + \frac{c_n}{l_n^2}\frac{(b_n-a_n)^3}{6} =$$

$$l_n + (\frac{l_{n+1}}{l_n} - 1)l_n = l_{n+1} = q(t_0+m+(n+1)a).$$

Consequently,

$$\int_0^x G(s)ds = \mu(C\cap[0,x]) + \int_{[0,x]\setminus C} G(s)ds =$$

$$\mu(C\cap[0,x]) + \sum_{\{n,m\in\mathbb{Z}: I_{n,m}\subset[0,x]\}} \int_{I_{n,m}} G(s)ds =$$

$$\mu(C\cap[0,x])+\sum_{\{n,m\in\mathbb{Z}:I_{n,m}\subset[0,x]\}}q(t_0+m+an+a)=$$

$$(1-\rho)H(x)+\sum_{0\le s\le H(x)}q(s+a).$$

Therefore,

$$F(x)=J(a)+\int_0^x G(s)ds=$$

$$(1-\rho)a+\sum_{0\le s\le a}q(s)+(1-\rho)H(x)+\sum_{0\le s\le H(x)}q(s+a)=$$

$$(1-\rho)(H(x)+a)+\sum_{0\le s\le a}q(s)+\sum_{a\le s\le H(x)+a}q(s)=$$

$$(1-\rho)(H(x)+a)+\sum_{0\le s\le H(x)+a}q(s)=J(H(x)+a),$$

since $q(a)=0$. Similarly, $F(x)=J(H(x)+a)$ for every $x\in J(\mathbb{R})$ with $x<0$. Because of 3.5.2, it remains to prove that $F(x+1)=F(x)+1$ for every $x\in\mathbb{R}$. We consider cases. If $x\in J(\mathbb{R})$, then $F(x+1)=J(H(x+1)+a)=J(H(x)+1+a)=J(H(x)+a)+1=F(x)+1$ and from the continuity of F the same is true for $x\in C$. If $x\in I_{n,m}$ for some $n,m\in\mathbb{Z}$, then

$$F(J(t_0+m+an)-l_n)=J(a)+\int_0^{J(t_0+m+an)-l_n}G(s)ds=$$

$$J(a)+\int_0^{J(t_0+m+an)}G(s)ds-\int_{I_{n,m}}G(s)ds=F(J(t_0+m+an))-\int_{I_{n,m}}F'_{n,m}(s)ds=$$

$$J(H(J(t_0+m+an))+a)-\int_{I_{n,m}}F'_{n,m}(s)ds=J(t_0+m+an+a)-\int_{I_{n,m}}F'_{n,m}(s)ds=$$

$$J(t_0+m+an+a)-l_{n+1}=a_{n+1}$$

and so

$$F(x)=a_{n+1}+\int_{a_n}^x F'_{n,m}(s)ds=F_{n,m}(x).$$

It follows that

$$F(x+1)=F_{n,m+1}(x+1)=R_{m+1}\circ F_{n,0}\circ R_{-m-1}(x+1)=$$

$$F_{n,0}(x-m)+m+1=F_{n,m}(x)+1=F(x)+1.\quad\square$$

The C^1 diffeomorphism $F:\mathbb{R}\to\mathbb{R}$ of 3.5.6 is increasing and $F(x+1)=F(x)+1$ for every $x\in\mathbb{R}$. Moreover, the set C is F-invariant, perfect, closed, totally disconnected, has Lebesgue measure $\mu(C\cap[0,1])=1-\rho$ and every orbit of F in C is dense in it.

If we define $f:S^1\to S^1$ by

$$f(e^{2\pi it})=e^{2\pi iF(t)}$$

then f is an orientation preserving C^1 diffeomorphism with lift F and has rotation number $\rho(f) = e^{2\pi ia}$, since $H \circ F = R_a \circ H$. The set $K = p(C)$, is a minimal Cantor set of f, because $\{F^n(x) + m : n, m \in \mathbb{Z}\}$ is dense in C for every $x \in C$. The normalized Lebesgue measure of K is $\mu(K) = \mu(C \cap [0,1]) = 1 - \rho$, but μ is not f-invariant.

3.6 Arnold families of circle diffeomorphisms

As we saw in Theorem 3.4.2, if $f : S^1 \to S^1$ is an orientation preserving C^2 diffeomorphism with irrational rotation number $e^{2\pi a}$ then there is an orientation preserving homeomorphism $h : S^1 \to S^1$ which conjugates f to the rotation r_a by the angle $2\pi a$. The question arises whether h can be chosen to be a C^1 diffeomorphism. This section is devoted to giving a negative answer to this question. Actually, we shall prove that even in the case where f is real analytic, it may be impossible to choose an absolutely continuous conjugation h.

Let $\mathcal{D}$ denote the set of increasing homeomorphisms of $\mathbb{R}$ which commute with integer translations. For every $a \in \mathbb{R}$ we denote by $T_a : \mathbb{R} \to \mathbb{R}$ the translation by a. Obviously, $T_a \in \mathcal{D}$. Every $\tilde{f} \in \mathcal{D}$ induces an orientation preserving homeomorphism $f : S^1 \to S^1$. In particular, the translation T_a induces the rotation $r_a : S^1 \to S^1$ by the angle $2\pi a$. Let $\mathcal{H}_+$ denote the set of orientation preserving homeomorphisms of the circle S^1. Every $f \in \mathcal{H}_+$ has a lift $\tilde{f} \in \mathcal{D}$ and any two such lifts of f differ by an integer translation.

We fix an element $\tilde{f} \in \mathcal{D}$ and the induced element $f \in \mathcal{H}_+$. Let $\rho : \mathbb{R} \to \mathbb{R}$ be the function defined by

$$\rho(a) = \tau(T_a \circ \tilde{f}),$$

where $\tau : \mathcal{D} \to \mathbb{R}$ is the translation number function. If on $\mathcal{D}$ we consider the distance

$$d(F, G) = \sup\{|F(t) - G(t)| : t \in \mathbb{R}\} = \sup\{|F(t) - G(t)| : t \in [0,1]\},$$

then τ is continuous. Since $|(T_a \circ \tilde{f})(t) - (T_b \circ \tilde{f})(t)| = |a - b|$ for every $t \in \mathbb{R}$ and a, $b \in \mathbb{R}$, it follows that ρ is continuous.

Note that $\tilde{f}_a = T_a \circ \tilde{f}$ is a lift of $r_a \circ f$. The following elementary observation will be of fundamental importance in the sequel.

3.6.1. Lemma. *If $a > 0$, then $\tilde{f}_a^n(t) \geq \tilde{f}^n(t) + a$ for every $t \in \mathbb{R}$ and $n \in \mathbb{N}$.*

Proof. This is true by definition for $n = 1$. Inductively, suppose that it holds for $n - 1$. Then, for every $t \in \mathbb{R}$ we have

$$\tilde{f}_a^n(t) \geq \tilde{f}_a(\tilde{f}^{n-1}(t) + a) > \tilde{f}_a(\tilde{f}^{n-1}(t)) = \tilde{f}(\tilde{f}^{n-1}(t)) + a = \tilde{f}^n(t) + a. \quad \square$$

It follows from Lemma 3.6.1 that ρ is increasing, because for $a > 0$ we have

$$\rho(a) = \lim_{n \to +\infty} \frac{\tilde{f}_a^n(t)}{n} \geq \lim_{n \to +\infty} \frac{\tilde{f}^n(t) + a}{n} = \lim_{n \to +\infty} \frac{\tilde{f}^n(t)}{n} = \rho(0).$$

If $a, b \in \mathbb{R}$ with $a > 0$, applying the above for $\tilde{f}_b$ in place of $\tilde{f}$ we get $\rho(a+b) \geq \rho(b)$, which means that ρ is increasing.

Suppose now that $\rho(0) \in \mathbb{R} \setminus \mathbb{Q}$. This means that f has no periodic point and has a unique minimal set $K \subset S^1$, which is the non-wandering set of f, such that either $K = S^1$ or K is topologically a Cantor set. In any case K is uncountable and so is $\tilde{K} = \exp^{-1}(K)$, where $\exp : \mathbb{R} \to S^1$ is the universal covering projection $\exp(t) = e^{2\pi i t}$. Since K is uncountable, there exists some $t_0 \in \tilde{K}$ which can be approximated from both sides by other points of $\tilde{K}$. So, there exist a sequence $(n_k)_{k\in\mathbb{N}}$ of positive integers with $n_k \to +\infty$ and a sequence $(m_k)_{k\in\mathbb{N}}$ of integers such that $\lim_{k\to+\infty} (\tilde{f}^{n_k}(t_0) - m_k - t_0) = 0$ and $\tilde{f}^{n_k}(t_0) < t_0 + m_k$ for every $k \in \mathbb{N}$. If $a > 0$, there exists some $n_k > 1$ such that $0 < t_0 + m_k - \tilde{f}^{n_k}(t_0) < a$ and therefore $t_0 + m_k < \tilde{f}^{n_k}(t_0) + a < \tilde{f}_a^{n_k}(t_0)$, by Lemma 1. In other words,

$$\tilde{f}^{n_k}(t_0) - t_0 - m_k < 0 < \tilde{f}_a^{n_k}(t_0) - t_0 - m_k$$

and by the intermediate value theorem there exists some $0 < b < a$ such that $\tilde{f}_b^{n_k}(t_0) = t_0 + m_k$. Thus, $\rho(b) \in \mathbb{Q}$ and since $\rho(0) \in \mathbb{R} \setminus Q$, we must have $\rho(0) < \rho(b) \leq \rho(a)$. As before, this implies that if $x \in \mathbb{R}$ is such that $\rho(x) \in \mathbb{R} \setminus \mathbb{Q}$, then $\rho(x) < \rho(y)$ for every $x < y$. Thus ρ is strictly increasing at points of $\mathbb{R}$ where it takes on irrational values.

Since ρ is continuous and increasing, for each $s \in \mathbb{R}$ the set $\rho^{-1}(s)$ is either a singleton or a closed interval. It is certainly not empty, because ρ is onto $\mathbb{R}$, as $\rho(x + k) = \rho(x) + k$ for every $x \in \mathbb{R}$ and $k \in \mathbb{Z}$.

3.6.2. Lemma. *If $f_a^n \neq id$ for every $a \in \mathbb{R}$ and every $n \in \mathbb{N}$, then the following hold.*
(a) If $\frac{p}{q} \in \mathbb{Q}$, where $p \in \mathbb{Z}$ and $q \in \mathbb{N}$ are such that $gcd(p, q) = 1$, then $\rho^{-1}(\frac{p}{q})$ is a closed interval and has non-empty interior.
(b) The set $\mathcal{R} = \rho^{-1}(\mathbb{R} \setminus \mathbb{Q})$ is nowhere dense in $\mathbb{R}$.

Proof. (a) If $a \in \rho^{-1}(\frac{p}{q})$, there exists some $t_0 \in \mathbb{R}$ such that $\tilde{f}_a^q(t_0) = t_0 + p$. We define the sets

$$K_{p/q}^+ = \{a \in \rho^{-1}(\frac{p}{q}) : \tilde{f}_a^q(t) \geq t + p \quad \text{for all} \quad t \in \mathbb{R}\},$$

$$K_{p/q}^- = \{a \in \rho^{-1}(\frac{p}{q}) : \tilde{f}_a^q(t) \leq t + p \quad \text{for all} \quad t \in \mathbb{R}\}.$$

Our assumption $f_a^n \neq id$ for every $a \in \mathbb{R}$ and every $n \in \mathbb{N}$ implies that $K_{p/q}^+ \cap K_{p/q}^- = \varnothing$. If $a \in K_{p/q}^+$ and $b > a$, then $\rho(b) > \rho(a)$, because $\tilde{f}_b = T_{b-a} \circ \tilde{f}_a$ and so

$$\tilde{f}_b^n(t) \geq \tilde{f}_a^n(t) + (b - a) \geq t + p + (b - a) > t + p$$

for every $t \in \mathbb{R}$ and $n \in \mathbb{N}$, by Lemma 3.6.1. This implies that $\rho(b) \neq \frac{p}{q}$. Similarly, if $a \in K_{p/q}^-$ and $b < a$, then $\rho(b) < \rho(a)$. These observations imply that $K_{p/q}^+ \cup K_{p/q}^- \subset \partial\rho^{-1}(\frac{p}{q})$.

On the other hand, the set

$$U_{p/q} = \{a \in \mathbb{R} : \tilde{f}_a^q(t_1) < t_1 + p \quad \text{and} \quad \tilde{f}_a^q(t_2) > t_2 + p \quad \text{for some} \quad t_1, t_2 \in \mathbb{R}\}$$

is open in $\mathbb{R}$ and is contained in $\rho^{-1}(\frac{p}{q})$, by continuity and the intermediate value theorem. Since $\rho^{-1}(\frac{p}{q})$ is the disjoint union of $K^+_{p/q}$, $K^-_{p/q}$ and $U_{p/q}$, it follows that $U_{p/q}$ id the interior of $\rho^{-1}(\frac{p}{q})$ and $K^+_{p/q} \cup K^-_{p/q} = \partial\rho^{-1}(\frac{p}{q})$. Thus, $\rho^{-1}(\frac{p}{q})$ is a closed interval with endpoints $K^+_{p/q}$ and $K^-_{p/q}$.

(b) Suppose on the contrary that the set $\overline{\mathcal{R}}$ contains an open interval (b, c), where $b < c$. Then there exists $b < a < c$ such that $\rho(a) \in \mathbb{R} \setminus \mathbb{Q}$ and so $\rho(a) < \rho(c)$. Let $r \in \mathbb{Q}$ be such that $\rho(a) < r < \rho(c)$. By (a), $\rho^{-1}(r)$ is a closed interval contained in (a, c). If J denotes its interior, then $\varnothing \neq J \subset (a, c) \subset \overline{\mathcal{R}}$, which implies that $J \cap \mathcal{R} \neq \varnothing$, and at the same time $J \subset \rho^{-1}(\mathbb{Q})$. This contradiction proves the assertion. □

In the sequel we shall give an example of a family which satisfies the assumption of Lemma 3.6.2. We shall assume now that we have such a family.

A function with the properties of $\rho : \mathbb{R} \to \mathbb{R}$ is often called a *devil's staircase*. The sets $\mathcal{R}$ and $\overline{\mathcal{R}}$ have interesting topological properties, the description of which will occupy the rest of this section.

It follows from Lemma 3.6.2(b) that the set $\overline{\mathcal{R}}$ is closed and totally disconnected. It is also perfect, because if $a \in \overline{\mathcal{R}}$ is an isolated point of $\overline{\mathcal{R}}$, there exists some $\delta > 0$ such that $(a - \delta, a + \delta) \cap \overline{\mathcal{R}} = \{a\}$ and so $(a - \delta, a) \cup (a, a + \delta) \subset \rho^{-1}(\mathbb{Q})$. Since ρ is continuous and $\mathbb{Q}$ is totally disconnected, there exists $r \in \mathbb{Q}$ such that $[a-\delta, a] \subset \rho^{-1}(r)$. Hence $[a-\delta, a+\delta] \subset \rho^{-1}(r)$, which is impossible, because $a \in \overline{\mathcal{R}}$. The same reasoning shows that $\overline{\mathcal{R}}$ is also perfect.

3.6.3. Lemma. *The set $\overline{\mathcal{R}} \setminus \mathcal{R}$ is countable.*

Proof. If $a \in \overline{\mathcal{R}} \setminus \mathcal{R}$, then $\rho(a) \in \mathbb{Q}$ and $I_a = \rho^{-1}(\rho(a))$ is a closed interval, by Lemma 1.2(a). Since $a \in \overline{\mathcal{R}} \setminus \mathcal{R}$, a is an endpoint of I_a. Thus for each $a \in \overline{\mathcal{R}} \setminus \mathcal{R}$ there exists a closed interval I_a such that a is an endpoint of I_a and the interior of I_a does not intersect $\overline{\mathcal{R}} \setminus \mathcal{R}$. This implies that $\overline{\mathcal{R}} \setminus \mathcal{R}$ is countable. □

Note that $\overline{\mathcal{R}} \setminus \mathcal{R} = \overline{\mathcal{R}} \cap \rho^{-1}(\mathbb{Q})$ is the set of the endpoints of the connected components of $\rho^{-1}(\mathbb{Q})$. In the notation of the proof of Lemma 3.6.2,

$$\overline{\mathcal{R}} \setminus \mathcal{R} = \bigcup_{r \in \mathbb{Q}} K_r^+ \cup K_r^-.$$

3.6.4. Lemma. *If B is a dense subset of $\mathbb{R}$, then $\overline{\mathcal{R}} \cap \rho^{-1}(B)$ is dense in $\overline{\mathcal{R}}$.*

Proof. Suppose on the contrary that $\overline{\mathcal{R}} \cap \rho^{-1}(B)$ is not dense in $\overline{\mathcal{R}}$. Then there exist real numbers $a < b$ such that $\overline{\mathcal{R}} \cap \rho^{-1}(B) \cap (a, b) = \varnothing$ and $\mathcal{R} \cap (a, b) \neq \varnothing$. Thus, $\rho(a) < \rho(b)$ and since B is assumed to be dense in $\mathbb{R}$, there exists some $s \in \rho^{-1}(B)$ such that $\rho(a) < \rho(s) < \rho(b)$. There exists a strictly increasing sequence $(a_n)_{n \in \mathbb{N}}$ in

$\mathbb{R}\setminus\mathbb{Q}$ converging to $\rho(s)$. If $(\rho^{-1}(a_n))_{n\in\mathbb{N}}$ converges to s, then $s \in \overline{\mathcal{R}}\cap\rho^{-1}(B)\cap(a,b)$, contradiction. So, there exists some $t < s$ such that $(\rho^{-1}(a_n))_{n\in\mathbb{N}}$ converges to t. Then, necessarily $\lim_{n\to+\infty} a_n = \rho(t) = \rho(s) \in \mathbb{Q}$ and t is the left endpoint of the interval $\rho^{-1}(\rho(s))$, according to Lemma 3.6.2(a). But now we arrive again at the contradiction $t \in \overline{\mathcal{R}} \cap \rho^{-1}(B) \cap (a,b)$. □

3.6.5. Corollary. *The set $\overline{\mathcal{R}} \setminus \mathcal{R}$ is dense in $\overline{\mathcal{R}}$.* □

Since $\overline{\mathcal{R}}$ is a perfect complete metric space, hence a Baire Hausdorff space, and $\overline{\mathcal{R}} \setminus \mathcal{R}$ is countable, it follows from the following general remark that $\mathcal{R}$ is a Baire metric space.

3.6.6. Remark. If X is a Baire Hausdorff space and $A \subset X$ is a countable set such that $X \setminus A$ is dense in X, then $X \setminus A$ is a Baire space. To see this, first observe that $X \setminus \{a_n\}$ is open and dense in X for every $n \in \mathbb{N}$, where $A = \{a_1, a_2, ..., a_n, ...\}$ is an enumeration of A. Let V_n, $n \in \mathbb{N}$ be a countable family of open subsets of X such that $V_n \cap (X \setminus A)$ is dense in $X \setminus A$ for every $n \in \mathbb{N}$. Then, $X \setminus A \subset \overline{V}_n$ and therefore V_n is dense in X. Moreover, $V_n \cap (X \setminus \{a_n\})$ is an open and dense subset of X for every $n \in \mathbb{N}$. From the Baire property of X we conclude that

$$\bigcap_{n=1}^{\infty} V_n \cap (X \setminus \{a_n\}) = (X \setminus A) \cap \bigcap_{n=1}^{\infty} V_n$$

is dense in X and in $X \setminus A$ as well. This proves that $X \setminus A$ is a Baire space.

In particular, if X is a perfect Baire Hausdorff space and $A \subset X$ is a countable set, then $X \setminus A$ is dense in X and therefore a Baire space.

Let now $f : \mathbb{C} \to \mathbb{C}$ be a holomorphic function. If f vanishes on the unit circle S^1, then f vanishes everywhere on $\mathbb{C}$. Indeed, the mean value property of holomorphic functions gives

$$f(z) = \frac{1}{2\pi i} \int_{S^1} \frac{f(\zeta)}{\zeta - z} d\zeta = 0$$

for $|z| < 1$ and in particular $f^{(n)}(0) = 0$ for every integer $n \geq 0$. Hence $f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} z^n = 0$ for every $z \in \mathbb{C}$.

3.6.7 Proposition. *Let $f : S^1 \to S^1$ be an analytic orientation preserving diffeomorphism, which has an extension to a holomorphic map $f : \mathbb{C} \to \mathbb{C}$. Let $\tilde{f}$ be a lift of f. If $f^q = id$ on S^1 and $\tau(\tilde{f}) = \frac{p}{q}$, where p and q are integers such that $q > 1$ and $gcd(p,q) = 1$, then f is the rational rotation by the angle $2\pi\frac{p}{q}$.*

Proof. Our assumption and the initial remark imply that $f^q = id$ on $\mathbb{C}$. Since $f^{q-1} \circ f = id$, it follows that f is biholomorphic. This means that there exist a, $b \in \mathbb{C}$ with $a \neq 0$ such that $f(z) = az + b$ for every $z \in \mathbb{C}$. However, necessarily $b = 0$ and $a \in S^1$, because $f(S^1) = S^1$. Thus f is a rotation and the conclusion follows. □

For each $\lambda \in \mathbb{R}$ with $0 < |\lambda| < \dfrac{1}{2\pi}$ let $f_\lambda : S^1 \to S^1$ be the analytic orientation preserving diffeomorphism induced by $\tilde{f}_\lambda : \mathbb{R} \to \mathbb{R}$ given by the formula

$$\tilde{f}_\lambda(t) = t + \lambda \sin 2\pi t.$$

Then we get the family $\tilde{f}_{\lambda,a}$, $a \in \mathbb{R}$, given by the formula

$$\tilde{f}_{\lambda,a}(t) = a + t + \lambda \sin 2\pi t.$$

It follows from Proposition 3.6.7 that the induced family $\{f_{\lambda,a} : a \in \mathbb{R}\}$ of analytic orientation preserving circle diffeomorphisms satisfy the hypothesis of Lemma 3.6.2 and therefore gives a devil's staircase $\rho_\lambda : \mathbb{R} \to \mathbb{R}$ for each $0 < |\lambda| < \dfrac{1}{2\pi}$.

We shall prove now that there exists a dense subset D of $\mathcal{R}$ such that f_a is not topologically conjugate to the rotation by the angle $2\pi\rho(a)$ by an absolutely continuous homeomorphism for every $a \in D$. We shall need the following elementary observation.

3.6.8. Lemma. *Let $H : \mathbb{R} \to \mathbb{R}$ be an increasing homeomorphism such that for every $k \in \mathbb{N}$ there exists an open set $A_k \subset \mathbb{R}$ with Lebesgue measure $\lambda(A_k) < \dfrac{1}{k}$ and $\lambda(H(A_k)) \geq \dfrac{1}{2}$. Then, H is not absolutely continuous.*

Proof. Let $\delta > 0$ be any and let $k \in \mathbb{N}$ be such that $\dfrac{1}{k} < \delta$. The corresponding open set A_k of the hypothesis is the disjoint union of at most countably many open intervals (a_n, b_n), $n \in \mathbb{Z}$. Thus,

$$\sum_{n\in\mathbb{Z}} (b_n - a_n) < \delta.$$

On the other hand, by assumption,

$$\sum_{n\in\mathbb{Z}} (H(b_n) - H(a_n)) = \lambda(H(A_k)) \geq \frac{1}{2}$$

and so there exists some $N \in \mathbb{N}$ such that

$$\sum_{|n|\leq N} (H(b_n) - H(a_n)) > \frac{1}{3}.$$

This means that H is not absolutely continuous. □

3.6.9. Theorem. *There exists a dense subset D of $\mathcal{R}$ such that for each $a \in D$ there exists no absolutely continuous homeomorphism $h : S^1 \to S^1$ which conjugates the analytic diffeomorphism f_a to the rotation by the angle $2\pi\rho(a)$.*

Proof. For every $k \in \mathbb{N}$ we consider the open subset

$$D_k = \{a \in \mathcal{R} : \quad \text{there exist an open set} \quad A \subset S^1 \quad \text{with} \quad \lambda(A) < \frac{1}{k}$$

$$\text{and some} \quad N \in \mathbb{N} \quad \text{with} \quad f_a^N(S^1 \setminus A) \subset A\}.$$

of $\mathcal{R}$. By the Baire property of $\mathcal{R}$ and Lemma 3.6.8, it suffices to prove that each D_k is dense in $\mathcal{R}$. Indeed, if $a \in D$, there exists a sequence $(A_k)_{k\in\mathbb{N}}$ of open subsets of S^1 such that $\lambda(A_k) < \frac{1}{k}$ and a sequence $(N_k)_{k\in\mathbb{N}}$ of positive integers such that $f_a^{N_k}(S^1 \setminus A_k) \subset A_k$ for every $k \in \mathbb{N}$. If $H : \mathbb{R} \to \mathbb{R}$ be an increasing homeomorphism such that $H \circ \tilde{f}_a \circ H^{-1} = T_{\rho(a)}$, then $\lambda(H(A_k)) = \lambda(T^n_{\rho(a)}(H(A_k))) = \lambda(H(\tilde{f}^n_a(A_k)))$ for every $n \in \mathbb{Z}$. In particular, we get

$$1 - \lambda(H(A_k)) = 1 - \lambda(H(\tilde{f}^n_a(A_k))) = \lambda(H(\tilde{f}^n_a(S^1 \setminus A_k))) \leq \lambda(H(A_k))$$

and so $\lambda(H(A_k)) \geq \frac{1}{2}$. By Lemma 3.6.8, H is not absolutely continuous.

It remains to prove that each D_k is dense in $\mathcal{R}$. Let $a \in \mathcal{R}$. Arbitrarily close to a we can find some $b \in K^{\pm}_{p/q}$, for some $\frac{p}{q} \in \mathbb{Q}$. Since f_b is analytic, it has a finite number of periodic points z_1,..., z_m, all of period q, and the orbit $\mathcal{O}_{f_b}(z_j)$ of each z_j is attracting from one side and repelling from the other. Also the positive and the negative limit set of any other orbit of f_b is contained in $\mathcal{O}_{f_b}(z_1) \cup \cdots \cup \mathcal{O}_{f_b}(z_m)$. If A_k is any open neighbourhood of the finite set $\mathcal{O}_{f_b}(z_1) \cup \cdots \cup \mathcal{O}_{f_b}(z_m)$ with $\lambda(A_k)) < \frac{1}{k}$, then there exists some $N_k \in \mathbb{N}$ such that $f_b^n(S^1 \setminus A_k) \subset A_k$ for all $n \geq N_k$. Since A_k is open and $S^1 \setminus A_k$ is compact, we have $f_c^{N_k}(S^1 \setminus A_k) \subset A_k$ for all c sufficiently close to b, by continuity. Thus, arbitrarily close to a there are points $c \in D_k$. This proves that D_k is dense in $\mathcal{R}$. □

For every $a \in \mathcal{R}$ the diffeomorphism f_a has a unique invariant Borel probability measure μ_a. If $h : S^1 \to S^1$ is an orientation preserving homeomorphism which conjugates f_a to the rotation by the angle $2\pi\rho(a)$, then $h_*\mu_a$ is the (normalized) Lebesgue measure on S^1 and h is absolutely continuous if and only if μ_a is absolutely continuous to the Lebesgue measure, because h lifts to an increasing homeomorphism of $\mathbb{R}$.

Chapter 4

Ergodicity

4.1 Ergodic endomorphisms

Let $(X, \mathcal{A}, \mu)$ be a probability space. An endomorphism $T : X \to X$ is called *ergodic* if for any $A \in \mathcal{A}$ such that $\mu(A \triangle T^{-1}(A)) = 0$ we have $\mu(A) = 0$ or 1.

4.1.1. Proposition. *Let $(X, \mathcal{A}, \mu)$ be a probability space. For an endomorphism $T : X \to X$ the following assertions are equivalent.*

(i) T is ergodic.

(ii) $\mu(A) = 0$ or 1, for every $A \in \mathcal{A}$ such that $T^{-1}(A) = A$.

(iii) For every A, $B \in \mathcal{A}$ such that $\mu(A) > 0$ and $\mu(B) > 0$ there exists $n \in \mathbb{Z}^+$ such that $\mu(T^{-n}(A) \cap B) > 0$.

Proof. It is trivial that (i) implies (ii). For the converse, let $A \in \mathcal{A}$ be such that $\mu(A \triangle T^{-1}(A)) = 0$. It suffices to find $B \in \mathcal{A}$ such that $\mu(A \triangle B) = 0$ and $T^{-1}(B) = B$, because then from our assumption we have $\mu(A) = 0$ or 1. Set

$$B = \bigcap_{n=0}^{\infty} \bigcup_{k=n}^{\infty} T^{-k}(A).$$

Then clearly $T^{-1}(B) = B$, and

$$T^{-k}(A) \triangle A \subset \bigcup_{i=0}^{k-1} T^{-(i+1)}(A) \triangle T^{-i}(A) = \bigcup_{i=0}^{k-1} T^{-i}(T^{-1}(A) \triangle A).$$

Thus,

$$\mu(T^{-k}(A) \triangle A) \leq \sum_{i=0}^{k-1} \mu(T^{-i}(T^{-1}(A) \triangle A)) = k\mu(T^{-1}(A) \triangle A) = 0$$

for every $k \in \mathbb{Z}^+$. Moreover,

$$\Big(\bigcup_{k=n}^{\infty} T^{-k}(A) \Big) \triangle A \subset \bigcup_{k=n}^{\infty} T^{-k}(A) \triangle A$$

and therefore

$$\mu\bigg(\Big(\bigcup_{k=n}^{\infty} T^{-k}(A)\Big)\triangle A\bigg) = 0$$

for every $n \in \mathbb{Z}^+$. It follows that

$$\mu(A\triangle B) = \lim_{n\to+\infty} \mu\bigg(\Big(\bigcup_{k=n}^{\infty} T^{-k}(A)\Big)\triangle A\bigg) = 0.$$

To prove that (i) implies (iii) let A, $B \in \mathcal{A}$ be such that $\mu(A) > 0$ and $\mu(B) > 0$, but $\mu(T^{-n}(A) \cap B) = 0$ for every $n \in \mathbb{Z}^+$. If $C = \cup_{n=1}^{\infty} T^{-n}(A)$, then $T^{-1}(C) \subset C$ and $\mu(C \cap B) = 0$. Hence $\mu(C) = 0$ or 1, because $\mu(C\triangle T^{-1}(C)) = \mu(C) - \mu(T^{-1}(C)) = 0$. On the other hand, $\mu(C) \geq \mu(T^{-1}(A)) = \mu(A) > 0$, and we must have $\mu(C) = 1$. But now $\mu(C \cup B) = 1$ also, and therefore

$$1 = \mu(C \cup B) = \mu(C) + \mu(B) - \mu(C \cap B) = 1 + \mu(B),$$

that is $\mu(B) = 0$, which contradicts the assumption. Finally, we prove that (iii) implies (ii). Let $A \in \mathcal{A}$ be such that $T^{-1}(A) = A$ with $0 < \mu(A) < 1$. If (iii) is true and since $\mu(A) > 0$ and $\mu(X \setminus A) > 0$, there exists $n \in \mathbb{Z}^+$ such that $\mu(T^{-n}(A) \cap (X \setminus A)) > 0$. This is impossible, because $T^{-n}(A) = A$ for every $n \in \mathbb{Z}^+$. □

Let $(X, \mathcal{A}, \mu)$ be a probability space and $T : X \to X$ be an endomorphism. A measurable function $f : X \to \mathbb{R}$ is called *T-invariant μ-almost everywhere* if $f = f \circ T$ μ-almost everywhere. If $f = f \circ T$ everywhere on X, then f is called *T-invariant*. If f is T-invariant μ-almost everywhere, there is a measurable T-invariant function $g : X \to \mathbb{R}$ such that $g = f$ μ-almost everywhere. Indeed, define $g = f$ on the set

$$\bigcap_{k=0}^{\infty} T^{-k}(\{x \in X : f(x) = f(T(x))\})$$

and $g = 0$ everywhere else. If T is an automorphism, we take the intersection from $-\infty$ to $+\infty$.

4.1.2. Proposition. *Let $(X, \mathcal{A}, \mu)$ be a probability space and $T : X \to X$ be an endomorphism. The following are equivalent.*

(i) T is ergodic.

(ii) Every measurable T-invariant μ-almost everywhere function is constant μ-almost everywhere.

(iii) Every measurable T-invariant μ-almost everywhere function in $L^2(\mu)$ is constant μ-almost everywhere.

Proof. To prove that (i) implies (ii) let $f : X \to \mathbb{R}$ be a measurable T-invariant μ-almost everywhere function. For every $k \in \mathbb{Z}$ and $n \in \mathbb{Z}^+$, let

$$A(k,n) = \{x \in X : \frac{k}{2^n} \leq f(x) < \frac{k+1}{2^n}\}.$$

By the invariance of f, we have $\mu(A(k,n)\triangle T^{-1}(A(k,n)) = 0$, and therefore $\mu(A(k,n)) = 0$ or 1, by the ergodicity of T. Since for every fixed $n \in \mathbb{Z}^+$ the

family $\{A(k,n) : k \in \mathbb{Z}\}$ is a partition of X, there exists some $k_n \in \mathbb{Z}$ such that $\mu(A(k_n,n)) = 1$. If $Y = \cap_{n=1}^{\infty} A(k_n,n)$, then $\mu(Y) = 1$ and f is constant on Y, because for any x, $y \in Y$ we have $|f(x) - f(y)| < 2^{-n}$ for every $n \in \mathbb{Z}^+$. It remains to prove that (iii) implies (i). Let $A \in \mathcal{A}$ be such that $\mu(A \triangle T^{-1}(A)) = 0$. Then, $\chi_A \in L^2(\mu)$ and is T-invariant μ-almost everywhere. If (iii) is true, then χ_A is constant μ-almost everywhere, which means exactly that $\mu(A) = 0$ or 1. $\square$

In the above we have fixed a probability space and defined when an endomorphism will be called ergodic. Suppose that X is a compact metrizable space and $T : X \to X$ is a continuous onto map. An element $\mu \in \mathcal{M}_T(X)$ is called *ergodic* if T is ergodic with respect to μ.

4.1.3. Proposition. *If $T : X \to X$ is a continuous, onto map of the compact metrizable space X, then the ergodic measures are precisely the extreme points of $\mathcal{M}_T(X)$.*

Proof. Let $\mu \in \mathcal{M}_T(X)$ be non-ergodic. Then, there exists a Borel set A such that $T^{-1}(A) = A$ and $0 < \mu(A) < 1$. The Borel measures μ_1 and μ_2 defined by

$$\mu_1(B) = \frac{\mu(A \cap B)}{\mu(A)}, \quad \mu_2(B) = \frac{\mu((X \setminus A) \cap B)}{\mu(X \setminus A)}$$

belong to $\mathcal{M}_T(X)$, are different, and $\mu = \mu(A)\mu_1 + (1 - \mu(A))\mu_2$. Hence μ is not an extreme point of $\mathcal{M}_T(X)$. For the converse, suppose that $\mu \in \mathcal{M}_T(X)$ is ergodic, but is not an extreme point, that is there exist μ_1, $\mu_2 \in \mathcal{M}_T(X)$ such that $\mu_1 \neq \mu_2$ and $\mu = t\mu_1 + (1-t)\mu_2$, for some $0 < t < 1$. Obviously, μ_1 is absolutely continuous with respect to μ, and so the Radon-Nikodym derivative $d\mu_1/d\mu$ exists. Let

$$A_1 = \{x \in X : \frac{d\mu_1}{d\mu}(x) < 1\}, \quad A_2 = \{x \in X : \frac{d\mu_1}{d\mu}(x) > 1\}.$$

Since

$$\int_{A_1 \cap T^{-1}(A_1)} \frac{d\mu_1}{d\mu} d\mu + \int_{A_1 \setminus T^{-1}(A_1)} \frac{d\mu_1}{d\mu} d\mu = \mu_1(A_1) =$$

$$\mu_1(T^{-1}(A_1)) = \int_{A_1 \cap T^{-1}(A_1)} \frac{d\mu_1}{d\mu} d\mu + \int_{T^{-1}(A_1) \setminus A_1} \frac{d\mu_1}{d\mu} d\mu,$$

we have

$$\int_{A_1 \setminus T^{-1}(A_1)} \frac{d\mu_1}{d\mu} d\mu = \int_{T^{-1}(A_1) \setminus A_1} \frac{d\mu_1}{d\mu} d\mu.$$

Moreover,

$$\mu(A_1 \setminus T^{-1}(A_1)) = \mu(A_1) - \mu(A_1 \cap T^{-1}(A_1)) =$$

$$\mu(T^{-1}(A_1)) - \mu(A_1 \cap T^{-1}(A_1)) = \mu(T^{-1}(A_1) \setminus A_1).$$

If now $\mu(A_1 \setminus T^{-1}(A_1)) > 0$, then

$$\mu(A_1 \setminus T^{-1}(A_1)) > \int_{A_1 \setminus T^{-1}(A_1)} \frac{d\mu_1}{d\mu} d\mu = \int_{T^{-1}(A_1) \setminus A_1} \frac{d\mu_1}{d\mu} d\mu \geq \mu(T^{-1}(A_1) \setminus A_1).$$

This contradiction shows that we must necessarily have $\mu(A_1 \setminus T^{-1}(A_1)) = \mu(T^{-1}(A_1) \setminus A_1) = 0$ or in other words $\mu(A_1 \triangle T^{-1}(A_1)) = 0$. Hence $\mu(A_1) = 0$ or 1, since μ is assumed to be ergodic. If $\mu(A_1) = 1$, then

$$1 = \mu_1(X) = \int_{A_1} \frac{d\mu_1}{d\mu} d\mu < \mu(A_1) = 1.$$

This contradiction shows that $\mu(A_1) = 0$. Similarly, $\mu(A_2) = 0$. It follows that

$$\frac{d\mu_1}{d\mu} = 1$$

μ-almost everywhere, and therefore $\mu = \mu_1$, contradiction. □

4.1.4. Corollary. *If the continuous, onto map $T : X \to X$ of the compact metrizable space X is uniquely ergodic, then it is ergodic with respect to the unique T-invariant Borel probability measure on X.*

This gives our first example of an ergodic endomorphism. Namely, a left translation of a compact, metrizable topological group, which has a dense orbit is ergodic with respect to the Haar measure. In particular, from Kronecker's theorem we have the following.

4.1.5. Corollary. *If the real numbers 1, a_1,...,a_k are linearly independent over $\mathbb{Q}$, the translation*

$$T(e^{2\pi i x_1}, ..., e^{2\pi i x_k}) = (e^{2\pi i (x_1 + a_1)}, ..., e^{2\pi i (x_n + a_k)})$$

of the k-torus is ergodic with respect to the Haar measure.

Recall that if X is a non-empty set, a class $\mathcal{S}$ of subsets of X is called a *semialgebra* if (i) $\varnothing \in \mathcal{S}$, (ii) $A \cap B \in \mathcal{S}$ for every A, $B \in \mathcal{S}$ and (iii) for every $A \in \mathcal{S}$ there exist mutually disjoint E_1,...,$E_n \in \mathcal{S}$ such that $X \setminus A = E_1 \cup ... \cup E_n$. A class of subsets of X is called an *algebra* if it contains $\varnothing$ and is closed under finite unions and complements. The intersection of algebras is an algebra. The smallest algebra which contains a class $\mathcal{C}$ of subsets of X is called the algebra generated by $\mathcal{C}$. The algebra generated by a semialgebra $\mathcal{S}$ consists of sets of the form $A_1 \cup ... \cup A_n$, where A_1,...,$A_n \in \mathcal{S}$ are mutually disjoint. If $(X, \mathcal{A}, \mu)$ is a probability space and the σ-algebra $\mathcal{A}$ is generated by an algebra $\mathcal{E}$, then for every $A \in \mathcal{A}$ and $\epsilon > 0$ there exists $E \in \mathcal{E}$ such that $\mu(A \triangle E) < \epsilon$.

4.1.6. Lemma. *Let $(X, \mathcal{A}, \mu)$ be a probability space and $T : X \to X$ be an endomorphism. Then, T is ergodic, if there is a semialgebra $\mathcal{S}$ which generates the σ-algebra $\mathcal{A}$ such that*

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

for every A, $B \in \mathcal{S}$.

Proof. It suffices to prove that the hypothesis extends to all members of $\mathcal{A}$, because then the conclusion is an immediate consequense of 4.1.1. Let $\mathcal{E}$ be the algebra generated by $\mathcal{S}$. It is clear that the hypothesis is extended to all members of $\mathcal{E}$. It is also clear that $\mathcal{E}$ generates $\mathcal{A}$. Let A, $B \in \mathcal{A}$ and $\epsilon > 0$. There exist E, $G \in \mathcal{E}$ such that $\mu(A \triangle E) < \epsilon$ and $\mu(B \triangle G) < \epsilon$. For every $k \in \mathbb{Z}^+$ we have

$$(T^{-k}(A) \cap B) \triangle (T^{-k}(E) \cap G) \subset (T^{-k}(A) \triangle T^{-k}(E)) \cup (B \triangle G),$$

and therefore

$$|\mu(T^{-k}(A) \cap B) - \mu(T^{-k}(E) \cap G)| \leq \mu((T^{-k}(A) \cap B) \triangle (T^{-k}(E) \cap G)) < 2\epsilon.$$

It follows that

$$\big|\frac{1}{n}\sum_{k=0}^{n-1} \mu(T^{-k}(A) \cap B) - \mu(A)\mu(B)\big| \leq$$

$$\frac{1}{n}\big|\sum_{k=0}^{n-1} \mu(T^{-k}(A) \cap B) - \sum_{k=0}^{n-1} \mu(T^{-k}(E) \cap G)\big| + \big|\frac{1}{n}\sum_{k=0}^{n-1} \mu(T^{-k}(E) \cap G) - \mu(E)\mu(G)\big|$$

$$+\big|\mu(E)\mu(G) - \mu(A)\mu(B)\big| <$$

$$2\epsilon + \big|\frac{1}{n}\sum_{k=0}^{n-1} \mu(T^{-k}(E) \cap G) - \mu(E)\mu(G)\big| + \big|\mu(E)\mu(G) - \mu(A)\mu(G)\big|$$

$$+\big|\mu(A)\mu(G) - \mu(A)\mu(B)\big| <$$

$$2\epsilon + \big|\frac{1}{n}\sum_{k=0}^{n-1} \mu(T^{-k}(E) \cap G) - \mu(E)\mu(G)\big| + \mu(A \triangle E)\mu(G) + \mu(B \triangle G)\mu(A) <$$

$$4\epsilon + \big|\frac{1}{n}\sum_{k=0}^{n-1} \mu(T^{-k}(E) \cap G) - \mu(E)\mu(G)\big|.$$

Hence

$$\lim_{n \to +\infty} \big|\frac{1}{n}\sum_{k=0}^{n-1} \mu(T^{-k}(A) \cap B) - \mu(A)\mu(B)\big| < 4\epsilon. \quad \square$$

Let now $(E, \mathcal{F}, \mu)$ be a probability space and $\tau : E^{\mathbb{Z}^+} \to E^{\mathbb{Z}^+}$ be the shift. The set $\mathcal{S}$ of all cylinders $\pi_{i_1}^{-1}(A_{i_1}) \cap ... \cap \pi_{i_n}^{-1}(A_{i_n})$, A_{i_1},...,$A_{i_n} \in \mathcal{F}$, $n \in \mathbb{Z}^+$, is a semialgebra. If $A = \pi_{i_1}^{-1}(A_{i_1}) \cap ... \cap \pi_{i_n}^{-1}(A_{i_n})$ and $B = \pi_{j_1}^{-1}(B_{j_1}) \cap ... \cap \pi_{j_m}^{-1}(B_{j_m})$, there exists some $k_0 \in \mathbb{Z}^+$ such that

$$\max\{j_1, ..., j_m\} < k_0 + \min\{i_1, ..., i_n\},$$

and for every $k \geq k_0$ we have

$$\tau^{-k}(A) \cap B = \pi_{i_1+k}^{-1}(A_{i_1}) \cap ... \cap \pi_{i_n+k}^{-1}(A_{i_n}) \cap \pi_{j_1}^{-1}(B_{j_1}) \cap ... \cap \pi_{j_m}^{-1}(B_{j_m}).$$

Obviously, $\mu^{\mathbb{Z}^+}(\tau^{-k}(A) \cap B) = \mu^{\mathbb{Z}^+}(A)\mu^{\mathbb{Z}^+}(B)$, for every $k \geq k_0$. Hence

$$\frac{1}{n}\sum_{k=0}^{n-1} \mu^{\mathbb{Z}^+}(\tau^{-k}(A) \cap B) = \frac{1}{n}\sum_{k=0}^{k_0-1} \mu^{\mathbb{Z}^+}(\tau^{-k}(A) \cap B) + \frac{1}{n}\sum_{k=k_0}^{n-1} \mu^{\mathbb{Z}^+}(A)\mu^{\mathbb{Z}^+}(B) =$$

$$\frac{1}{n}\sum_{k=0}^{k_0-1}\mu^{\mathbb{Z}^+}(\tau^{-k}(A)\cap B)+\frac{n-k_0}{n}\mu^{\mathbb{Z}^+}(A)\mu^{\mathbb{Z}^+}(B),$$

from which follows that

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{k=0}^{n-1}\mu^{\mathbb{Z}^+}(\tau^{-k}(A)\cap B)=\mu^{\mathbb{Z}^+}(A)\mu^{\mathbb{Z}^+}(B).$$

From 4.1.6 we conclude that the shift τ is an ergodic endomorphism of the product probability space $(E^{\mathbb{Z}^+},\mathcal{F}^{\mathbb{Z}^+},\mu^{\mathbb{Z}^+})$.

4.2 The ergodic theorem

In physics the orbit of a point under an endomorphism $T:X\to X$ of a probability space $(X,\mathcal{A},\mu)$ represents the history of a phase of the studied physical system. The σ-algebra $\mathcal{A}$ describes all the observable events and μ the probability of occurence of each event. The measurememt of a physical parameter is represented mathematically by a measurable function $f:X\to\mathbb{R}$. The measurement is carried out in many succesive times and the time average

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

is of interest for large n. A basic problem is whether this average has a limit for $n\to+\infty$. If the limit exists, it is taken as the central value of f. In this section we shall prove the celebrated ergodic theorem of G. Birkhoff, which assures the existence of the limits of the time averages μ-almost everywhere.

4.2.1. Theorem (Ergodic theorem of Birkhoff). *Let $(X,\mathcal{A},\mu)$ be a probability space and $T:X\to X$ be an endomorphism. Then, for every $f\in L^1(\mu)$ the limit*

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

exists and is T-invariant μ-almost everywhere. Moreover, $f^*\in L^1(\mu)$ and

$$\int_X f^*d\mu=\int_X fd\mu.$$

In particular for ergodic endomorphisms we have the following corollary, known as the ergodic hypothesis in 19th century physics.

4.2.2. Corollary. *Let $(X,\mathcal{A},\mu)$ be a probability space and $T:X\to X$ be an ergodic endomorphism. Then, for every $f\in L^1(\mu)$,*

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{k=0}^{n-1}f(T^k(x))=\int_X fd\mu$$

μ-almost everywhere.

For convenience in the sequel we set

$$S_nf(x) = \sum_{k=0}^{n-1} f(T^k(x)),$$

$S_0f = 0$ and $S_n^*f = \max\{S_kf : 0 \leq k \leq n\}$. We shall need the following.

4.2.3. Lemma (Maximal ergodic theorem). *Let $(X, \mathcal{A}, \mu)$ be a probability space, $T : X \to X$ be an endomorphism and $f \in L^1(\mu)$. Then, for every $n \in \mathbb{Z}^+$ we have*

$$\int_{\{x\in X : S_n^*f(x)>0\}} f d\mu \geq 0.$$

Proof. Obviously, S_nf, $S_n^*f \in L^1(\mu)$, and for every $0 \leq k \leq n$ we have $S_n^*f \geq S_kf$ and $S_n^*f(T(x)) \geq S_kf(T(x)) = S_{k+1}f(x) - f(x)$. Thus, if $S_n^*f(x) > 0$, then

$$S_n^*f(T(x)) + f(x) \geq \max\{S_kf(x) : 1 \leq k \leq n\} = S_n^*f(x).$$

It follows that

$$\int_{\{x\in X : S_n^*f(x)>0\}} f d\mu \geq \int_{\{x\in X : S_n^*f(x)>0\}} S_n^*f d\mu - \int_{\{x\in X : S_n^*f(x)>0\}} (S_n^*f) \circ T d\mu =$$

$$\int_X S_n^*f d\mu - \int_{\{x\in X : S_n^*f(x)>0\}} (S_n^*f) \circ T d\mu \geq \int_X S_n^*f d\mu - \int_X (S_n^*f) \circ T d\mu = 0. \quad \square$$

4.2.4. Corollary. *Let $(X, \mathcal{A}, \mu)$ be a probability space, $T : X \to X$ be an endomorphism and $f \in L^1(\mu)$. If*

$$B_a = \{x \in X : \sup\{\frac{1}{n} S_nf(x) : n \geq 1\} > a\},$$

then for every $A \in \mathcal{A}$ such that $T^{-1}A = A$ we have

$$a\mu(A \cap B_a) \leq \int_{A\cap B_a} f d\mu.$$

Proof. Suppose first that $A = X$. If $g = f - a$ and $G_n = \{x \in X : S_n^*g(x) > 0\}$, then $G_n \subset G_{n+1}$ and $B_a = \cup_{n=1}^{\infty} G_n$. From 4.2.3 we have

$$\int_{G_n} g d\mu \geq 0$$

for every $n \geq 1$ and therefore

$$\int_{B_a} g d\mu = \lim_{n\to+\infty} \int_{G_n} g d\mu \geq 0.$$

The general case follows by applying the above to $T|A$. $\square$

Proof of 4.2.1. Let

$$f^*(x) = \limsup_{n\to+\infty} \frac{1}{n} S_n f(x) \quad \text{and} \quad f_*(x) = \liminf_{n\to+\infty} \frac{1}{n} S_n f(x).$$

Let also $E = \{x \in X : f^*(x) \neq f_*(x)\}$ and for every a, $b \in \mathbb{Q}$ with $a > b$ let $E(a,b) = \{x \in X : f_*(x) < b \text{ and } f^*(x) > a\}$. Then

$$E = \bigcup_{a,b\in\mathbb{Q}, a>b} E(a,b)$$

and to prove that the limit of the averages exists μ-almost everywhere, it suffices to prove that $\mu(E(a,b)) = 0$ for every a, $b \in \mathbb{Q}$ with $a > b$. Since

$$\frac{1}{n}|S_{n+1}f(x) - S_n f(T(x))| = \frac{|f(x)|}{n},$$

it is obvious that $f^* \circ T = f^*$ and $f_* \circ T = f_*$. It follows that $E(a,b) = T^{-1}(E(a,b))$ and from 4.2.4 we have

$$a\mu(E(a,b)) = a\mu(E(a,b) \cap B_a) \leq \int_{E(a,b)} f d\mu,$$

where B_a is defines as in 4.2.4, since $E(a,b) \subset B_a$. Applying this to $-f$ and $-b$, $-a$, we also have

$$-b\mu((E(a,b)) \leq \int_{E(a,b)} (-f) d\mu.$$

Therefore, $(a-b)\mu(E(a,b)) \leq 0$ and so necessarily $\mu(E(a,b)) = 0$. That $f^* \in L^1(\mu)$ follows now from Fatou's lemma, since

$$\int_X |f^*| d\mu = \int_X |f_*| d\mu = \int_X \big(\liminf_{n\to+\infty} \frac{1}{n}|S_n f|\big) d\mu \leq$$

$$\liminf_{n\to+\infty} \frac{1}{n} \int_X |S_n f| d\mu \leq \liminf_{n\to+\infty} \frac{1}{n} \sum_{k=0}^{n-1} \int_X |f \circ T^k| d\mu = \|f\|_1.$$

It remains to prove that the integral of f^* is equal to the integral of f. For $k \in \mathbb{Z}$ and $n \in \mathbb{N}$ let

$$A(k,n) = \{x \in X : \frac{k}{n} \leq f^*(x) < \frac{k+1}{n}\}.$$

Then, for every $\epsilon > 0$ we have $A(k,n) \subset B_{\frac{k}{n}-\epsilon}$, and from 4.2.4,

$$\int_{A(k,n)} f d\mu \geq \big(\frac{k}{n} - \epsilon\big)\mu(A(k,n)).$$

It follows that

$$\int_{A(k,n)} f d\mu \geq \frac{k}{n}\mu(A(k,n)).$$

On the other hand, from the very definition of $A(k,n)$ we have

$$\int_{A(k,n)} f^* d\mu \leq \frac{k+1}{n}\mu(A(k,n)) \leq \int_{A(k,n)} f d\mu + \frac{1}{n}\mu(A(k,n)).$$

Summing up over all $k \in \mathbb{Z}$, we get

$$\int_X f^* d\mu \leq \int_X f d\mu + \frac{1}{n}$$

for every $n \in \mathbb{N}$, and hence

$$\int_X f^* d\mu \leq \int_X f d\mu.$$

Applying this to $-f$ we also have

$$\int_X (-f^*) d\mu = \int_X (-f_*) d\mu = \int_X (-f)^* d\mu \leq \int_X (-f) d\mu,$$

and therefore

$$\int_X f^* d\mu = \int_X f d\mu. \quad \square$$

If in the ergodic theorem we start with a function $f \in L^p(\mu)$, $p \geq 1$, then $f \in L^1(\mu)$ and the limit

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

exists and is T-invariant μ-almost everywhere. According to the following, the limit exists also in $L^p(\mu)$ and is the same μ-almost everywhere.

4.2.5. Corollary (L^p ergodic theorem of von Neumann). *Let $(X, \mathcal{A}, \mu)$ be a probability space, $T : X \to X$ be an endomorphism and $f \in L^p(\mu)$, $p \geq 1$. Then there exists a T-invariant μ-almost everywhere $f^* \in L^p(\mu)$, such that*

$$\lim_{n\to+\infty} \|\frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k - f^*\|_p = 0.$$

Proof. Suppose first that $f \in L^\infty(\mu)$. Then, $f \in L^p(\mu)$ for every $p \geq 1$, and by the ergodic theorem, there exists $f^* \in L^1(\mu)$ such that

$$\lim_{n\to+\infty} \frac{1}{n} S_n f = f^*$$

μ-almost everywhere. Of course, $f^* \in L^\infty(\mu)$. From the bounded convergence theorem of Lebesgue we have

$$\lim_{n\to+\infty} \|\frac{1}{n} S_n f - f^*\|_p = 0.$$

Thus, for every $f \in L^\infty(\mu)$ and $\epsilon > 0$ there exists some $N(\epsilon, f) \in \mathbb{N}$ such that

$$\|\frac{1}{n}S_n f - \frac{1}{n+k}S_{n+k}f\|_p < \epsilon$$

for $n \geq N(\epsilon, f)$ and $k \in \mathbb{Z}^+$. Let now $f \in L^p(\mu)$ and $\epsilon > 0$. There exists $g \in L^\infty(\mu)$ such that $\|g - f\|_p < \epsilon$. For $n \geq N(\epsilon/3, g)$ and $k \in \mathbb{Z}^+$ we have

$$\|\frac{1}{n}S_n f - \frac{1}{n+k}S_{n+k}f\|_p \leq$$

$$\|\frac{1}{n}S_n g - \frac{1}{n+k}S_{n+k}g\|_p + \frac{1}{n}\|S_n g - S_n f\|_p + \frac{1}{n+k}\|S_{n+k}f - S_{n+k}g\|_p \leq$$

$$\frac{\epsilon}{3} + \frac{1}{n}\|f - g\|_p + \|f - g\|_p < \epsilon.$$

By completeness of $L^p(\mu)$, there exists $f^* \in L^p(\mu)$ such that

$$\lim_{n\to+\infty} \|\frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k - f^*\|_p = 0.$$

The T-invariance follows from the observation that

$$\frac{1}{n}\|S_{n+1}f - S_n f \circ T\|_p = \frac{1}{n}\|f\|_p$$

for every $n \in \mathbb{N}$, and so the sequences $(\frac{1}{n}S_n f)_{n\in\mathbb{N}}$ and $(\frac{1}{n}S_n f \circ T)_{n\in\mathbb{N}}$ must have the same limits in $L^p(\mu)$. □

4.2.6. Corollary (Strong law of large numbers). *Let $(X, \mathcal{A}, \mu)$ be a probability space and $f_n : X \to \mathbb{R}$, $n \in \mathbb{Z}^+$, be a sequence of independent and identically distributed random variables. Then,*

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{k=0}^{n-1} f_k = \int_X f_0 d\mu$$

μ-almost everywhere.

Proof. Let $\nu = (f_0)_*\mu$ be the common distribution of f_n, $n \in \mathbb{Z}^+$. The product measure $\nu^{\mathbb{Z}^+}$ on $(\mathbb{R}^{\mathbb{Z}^+}, \mathcal{B}^{\mathbb{Z}^+})$ is invariant by the shift, where $\mathcal{B}$ denotes the σ-algebra of Borel subsets of $\mathbb{R}$. From the ergodicity of the shift and the ergodic theorem, if

$$K = \{y \in \mathbb{R}^{\mathbb{Z}^+} : \lim_{n\to+\infty} \frac{1}{n}\sum_{k=0}^{n-1} \pi_0(\tau^k(y)) = \int_{\mathbb{R}^{\mathbb{Z}^+}} \pi_0 d\nu^{\mathbb{Z}^+}\},$$

then $\nu^{\mathbb{Z}^+}(K) = 1$, where π_i is the projection to the i-th term. But $\pi_0 \circ \tau^k = \pi_k$ and

$$\int_{\mathbb{R}^{\mathbb{Z}^+}} \pi_0 d\nu^{\mathbb{Z}^+} = \int_{\mathbb{R}} id_{\mathbb{R}} d\nu = \int_{\mathbb{R}} id_{\mathbb{R}} d(f_0)_*\mu = \int_X f_0 d\mu.$$

Hence

$$K = \{y \in \mathbb{R}^{\mathbb{Z}^+} : \lim_{n\to+\infty} \frac{1}{n}\sum_{k=0}^{n-1} \pi_k(y) = \int_X f_0 d\mu\}.$$

If now $f = (f_n)_{n\in\mathbb{Z}^+}$, then $f : X \to \mathbb{R}^{\mathbb{Z}^+}$ is measurable and

$$f^{-1}(K) = \{x \in X : \lim_{n\to+\infty} \frac{1}{n}\sum_{k=0}^{n-1} f_k(x) = \int_X f_0 d\mu\},$$

while $\mu(f^{-1}(K)) = (f_*\mu)(K) = \nu^{\mathbb{Z}^+}(K) = 1$. □

There is a class of homeomorphisms of compact metrizable spaces, which satisfy a strong form of the ergodic theorem for continuous functions. Let X be a compact metrizable space and d be a compatible metric on X. A homeomorphism $h : X \to X$ is called *regular* if its iterates $\{h^n : n \in \mathbb{Z}\}$ form an equicontinuous family. This property is independent of the choice of metric, since X is compact, and is equivalent to saying that h is an isometry with respect to some compatible metric on X. Such a metric can be defined by

$$d^*(x, y) = \sup\{d(h^n(x), h^n(y)) : n \in \mathbb{Z}\}.$$

It is obvious that if h is regular, then $y \in L^+(x)$ if and only if $x \in L^-(y)$. This implies that the orbit closure of each point $x \in X$ coincides with $L^+(x)$ and $L^-(x)$, and is a minimal set. It follows from 3.2.1 and 3.2.4 that the time averages of every continuous function converge pointwise. In fact more is true.

4.2.7. Theorem. *Let X be a compact metrizable space and $h : X \to X$ be a regular homeomorphism. Then for every continuous function $f : X \to \mathbb{R}$, there exists a continuous function $f^* : X \to \mathbb{R}$ such that*

$$f^* = \lim_{n\to+\infty} \frac{1}{n}\sum_{k=0}^{n-1} f \circ h^k$$

uniformly on X.

Proof. As we observed above, there exists a function $f^* : X \to \mathbb{R}$ such that

$$f^* = \lim_{n\to+\infty} \frac{1}{n}\sum_{k=0}^{n-1} f \circ h^k$$

pointwise on X. Since h is regular, the time averages form an equicontinuous family of functions, which is also uniformly bounded by $\|f\|$. Thus, by Ascoli's theorem, some subsequence converges uniformly to f^*. In particular f^* is uniformly continuous. It remains to prove uniform convergence. Let $\epsilon > 0$. By the uniform continuity of f^* and the equicontinuity of the time averages, there exists $\delta > 0$ such that if $d(x, y) < \delta$, then

$$\big|\frac{1}{n}S_n f(x) - \frac{1}{n}S_n f(y)\big| < \epsilon$$

and $|f^*(x) - f^*(y)| < \epsilon$. Since X is compact, there are x_1,...,$x_m \in X$ such that $X = S(x_1, \delta) \cup ... \cup S(x_m, \delta)$. There exists some $n_0 \in \mathbb{N}$ such that

$$\big|\frac{1}{n} S_n f(x_i) - f^*(x_i)\big| < \epsilon$$

for every $1 \leq i \leq m$ and $n \geq n_0$. If now $x \in X$, there exists some $1 \leq i \leq m$ such that $x \in S(x_i, \delta)$ and therefore

$$\big|\frac{1}{n} S_n f(x) - f^*(x)\big| \leq$$

$$\frac{1}{n}\big|S_n f(x) - S_n f(x_i)\big| + \big|\frac{1}{n} S_n f(x_i) - f^*(x_i)\big| + \big|f^*(x_i) - f^*(x)\big| < 3\epsilon$$

for every $n \geq n_0$. □

4.3 Ergodic decomposition of invariant measures

Let X be a compact metrizable space and $T : X \to X$ be a continuous onto map. A Borel set $E \subset X$ is called of *zero probability* if $\mu(E) = 0$ for every $\mu \in \mathcal{M}_T(X)$ and of *maximum probability* if $\mu(E) = 1$ for every $\mu \in \mathcal{M}_T(X)$. A point $x \in X$ is called *quasi-regular* if for every $f \in C(X)$ the limit

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

exists in $\mathbb{R}$.

4.3.1. Theorem. *The set Q of all quasi-regular points is Borel, T-invariant and of maximum probability.*

Proof. The T-invariance of Q is obvious, since

$$\big|\frac{1}{n} \sum_{k=1}^{n} f(T^k(x)) - \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(x))\big| = \frac{1}{n}|f(T^n(x)) - f(x)| \leq \frac{2\|f\|}{n}.$$

Let $\{f_n : n \in \mathbb{N}\}$ be a countable dense subset of $C(X)$ and for every $n \in \mathbb{N}$ let

$$E_n = \{x \in X : \lim_{r \to +\infty} \frac{1}{r} \sum_{k=0}^{r-1} f_n(T^k(x)) \text{ does not exist in } \mathbb{R}\}.$$

For every n, m, $l \in \mathbb{N}$, the set

$$E_{n,m,l} = \{x \in X : \big|\frac{1}{n_1} \sum_{k=0}^{n_1-1} f_n(T^k(x)) - \frac{1}{n_2} \sum_{k=0}^{n_2-1} f_n(T^k(x))\big| \leq \frac{1}{m} \text{ for every } n_1, n_2 \geq l\}$$

is closed and

$$X \setminus E_n = \bigcap_{m=1}^{\infty} \bigcup_{l=1}^{\infty} E_{n,m.l}.$$

Thus, E_n is Borel for every $n \in \mathbb{N}$. It is also clear that if $E = \cup_{n=1}^{\infty} E_n$, then $Q \subset X \setminus E$. From the ergodic theorem we have $\mu(E_n) = 0$ for every $n \in \mathbb{N}$ and $\mu \in \mathcal{M}_T(X)$, and therefore E is a set of zero probability. So, it suffices to prove that $X \setminus E \subset Q$. Let $x \in X \setminus E$, $f \in C(X)$ and $\epsilon > 0$. There exists $n \in \mathbb{N}$ such that $\|f - f_n\| < \epsilon/3$. Since $x \in X \setminus E_n$, there exists $l \in \mathbb{N}$ such that

$$\Big|\frac{1}{n_1}\sum_{k=0}^{n_1-1} f_n(T^k(x)) - \frac{1}{n_2}\sum_{k=0}^{n_2-1} f_n(T^k(x))\Big| < \frac{\epsilon}{3}$$

for every $n_1, n_2 \geq l$, and then

$$\Big|\frac{1}{n_1}\sum_{k=0}^{n_1-1} f(T^k(x)) - \frac{1}{n_2}\sum_{k=0}^{n_2-1} f(T^k(x))\Big| \leq$$

$$\Big|\frac{1}{n_1}\sum_{k=0}^{n_1-1} f(T^k(x)) - \frac{1}{n_1}\sum_{k=0}^{n_1-1} f_n(T^k(x))\Big| + \Big|\frac{1}{n_1}\sum_{k=0}^{n_1-1} f_n(T^k(x)) - \frac{1}{n_2}\sum_{k=0}^{n_2-1} f_n(T^k(x))\Big| +$$

$$\Big|\frac{1}{n_2}\sum_{k=0}^{n_2-1} f_n(T^k(x)) - \frac{1}{n_2}\sum_{k=0}^{n_2-1} f(T^k(x))\Big| < \epsilon.$$

this shows that the limit

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

exists in $\mathbb{R}$. $\square$

It is obvious that for every $x \in Q$ the formula

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

defines a T-invariant positive linear functional $\mu_x : C(X) \to \mathbb{R}$ with $\mu_x(1) = 1$. In other words $\mu_x \in \mathcal{M}_T(X)$ for every $x \in Q$. We shall examine how μ_x, $x \in Q$, are related to each other and to the other elements of $\mathcal{M}_T(X)$. Of course, $\mu_x = \mu_y$, if $y = T^k(x)$, for some $k \in \mathbb{Z}^+$.

4.3.2. Lemma. *If $f \in C(X)$ and $\mu \in \mathcal{M}_T(X)$, then*

$$\int_X f d\mu = \int_Q \Big(\int_X f d\mu_x\Big) d\mu.$$

Proof. The function $g : Q \to \mathbb{R}$ with $g(x) = \mu_x(f)$ is measurable, as it is the pointwise limit of continuous functions. For every $n \in \mathbb{N}$ we have

$$\int_X f d\mu = \int_X \Big(\frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k\Big) d\mu$$

and by 4.3.1 and dominated convergence

$$\int_Q \left(\int_X f d\mu_x \right) d\mu = \lim_{n \to +\infty} \int_Q \left(\frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k \right) d\mu = \int_X f d\mu. \quad \square$$

Let $\mu \in \mathcal{M}_T(X)$. By the ergodic theorem, for every bounded measurable function $f : Q \to \mathbb{R}$, the limit

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

exists μ-almost everywhere. The set $E(\mu)$ of all bounded, measurable functions $f : Q \to \mathbb{R}$ such that

$$\int_X f d\mu_x = \tilde{f}(x)$$

μ-almost everywhere is a vector space and contains $C(X)$, by 4.3.1. In order to extent 4.3.2 to bounded, measurable functions, we shall need a series of lemmas.

4.3.3. Lemma. *If $(f_n)_{n \in \mathbb{N}}$ is a uniformly bounded sequence of elements of $E(\mu)$ and $f_n \to f$ pointwise, then $f \in E(\mu)$.*

Proof. By dominated convergence we have

$$\int_X f d\mu_x = \lim_{n \to +\infty} \int_X f_n d\mu_x = \lim_{n \to +\infty} \tilde{f}_n(x)$$

and from the ergodic theorem

$$\int_X |\tilde{f} - \tilde{f}_n| d\mu \leq \int_X \widetilde{|f - f_n|} d\mu = \int_X |f - f_n| d\mu \to 0$$

as $n \to +\infty$. Hence $\tilde{f}_n \to \tilde{f}$ in $L^1(\mu)$ and there exists a subsequence $(f_{n_k})_{k \in \mathbb{N}}$ such that $\tilde{f}_{n_k} \to \tilde{f}$ μ-almost everywhere. It follows that

$$\int_X f d\mu_x = \lim_{k \to +\infty} \int_X f_{n_k} d\mu_x = \lim_{k \to +\infty} \tilde{f}_{n_k}(x) = \tilde{f}(x)$$

μ-almost everywhere. $\square$

4.3.4. Lemma. *If $A \subset X$ is closed, then $\chi_A \in E(\mu)$.*

Proof. Since A is closed, there is a sequence of continuous functions $f_n : X \to [0, 1]$, $n \in \mathbb{N}$, such that $f_n \to \chi_A$ pointwise and the conclusion is immediate from 4.3.3. $\square$

4.3.5. Lemma. *If $A \subset X$ is a Borel set, then $\chi_A \in E(\mu)$.*

Proof. By the regularity of μ, there is a sequence of closed sets $A_1 \subset A_2 \subset ... \subset A$ such that

$$\mu(A \setminus \bigcup_{n=1}^{\infty} A_n) = 0.$$

Thus, $\chi_{A_n} \to \chi_A$ μ-almost everywhere, and the sequence $(\chi_{A_n})_{n\in\mathbb{N}}$ is dominated by χ_A. From the ergodic theorem

$$\int_X |\tilde{\chi}_{A_n} - \tilde{\chi}_A| d\mu \leq \int_X |\widetilde{\chi_{A_n} - \chi_A}| d\mu = \int_X |\chi_{A_n} - \chi_A| d\mu \to 0$$

as $n \to +\infty$. Hence $\tilde{\chi}_{A_n} \to \tilde{\chi}_A$ in $L^1(\mu)$ and there is a subsequence $(\chi_{A_{n_k}})_{k\in\mathbb{N}}$ such that $\tilde{\chi}_{A_{n_k}} \to \tilde{\chi}_A$ μ-almost everywhere. By dominated convergence we have

$$\int_X \chi_A d\mu_x \geq \limsup_{n\to+\infty} \int_X \chi_{A_n} d\mu_x = \limsup_{n\to+\infty} \tilde{\chi}_{A_n}(x) \geq \tilde{\chi}_A(x)$$

μ-almost everywhere. Similarly we have

$$\int_X \chi_{X\setminus A} d\mu_x \geq \tilde{\chi}_{X\setminus A}(x)$$

μ-almost everywhere. Hence

$$\int_X \chi_A d\mu_x = 1 - \int_X \chi_{X\setminus A} d\mu_x \leq 1 - \tilde{\chi}_{X\setminus A}(x) = \tilde{\chi}_A(x)$$

μ-almost everywhere. □

4.3.6. Proposition. *If $\mu \in \mathcal{M}_T(X)$ and $f : X \to \mathbb{R}$ is a bounded, measurable function, then*

$$\int_X f d\mu_x = \tilde{f}(x)$$

μ-almost everywhere on X.

Proof. It suffices to prove that $E(\mu)$ coincides with the space of all bounded measurable real functions. Indeed, every positive, bounded, measurable function is the pointwise limit of a sequence of linear combinations of characteristic functions of Borel subsets of X, and therefore belongs to $E(\mu)$, by 4.3.3 and 4.3.5. Finally, every bounded, measurable function is the difference of two positive, bounded, measurable functions. □

4.3.7. Corollary. *If $\mu \in \mathcal{M}_T(X)$ and $f : X \to \mathbb{R}$ is a bounded, measurable function, then*

$$\int_X f d\mu = \int_Q \left(\int_X f d\mu_x \right) d\mu.$$

4.3.8. Proposition. *If $\mu \in \mathcal{M}_T(X)$ and $f \in C(X)$, then*

$$\int_X |\tilde{f} - \tilde{f}(x)|^2 d\mu_x = 0$$

μ-almost everywhere on Q.

Proof. Since $\mu_x \in \mathcal{M}_T(X)$ for $x \in Q$, we have

$$\int_X \tilde{f} d\mu_x = \lim_{n\to+\infty} \int_X \left(\frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k\right) d\mu_x = \lim_{n\to+\infty} \int_X f d\mu_x = \tilde{f}(x).$$

Therefore,

$$\int_X |\tilde{f} - \tilde{f}(x)|^2 d\mu_x = \tilde{f}(x)^2 - 2\tilde{f}(x)\int_X \tilde{f} d\mu_x + \int_X \tilde{f}^2 d\mu_x = \int_X \tilde{f}^2 d\mu_x - \tilde{f}(x)^2.$$

Integrating with respect to μ, we get

$$\int_Q \left(\int_X |\tilde{f} - \tilde{f}(x)|^2 d\mu_x\right) d\mu = \int_Q \left(\int_X \tilde{f}^2 d\mu_x\right) d\mu - \int_Q \tilde{f}^2 d\mu.$$

Since $\tilde{f} : Q \to \mathbb{R}$ is bounded and measurable, so is $\tilde{f}^2$. Thus, from 4.3.7 we have

$$\int_Q \left(\int_X |\tilde{f} - \tilde{f}(x)|^2 d\mu_x\right) d\mu = 0$$

and therefore

$$\int_X |\tilde{f} - \tilde{f}(x)|^2 d\mu_x = 0$$

μ-almost everywhere on Q. □

Let now $U = \{x \in Q : \mu_x = \mu_y,\ \mu_x\text{-almost for every } y \in Q\}$. Then,

$$U = \{x \in Q : \int_X |\tilde{f} - \tilde{f}(x)|^2 d\mu_x = 0, \text{ for every } f \in C(X)\}.$$

4.3.9. Theorem. *The set U is T-invariant, Borel and of maximum probability.*

Proof. The T-invariance of U is obvious. Let $\{f_n : n \in \mathbb{N}\}$ be a countable dense subset of $C(X)$. The set

$$E_n = \{x \in Q : \int_X |\tilde{f}_n - \tilde{f}_n(x)|^2 d\mu_x > 0\}$$

is Borel and of zero probability, for every $n \in \mathbb{N}$, by 4.3.8 and so is the set $E = \cup_{n=1}^{\infty} E_n$. Clearly, $U \subset Q \setminus E$ and it suffices to prove that $Q \setminus E \subset U$. Let $x \in Q \setminus E$, $f \in C(X)$ and $\epsilon > 0$. There esists $n \in \mathbb{N}$ such that $\|f - f_n\| < \epsilon$. Since $x \in Q \setminus E_n$, we have

$$\int_X |\tilde{f}_n - \tilde{f}_n(x)|^2 d\mu_x = 0,$$

while

$$|\tilde{f} - \tilde{f}(x)|^2 \le |\tilde{f} - \tilde{f}_n|^2 + |\tilde{f}_n - \tilde{f}_n(x)|^2 + |\tilde{f}_n(x) - \tilde{f}(x)|^2.$$

It follows that

$$\int_X |\tilde{f} - \tilde{f}(x)|^2 d\mu_x < 2\epsilon^2. \quad \square$$

4.3.10. Theorem. *If $x \in U$, then μ_x is ergodic.*

Proof. Let $x \in U$. For every bounded, measurable function $f : X \to \mathbb{R}$ we have

$$\tilde{f}(x) = \int_X f d\mu_x = \int_X f d\mu_y = \tilde{f}(y)$$

μ_x-almost for every $y \in Q$. Let now $A \subset X$ be a Borel set with $A = T^{-1}(A)$. Then,

$$\tilde{f}(x)\mu_x(A) = \int_X \chi_A \tilde{f}(x) d\mu_x = \int_X \chi_A \tilde{f} d\mu_x =$$

$$\int_X \chi_A \big(\lim_{n\to+\infty} \frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k \big) d\mu_x = \lim_{n\to+\infty} \frac{1}{n} \sum_{k=0}^{n-1} \int_X (\chi_A \circ T^k)(f \circ T^k) d\mu_x =$$

$$\lim_{n\to+\infty} \frac{1}{n} \sum_{k=0}^{n-1} \int_X \chi_A f d\mu_x = \int_A f d\mu_x,$$

because $\chi_A \circ T = \chi_A$. In particular, for $f = \chi_A$ we get

$$(\mu_x(A))^2 = \tilde{\chi}_A(x)\mu_x(A) = \int_A \chi_A d\mu_x = \mu_x(A).$$

Hence $\mu_x(A) = 0$ or 1. □

4.3.11. Proposition. *The set $D = \{x \in Q : x \in \mathrm{supp}\mu_x\}$ is T-invariant, Borel and of maximum probability.*

Proof. The T-invariance of D follows immediately from the continuity of T and the T-invariance of μ_x. Let d be a compatible metric on X. For every $m \in \mathbb{N}$ there exist $x_{1,m},...,x_{k_m,m} \in X$ such that

$$X = S(x_{1,m}, \frac{1}{m}) \cup ... \cup S(x_{k_m,m}, \frac{1}{m}).$$

There are continuous functions $f_{n,m} : X \to [0,1]$ such that

$$f_{n,m}^{-1}(0) = X \setminus S(x_{n,m}, \frac{2}{m}) \quad \text{and} \quad f_{n,m}^{-1}(1) = \overline{S(x_{n,m}, \frac{1}{m})},$$

for $1 \le n \le k_m$. Each set $E_{n,m} = \{x \in Q : \tilde{f}_{n,m}(x) = 0\}$ is T-invariant and Borel, because $\tilde{f}_{n,m} : Q \to [0,1]$ is T-invariant and measurable. For every $\mu \in \mathcal{M}_T(X)$ we have

$$0 = \int_{E_{n,m}} \tilde{f}_{n,m} d\mu = \lim_{n\to+\infty} \frac{1}{n} \sum_{k=0}^{n-1} \int_{E_{n,m}} (f_{n,m} \circ T^k) d\mu = \int_{E_{n,m}} f_{n,m} d\mu \ge$$

$$\mu(E_{n,m} \cap S(x_{n,m}, \frac{1}{m})).$$

Consequently, the Borel set

$$E = Q \setminus \bigcup_{m=1}^{\infty} \bigcup_{n=1}^{k_m} E_{n,m} \cap S(x_{n,m}, \frac{1}{m})$$

is of maximum probability. We shall prove that $D = E$. First let $x \in D$. If $x \in S(x_{n,m}, \frac{1}{m})$ for some $1 \leq n \leq k_m$ and $m \in \mathbb{N}$, there exists $\epsilon > 0$ such that $S(x, \epsilon) \subset S(x_{n,m}, \frac{1}{m})$ and $\mu_x(S(x, \epsilon)) > 0$. Therefore,

$$\tilde{f}_{n,m}(x) = \int_X f_{n,m} d\mu_x > \mu_x(S(x_{n,m}, \frac{1}{m})) \geq \mu_x(S(x, \epsilon)) > 0,$$

which means that $x \in Q \setminus E_{n,m} \cap S(x_{n,m}, \frac{1}{m})$. This shows that $D \subset E$. Conversely, if $x \notin Q$, there exists $\epsilon > 0$ such that $\mu_x(S(x, \epsilon)) = 0$. There is some $m \in \mathbb{N}$ and some $1 \leq n \leq k_m$ such that

$$x \in S(x_{n,m}, \frac{1}{m}) \subset S(x_{n,m}, \frac{2}{m}) \subset S(x, \epsilon).$$

Hence

$$\tilde{f}_{n,m}(x) = \int_X f_{n,m} d\mu_x \leq \mu_x(S(x, \epsilon)) = 0,$$

which means that $x \in E_{n,m} \cap S(x_{n,m}, \frac{1}{m})$. □

So far we have proved that the set $R = U \cap D$ is T-invariant, Borel and of maximum probability. The points of R are called *regular*. So, if $x \in Q$ is a regular point, then μ_x is ergodic and $x \in \text{supp}\mu_x$.

4.3.12. Theorem. *If $\mu \in \mathcal{M}_T(X)$, then every $f \in L^1(\mu)$ is μ_x-integrable for μ-almost every $x \in R$ and*

$$\int_X f d\mu = \int_X \left(\int_X f d\mu_x \right) d\mu.$$

Proof. If $f \in L^1(\mu)$ is non-negative, then it is the pointwise limit of an increasing sequence $(f_n)_{n \in \mathbb{N}}$ of bounded, measurable functions. Moreover,

$$\int_X f d\mu_x = \lim_{n \to +\infty} \int_X f_n d\mu_x = \lim_{n \to +\infty} \tilde{f}_n(x)$$

μ-almost everywhere on X. The sequence $(\tilde{f}_n)_{n \in \mathbb{N}}$ is also increasing and from 4.3.7 and monotone convergence we have

$$\int_X \left(\int_X f d\mu_x \right) d\mu = \lim_{n \to +\infty} \int_X \tilde{f}_n d\mu = \lim_{n \to +\infty} \int_X f_n d\mu = \int_X f d\mu.$$

If f is not non-negative, it is the difference of two non-negative elements of $L^1(\mu)$ and the theorem follows. □

4.3.13. Corollary. *If $A \subset X$ is a Borel set and $\mu \in \mathcal{M}_T(X)$, then*

$$\mu(A) = \int_X \mu_x(A) d\mu.$$

4.3.14. Corollary. *A Borel set $E \subset X$ is of maximum probability if and only if $\mu(E) = 1$ for every ergodic $\mu \in \mathcal{M}_T(X)$.*

Proof. If $\mu(E) = 1$ for every ergodic $\mu \in \mathcal{M}_T(X)$, then $\mu_x(E) = 1$ for every $x \in R$. For any $\mu \in \mathcal{M}_T(X)$ now we have

$$\mu(E) = \int_R \mu_x(E) d\mu = 1,$$

by 4.3.13. □

4.3.15. Corollary. *If $\mu \in \mathcal{M}_T(X)$ is ergodic, there exists a T-invariant, Borel set $E \subset R$ such that $\mu(E) = 1$ and $\mu = \mu_x$ for every $x \in E$.*

Proof. The set $F = \text{supp}\mu$ is closed, T-invariant and $\mu(F) = 1$. Let $\{f_n : n \in \mathbb{N}\}$ be a countable dense subset of $C(X)$. The function $\tilde{f}_n : R \to \mathbb{R}$ is measurable and T-invariant for every $n \in \mathbb{N}$. Thus, $\tilde{f}_n$ is constant μ-almost everywhere, since μ is ergodic. This means that there is a T-invariant, Borel set $E_n \subset F \cap R$ such that $\mu(E_n) = 1$ and $\tilde{f}_n$ is constant on E_n. The set $E = \cap_{n=1}^{\infty} E_n$ is also T-invariant, Borel and $\mu(E) = 1$. Let now $f \in C(X)$ and $\epsilon > 0$. There exists $n \in \mathbb{N}$ such that $\|f - f_n\| < \epsilon/2$. For every x, $y \in E$ we have

$$|\tilde{f}(x) - \tilde{f}(y)| \leq |\tilde{f}(x) - \tilde{f}_n(x)| + |\tilde{f}(y) - \tilde{f}_n(y)| < \epsilon.$$

So, $\tilde{f}$ is constant on E, and since μ is ergodic, for every $x \in E$ we have

$$\int_X f d\mu_x = \tilde{f}(x) = \int_X f d\mu. \quad \square$$

4.3.16. Example. Let $T : [0,1] \to [0,1]$ be the continuous onto map defined by

$$T(x) = \frac{1}{2}(x + x^2).$$

For every $0 \leq x < 1$ we have $\lim_{n\to+\infty} T^n(x) = 0$ and $T(0) = 0$, $T(1) = 1$. So, for every $f \in C([0,1])$ and $0 \leq x < 1$ we have

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

and obviously $\tilde{f}(1) = f(1)$. Therefore, $Q = [0,1]$ and $\mu_x = \delta_0$ for every $0 \leq x < 1$, while $\mu_1 = \delta_1$. Moreover, $U = [0,1]$ and $R = \{0,1\}$. From 4.3.15, the Dirac point measures δ_0 and δ_1 are the only ergodic elements of $\mathcal{M}_T([0,1])$. The latter is thus the line segment in $\mathcal{M}([0,1])$ with endpoints δ_0 and δ_1.

4.4 Ergodicity of flows

A measurable flow $(\phi_t)_{t\in\mathbb{R}}$ on a probability space $(X, \mathcal{A}, \mu)$ is called *ergodic* if for every $A \in \mathcal{A}$ such that $\mu(A \triangle \phi_t(A)) = 0$ for every $t \in \mathbb{R}$ we have $\mu(A) = 0$ or 1. The theory of ergodic flows is similar to the theory of ergodic automorphisms. A technical difficulty we face now is the fact that the time parameter varies in an uncountable set.

A measurable function $f : X \to \mathbb{R}$ is called *ϕ-invariant μ-almost everywhere*, if for every $t \in \mathbb{R}$ we have $f \circ \phi_t = f$ μ-almost everywhere, and is called *ϕ-invariant* if $f(\phi_t(x)) = f(x)$ for every $x \in X$.

4.4.1. Proposition. *Let $(\phi_t)_{t\in\mathbb{R}}$ be a measurable flow on a complete probability space $(X, \mathcal{A}, \mu)$. If $f : X \to \mathbb{R}$ is a ϕ-invariant μ-almost everywhere function, then there exists a ϕ-invariant, measurable function $g : X \to \mathbb{R}$ such that $g = f$ μ-almost everywhere.*

Proof. Let $E = \{(t, x) \in \mathbb{R} \times X : f(\phi_t(x)) \neq f(x)\}$. If we denote by dt the Lebesgue measure on $\mathbb{R}$, then

$$\int_X \left(\int_{\mathbb{R}} \chi_{E_x}(t)dt \right) d\mu = \int_{\mathbb{R}} \left(\int_X \chi_{E_t}(x)d\mu \right) dt = 0,$$

by Fubini's theorem, where $E_x = \{t \in \mathbb{R} : (t, x) \in E\}$ and $E_t = \{x \in X : (t, x) \in E\}$. So, there exists some $N \in \mathcal{A}$ such that $\mu(N) = 0$ and

$$\int_{\mathbb{R}} \chi_{E_x}(t)dt = 0$$

for every $x \in X \setminus N$. It follows that for every $x \in X \setminus N$ there exists a Borel set $N_x \subset \mathbb{R}$ of Lebesgue measure zero such that $\chi_E(t, x) = \chi_{E_x}(t) = 0$ for every $t \in \mathbb{R} \setminus N_x$, or in other words $f(\phi_t(x)) = f(x)$. Let now $x, y \in X \setminus N$ be such that $y = \phi_t(x)$ for some $t \in \mathbb{R}$. The Borel set $(N_x - t) \cup N_y$ has Lebesgue measure zero and so there is some $s \in \mathbb{R} \setminus (N_x - t) \cup N_y$. Then,

$$f(x) = f(\phi_{s+t}(x)) = f(\phi_s(y)) = f(y).$$

We define the function $g : X \to \mathbb{R}$ as follows. If $x \in X \setminus N$, we put $g(x) = f(x)$. If the orbit of $x \in N$ is contained entirely in N, we put $g(x) = 0$. If $x \in N$ and there exists some $t \in \mathbb{R}$ such that $y = \phi_t(x) \in X \setminus N$, we put $g(x) = f(y)$. From the above follows that in this case the definition of $g(x)$ does not depend on the choice of y. Evidently, g is ϕ-invariant and it is measurable since the measure is assumed to be complete. □

The proof of the following characterization of ergodic measurable flows is the same as of 4.1.2.

4.4.2. Proposition. *Let $(X, \mathcal{A}, \mu)$ be a probability space and $(\phi_t)_{t\in\mathbb{R}}$ be a measurable flow on X. The following are equivalent.*

(i) $(\phi_t)_{t\in\mathbb{R}}$ is ergodic.

(ii) Every measurable ϕ-invariant μ-almost everywhere function is constant μ-almost everywhere.

(iii) Every measurable ϕ-invariant μ-almost everywhere function in $L^2(\mu)$ is constant μ-almost everywhere.

There is also a version of the ergodic theorem for flows, which is actually a consequence of the ergodic theorem for endomorphisms.

4.4.3. Theorem (Ergodic theorem of Birkhoff for flows). *Let $(X, \mathcal{A}, \mu)$ be a probability space and $(\phi_t)_{t\in\mathbb{R}}$ be a measurable flow on X. Then, for every $f \in L^1(\mu)$ the limit*

$$f^*(x) = \lim_{t\to+\infty} \frac{1}{t}\int_0^t f(\phi_s(x))ds$$

exists and is ϕ-invariant μ-almost everywhere. Moreover, $f^* \in L^1(\mu)$ and

$$\int_X f^* d\mu = \int_X f d\mu.$$

If $f \in L^p(\mu)$, $p \geq 1$, then we have convergence also in $L^p(\mu)$.

Proof. First observe that by Fubini's theorem

$$\int_X \left(\int_0^t |f(\phi_s(x))|ds\right) d\mu = t\|f\|_1$$

for every $t > 0$. Thus for every $n \in \mathbb{N}$ there exists $A_n \in \mathcal{A}$ such that $\mu(A_n) = 1$ and

$$\int_0^n |f(\phi_s(x))|ds < +\infty$$

for every $x \in A_n$. If now $A = \cap_{n=1}^{\infty} A_n$, then

$$\int_0^t |f(\phi_s(x))|ds \leq \int_0^{[t]+1} |f(\phi_s(x))|ds < +\infty$$

for every $x \in A$. This shows that

$$\int_0^t f(\phi_s(x))ds$$

is well defined for every $t > 0$ and $x \in A$. Let $F : X \to \mathbb{R}$ be defined by

$$F(x) = \int_0^1 f(\phi_s(x))ds$$

for $x \in A$, and $F(x) = 0$ for $x \in X \setminus A$. Then,

$$\int_X |F|d\mu \leq \int_X \left(\int_0^1 |f(\phi_s(x))|ds\right) d\mu = \|f\|_1$$

and thus $F \in L^1(\mu)$. From the ergodic theorem for endomorphisms, the limit

$$f^*(x) = \lim_{n\to+\infty} \frac{1}{n} \sum_{k=0}^{n-1} F(\phi_k(x)) = \lim_{n\to+\infty} \frac{1}{n} \int_0^n f(\phi_s(x))ds$$

exists μ-almost everywhere and $f^* \in L^1(\mu)$. Similarly, the limit

$$|f|^*(x) = \lim_{n\to+\infty} \frac{1}{n} \int_0^n |f(\phi_s(x))|ds$$

exists μ-almost everywhere. For every $t > 0$ we have

$$\left| \int_0^t f(\phi_s(x))ds - \int_0^{[t]} f(\phi_s(x))ds \right| \leq \int_0^{[t]+1} |f(\phi_s(x))|ds - \int_0^{[t]} |f(\phi_s(x))|ds.$$

It follows that

$$f^*(x) = \lim_{n\to+\infty} \frac{1}{n} \int_0^n f(\phi_s(x))ds = \lim_{t\to+\infty} \frac{1}{t} \int_0^t f(\phi_s(x))ds.$$

For the ϕ-invariance of f^* we observe that for every $\tau > 0$ we have

$$\frac{1}{t} \left| \int_0^t f(\phi_{s+\tau}(x))ds - \int_0^{t+\tau} f(\phi_s(x))ds \right| \leq \frac{1}{t} \int_0^\tau |f(\phi_s(x))|ds,$$

which tends to zero as $t \to +\infty$, μ-almost everywhere. Hence

$$f^*(\phi_\tau(x)) = \lim_{t\to+\infty} \frac{1}{t} \int_0^t f(\phi_{s+\tau}(x))ds = \lim_{t\to+\infty} \frac{1}{t} \int_0^{t+\tau} f(\phi_s(x))ds = f^*(x),$$

μ-almost everywhere. Similarly for $\tau < 0$. It remains to prove that if $f \in L^p(\mu)$, $p \geq 1$, then we have convergence in $L^p(\mu)$. From this it will follow that f^* and f have the same μ-integral over X, because for $p = 1$ we will have

$$\int_X f^* d\mu = \int_X \left(\lim_{t\to+\infty} \frac{1}{t} \int_0^t f(\phi_s(x))ds \right) d\mu = \lim_{t\to+\infty} \frac{1}{t} \int_X \left(\int_0^t f(\phi_s(x))ds \right) d\mu =$$

$$\lim_{t\to+\infty} \frac{1}{t} \int_0^t \left(\int_X f(\phi_s(x))d\mu \right) ds = \lim_{t\to+\infty} \frac{1}{t} \int_0^t \left(\int_X f d\mu \right) ds = \int_X f d\mu,$$

the second equality being due to $L^1(\mu)$-convergence. To prove $L^p(\mu)$-convergence, we observe that

$$\left(\int_X \left| \int_0^1 f(\phi_s(x))ds \right|^p d\mu \right)^{\frac{1}{p}} \leq \int_0^1 \left(\int_X |f(\phi_s(x))|^p d\mu \right)^{\frac{1}{p}} ds = \|f\|_p$$

by the generalized Minkowski inequality, and therefore $F \in L^p(\mu)$. From the L^p ergodic theorem for endomorphisms we have

$$\lim_{n\to+\infty} \| \frac{1}{n} \int_0^n (f \circ \phi_s)ds - f^* \|_p = 0.$$

However,

$$\left|\frac{1}{t}\int_0^t f(\phi_s(x))ds - \frac{1}{[t]}\int_0^{[t]} f(\phi_s(x))ds\right| \leq$$

$$\frac{[t]}{t}\left(\frac{1}{[t]+1}\int_0^{[t]+1}|f(\phi_s(x))|ds - \frac{1}{[t]}\int_0^{[t]}|f(\phi_s(x))|ds\right)+$$

$$\frac{1}{t}\frac{1}{[t]+1}\int_0^{[t]+1}|f(\phi_s(x))|ds + (1-\frac{[t]}{t})\frac{1}{[t]}\int_0^{[t]}|f(\phi_s(x))|ds.$$

Using the Minkowski inequality we have

$$\|\frac{1}{t}\int_0^t (f\circ\phi_s)ds - \frac{1}{[t]}\int_0^{[t]}(f\circ\phi_s)ds\|_p \leq$$

$$\frac{[t]}{t}\|\frac{1}{[t]+1}\int_0^{[t]+1}|f\circ\phi_s|ds - \frac{1}{[t]}\int_0^{[t]}|f\circ\phi_s|ds\|_p+$$

$$\frac{1}{t}\|\frac{1}{[t]+1}\int_0^{[t]+1}|f\circ\phi_s|ds\|_p + (1-\frac{[t]}{t})\|\frac{1}{[t]}\int_0^{[t]}|f\circ\phi_s|ds\|_p.$$

It follows from the above that

$$\lim_{t\to+\infty}\|\frac{1}{t}\int_0^t (f\circ\phi_s)ds - \frac{1}{[t]}\int_0^{[t]}(f\circ\phi_s)ds\|_p = 0.$$

This completes the proof. □

Chapter 5

Geodesic flows of hyperbolic surfaces

5.1 The hyperbolic plane

Let $\mathbb{H}^2 = \{z \in \mathbb{C} : \mathrm{Im}z > 0\}$ be the upper half plane. We shall use complex numbers to denote the points of $\mathbb{H}^2$, as well as its tangent vectors. The *hyperbolic* metric on $\mathbb{H}^2$ is the complete Riemannian metric of constant negative curvature -1 defined by

$$\langle u+iv, u'+iv' \rangle_z = \mathrm{Re}\frac{(u+iv)(u'-iv')}{(\mathrm{Im}z)^2},$$

where $u+iv$, $u'+iv' \in T^1_z\mathbb{H}^2$ and $z \in \mathbb{H}^2$. Thus,

$$\|u+iv\|^2_z = \frac{1}{(\mathrm{Im}z)^2}(u^2+v^2),$$

and angles in the hyperbolic sense are the same as the euclidean. The hyperbolic geodesics are either euclidean half lines orthogonal to the real axis or euclidean semicircles with center on the real axis.

For every a, b, c, $d \in \mathbb{R}$ with $ad-bc=1$, the Möbius transformation of the Riemann sphere defined by

$$T(z) = \frac{az+b}{cz+d}$$

has complex derivative

$$T'(z) = \frac{1}{(cz+d)^2}$$

and $\mathrm{Im}T(z) = |T'(z)|\mathrm{Im}z$. Hence $T(\mathbb{H}^2) = \mathbb{H}^2$. Moreover, T is a hyperbolic isometry, because for every $z \in \mathbb{H}^2$ and $u+iv$, $u'+iv' \in T^1_z\mathbb{H}^2$ we have

$$\langle T'(z)(u+iv), T'(z)(u'+iv') \rangle_{T(z)} = \mathrm{Re}\frac{T'(z)(u+iv)\overline{T'(z)(u'+iv')}}{(\mathrm{Im}T(z))^2} =$$

$$\frac{T'(z)\overline{T'(z)}}{|T'(z)|^2}\mathrm{Re}\frac{(u+iv)(u'-iv')}{(\mathrm{Im}z)^2} = \langle u+iv, u'+iv' \rangle_z.$$

The group of real Möbius transformations is precisely the group of the orientation preserving hyperbolic isometries or in other words is the connected component of the identity of the group of isometries of $\mathbb{H}^2$ endowed with the compact-open topology, and is isomorphic as a Lie group to $PSL(2,\mathbb{R})$. Apart from itself, it has only one other coset in the group of hyperbolic isometries, the one represented by the reflection through the imaginary semiaxis. We shall identify the group of orientation preserving hyperbolic isometries with $PSL(2,\mathbb{R})$.

5.1.1. Proposition. *(a) If I is the imaginary semiaxis, then for every hyperbolic geodesic C there is some $T \in PSL(2,\mathbb{R})$ such that $T(I) = C$.*

(b) For every $z_0 \in \mathbb{H}^2$ and every $v \in T_{z_0}\mathbb{H}^2$ with $\|v\|_{z_0} = 1$ there is some $T \in PSL(2,\mathbb{R})$ such that $T(i) = z_0$ and $T'(i)i = v$.

Proof. (a) If $C = \{z \in \mathbb{H}^2 : \mathrm{Re}z = b\}$, for some $b \in \mathbb{R}$, it suffices to take $T(z) = z + b$. Suppose that C is a hyperbolic geodesic with endpoints x, $x + r \in \mathbb{R}$. The Möbius transformation

$$T_1(z) = \frac{z}{z+1}$$

maps I onto the hyperbolic geodesic with endpoints 0 and 1. Thus, if $T_2(z) = rz$ and $T_3(z) = z + x$, then it suffices to take $T = T_3 \circ T_2 \circ T_1$.

(b) There exists a unique parametrized hyperbolic geodesic C through z_0 with velocity v at z_0. By (a), there exists $T_1 \in PSL(2,\mathbb{R})$ such that $T_1(I) = C$. Let $a > 0$ be such that $T_1(ai) = z_0$. Then $(T_1^{-1})'(z_0)v = \pm ai$. If $(T_1^{-1})'(z_0)v = ai$, let $T_0(z) = az$. Then $T_0(i) = ai$ and $T_0'(ai)i = ai$. If $(T_1^{-1})'(z_0)v = -ai$, let $T_0(z) = -a/z$. In both cases it suffices to take $T = T_1 \circ T_0$. □

Thus, $PSL(2,\mathbb{R})$ acts transitively on the unit tangent bundle $T^1\mathbb{H}^2 \cong \mathbb{H}^2 \times S^1$, and as we see easily, the isotropy group of (i,i) is trivial. It follows that the smooth map $\psi : PSL(2,\mathbb{R}) \to T^1\mathbb{H}^2$ defined by $\psi(T) = (T(i), T'(i)i)$ is one-to-one, onto, and the proof of 5.1.1 shows that it is a diffeomorphism. An analytical formula of its inverse can be given using the Iwasawa decomposition of $SL(2,\mathbb{R})$. We consider the following one parameter subgroups of $SL(2,\mathbb{R})$:

$$K = \{k_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} : 0 \le \theta < 2\pi\} \cong S^1,$$

$$A = \{a_t = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} : t \in \mathbb{R}\} \cong \mathbb{R},$$

$$N = \{n_t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} : t \in \mathbb{R}\} \cong \mathbb{R}.$$

Observe that $n_t a_s = a_s n_{te^{-2s}}$ for every t, $s \in \mathbb{R}$. Therefore, $NA = AN$ is a subgroup of $SL(2,\mathbb{R})$ consisting of upper triangular matrices and $N \lhd NA$. Obviously, $A \cap N$ and $K \cap NA$ are trivial.

5.1.2. Lemma. *For every $g \in SL(2,\mathbb{R})$ there exist unique $k_\theta \in K$, $a_s \in A$ and $n_t \in N$ such that $g = n_t a_s k_\theta$.*

Proof. The equation

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} e^s & 0 \\ 0 & e^{-s} \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

is equivalent to the system of equations

$$a = e^s \cos\theta - te^{-s} \sin\theta$$
$$b = e^s \sin\theta + te^{-s} \cos\theta$$
$$c = -e^{-s} \sin\theta$$
$$d = e^{-s} \cos\theta.$$

From the last two we get $s = -\frac{1}{2}\log(c^2 + d^2)$, and then from the first two we have

$$t = e^{2s}(ca + db) = \frac{ac + bd}{c^2 + d^2}$$

and

$$\cos\theta = \frac{d}{\sqrt{c^2 + d^2}}, \qquad \sin\theta = -\frac{c}{\sqrt{c^2 + d^2}}.$$

For the uniqueness, if $nak = n'a'k'$, where $k, k' \in K$, $a, a' \in A$ and $n, n' \in N$, then $k'k^{-1} = (n'a')^{-1}(na) \in K \cap NA$ which is trivial. Hence $k = k'$ and $na = n'a'$. But then $a'a^{-1} = n(n')^{-1} \in A \cap N$ which is also trivial, and so $a = a'$ and $n = n'$. □

From 5.1.2 follows that the map $\chi : T^1\mathbb{H}^2 \to SL(2, \mathbb{R})$ defined by

$$\chi(z, (\mathrm{Im}z)e^{i\theta}) = n_{\mathrm{Re}z} \cdot a_{\frac{1}{2}\log \mathrm{Im}z} \cdot k_\theta$$

is a smooth diffeomorphism. The quotient map $p : SL(2, \mathbb{R}) \to PSL(2, \mathbb{R})$ is a double covering. If $p(g) = T$, then

$$p^{-1}(T) = \{g, g \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}\}.$$

For $g = nak_\theta$, $0 \le \theta < 2\pi$, we have $p^{-1}(T) = \{nak_\theta, nak_{\theta+\pi}\}$. Consider now the smooth, one-to-one, onto map $\phi : T^1\mathbb{H}^2 \to PSL(2, \mathbb{R})$ defined by

$$\phi(z, (\mathrm{Im}z)e^{i\theta}) = p(n_{\mathrm{Re}z} \cdot a_{\frac{1}{2}\log \mathrm{Im}z} \cdot k_{\theta/2}).$$

Let $z_0 = x + iy$, $x \in \mathbb{R}$ and $y > 0$, and $T = \phi(z_0, e^{i\theta})$, that is

$$T(z) = y \cdot \frac{(\cos\frac{\theta}{2})z + \sin\frac{\theta}{2}}{(-\sin\frac{\theta}{2})z + \cos\frac{\theta}{2}} + x$$

and

$$T'(z) = \frac{y}{[(-\sin\frac{\theta}{2})z + \cos\frac{\theta}{2}]^2}.$$

Therefore, $T(i) = z_0$ and

$$T'(i)i = y[\cos(\theta + \frac{\pi}{2}) + i\sin(\theta + \frac{\pi}{2})].$$

It follows that $T = \psi^{-1}(z_0, ye^{i(\theta+\frac{\pi}{2})})$ and so $(\psi \circ \phi)(z_0, ye^{i\theta}) = (z_0, ye^{i(\theta+\frac{\pi}{2})})$. Hence ψ and ϕ are smooth diffeomorphisms.

5.2 The Haar measure on $PSL(2,\mathbb{R})$

The positive linear functional $\mu : C_c(SL(2,\mathbb{R})) \to \mathbb{R}$ defined by

$$\mu(f) = \int_{\mathbb{R}^3} \frac{1}{|\delta|} f \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} d\beta d\gamma d\delta,$$

where $\delta \neq 0$ and therefore $\alpha = \dfrac{1+\beta\gamma}{\delta}$, defines a Borel measure on $SL(2,\mathbb{R})$. If

$$\begin{pmatrix} \alpha' & \beta' \\ \gamma' & \delta' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$$

then

$$(\beta', \gamma', \delta') = (a\beta + b\delta, \frac{c(1+\beta\gamma)}{\delta} + \gamma d, c\beta + \delta d).$$

The Jacobian matrix of this transformation is

$$\begin{pmatrix} a & 0 & b \\ \dfrac{c\gamma}{\delta} & \dfrac{c\beta}{\delta} + d & -\dfrac{c(1+\beta\gamma)}{\delta^2} \\ c & 0 & d \end{pmatrix}$$

and has determinant equal to

$$(\frac{c\beta}{\delta} + d)(ad - bc) = \frac{c\beta + \delta d}{\delta} = \frac{\delta'}{\delta}.$$

Consequently,

$$\frac{1}{|\delta'|} d\beta' d\gamma' d\delta' = \frac{1}{|\delta|} d\beta d\gamma d\delta$$

and μ is invariant by left translations in $SL(2,\mathbb{R})$. Similarly, it is invariant by right translations also. Hence μ is the Haar measure on $SL(2,\mathbb{R})$, modulo a constant, and $SL(2,\mathbb{R})$ is a unimodular connected Lie group. Moreover, μ projects to the Haar measure on $PSL(2,\mathbb{R})$, which we shall also denote by μ. If $f : PSL(2,\mathbb{R}) \to \mathbb{R}$ is a continuous function with compact support, then the μ-integral of f over $PSL(2,\mathbb{R})$ is the integral of $f \circ p$ over a fundamental domain of the double covering map $p : SL(2,\mathbb{R}) \to PSL(2,\mathbb{R})$. Considering the Iwasawa decomposition of $SL(2,\mathbb{R})$, such a fundamental domain consists of all the elements of the form

$$\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y^{1/2} & 0 \\ 0 & y^{-1/2} \end{pmatrix} \begin{pmatrix} \cos\dfrac{\theta}{2} & \sin\dfrac{\theta}{2} \\ -\sin\dfrac{\theta}{2} & \cos\dfrac{\theta}{2} \end{pmatrix}$$

where x, $y \in \mathbb{R}$, $y > 0$ and $0 \leq \theta < 2\pi$. Then,

$$(\beta, \gamma, \delta) = (y^{1/2} \sin\frac{\theta}{2} + xy^{-1/2} \cos\frac{\theta}{2}, -y^{-1/2} \sin\frac{\theta}{2}, y^{-1/2} \cos\frac{\theta}{2})$$

and the Jacobian determinant of this transformation is $(-\cos\dfrac{\theta}{2}) \cdot \frac{1}{4} y^{-5/2}$. It follows that

$$\frac{1}{|\delta|} d\beta d\gamma d\delta = \frac{y^{1/2}}{\cos\dfrac{\theta}{2}} \cdot \cos\frac{\theta}{2} \cdot \frac{1}{4y^{5/2}} dx dy d\theta = \frac{1}{4y^2} dx dy d\theta.$$

This means that the Haar measure on $PSL(2,\mathbb{R})$ is transformed by the smooth diffemorphism $\phi^{-1} : PSL(2,\mathbb{R}) \to T^1\mathbb{H}^2$ of the preceding section to the Liouville measure on $T^1\mathbb{H}^2$, modulo a constant. In the rest of this section we shall give alternative descriptions of the Haar measure on $PSL(2,\mathbb{R})$ using other coordinate systems on $T^1\mathbb{H}^2 \cong \mathbb{H}^2 \times S^1$.

Let $x+iy \in \mathbb{H}^2$, $0 \le \theta < 2\pi$ and $\xi \in \mathbb{R} \cup \{\infty\} \approx S^1$ be the positive end of the parametrized hyperbolic geodesic C through $x+iy$, whose velocity at this point is the unit tangent vector making angle θ with the vertical half line at $x+iy$. Then,

$$\tan\frac{\theta}{2} = \frac{y}{x-\xi} \qquad \text{and} \qquad \xi = x - y\cot\frac{\theta}{2}.$$

Therefore,

$$\frac{1}{y^2}dxdyd\theta = 2\sin^2\frac{\theta}{2}\cdot\frac{1}{y^3}dxdyd\xi = \frac{2}{y|(x+iy)-\xi|^2}dxdyd\xi.$$

Let now η be the negative end of C and s be the signed hyperbolic distance of $x+iy$ from the highest point of C. Obviously,

$$\eta = x + y\cot(\frac{\pi-\theta}{2}) = x + y\tan\frac{\theta}{2}.$$

Let $T = \phi(x+iy, ye^{i\theta})$, that is

$$T(z) = y\cdot\frac{(\cos\frac{\theta}{2})z + \sin\frac{\theta}{2}}{(-\sin\frac{\theta}{2})z + \cos\frac{\theta}{2}} + x$$

and in particular $T(0) = \eta$, $T(i) = x+iy$ and $T(\infty) = \xi$. Thus, T maps the imaginary semiaxis onto C and for every $z \in C$ there is a unique $t > 0$ such that $T(it) = z$. The hyperbolic distance of z from $x+iy$ is equal to the hyperbolic distance of it from i, and therefore equal to $|\log t|$. If

$$T(z) = \frac{\alpha z + \beta}{\gamma z + \delta},$$

then $\xi = \alpha/\gamma$, $\eta = \beta/\delta$ and

$$T(it) = \frac{(\alpha\gamma t^2 + \beta\delta) + it}{\gamma^2 t^2 + \delta^2}.$$

Hence

$$\mathrm{Im}T(it) = \frac{t}{\gamma^2 t^2 + \delta^2},$$

which takes its maximum value for $t = |\delta/\gamma|$. It follows that $s = \log|\delta/\gamma|$ and so

$$(\eta, \xi, s) = (\frac{\beta}{\delta}, \frac{1}{\gamma\delta} + \frac{\beta}{\delta}, \log|\frac{\delta}{\gamma}|).$$

The Jacobian matrix of this trasformation is

$$\begin{pmatrix} \frac{1}{\delta} & 0 & -\frac{\beta}{\delta^2} \\ \frac{1}{\delta} & -\frac{1}{\delta\gamma^2} & -\frac{1+\beta\gamma}{\gamma\delta^2} \\ 0 & -\frac{1}{\gamma} & \frac{1}{\delta} \end{pmatrix}$$

and has determinant $-2/\delta^3\gamma^2$. Consequently,

$$\frac{1}{|\eta - \xi|^2} d\eta d\xi ds = \frac{2}{|\delta|} d\beta d\gamma d\delta.$$

Finally, consider the horocycle that passes through ξ and $x + iy$. Let $r > 0$ be its euclidean radius and u be the signed hyperbolic length of the arc on the horocycle from $x + iy$ to the highest point of the horocycle.

Since $T(i) = x + iy$ and $T(\infty) = \xi$, the horocycle is the image of the horocycle $\{z \in \mathbb{H}^2 : \mathrm{Im}z = 1\}$ by T. In other words, the horocycle is the set $\{T(t+i) : t \in \mathbb{R}\}$. Now

$$\mathrm{Im}T(t+i) = \frac{1}{(\gamma t + \delta)^2 + \gamma^2},$$

which takes its maximum value $1/\gamma^2$ for $t = -\delta/\gamma$. Therefore $r = 1/2\gamma^2$. On the other hand, since T is a hyperbolic isometry, u is equal to the signed hyperbolic length of the euclidean line segment from i to $i - \dfrac{\delta}{\gamma}$, and this is δ/γ. Thus,

$$(\xi, r, u) = (\frac{1+\beta\gamma}{\gamma\delta}, \frac{1}{2\gamma^2}, \frac{\delta}{\gamma})$$

and the Jacobian matrix of this transformation is

$$\begin{pmatrix} \frac{1}{\delta} & -\frac{1}{\delta\gamma^2} & -\frac{1+\beta\gamma}{\gamma\delta^2} \\ 0 & -\frac{1}{\gamma^3} & 0 \\ 0 & -\frac{\delta}{\gamma^2} & \frac{1}{\gamma} \end{pmatrix}$$

which has determinant $-\frac{1}{\delta\gamma^4}$. Hence

$$\frac{4}{|\delta|}d\beta d\gamma d\delta = 4\gamma^4 d\xi dr du = \frac{1}{r^2}d\xi dr du.$$

5.3 The geodesic flow of the hyperbolic plane

Using the notations of section 5.1, the set $p(A)$ is a one parameter subgroup of $PSL(2,\mathbb{R})$. For every $t \in \mathbb{R}$ we have $p(a_{t/2})(z) = e^t z$ for every $z \in \mathbb{C} \cup \{\infty\}$. The formula

$$g_t(z,(\mathrm{Im}z)e^{i\theta}) = \phi^{-1}(\phi(z,(\mathrm{Im}z)e^{i\theta}) \circ p(a_{t/2})), \quad t \in \mathbb{R}, \quad (z,(\mathrm{Im}z)e^{i\theta}) \in T^1\mathbb{H}^2$$

defines a smooth flow on the unit tangent bundle of $\mathbb{H}^2$. If we set $T = \phi(z,(\mathrm{Im}z)e^{i\theta})$, then $T(0) = \eta$, $T(\infty) = \xi$ and $T(i) = z$, according to the notations of section 5.2. Thus, $g_t(z,(\mathrm{Im}z)e^{i\theta})$ defines the same parametrized hyperbolic geodesic as $(z,(\mathrm{Im}z)e^{i\theta})$. Moreover, the hyperbolic distance of $(T \circ p(a_{t/2}))(i) = T(e^t i)$ from z is $|t|$. This means that the point of application of the tangent vector $g_t(z,(\mathrm{Im}z)e^{i\theta})$ is the point on the parametrized geodesic defined by $(z,(\mathrm{Im}z)e^{i\theta})$ after time t. This shows that $(g_t)_{t\in\mathbb{R}}$ is the geodesic flow of $\mathbb{H}^2$. It is clear that in the coordinates (η,ξ,s) for $T^1\mathbb{H}^2$ of section 5.2, the geodesic flow is the parallel flow given by

$$g_t(\eta,\xi,s) = (\eta,\xi,s+t)$$

and has global section the set of points of the form $(\eta,\xi,0)$, $\eta \neq \xi$, which is smoothly diffeomorphic to a cylinder.

On $PSL(2,\mathbb{R})$ there is also the one parameter group $p(N)$. For every $t \in \mathbb{R}$ we have $p(n_t)(z) = z + t$ for every $z \in \mathbb{C} \cup \{\infty\}$. The formula

$$h_t(z,(\mathrm{Im}z)e^{i\theta}) = \phi^{-1}(\phi(z,(\mathrm{Im}z)e^{i\theta}) \circ p(n_t)), \quad t \in \mathbb{R}, \quad (z,(\mathrm{Im}z)e^{i\theta}) \in T^1\mathbb{H}^2$$

defines a smooth flow on the unit tangent bundle of $\mathbb{H}^2$. Now $(z,(\mathrm{Im}z)e^{i\theta})$ and $h_t(z,(\mathrm{Im}z)e^{i\theta})$ determine parametrized hyperbolic geodesics which are positively asymptotic at ξ. Moreover, since $(T \circ p(n_t))(i) = T(i+t)$, the points of application of the tangent vectors $(z,(\mathrm{Im}z)e^{i\theta})$ and $h_t(z,(\mathrm{Im}z)e^{i\theta})$ are on the horocycle

$$T(\{w \in \mathbb{H}^2 : \mathrm{Im}w = 1\})$$

and the vectors are orthogonal to the horocycle. The hyperbolic length of the arc on the horocycle from z to the point of application of $h_t(z,(\mathrm{Im}z)e^{i\theta})$ is equal to the hyperbolic length of the euclidean line segment from i to $i+t$, which is $|t|$. Thus, $h_t(z,(\mathrm{Im}z)e^{i\theta})$ is taken by translating $(z,(\mathrm{Im}z)e^{i\theta})$ along the horocycle it determines with ξ, keeping it orthogonal to the horocycle, by a signed hyperbolic length t. The flow $(h_t)_{t\in\mathbb{R}}$ is called the *horocycle flow* of the hyperbolic plane. It is clear that in the coordinates (ξ,r,u) of section 5.2 the horocycle flow is the parellel flow given by

$$h_t(\xi,r,u) = (\xi,r,u+t)$$

and has global section the set of points of the form $(\xi, r, 0)$ which is diffeomorphic to a cylinder. Since $n_t a_s = a_s n_{te^{-2s}}$, we have

$$g_s \circ h_t = h_{te^{-s}} \circ g_s$$

for every t, $s \in \mathbb{R}$.

On $PSL(2,\mathbb{R})$ there is also the one parameter group $p(K)$, which defines on the unit tangent bundle of $\mathbb{H}^2$ the smooth flow $R_t(z,(\mathrm{Im}z)e^{i\theta}) = \phi^{-1}(\phi(z,(\mathrm{Im}z)e^{i\theta}) \circ p(k_{t/2})) = (z,(\mathrm{Im}z)e^{i(\theta+t)})$, $t \in \mathbb{R}$, $(z,(\mathrm{Im}z)e^{i\theta}) \in T^1\mathbb{H}^2$. The three flows are related as follows.

5.3.1. Proposition. *If $t \in \mathbb{R}$, then*

$$h_s = R_{\pi-a} \circ g_t \circ R_{2\pi-a},$$

where $\cot a = \dfrac{s}{2}$, $0 < a < \pi$ *and* $s = 2\sinh\dfrac{t}{2}$.

Proof. Since $\sinh\dfrac{t}{2} = \cot a$, we have

$$\tan^2\frac{a}{2} + (e^{t/2} - e^{-t/2})\tan\frac{a}{2} - 1 = 0$$

from which follows that $\tan\dfrac{a}{2} = e^{-t/2}$, because $0 < a < \pi$. On the other hand, since $(R_{\pi-a})^{-1} = R_{\pi+a}$, it suffices to prove that $R_{\pi+a} \circ h_s = g_t \circ R_{2\pi-a}$. Now $R_{\pi+a} \circ h_s$ is represented by the matrix

$$\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}\begin{pmatrix} -\sin\frac{a}{2} & \cos\frac{a}{2} \\ -\cos\frac{a}{2} & -\sin\frac{a}{2} \end{pmatrix} = \begin{pmatrix} -\sin\frac{a}{2} - s\cos\frac{a}{2} & \cos\frac{a}{2} - s\sin\frac{a}{2} \\ -\cos\frac{a}{2} & -\sin\frac{a}{2} \end{pmatrix}$$

and $g_t \circ R_{2\pi-a}$ is represented by

$$\begin{pmatrix} -\cos\frac{a}{2} & \sin\frac{a}{2} \\ -\sin\frac{a}{2} & -\cos\frac{a}{2} \end{pmatrix}\begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix} = \begin{pmatrix} -e^{t/2}\cos\frac{a}{2} & e^{-t/2}\sin\frac{a}{2} \\ -e^{t/2}\sin\frac{a}{2} & -e^{-t/2}\cos\frac{a}{2} \end{pmatrix}.$$

Since $\tan\dfrac{a}{2} = e^{-t/2}$ and $\cot a = \dfrac{s}{2}$, the two matrices are equal. □

5.4 The Poincaré disc model

Let $\mathbb{D}^2 = \{z \in \mathbb{C} : |z| < 1\}$. The Cayley transformation

$$F(z) = \frac{iz+1}{z+i}$$

maps $\mathbb{H}^2$ onto $\mathbb{D}^2$ and has inverse

$$F^{-1}(z) = \frac{-iz+1}{z-i}$$

with complex derivative

$$(F^{-1})'(z) = \frac{-2}{(z-i)^2}.$$

If we trasform the hyperbolic Riemannian metric on $\mathbb{H}^2$ by F, we get on $\mathbb{D}^2$ the Riemannian metric

$$\langle v, w\rangle_z = \frac{4}{(1-|z|^2)^2}\mathrm{Re}(v\bar{w})$$

where v, $w \in T_z\mathbb{D}^2$ and $z \in \mathbb{D}^2$. The unit disc equipped with this metric is an alternative model for the plane hyperbolic geometry. The hyperbolic area of a Borel set $A \subset \mathbb{D}^2$ is

$$\int_A \frac{4}{(1-|z|^2)^2} dxdy.$$

The hyperbolic geodesics in $\mathbb{D}^2$ are the images by the Cayley transformation of the hyperbolic geodesics in $\mathbb{H}^2$, and are either diameters of $\mathbb{D}^2$ or euclidean circular arcs orhogonal to the boundary circle $\partial\mathbb{D}^2$. The orientation preserving hyperbolic isometries of $\mathbb{D}^2$ are of the form $F \circ T \circ F^{-1}$, where T is an orientation preserving hyperbolic isometry of $\mathbb{H}^2$. This is precisely the set $\mathcal{M}$ of Möbius transformations that preserve the unit disc. Thus, the elements of $\mathcal{M}$ are of the form

$$S(z) = \frac{az+\bar{c}}{cz+\bar{a}}$$

where a, $c \in \mathbb{C}$ are such that $|a|^2 - |c|^2 = 1$. The fact that S is a hyperbolic isometry of $\mathbb{D}^2$ is expressed by the equality

$$\frac{|S'(z)|}{1-|S(z)|^2} = \frac{1}{1-|z|^2}$$

which is easily verified. The elements of $\mathcal{M}$ have the following useful properties.

5.4.1. Proposition. *If $S \in \mathcal{M}$, then*
(i) $|S(z) - S(w)| = |S'(z)|^{1/2}|S'(w)|^{1/2}|z-w|$, for every z, $w \in \mathbb{C}$, and
(ii)

$$|(S^{-1})'(\xi)| = \frac{1-|S(0)|^2}{|\xi - S(0)|^2}$$

for every $\xi \in \partial\mathbb{D}^2$.

Proof. (i) Since

$$S'(z) = \frac{1}{(cz + \bar{a})^2},$$

we have

$$|S(z) - S(w)| = \frac{|z - w|}{|cz + \bar{a}| \cdot |cw + \bar{a}|} = |S'(z)|^{1/2} |S'(w)|^{1/2} |z - w|.$$

(ii) Applying (i) for $z = S^{-1}(\xi)$ and $w = 0$, we have

$$|\xi - S(0)|^2 = |S'(S^{-1}(\xi))| \cdot |S'(0)| \cdot |S^{-1}(\xi)|^2 = \frac{1}{|(S^{-1})'(\xi)|} \cdot (1 - |S(0)|^2),$$

since $|S'(0)| = 1 - |S(0)|^2$, because S is a hyperbolic isometry of $\mathbb{D}^2$. □

The elements of the unit tangent bundle $T^1\mathbb{D}^2$ are determined by triples $(x, y, e^{i\theta})$, where $x + iy$ is the point of application and θ is the angle the tangent vector forms with the horizontal axis. The derivative of the Cayley transformation transfers the geodesic and horocycle flows of $\mathbb{H}^2$ to those of $\mathbb{D}^2$. The invariant Liouville measure on $T^1\mathbb{D}^2$ is

$$\frac{4}{[1 - (x^2 + y^2)]^2} dx dy d\theta.$$

5.5 Ergodicity of geodesic flows of hyperbolic surfaces

A *hyperbolic surface* M is the quotient space of $\mathbb{H}^2$ by a subgroup G of $PSL(2, \mathbb{R})$ which acts freely and properly discontinuously by hyperbolic isometries on $\mathbb{H}^2$. Then, M is an orientable, connected 2-manifold, and can be equipped with a Riemannian metric that makes the quotient map $q : \mathbb{H}^2 \to M$ a local isometry. Moreover, q is the universal covering of M and $\pi_1(M) \cong G$. Recall that any orientable, connected, complete Riemannian 2-manifold of constant negative curvature -1 arises in this way.

The action of G on $\mathbb{H}^2$ induces an action on $T^1\mathbb{H}^2$ in the obvious way, which is also free and properly discontinuous. The quotient space of $T^1\mathbb{H}^2$ by this action is precisely T^1M and the quotient map is the derivative of q. It is clear that T^1M is smoothly diffeomorphic to the homogeneous space $G\backslash PSL(2, \mathbb{R})$ of right cosets of G in $PSL(2, \mathbb{R})$.

The geodesics in M are images of the hyperbolic geodesics in $\mathbb{H}^2$ by q. Since the geodesic flow of $\mathbb{H}^2$ is invariant under hyperbolic isometries, it projects by q to the geodesic flow of M. Similarly, q maps the horocycle flow of $\mathbb{H}^2$ to a flow on T^1M, which we shall call horocycle flow of M. In terms of $PSL(2, \mathbb{R})$, the geodesic and horocycle flow of M can be described as follows, using the notations of section 5.1. The right action of the one parameter subgroup $p(A)$ on $PSL(2, \mathbb{R})$ commutes with the left action of G on $PSL(2, \mathbb{R})$, and therefore induces a smooth flow on $G\backslash PSL(2, \mathbb{R})$, which is precisely (conjugate to) the geodesic flow of M. In the same way, the horocycle flow of M is (conjugate to) the smooth flow on $G\backslash PSL(2, \mathbb{R})$ that is induced by the right action of the one parameter subgroup $p(N)$ of $PSL(2, \mathbb{R})$.

The Liouville measure μ on T^1M is obtained from the Liouville measure on $T^1\mathbb{H}^2$. If P is a Dirichlet polygon of G, then $\mu(A)$ is the Liouville measure of $(Dq)^{-1}(A) \cap (P \times S^1)$ in $T^1\mathbb{H}^2$, for every Borel set $A \subset T^1M$. If M is compact, then P is a finite hyperbolic polygon, and so it has finite hyperbolic area. Thus, in this case the Liouville measure on T^1M is finite.

In the sequel, we shall assume that M is compact and we shall denote by $(g_t)_{t\in\mathbb{R}}$ the geodesic and by $(h_t)_{t\in\mathbb{R}}$ the horocycle flow of M. Note that $g_t \circ h_s = h_{se^{-t}} \circ g_t$, by the corresponding property of the geodesic and the horocycle flow of $\mathbb{H}^2$. We shall also assume that μ is the normalized Liouville measure on T^1M, that is $\mu(T^1M) = 1$.

5.5.1. Lemma. *Let $f : T^1M \to \mathbb{R}$ be in $L^1(\mu)$. If f is invariant by the geodesic flow, then it is μ-almost everywhere invariant by the horocycle flow.*

Proof. Let $s \in \mathbb{R}$. Since $C(T^1M)$ is dense in $L^1(\mu)$, there exists a sequence of continuous functions $f_n : T^1M \to \mathbb{R}$, $n \in \mathbb{N}$, such that

$$\lim_{n\to+\infty} \int_{T^1M} |f_n - f| d\mu = 0.$$

Thus, for every $\epsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$ we have

$$\int_{T^1M} |f_n - f| d\mu < \frac{\epsilon}{3}$$

and so

$$\int_{T^1M} |f_n \circ h_{se^{-t}} \circ g_t - f \circ h_{se^{-t}} \circ g_t| d\mu = \int_{T^1M} |f_n - f| d\mu < \frac{\epsilon}{3}$$

for every $t \in \mathbb{R}$. Since T^1M is compact, there exists $t_0 > 0$ such that

$$\int_{T^1M} |f_{n_0} \circ h_{se^{-t}} \circ g_t - f_{n_0} \circ g_t| d\mu < \frac{\epsilon}{3}$$

for every $t \geq t_0$. It follows from these that

$$\int_{T^1M} |f \circ h_{se^{-t}} \circ g_t - f \circ g_t| d\mu < \epsilon$$

for every $t \geq t_0$. But $f \circ h_{se^{-t}} \circ g_t = f \circ g_t \circ h_s = f \circ h_s$, because f is invariant by the geodesic flow. So we get

$$\int_{T^1M} |f \circ h_s - f| d\mu < \epsilon$$

for every $\epsilon > 0$, and the conclusion follows. □

Let now $f : T^1M \to \mathbb{R}$ be a measurable, invariant function by the geodesic flow, and therefore μ-almost everywhere invariant by the horocycle flow too. The same are true for $f \circ q_* : T^1\mathbb{H}^2 \to \mathbb{R}$, which is moreover invariant by the action of G. Let $\xi \in \mathbb{R} \cup \{\infty\}$. For every $x_0 + iy_0 \in \mathbb{H}^2$ there exists a unique hyperbolic geodesic passing through $x_0 + iy_0$, having positive end ξ, and a unique horocycle passing

through x_0+iy_0 and ξ. The hyperbolic geodesic passing through an arbitrary point $x'+iy' \in \mathbb{H}^2$ and having positive end ξ intersects the horocycle at a point $x+iy$, which depends only on x' and y'. Then,

$$(f \circ q_*)(x', y', \xi) = (f \circ q)(x, y, \xi) = (f \circ q)(x_0, y_0, \xi)$$

μ-almost for every $x_0 + iy_0$, $x' + iy' \in \mathbb{H}^2$, since f is invariant by the geodesic flow. Thus, we have a measurable function $\tilde{f} : \mathbb{R} \cup \{\infty\} \to \mathbb{R}$ defined by $\tilde{f}(\xi) = (f \circ q_*)(x_0, y_0, \xi)$, which is invariant by the action of G on $\mathbb{R} \cup \{\infty\}$.

5.5.2. Proposition. *If the action of G on $\mathbb{R} \cup \{\infty\}$ is ergodic with respect to the Lebesgue measure, then the geodesic flow is ergodic with respect to the Liouville measure.*

Proof. Let $f : T^1M \to \mathbb{R}$ be a measurable function, invariant by the geodesic flow. Let $\tilde{f}$ be the measurable function defined above. If the action of G on $\mathbb{R} \cup \{\infty\}$ is ergodic, then $\tilde{f}$ is constant almost everywhere with respect to the Lebesgue measure. As we saw in section 5.2 the Liouville measure has the form

$$\frac{2}{y|(x+iy) - \xi|^2} dx dy d\xi$$

modulo a constant. Consequently, $f \circ q_*$ must be almost everywhere constant with respect to the Liouville measure on $T^1\mathbb{H}^2$ and so must be f on T^1M. □

Thus, the ergodicity of the geodesic flow is reduced to the ergodicity of the action of G on $\mathbb{R} \cup \{\infty\} \approx S^1$. To study this, it is more convenient to work with the Poincaré disc model. Thus, in the sequel we assume that G is a subgroup of $\mathcal{M}$ which acts freely and properly discontinuously on $\mathbb{D}^2$, and its orbit space is compact, that is it has a Dirichlet polygon which is a compact subset of $\mathbb{D}^2$.

For every $\xi \in S^1 = \partial\mathbb{D}^2$, one can define the harmonic function $P(., \xi) : \mathbb{D}^2 \to \mathbb{R}^+$ by

$$P(z, \xi) = \frac{1 - |z|^2}{|z - \xi|^2},$$

which is called the *Poisson kernel*. For every $g \in \mathcal{M}$ we have

$$P(g(z), g(\xi)) = \frac{1 - |g(z)|^2}{|g(z) - g(\xi)|^2} = \frac{|g'(z)|(1 - |z|^2)}{|g'(z)||g'(\xi)||z - \xi|^2} = \frac{1}{|g'(\xi)|} P(z, \xi).$$

If $f : S^1 \to \mathbb{R}$ is in L^1 of the Lebesgue measure, the function $F : \mathbb{D}^2 \to \mathbb{R}$ defined by

$$F(z) = \frac{1}{2\pi} \int_0^{2\pi} P(z, e^{i\theta}) f(e^{i\theta}) d\theta$$

is the *Poisson integral* of f and is harmonic, because it is the real part of the function

$$\frac{1}{2\pi} \int_0^{2\pi} \frac{e^{i\theta} + z}{e^{i\theta} - z} f(e^{i\theta}) d\theta = \frac{1}{2\pi} \int_0^{2\pi} \frac{2}{e^{i\theta} - z} e^{i\theta} f(e^{i\theta}) d\theta - \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\theta}) d\theta$$

which is holomorphic in $\mathbb{D}^2$. For every $g \in \mathcal{M}$ and $z \in \mathbb{D}^2$ we have

$$F(g(z)) = \frac{1}{2\pi} \int_0^{2\pi} P(g(z), e^{i\theta}) f(e^{i\theta}) d\theta =$$

$$\frac{1}{2\pi}\int_0^{2\pi} P(z,g^{-1}(e^{i\theta}))|(g^{-1})'(e^{i\theta})|f(e^{i\theta})d\theta = \frac{1}{2\pi}\int_0^{2\pi} P(z,e^{i\theta})f(g(e^{i\theta}))d\theta.$$

It follows from this that if $f\circ g = f$, then $F\circ g = F$.

For every $0 \le r < 1$ let $F_r : S^1 \to \mathbb{R}$ be the function defined by $F_r(\xi) = F(r\xi)$. Then, $\|F\|_1 \le \|f\|_1$ and

$$\lim_{r\to 1}\|F_r - f\|_1 = 0.$$

5.5.3. Theorem. *Let G be a subgroup of $\mathcal{M}$, which acts freely and properly discontinuously on $\mathbb{D}^2$. If the orbit space of the action is compact, then G acts ergodically on $S^1 = \partial\mathbb{D}^2$ with respect to the Lebesgue measure.*

Proof. Let $f : S^1 \to \mathbb{R}$ be a measurable function, invariant by the action of G. Then, the Poisson integral

$$F(z) = \frac{1}{2\pi}\int_0^{2\pi} P(z,e^{i\theta})f(e^{i\theta})d\theta$$

is a harmonic function on $\mathbb{D}^2$ invariant under the action of G. Since the orbit space of G is compact, G has a compact Dirichlet polygon Q in $\mathbb{D}^2$. It follows that F takes on extreme values on Q, and therefore in $\mathbb{D}^2$, because it is invariant under G. By the Maximum Principle, F must be constant and so

$$\int_{S^1}|F(0) - f| = \lim_{r\to 1}\|F_r - f\|_1 = 0.$$

Hence $f = F(0)$ almost everywhere with respect to the Lebesgue measure. □

5.5.4. Corollary. *The geodesic flow of a compact hyperbolic surface is ergodic with respect to the Liouville measure.*

We shall now investigate the ergodicity of the horocycle flow. We shall need the following.

5.5.5. Lemma. *If $f \in L^1(\mu)$, then $\lim_{a\to 0}\|f\circ R_a - f\|_1 = 0$.*

Proof. Since $C(T^1M)$ is dense in $L^1(\mu)$, there exists a sequence of continuous functions $f_n : T^1M \to \mathbb{R}$, $n \in \mathbb{N}$, such that $\lim_{n\to+\infty}\|f_n - f\|_1 = 0$. Let $\epsilon > 0$. There exists $n_0 \in \mathbb{N}$ such that $\|f_n - f\|_1 < \epsilon/3$, for $n \ge n_0$. Since T^1M is compact, each f_n is uniformly continuous, and therefore $\lim_{a\to 0}\|f_n\circ R_a - f_n\| = 0$. Let $\delta > 0$ be such that $\|f_{n_0}\circ R_a - f_{n_0}\| < \epsilon/3$ for $|a| < \delta$. Then,

$$\|f\circ R_a - f\|_1 \le \|f\circ R_a - f_{n_0}\circ R_a\|_1 + \|f_{n_0}\circ R_a - f_{n_0}\|_1 + \|f_{n_0} - f\|_1 <$$

$$\|(f - f_{n_0})\circ R_a\|_1 + \|f_{n_0}\circ R_a - f_{n_0}\|_1 + \frac{\epsilon}{3} <$$

$$\|f - f_{n_0}\|_1 + \frac{2\epsilon}{3} < \epsilon. \quad \square$$

5.5.6. Theorem. *The horocycle flow of a compact hyperbolic surface is ergodic with respect to the Liouville measure.*

Proof. Let $f \in L^1(\mu)$ be invariant by the horocycle flow $(h_s)_{s\in\mathbb{R}}$. Then, $f = f\circ h_s = f \circ R_{\pi-a} \circ g_t \circ R_{2\pi-a}$, where $\cot a = s/2$, $0 < a < \pi$, and $s = 2\sinh(t/2)$, by 5.3.1. Thus, $f \circ R_a \circ g_{-t} = f \circ R_{\pi-a}$. For every $l \in L^1(\mu)$ we have

$$\int_{T^1M} (f \circ R_{\pi-a}) \cdot l d\mu = \int_{T^1M} (f \circ R_a) \cdot (l \circ g_t) d\mu$$

and taking the limit we get

$$\lim_{t\to+\infty} \int_{T^1M} f \cdot (l \circ g_t) d\mu = \int_{T^1M} (f \circ R_\pi) \cdot l d\mu,$$

because $\lim_{t\to+\infty} a = 0$. Therefore,

$$\lim_{t\to+\infty} \frac{1}{t} \int_0^t \left(\int_{T^1M} f \cdot (l \circ g_s) d\mu \right) ds = \int_{T^1M} (f \circ R_\pi) \cdot l d\mu.$$

Since the geodesic flow is ergodic, by Fubini's theorem and dominated convergence, we have

$$\lim_{t\to+\infty} \frac{1}{t} \int_0^t \left(\int_{T^1M} f \cdot (l \circ g_s) d\mu \right) ds = \left(\int_{T^1M} f d\mu \right) \left(\int_{T^1M} l d\mu \right).$$

It follows that

$$\int_{T^1M} l \cdot \left(f \circ R_\pi - \int_{T^1M} f d\mu \right) d\mu = 0$$

for every $l \in L^1(\mu)$. Consequently,

$$f \circ R_\pi = \int_{T^1M} f d\mu$$

μ-almost everywhere, and f is μ-almost everywhere constant. □

Chapter 6

Horocycle flows of hyperbolic surfaces

6.1 Horocycle flows and discrete subgroups of $SL(2,\mathbb{R})$

Let M be a compact hyperbolic surface, that is M is the quotient space of $\mathbb{H}^2$ by a subgroup G of $PSL(2,\mathbb{R})$, which acts freely and properly discontinuously by hyperbolic isometries on $\mathbb{H}^2$. As we saw in section 5.5, the unit tangent bundle T^1M is smoothly diffeomorphic to the homogeneous space $G\backslash PSL(2,\mathbb{R})$ of right cosets of G in $PSL(2,\mathbb{R})$ and the Liouville measure is the projection of the Haar measure on $PSL(2,\mathbb{R})$, modulo a constant. If $p : SL(2,\mathbb{R}) \to PSL(2,\mathbb{R})$ is the double covering map, then $\Gamma = p^{-1}(G)$ is a discrete subgroup of $SL(2,\mathbb{R})$ and $\Gamma\backslash SL(2,\mathbb{R})$ is smoothly diffeomorphic to $G\backslash PSL(2,\mathbb{R})$, and the Liouville measure corresponds to the induced by the Haar measure on $SL(2,\mathbb{R})$. The latter is the unique Borel measure on $\Gamma\backslash SL(2,\mathbb{R})$ which is invariant by the right action of $SL(2,\mathbb{R})$ such that if $f \in C_c(SL(2,\mathbb{R}))$ and $f^\Gamma \in C(\Gamma\backslash SL(2,\mathbb{R}))$ is defined by

$$f^\Gamma(\Gamma g) = \sum_{\gamma\in\Gamma} f(\gamma g),$$

then the integral of f^Γ over $\Gamma\backslash SL(2,\mathbb{R})$ is equal to the integral of f over $SL(2,\mathbb{R})$.

The above diffeomorphism gives an isomorphism of the geodesic flow of M with the right action of the one parameter subgroup A of $SL(2,\mathbb{R})$ on $\Gamma\backslash SL(2,\mathbb{R})$ defined by

$$g_t(\Gamma g) = \Gamma(g a_{t/2}), \quad t \in \mathbb{R}$$

and an isomorphism of the horocycle flow with the right action of the one parameter subgroup N of $SL(2,\mathbb{R})$ on $\Gamma\backslash SL(2,\mathbb{R})$ defined by

$$h_t(\Gamma g) = \Gamma(g n_t), \quad t \in \mathbb{R}.$$

Recall that $g_t \circ h_s = h_{se^{-t}} \circ g_t$ for every t, $s \in \mathbb{R}$.

In the next section we shall prove that the horocycle flow of a compact hyperbolic surface is minimal. The compactness assumption is essential for minimality. If M has merely finite hyperbolic area, but is not compact, then G contains at least one

parabolic element of the form $p(gn_tg^{-1})$, for some $t \in \mathbb{R}$, $t \neq 0$, and $g \in SL(2,\mathbb{R})$, and then

$$h_t(\Gamma g) = \Gamma(gn_tg^{-1}g) = \Gamma g.$$

In other words, Γg is a periodic point.

The geodesic flow is not minimal, even when M is compact. If T is a hyperbolic element of G, it has two different fixed points $\xi, \eta \in \partial\mathbb{H}^2$ and it leaves the hyperbolic geodesic with positive end ξ and negative end η invariant. It follows that for every $s \in \mathbb{R}$ the point (ξ, η, s) in $T^1\mathbb{H}^2$ projects to a periodic point of the geodesic flow on T^1M.

6.1.1. Lemma. *Let $g \in SL(2,\mathbb{R})$. The orbit of Γg under the horocycle flow is dense in $\Gamma\backslash SL(2,\mathbb{R})$ if and only if the orbit of gN under the left action of Γ on the homogeneous space $SL(2,\mathbb{R})/N$ of left cosets of N in $SL(2,\mathbb{R})$ is dense in $SL(2,\mathbb{R})/N$.*

Proof. We observe that Γg has a dense orbit in $\Gamma\backslash SL(2,\mathbb{R})$ under the horocycle flow if and only if $\overline{\Gamma gN} = SL(2,\mathbb{R})$, because the quotient projection of $SL(2,\mathbb{R})$ onto $\Gamma\backslash SL(2,\mathbb{R})$ is a continuous, open map. For a similar reason, gN has a dense orbit under the left action of Γ if and only if $\overline{\Gamma gN} = SL(2,\mathbb{R})$. □

Thus, the horocycle flow is minimal if and only if the left action of Γ on $SL(2,\mathbb{R})/N$ is minimal. We examine this action more closely. The natural action of $SL(2,\mathbb{R})$ on $\mathbb{R}^2$ by evaluation has only two orbits, namely $\{(0,0)\}$ and $\mathbb{R}^2 \setminus \{(0,0)\}$. The isotropy group of the vector e_1 is N. Therefore, the map $\phi : SL(2,\mathbb{R})/N \to \mathbb{R}^2 \setminus \{(0,0)\}$ defined by $\phi(gN) = g(e_1)$ is smooth, one-to-one and onto. Actually, it is a diffeomorphism, since its inverse ψ is defined by

$$\psi(x,y) = \begin{pmatrix} x & \frac{-y}{x^2+y^2} \\ y & \frac{x}{x^2+y^2} \end{pmatrix} N.$$

Moreover, ϕ is an isomorphism of the left action of $SL(2,\mathbb{R})$ on $SL(2,\mathbb{R})/N$ to the action of $SL(2,\mathbb{R})$ on $\mathbb{R}^2\setminus\{(0,0)\}$ by evaluation. Thus, we arrive at the following.

6.1.2. Proposition. *The horocycle flow is minimal if and only if the natural action of Γ on $\mathbb{R}^2 \setminus \{(0,0)\}$ by evaluation is minimal.*

6.2 Dynamics of discrete subgroups of $SL(2,\mathbb{R})$

Let Γ be a discrete subgroup of $SL(2,\mathbb{R})$ such that $-I_2 \in \Gamma$, where I_2 is the identity 2×2 matrix, and $\Gamma\backslash SL(2,\mathbb{R})$ is compact. The first condition implies that if $g \in \Gamma$, then $-g \in \Gamma$ also. We shall denote by μ the normalized measure on $\Gamma\backslash SL(2,\mathbb{R})$ which is induced by the Haar measure on $SL(2,\mathbb{R})$.

6.2.1. Lemma. *For every $g \in SL(2,\mathbb{R})$ and every open neighbourhood V of I_2 in $SL(2,\mathbb{R})$ there exists $n \in \mathbb{N}$ such that $\Gamma \cap Vg^nV^{-1} \neq \varnothing$.*

Proof. Since the quotient map $q : SL(2,\mathbb{R}) \to \Gamma\backslash SL(2,\mathbb{R})$ is a continuous, open, onto map, and the Haar measure is positive on open sets, we have

$\mu(q(Vg^k)) = \mu(q(V)) > 0$ for every $k \in \mathbb{Z}$. Hence there are k, $l \in \mathbb{Z}$ with $k > l$ such that $q(Vg^k) \cap q(Vg^l) \neq \varnothing$, because μ is finite. This means that there are some x, $y \in V$ such that $xg^{k-l}y^{-1} \in \Gamma$. $\square$

6.2.2. Lemma. *For every $\epsilon > 0$ and $t > 0$ there exists an open neighbourhood V of I_2 in $SL(2,\mathbb{R})$ such that for every $s \geq t$ every element of a_sV has real eigenvalues $\lambda_1 > 1$ and $\lambda_2 = 1/\lambda_1$ with respective eigenvectors x_1 and x_2 such that $\|x_j\| = 1$ and $\|x_j - e_j\| < \epsilon$, $j = 1, 2$.*

Proof. For every $\delta > 0$ we consider the set

$$V_\delta = \{\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{R}) : \max\{|a-1|, |b|, |c|, |d-1|\} < \delta\}.$$

The family $\{V_\delta : 0 < \delta < 1\}$ is a basis of open neighbourhoods of I_2 in $SL(2,\mathbb{R})$. Let

$$0 < \delta < \frac{(e^t-1)^2}{e^{2t}+1} < \frac{e^t-1}{e^t+1} < 1,$$

so that $1 + \delta < e^t(1-\delta)$. If

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in V_\delta,$$

then a, $d > 0$ and

$$a_sg = \begin{pmatrix} ae^s & be^s \\ ce^{-s} & de^{-s} \end{pmatrix}$$

for every $s \in \mathbb{R}$, which has eigenvalues $\lambda_{1,2} = \frac{1}{2}(ae^s + de^{-s} \pm \sqrt{(ae^s + de^{-s})^2 - 4})$. The quantity $ae^s + de^{-s}$ is increasing with $s \geq t$, because

$$\sqrt{\frac{d}{a}} < \sqrt{\frac{1+\delta}{1-\delta}} < \frac{1+\delta}{1-\delta} < e^t.$$

So, for every $s \geq t$ we have $ae^s + de^{-s} \geq ae^t + de^{-t} > (1-\delta)(e^t + e^{-t}) > 2$. It follows that $\lambda_{1,2} \in \mathbb{R}$ and $\lambda_1 > 1$, $0 < \lambda_2 = 1/\lambda_1 < 1$. An eigenvector of a_sg corresponding to λ_1 is $y_1 = (\lambda_1 - de^{-s}, ce^{-s})$. If $x_1 = y_1/\|y_1\|$, then

$$\|x_1 - e_1\|^2 = 2(1 - \frac{1}{\|y_1\|}\langle y_1, e_1\rangle) = 2(1 - \frac{\lambda_1 - de^{-s}}{\sqrt{(\lambda_1 - de^{-s})^2 + (ce^{-s})^2}}) \leq$$

$$2(1 - \frac{\lambda_1 - de^{-s}}{(\lambda_1 - de^{-s}) + |c|e^{-s}}) = \frac{2|c|e^{-s}}{(\lambda_1 - de^{-s}) + |c|e^{-s}}) \leq \frac{2|c|}{e^s(\lambda_1 - de^{-s})},$$

because $\lambda_1 - de^{-s} > \frac{1}{2}[(1-\delta)e^s - (1+\delta)e^{-s}] > 0$. But

$$e^s(\lambda_1 - de^{-s}) = \frac{1}{2}e^s(ae^s - de^{-s} + \sqrt{ae^s + de^{-s})^2 - 4}) \geq \frac{1}{2}(ae^{2s} - d) \geq$$

$$\frac{1}{2}(ae^{2t} - d) \geq \frac{1}{2}[(1-\delta)e^{2t} - (1+\delta)] = \frac{1}{2}[(e^{2t}-1) - \delta(1+e^{2t})] > e^t - 1.$$

Hence

$$\|x_1 - e_1\|^2 < \frac{2|c|}{e^t - 1} < \frac{2\delta}{e^t - 1}.$$

It is obvious now that for every $\epsilon > 0$ there exists $\delta > 0$ such that $\|x_1 - e_1\| < \epsilon$ for every $g \in V_\delta$ and $s \geq t$. Similarly, there is a unit eigenvector x_2 for λ_2 such that $\|x_2 - e_2\| < \epsilon$. □

6.2.3. Lemma. *Let $g \in SL(2,\mathbb{R})$ be an element with two real eigenvalues $\lambda_1 > 1$ and $0 < \lambda_2 = 1/\lambda_1 < 1$ with corresponding unit eigenvectors z_1 and z_2. Then for every $\epsilon > 0$ there exists an open neighbourhood V of I_2 in $SL(2,\mathbb{R})$ such that for every $n \in \mathbb{N}$ every element of $g^n V$ has two real eigenvalues $\tilde{\lambda}_1 > 1$ and $0 < \tilde{\lambda}_2 = 1/\tilde{\lambda}_1 < 1$ with corresponding unit eigenvectors $\tilde{z}_1$ and $\tilde{z}_2$ such that $\|z_j - \tilde{z}_j\| < \epsilon$, $j = 1, 2$.*

Proof. Let $0 < \epsilon < 1$. We observe first that there are $T \in SL(2,\mathbb{R})$ and $t > 0$ such that $TgT^{-1} = a_t$. Thus, $z_j = c_j T^{-1}(e_j)$, where $c_j = \|T^{-1}(e_j)\|^{-1}$, $j = 1, 2$. Let $c = \max\{c_1, c_2\}$. There exists $0 < \delta < \epsilon$ such that $\|T^{-1}(y) - T^{-1}(e_j)\| < \epsilon/2c$, whenever $y \in \mathbb{R}^2 \setminus \{(0,0)\}$ and $\|y - e_j\| < \delta$, $j = 1, 2$. By 6.2.2, there exists an open neighbourhood W of I_2 in $SL(2,\mathbb{R})$ such that for every $n \in \mathbb{N}$ every element of $(Tg^nT^{-1})W$ has eigenvalues $\tilde{\lambda}_1 > 1$ and $0 < \tilde{\lambda}_2 = 1/\tilde{\lambda}_1 < 1$ with corresponding unit eigenvectors y_1 and y_2 such that $\|y_j - e_j\| < \delta$, $j = 1, 2$. The set $V = T^{-1}WT$ is an open neighbourhood of I_2 and $(Tg^nT^{-1})W = T(g^nV)T^{-1}$ for every $n \in \mathbb{N}$. Thus, for every $h \in g^nV$ there exists some $h' \in (Tg^nT^{-1})W$ such that $h = T^{-1}h'T$, and h has the same eigenvalues as h' with corresponding eigenvectors $x_j = c_j T^{-1}(y_j)$, $j = 1, 2$. Therefore,

$$\|x_j - z_j\| = c_j \|T^{-1}(y_j) - T^{-1}(e_j)\| < c_j \frac{\epsilon}{2c} \leq \frac{\epsilon}{2}.$$

If now $\tilde{z}_j = x_j/\|x_j\|$, $j = 1, 2$, then

$$\|\tilde{z}_j - z_j\| < 2\|x_j - z_j\| \leq \epsilon. \quad \square$$

6.2.4. Lemma. *Let $g \in SL(2,\mathbb{R})$ be an element with two real eigenvalues $\lambda_1 > 1$ and $0 < \lambda_2 = 1/\lambda_1 < 1$ with corresponding unit eigenvectors z_1 and z_2. Then for every $\epsilon > 0$ there exists an open neighbourhood W of I_2 in $SL(2,\mathbb{R})$ such that for every $n \in \mathbb{N}$ every element of Wg^nW^{-1} satisfies the conclusion of 6.2.3.*

Proof. Let $\delta > 0$ be such that $\delta(\|z_j\| + \delta) + \delta < \epsilon/2$, $j = 1, 2$, and let V be the open neighbourhood given by 6.2.3 for δ. Let W be an open neighbourhood of I_2 in $SL(2,\mathbb{R})$ such that $W = W^{-1}$, $W \cdot W \subset V$ and $\|h - I_2\| < \delta$ for every $h \in W$. If now $h \in Wg^nW^{-1}$, there are $h_1, h_2 \in W$ such that $h = h_1 g^n h_2^{-1} = h_1(g^n h_2^{-1} h_1) h_1^{-1}$, and h has the same eigenvalues with $g^n h_2^{-1} h_1 \in g^n V$. If y_1 and y_2 are corresponding unit eigenvectors of $g^n h_2^{-1} h_1$, then $x_j = h_1(y_j)$, $j = 1, 2$, are corresponding eigenvectors of h, and

$$\|x_j - z_j\| \leq \|h_1(y_j) - y_j\| + \|y_j - z_j\| \leq \|h - I_2\|\|y_j\| + \delta < \delta(\|z_j\| + \delta) + \delta < \epsilon/2. \quad \square$$

6.2.5. Proposition. *For any pair of non-empty open sets W_1, $W_2 \subset \mathbb{R}^2 \setminus \{(0,0)\}$ there exists $\gamma \in \Gamma$ having two real different eigenvalues with corresponding eigenvectors $x_1 \in W_1$ and $x_2 \in W_2$.*

Proof. There are $z_j \in W_j$, $j = 1, 2$, such that $\{z_1, z_2\}$ is a basis of the linear space $\mathbb{R}^2$. Let $\epsilon > 0$ be such that $S(z_j, \epsilon) \subset W_j$, $j = 1, 2$. For every $r > 1$ there exists $g \in SL(2,\mathbb{R})$ with eigenvalues r and $1/r$, and corresponding eigenvectors z_1, z_2. Let W be the open neighbourhood of I_2 given by 6.2.4. From 6.2.1 there exists $\gamma \in \Gamma \cap W g^n W^{-1}$ for some $n \in \mathbb{N}$. Therefore γ has two real different eigenvalues with corresponding eigenvectors x_1 and x_2 such that $\|x_j - z_j\| < \epsilon$, $j = 1, 2$. Hence $x_1 \in W_2$ and $x_2 \in W_2$. □

6.2.6. Theorem. *The set $D = \{g \in SL(2,\mathbb{R}) : g\Gamma g^{-1} \cap A \neq \{I_2\}\}$ is dense in $SL(2,\mathbb{R})$.*

Proof. The map $\psi : GL^+(2,\mathbb{R}) \to (\mathbb{R}^2 \setminus \{(0,0)\}) \times (\mathbb{R}^2 \setminus \{(0,0)\})$ defined by $\psi(g) = (g(e_1), g(e_2))$ is a topological embedding of $GL^+(2,\mathbb{R})$ onto an open subset of $(\mathbb{R}^2 \setminus \{(0,0)\}) \times (\mathbb{R}^2 \setminus \{(0,0)\})$. Let W_1, $W_2 \subset \mathbb{R}^2 \setminus \{(0,0)\}$ be two non-empty open sets such that $W_1 \times W_2 \subset \psi(GL^+(2,\mathbb{R}))$. By 6.2.5, there exists $\gamma \in \Gamma$ having two real different eigenvalues with corresponding eigenvectors $x_1 \in W_1$ and $x_2 \in W_2$. Let $g \in GL(2,\mathbb{R})$ be such that $g(e_j) = x_j$, $j = 1, 2$. Then $g^{-1}\gamma g$ has the same eigenvalues with γ and corresponding eigenvectors e_1 and e_2. Therefore $g^{-1}\gamma g \in A$. Consequently, $D_0 \cap (W_1 \times W_2) \neq \varnothing$, where $D_0 = \{g \in GL^+(2,\mathbb{R}) : g\Gamma g^{-1} \cap A \neq \{I_2\}\}$. This shows that D_0 is dense in $GL^+(2,\mathbb{R})$. Recall now that the homomorphism $r : GL^+(2,\mathbb{R}) \to SL(2,\mathbb{R})$ with $r(g) = (\det g)^{-1/2} g$ is a retraction. Hence $D = r(D_0)$ is dense in $SL(2,\mathbb{R})$. □

6.2.7. Corollary. *The geodesic flow on $\Gamma\backslash SL(2,\mathbb{R})$ has a dense set of periodic orbits.*

Proof. A point $\Gamma g \in \Gamma\backslash SL(2,\mathbb{R})$ is periodic with respect to the geodesic flow if and only if there exists $t \neq 0$ such that $g a_{t/2} g^{-1} \in \Gamma$ or equivalently $g^{-1}\Gamma g \cap A \neq \{I_2\}$. According to 6.2.6, the set of all such g is dense in $SL(2,\mathbb{R})$. Therefore its image in $\Gamma\backslash SL(2,\mathbb{R})$, which is the set of periodic points of the geodesic flow, is dense. □

By 6.2.6, for our purposes we may assume that $\Gamma \cap A \neq \{I_2\}$. Indeed, there exists some $g \in SL(2,\mathbb{R})$ such that $g\Gamma g^{-1} \cap A \neq \{I_2\}$, and $\Gamma' = g\Gamma g^{-1}$ is a discrete subgroup of $SL(2,\mathbb{R})$ which contains $-I_2$. Obviously, the action of Γ on $\mathbb{R}^2 \setminus \{(0,0)\}$ is minimal if and only if the action of Γ' is.

As a first step to the main theorem of this section, we shall prove that Γ acts minimaly on S^1. As action of Γ on S^1 we mean the restriction of the action of $SL(2,\mathbb{R})$ on S^1 defined by

$$g \cdot x = \frac{g(x)}{\|g(x)\|}.$$

Concerning this action we make two remarks. Let z_1, $z_2 \in S^1$ be linearly indepen-

dent. If we consider them as columns and $d = \det(z_1, z_2)$, then $g = (z_1, \frac{1}{d}z_2) \in SL(2,\mathbb{R})$ and $g \cdot e_1 = z_1$, $g \cdot e_2 = \pm z_2$.

If $z = (x, y) \in S^1$ with $y > 0$ and $t = -x/y$, then $n_t \cdot z = e_2$. If $y < 0$, then $n_t z = -e_2$.

6.2.8. Lemma. *If W_1, $W_2 \subset S^1$ are two non-empty open sets, then for every y_1, $y_2 \in S^1$ there exists $g \in SL(2,\mathbb{R})$ such that $\pm g \cdot y_1 \in W_1$ and $g^{-1} \cdot y_2 \in W_2$.*

Proof. There exists a rotation g_1 such that $e_1 \in g_1 \cdot W_2$ and $g_1 \cdot y_1 \neq \pm e_1$. From the second remark above, there exists $g_2 \in N$ such that $(g_2 g_1) \cdot y_1 = \pm e_2$. Since e_1 is fixed by N, we also have $e_1 \in (g_2 g_1) \cdot W_2$. Applying the first remark to y_2 and any other point $z \in W_1$, which is linearly independent to y_2, there exists $g_3 \in SL(2,\mathbb{R})$ such that $g_3 \cdot e_1 = y_2$ and $(g_3 g_2 g_1) \cdot y_1 = \pm z$. It follows that $y_2 \in g \cdot W_2$ and $\pm g \cdot y_1 \in W_1$, where $g = g_3 g_2 g_1$. □

6.2.9. Proposition. *The action of Γ on S^1 is minimal.*

Proof. Let $z \in S^1$ and $W \subset S^1$ be a non-empty open set. The set

$$J = \{(a, b) \in S^1 : a, b > \frac{1}{2}\}$$

is an open interval. By 6.2.8, there exists some $g \in SL(2,\mathbb{R})$ such that $\pm g \cdot e_1 \in W$ and $z \in g \cdot J$. Hence $z = a(g \cdot e_1) + b(g \cdot e_2)$ for some a, $b > 0$. By continuity, there exists an open neighbourhood V_j of $g \cdot e_j$ in S^1 such that every $z_1 \in V_1$ and $z_2 \in V_2$ are linearly independent and $z = cz_1 + dz_2$ for some c, $d > 0$. From 6.2.5, there exists $\gamma \in \Gamma$ with real eigenvalues $\lambda > 1$ and $0 < 1/\lambda < 1$, and corresponding eigenvectors $z_1 \in V_1$, $z_2 \in V_2$. Thus, there are c, $d > 0$ such that $z = cz_1 + dz_2$ and for every $n \in \mathbb{N}$ we have $\gamma^n(z) = c\lambda^n z_1 + d\lambda^{-n} z_2$. Therefore

$$\lim_{n \to +\infty} \gamma^n \cdot z = \lim_{n \to +\infty} \frac{c\lambda^n z_1 + d\lambda^{-n} z_2}{\|c\lambda^n z_1 + d\lambda^{-n} z_2\|} = z_1,$$

because $c > 0$ and $\lambda > 1$. If $g \cdot e_1 \in W$, then choosing $V_1 \subset W$ we have $\gamma^n \cdot z \in W$ eventually. If $-g \cdot e_1 \in W$, then we choose $-V_1 \subset W$, and so $\lim_{n \to +\infty} (-\gamma^n) \cdot z = -z_1 \in W$, while $-\gamma^n \in \Gamma$, since $-I_2 \in \Gamma$. □

6.2.10. Lemma. *If for every pair of non-empty open sets W_1, $W_2 \subset \mathbb{R}^2 \setminus \{(0,0)\}$ there exists $\gamma \in \Gamma$ such that $\gamma(W_1) \cap W_2 \neq \varnothing$, then there exists a dense orbit of Γ in $\mathbb{R}^2 \setminus \{(0,0)\}$.*

Proof. Let $\{V_n : n \in \mathbb{N}\}$ be a countable basis of open sets of $\mathbb{R}^2 \setminus \{(0,0)\}$. A point $x \in \mathbb{R}^2 \setminus \{(0,0)\}$ has a dense otbit under Γ if and only if $x \in \cap_{n=1}^{\infty} U_n$, where $U_n = \cup_{\gamma \in \Gamma} \gamma(V_n)$. Since U_n is non-empty, open, invariant by Γ and dense in $\mathbb{R}^2 \setminus \{(0,0)\}$, by our assumption, it follows from the Baire theorem that $\cap_{n=1}^{\infty} U_n$ is dense in $\mathbb{R}^2 \setminus \{(0,0)\}$. □

6.2.11. Proposition. *There exists at least one dense orbit of* Γ *in* $\mathbb{R}^2 \setminus \{(0,0)\}$.

Proof. Because of 6.2.10, it suffices to prove that for every pair of non-empty open sets V_1, V_2 in $\mathbb{R}^2 \setminus \{(0,0)\}$ there exists some $\gamma \in \Gamma$ such that $\gamma(V_1) \cap V_2 \neq \varnothing$. With no loss of generality we may assume that V_1, V_2 are open discs. Let $r : \mathbb{R}^2 \setminus \{(0,0)\} \to S^1$ be the retraction. Let $y \in V_2$. From 6.2.9, there exists some $\gamma_0 \in \Gamma$ such that $r(\gamma_0(y)) = r(\gamma_0(r(y))) \in r(V_1)$, and so $r(\gamma_0(y)) \in r(V_1) \cap r(\gamma_0(V_2))$, which means that $r(V_1) \cap r(\gamma_0(V_2)) \neq \varnothing$. Applying 6.2.5, there exists $\gamma \in \Gamma$ with two real eigenvalues $\lambda > 1$ and $0 < 1/\lambda < 1$ and corresponding eigenvectors x_1, $x_2 \in V_1$, such that $\{tx_1 : t > 0\} \cap \gamma_0(V_2) \neq \varnothing$. Denoting by $[x_1, x_2]$ the straight line segment with endpoints x_1 and x_2, we have $\gamma^n([x_1, x_2]) = [\gamma^n(x_1), \gamma^n(x_2)] = [\lambda^n x_1, \lambda^{-n} x_2]$ for every $n \in \mathbb{N}$. It follows that every point of the halfline $\{tx_1 : t > 0\}$ is the limit of some sequence $(y_n)_{n\in\mathbb{N}}$, where $y_n \in [\lambda^n x_1, \lambda^{-n} x_2]$, $n \in \mathbb{N}$. Since $\gamma_0(V_2)$ is open, this implies that there exists some $n_0 \in \mathbb{N}$ such that $\gamma_0(V_2) \cap \gamma^n([x_1, x_2]) \neq \varnothing$ for every $n \geq n_0$. Since $[x_1, x_2] \subset V_1$, because V_1 is convex, we conclude that $(\gamma_0^{-1}\gamma^n)(V_1) \cap V_2 \neq \varnothing$. □

6.2.12. Corollary. *There exists at least one point in* S^1 *whose orbit is dense in* $\mathbb{R}^2 \setminus \{(0,0)\}$.

Proof. If $x \in \mathbb{R}^2 \setminus \{(0,0)\}$ has a dense orbit, then so does $x/\|x\|$. □

6.2.13. Lemma. *The vector* e_1 *has a dense orbit in* $\mathbb{R}^2 \setminus \{(0,0)\}$.

Proof. By 6.2.12 there exists some $x_0 \in S^1$ with a dense orbit in $\mathbb{R}^2 \setminus \{(0,0)\}$, and by 6.2.9 there exists a sequence $(\gamma_n)_{n\in\mathbb{N}}$ in Γ such that $\lim_{n\to+\infty} r(\gamma_n(e_1)) = x_0$. Since we assume that $\Gamma \cap A \neq \{I_2\}$, there is some diagonal $\gamma_0 \in \Gamma \cap A$ with eigenvalues $\lambda > 1$ and $0 < 1/\lambda < 1$. For every $n \in \mathbb{N}$, there exists some $k_n \in \mathbb{Z}$ such that $1 \leq \lambda^{k_n}\|\gamma_n(e_1)\| \leq \lambda$, or in other words $1 \leq \|\gamma_n\gamma_0^{k_n}(e_1)\| \leq \lambda$. Passing to a subsequence if necessary, we may assume by compactness that there is some $y \in \mathbb{R}^2 \setminus \{(0,0)\}$ with $1 \leq \|y\| \leq \lambda$ such that $\lim_{n\to+\infty} \gamma_n\gamma_0^{k_n}(e_1)) = y$. It follows that

$$y = \lim_{n\to+\infty} \lambda^{k_n}\|\gamma_n(e_1)\| r(\gamma_n(e_1)) = \|y\| x_0.$$

Hence $\|y\|x_0 \in \overline{\Gamma(e_1)}$, and by linearity $\mathbb{R}^2 \setminus \{(0,0)\} = \overline{\Gamma(\|y\|x_0)} \subset \overline{\Gamma(e_1)}$. □

We consider now a $0 < \theta_0 < \pi/6$, and for every $0 < \epsilon \leq 1$ we set

$$J_\epsilon = \{(\cos\theta, \sin\theta) \in S^1 : |\theta| < \epsilon\theta_0\}.$$

The family $\{J_\epsilon : 0 < \epsilon \leq 1\}$ is a neighbourhood base of open intervals of e_1 in S^1.

6.2.14. Lemma. *If* $\gamma \in \Gamma \cap A$, $\gamma \neq I_2$, *then for every* $\epsilon > 0$ *there exists* $k(\epsilon) \in \mathbb{N}$ *such that* $r(\gamma^k(x)) \in J_\epsilon$ *for every* $x \in \mathbb{R}^2 \setminus \{(0,0)\}$ *with* $r(x) \in J_1$ *and* $k \geq k(\epsilon)$.

Proof. This follows easily, since γ is diagonal with eigenvalues some $\lambda > 1$ and $0 < 1/\lambda < 1$. Indeed, if $r(x) \in J_1$, then $r(x) = (\cos\theta, \sin\theta)$ for some $|\theta| < \theta_0$, and $\gamma^k(x) = \|x\|(\lambda^k \cos\theta, \lambda^{-k}\sin\theta) = (t_k \cos\theta_k, t_k \sin\theta_k)$, where $t_k > 0$ and

$\tan\theta_k = \lambda^{-2k}\tan\theta$. Therefore, $|\theta_k| \leq |\tan\theta_k| = \lambda^{-2k}|\tan\theta| \leq \lambda^{-2k}|\tan\theta_0|$. Consequently, there exists some $k(\epsilon) \in \mathbb{N}$ with $|\theta_k| < \epsilon\theta_0$ for every $k \geq k(\epsilon)$, and then $r(\gamma^k(x)) = (\cos\theta_k, \sin\theta_k) \in J_\epsilon$. □

6.2.15. Lemma. *Let $x \in S^1$ and $L_x = \{\|\gamma(x)\| : \gamma \in \Gamma$ is such that $r(\gamma(x)) \in J_1\}$. There exists a compact set $K_x \subset (0, +\infty)$ such that $(0, +\infty) = \{ts : t \in L_x, s \in K_x\}$.*

Proof. By 6.2.9, the action of Γ on S^1 is minimal and so $S^1 = \cup_{\gamma\in\Gamma}\gamma \cdot J_1$. By compactness of S^1, there is a finite set $F = \{\gamma_1, ..., \gamma_n\} \subset \Gamma$ such that $S^1 = \gamma_1 \cdot J_1 \cup ... \cup \gamma_n \cdot J_1$. Let $D(x, J_1) = \{\gamma \in \Gamma : \gamma \cdot x \in J_1\}$. For every $\gamma \in \Gamma$ there is some $\gamma_i \in F$ such that $\gamma(x) \in \gamma_i(J_1)$, that is $\gamma \in \gamma_i D(x, J_1)$. Thus, $\Gamma = FD(x, J_1)$. On the other hand, since $\Gamma\backslash SL(2, \mathbb{R})$ is compact, there exists a compact set $C \subset SL(2, \mathbb{R})$ such that $SL(2, \mathbb{R}) = C\Gamma = CFD(x, J_1)$. The set $C_0 = CF$ is compact, and for every $g \in SL(2, \mathbb{R})$ there are $\sigma \in C_0$ and $\gamma \in D(x, J_1)$ such that $g = \sigma\gamma$. So $\|g(x)\| = \|\gamma(x)\| \cdot \|\sigma(r(\gamma(x)))\|$. From the definitions $\|\gamma(x)\| \in L_x$, the set $K_x = \{\|\sigma(y)\| : \sigma \in C_0$ and $y \in \overline{J_1}\}$ is compact, and we have $\|g(x)\| \in L_xK_x$ for every $g \in SL(2, \mathbb{R})$. Observe now that for every $t > 0$ there exists $g \in SL(2, \mathbb{R})$ such that $g(x) = tx$ and so $t = \|g(x)\| \in L_xK_x$. □

6.2.16. Corollary. *For every $x \in S^1$ and every $\epsilon > 0$ there exists $\gamma \in \Gamma$ such that $r(\gamma(x)) \in J_1$ and $0 < \|\gamma(x)\| < \epsilon$.*

Proof. From 6.2.15 we have $\mathbb{R} = \log(L_x) + \log(K_x)$ and $\log(K_x)$ is compact. Hence $\inf L_x = 0$ and the conclusion is immediate from the definition of L_x. □

6.2.17. Theorem. *The natural action of Γ on $\mathbb{R}^2 \setminus \{(0, 0)\}$ by evaluation is minimal.*

Proof. By linearity of the action, it suffices to prove that $\overline{\Gamma(x)} = \mathbb{R}^2\backslash\{(0, 0)\}$ for every $x \in S^1$. Let $\gamma_0 \in \Gamma \cap A$, $\gamma \neq I_2$, be diagonal with eigenvalues $\lambda > 1$ and $0 < 1/\lambda < 1$. For every $x \in S^1$ and $\epsilon > 0$ there exists $k(\epsilon) \in \mathbb{N}$ such that $r(\gamma_0^k\gamma(x)) \in J_\epsilon$ for every $\gamma \in \Gamma$ with $r(\gamma(x)) \in J_1$ and $k \geq k(\epsilon)$, by 6.2.14. From 6.2.16, there exists some $\gamma \in \Gamma$ such that $r(\gamma(x)) \in J_1$ and $0 < \|\gamma(x)\| < \lambda^{-k(\epsilon)}$. Moreover, there exists $k \in \mathbb{Z}$ such that $1 \leq \lambda^k\|\gamma(x)\| < \lambda$. We have now $0 \leq k\log\lambda + \log\|\gamma(x)\| < \log\lambda$ and $\log\|\gamma(x)\| < -k(\epsilon)\log\lambda$, and therefore

$$k(\epsilon) < \frac{-\log\|\gamma(x)\|}{\log\lambda} \leq k.$$

If $\gamma_\epsilon = \gamma_0^k\gamma$, then $r(\gamma_\epsilon(x)) \in J_\epsilon$ and $\|\gamma_\epsilon(x)\| = \|\gamma(x)\| \cdot \|\gamma_0^k(r(\gamma(x)))\|$. On the other hand, if $r(\gamma(x)) = (z_1, z_2)$, then $1/2 < z_1 \leq 1$ and $|z_2| \leq 1/2$, from the definition of J_1. Now

$$\frac{1}{2}\lambda^k < \lambda^k z_1 \leq \|\gamma_0^k(r(\gamma(x)))\| = \sqrt{\lambda^{2k}z_1^2 + \lambda^{-2k}z_2^2} \leq \sqrt{\lambda^{2k} + \frac{1}{4}} < \lambda^k + 1,$$

and therefore

$$\frac{1}{2} \leq \frac{1}{2}\lambda^k\|\gamma(x)\| \leq \|\gamma_\epsilon(x)\| < \|\gamma(x)\|(\lambda^k + 1) \leq \lambda + \lambda^{-k(\epsilon)} < \lambda + 1.$$

This shows that for every $\epsilon > 0$, the point $\gamma_\epsilon(x)$ lies in the compact set $\{(t\cos\theta, t\sin\theta) : 1/2 \leq t \leq \lambda+1, |\theta| \leq \epsilon|\theta_0|\}$. So there are $\epsilon_n \searrow 0$ and $1/2 \leq s \leq \lambda+1$ such that $\lim_{n\to+\infty}\gamma_{\epsilon_n}(x) = se_1$. It follows that $se_1 \in \overline{\Gamma(x)}$, and consequently

$$\mathbb{R}^2 \setminus \{(0,0)\} = \overline{\Gamma(e_1)} = \frac{1}{s}\overline{\Gamma(se_1)} \subset \frac{1}{s}\overline{\Gamma(x)},$$

by 6.2.13. Hence $\mathbb{R}^2 \setminus \{(0,0)\} = \overline{\Gamma(x)}$. □

From 6.1.2 and 6.2.17 we get the main result of this section.

6.2.18. Theorem. *The horocycle flow of a compact hyperbolic surface is minimal.*

6.3 Inheritance of minimality

In this section we shall make a small digression to topological dynamics. More precisely, we shall examine whether the minimality of a continuous flow is inherited by some homeomorphism of its one parameter group. Let X be a compact metrizable space carrying a continuous flow $(\phi_t)_{t\in\mathbb{R}}$. Let $t_0 > 0$ and $S = t_0\mathbb{Z}$. Then, $\mathbb{R} = S + [-t_0, t_0]$. If $x \in X$, the set

$$S_x = \{s \in \mathbb{R} : \phi_s(x) \in \overline{Sx}\},$$

where $Sx = \{\phi_t(x) : t \in S\}$, is a closed monoid and $S \subset S_x$. Actually, S_x is a subgroup of $\mathbb{R}$. Indeed, if $s \in S_x$, there exist $s_1 \in S$ and $|t_1| \leq t_0$ such that $-s = s_1 + t_1$. Inductively, there exist sequences $(s_n)_{n\in\mathbb{N}}$ in S and $(t_n)_{n\in\mathbb{N}}$ in $[-t_0, t_0]$ such that $-s + t_n = s_{n+1} + t_{n+1}$ for every $n \in \mathbb{N}$. By compactness, there is a convergent subsequence $(t_{n_k})_{k\in\mathbb{N}}$. Now

$$-s + t_{n_k} - t_{n_{k+1}} = s_{n_k+1} + s + s_{n_k+2} + ... + s + s_{n_{k+1}-1},$$

and therefore

$$-s = \lim_{k\to+\infty} s_{n_k+1} + s + s_{n_k+2} + ... + s + s_{n_{k+1}-1} \in S_x,$$

because S_x is closed. Note also that $\overline{S_x x} = \overline{Sx}$.

6.3.1. Lemma. *If the flow is minimal, then $\overline{Sx}$ is minimal under ϕ_{t_0} for every $x \in X$.*

Proof. Let x, $y \in X$ be such that $y \in \overline{Sx}$. Since the flow is minimal, we have $x \in \overline{C(y)} = \overline{\phi([-t_0, t_0] \times Sy)}$, where ϕ is the flow map. By compactness and continuity of the flow, we have

$$\overline{\phi([-t_0,t_0]\times Sy)} \subset \overline{\phi([-t_0,t_0]\times\overline{Sy})} = \phi([-t_0,t_0]\times\overline{Sy}) = \phi(\overline{[-t_0,t_0]\times Sy})$$

$$\subset \overline{\phi([-t_0,t_0]\times Sy)}.$$

It follows that $x \in \phi([-t_0, t_0] \times \overline{Sy})$ and there exists $s \in [-t_0, t_0]$ such that $\phi_s(x) \in \overline{Sy} \subset \overline{Sx}$. In other words, $s \in S_x$ and $\overline{S\phi_s(x)} \subset \overline{Sy} \subset \overline{Sx}$. But

$$\overline{S\phi_s(x)} = \overline{\phi_s(Sx)} = \phi_s(\overline{Sx}) = \phi_s(\overline{S_x x}) = \overline{\phi_s(S_x x)} = \overline{S_x x} = \overline{Sx}.$$

Hence $\overline{Sy} = \overline{Sx}$. □

6.3.2. Theorem. *If $(\phi_t)_{t\in\mathbb{R}}$ is a minimal flow on a compact metrizable space X, there exists some $t > 0$ such that ϕ_t is a minimal homeomorphism of X.*

Proof. Suppose that ϕ_t is not minimal for every $t \in \mathbb{R}$. By 6.3.1, for every $t_0 > 0$ and $x \in X$, the set

$$A_{t_0}(x) = \overline{\{\phi_{nt_0}(x) : n \in \mathbb{Z}\}}$$

is minimal under ϕ_{t_0} and by our hypothesis $A_{t_0}(x) \neq X$. The family of $A_{t_0}(x)$, $x \in X$, is a decomposition of X into uncountably many ϕ_{t_0}-minimal sets. Indeed, for s_1, $s_2 \in \mathbb{R}$ we have $A_{t_0}(\phi_{s_1}(x)) = A_{t_0}(\phi_{s_2}(x))$ if and only if $s_1 - s_2 \in S_x$, that is $s_1 + S_x = s_2 + S_x$ in $\mathbb{R}/S_x$. Since $A_{t_0}(x) \neq X$ and S_x is a closed subgroup of $\mathbb{R}$, there exists $s_0 > 0$ such that $S_x = s_0\mathbb{Z}$, and therefore $\mathbb{R}/S_x$ is homeomorphic to S^1, which is uncountable. If now $s > 0$ and $A_s(x) = A_{t_0}(x)$, then $s \in S_x$ and so s is a rational multiple of t_0. Moreover, all such s are bounded away from zero. Thus,

$$s_0 = \min\{s > 0 : A_s(x) = A_{t_0}(x)\} > 0$$

and $A_{s_0}(x) = A_{t_0}(x)$. Let $f_{s_0} : C(x) \to S^1$ be the function defined by

$$f_{s_0}(\phi_T(x)) = e^{2\pi i T/s_0}.$$

If $y \in X$ and $(t_n)_{n\in\mathbb{N}}$, $(T_n)_{n\in\mathbb{N}}$ are such that

$$y = \lim_{n\to+\infty} \phi_{t_n}(x) = \lim_{n\to+\infty} \phi_{T_n}(x),$$

then the fractional part of $\dfrac{T_n - t_n}{s_0}$ tends to 0 or 1. For otherwise, we may assume, passing to a subsequence if necessary, that $y = \phi_\tau(z) = \phi_{\tau'}(z')$ for some z, $z' \in A_{s_0}(x)$, where

$$\tau = \lim_{n\to+\infty} s_0\Big(\frac{T_n}{s_0} - \Big[\frac{T_n}{s_0}\Big]\Big) \text{ and } \tau' = \lim_{n\to+\infty} s_0\Big(\frac{t_n}{s_0} - \Big[\frac{t_n}{s_0}\Big]\Big),$$

so that $0 < \tau - \tau' < s_0$. But then, $\phi_{\tau-\tau'}(z) = z'$, which means that $\tau - \tau' \in S_z = S_x$ and $\overline{S_z z} = \overline{S_x x}$. This contradicts the choice of s_0. Since S^1 is compact, we may thus extend f_{s_0} to a continuous function $f : X \to S^1$ such that

$$f(\phi_t(y)) = f(y)e^{2\pi i t/s_0}$$

for every $y \in X$ and $t \in \mathbb{R}$. Note that $f(y) = 1$ for $y \in A_{t_0}(x)$. Conversely, if $f(y) = 1$ and $y = \lim_{n\to+\infty} \phi_{t_n}(x)$, then

$$\lim_{n\to+\infty} s_0\Big(\frac{t_n}{s_0} - \Big[\frac{t_n}{s_0}\Big]\Big) = 0 \text{ or } s_0.$$

It follows that

$$\lim_{n\to+\infty} \phi_{s_0[t_n/s_0]}(x) = y \text{ or } \phi_{-s_0}(y).$$

Therefore, $f^{-1}(1) = A_{s_0}(x) = A_{t_0}(x)$. In this way, we have associated to each set $A_{t_0}(x)$ an eigenfunction of the flow, and from the above, in a one to one manner. Note that s_0 depends on x, but the set of all such $s_0(x)$, $x \in X$, is countable, as it is a subset of $\{t_0/n : n \in \mathbb{N}\}$. If f_j is the eigenfunction associated to $A_{t_0}(x_j)$, $j = 1, 2$, then

$$\|f_1 - f_2\| = \sup\{|f_1(x_1)e^{2\pi i(\frac{1}{s_0(x_1)} - \frac{1}{s_0(x_2)})t} - f_2(x_1)| : t \in \mathbb{R}\}.$$

If $s_0(x_1) \neq s_0(x_2)$, then $\|f_1 - f_2\| = 2$. So far we had a fixed t_0. Varying now t_0 in an uncountable set $I \subset \mathbb{R}$ such that $t\mathbb{Q} \cap t'\mathbb{Q} = \varnothing$ for t, $t' \in I$ with $t \neq t'$, we get an uncountable set of eigenfunctions of the flow with different eigenvalues. From the above, this set is discrete. This however contradicts the separability of $C(X)$. □

6.4 Unique ergodicity of horocycle flows

Let M be a compact hyperbolic surface and $(h_t)_{t\in\mathbb{R}}$ be the horocycle flow on T^1M, which is minimal by 6.2.18. In this section we shall prove that it is uniquely ergodic, the Liouville measure being the unique invariant measure. Since a continuous flow on a compact metrizable space is uniquely ergodic if and only if some reparametrization of it by a positive constant is, we may assume with no loss of generality that the time one map h_1 is a minimal smooth diffeomorphism of T^1M, by 6.3.2. Recall that if $(g_t)_{t\in\mathbb{R}}$ is the geodesic flow, then $g_t \circ h_s = h_{se^{-t}} \circ g_t$ for every t, $s \in \mathbb{R}$.

6.4.1. Lemma. *Let $t_n \to +\infty$ and $R_n : C(T^1M) \to C(T^1M)$, $n \in \mathbb{N}$, be the sequence of operators defined by*

$$R_nf(x) = \frac{1}{2^n}\int_0^{2^n} f(h_s(g_{-t_n}(x)))ds,$$

for $f \in C(T^1M)$ and $x \in T^1M$. If for every $f \in C(T^1M)$ the sequence $(R_nf)_{n\in\mathbb{N}}$ has a subsequence which converges uniformly to a constant, then the horocycle flow is uniquely ergodic.

Proof. Let $f \in C(T^1M)$. According to the hypothesis, there exist a constant $c \in \mathbb{R}$ and $n_k \to +\infty$ such that $R_{n_k}f \to c$ uniformly on T^1M. Thus for every $\epsilon > 0$ there is a $k_0 \in \mathbb{N}$ such that $|R_{n_k}f(x) - c| < \epsilon$ for every $x \in T^1M$ and $k \geq k_0$. For every $x \in T^1M$ and n, $m \in \mathbb{N}$ we have

$$\frac{1}{m}\sum_{j=0}^{m-1} R_nf(g_{t_n}(h_{2^nj}(x))) = \frac{1}{2^nm}\sum_{j=0}^{m-1}\int_0^{2^n} f(h_{2^nj+s}(x))ds =$$

$$\frac{1}{2^nm}\sum_{j=0}^{m-1}\int_{2^nj}^{2^n(j+1)} f(h_s(x))ds = \frac{1}{2^nm}\int_0^{2^nm} f(h_s(x))ds.$$

Let $k \geq k_0$. For every $t > 2^{n_k}$ such that $2^{n_k}\|f\| < t\epsilon$, there exists $0 \leq r < 2^{n_k}$ such that $t = 2^{n_k}m + r$, for some $m \in \mathbb{N}$. Now we have

$$\Big|\frac{1}{t}\int_0^t f(h_s(x))ds - c\Big| \leq \Big|\frac{1}{t}\int_0^{2^{n_k}m} f(h_s(x))ds - c\Big| + \Big|\frac{1}{t}\int_{2^{n_k}m}^t f(h_s(x))ds\Big| \leq$$

$$\Big|\frac{2^{n_k}}{t}\sum_{j=0}^{m-1} R_{n_k}f(g_{t_{n_k}}(h_{2^{n_k}j}(x))) - c\Big| + \frac{r}{t}\|f\| <$$

$$\Big|\frac{1}{m}(1-\frac{r}{t})\sum_{j=0}^{m-1} R_{n_k}f(g_{t_{n_k}}(h_{2^{n_k}j}(x))) - c\Big| + \epsilon \leq$$

$$\frac{1}{m}\sum_{j=0}^{m-1}|R_{n_k}f(g_{t_{n_k}}(h_{2^{n_k}j}(x))) - c| + \frac{1}{m}\cdot\frac{r}{t}\sum_{j=0}^{m-1}|R_{n_k}f(g_{t_{n_k}}(h_{2^{n_k}j}(x)))| + \epsilon <$$

$$\frac{1}{m}m\epsilon + \frac{1}{m}\frac{r}{t}m\|f\| + \epsilon < 3\epsilon.$$

This shows that

$$\lim_{t\to+\infty}\frac{1}{t}\int_0^t f(h_s(x))ds = c$$

uniformly for every $x \in T^1M$. Therefore, $(h_t)_{t\in\mathbb{R}}$ is uniquely ergodic. □

In the sequel we take $t_n = \log 2^n$, and so $t_{n+m} = t_n + t_m$ for every n, $m \in \mathbb{N}$.

6.4.2. Lemma. *For every n, $m \in \mathbb{N}$ and $f \in C(T^1M)$ we have*

$$R_{n+m}f = \frac{1}{2^m}\sum_{j=0}^{2^m-1} R_nf \circ h_j \circ g_{-t_m}.$$

Proof. From the definition of R_n, for every $x \in T^1M$ we have

$$R_nf(x) = \frac{1}{2^n}\int_0^{2^n} f(g_{-t_n}(h_{se^{-t_n}}(x)))ds = \frac{1}{2^n}\int_0^{2^n} f(g_{-t_n}(h_{s2^{-n}}(x)))ds =$$

$$\int_0^1 f(g_{-t_n}(h_s(x)))ds,$$

and

$$R_nf(h_j(g_{-t_m}(x))) = \int_0^1 f(g_{-t_n}(h_{s+j}(g_{-t_m}(x))))ds =$$

$$\int_j^{j+1} f(g_{-t_n}(h_s(g_{-t_m}(x))))ds = \int_j^{j+1} f(h_{se^{t_n}}(g_{-(t_n+t_m)}(x)))ds =$$

$$\frac{1}{2^n}\int_{2^nj}^{2^n(j+1)} f(h_s(g_{-(t_n+t_m)}(x)))ds.$$

Consequently,

$$\frac{1}{2^m}\sum_{j=0}^{2^m-1} R_n f(h_j(g_{-t_m}(x))) = \frac{1}{2^{n+m}}\sum_{j=0}^{2^m-1}\int_{2^n j}^{2^n(j+1)} f(h_s(g_{-(t_n+t_m)}(x)))ds =$$

$$\frac{1}{2^{n+m}}\int_0^{2^{n+m}} f(h_s(g_{-(t_n+t_m)}(x)))ds = R_{n+m}f(x). \quad \square$$

6.4.3. Proposition. *If for every $f \in C(T^1M)$ the sequence $(R_n f)_{n\in\mathbb{N}}$ is equicontinuous, then the horocycle flow of M is uniquely ergodic.*

Proof. By 6.4.1, it suffices to prove that for every $f \in C(T^1M)$ the sequence $(R_n f)_{n\in\mathbb{N}}$ has a subsequence which converges uniformly to a constant. Let c_n be the minimum value of $R_n f$ on T^1M. Then, $c_n \leq \|f\|$ and $R_{n+m}f(x) \geq c_n$ for every n, $m \in \mathbb{N}$ and $x \in T^1M$, by 6.4.2, that is $c_{n+m} \geq c_n$. This means that the sequence $(c_n)_{n\in\mathbb{N}}$ is nondecreasing and bounded from above, hence converges to a limit $c \in \mathbb{R}$. The sequence of continuous functions $(R_n f)_{n\in\mathbb{N}}$ is obviously uniformly bounded by $\|f\|$. It follows from this and the equicontinuity that there are $F \in C(T^1M)$ and $n_k \to +\infty$ such that $R_{n_k}f \to F$ uniformly on T^1M, by Ascoli's theorem. Thus, for every $\epsilon > 0$ there exists $k_0 \in \mathbb{N}$ such that $F(x) - \epsilon < R_{n_k}f(x) < F(x) + \epsilon$ for every $x \in T^1M$ and $k \geq k_0$. In particular, $c_{n_k} < F(x) + \epsilon$ for every $x \in T^1M$ and $k \geq k_0$, which implies that $c - \min\{F(x) : x \in T^1M\} \leq \epsilon$. On the other hand, $\min\{F(x) : x \in T^1M\} - \epsilon < R_{n_k}f(x)$ every $x \in T^1M$ and $k \geq k_0$, and therefore $\min\{F(x) : x \in T^1M\} - \epsilon \leq c$. Hence $|c - \min\{F(x) : x \in T^1M\}| \leq \epsilon$ for every $\epsilon > 0$, and so $c = \min\{F(x) : x \in T^1M\}$. If now $m \in \mathbb{N}$ and

$$F_m = \frac{1}{2^m}\sum_{j=0}^{2^m-1} F \circ h_j \circ g_{-t_m}$$

then $R_{n_k+m}f \to F_m$ uniformly on T^1M. Consequently,

$$\min\{F_m(x) : x \in T^1M\} = \lim_{k\to+\infty} \min\{R_{n_k+m}f(x) : x \in T^1M\} = c.$$

So, $F_m(y_m) = c$ for some $y_m \in T^1M$. This implies that $F(h_j(g_{-t_m}(y_m))) = c$ for every $0 \leq j \leq 2^m - 1$. Let $x_m = g_{-t_m}(y_m)$. By compactness of T^1M, the sequence $(x_m)_{m\in\mathbb{N}}$ has at least one limit point $z \in T^1M$. Then, $F(h_j(z)) = c$ for every $j \in \mathbb{Z}^+$. Since now h_1 is a minimal homeomorphism of the compact space T^1M, it follows that F is constant on T^1M. This proves the proposition. $\square$

Using the notations we have introduced above, in order to prove that the horocycle flow is uniquely ergodic, it suffices to prove that for every $f \in C(T^1M)$ the sequence $(R_n f)_{n\in\mathbb{N}}$ is equicontinuous, by 6.4.3. We shall prove this using a convenient local reparametrization of the horocycle flow, which we introduce first.

Let x, $y \in T^1\mathbb{H}^2$ and for every $z \in T^1\mathbb{H}^2$ let H_z denote the horocycle through the point of application of z, that is tangent to $\partial\mathbb{H}^2$ at the positive end of the hyperbolic geodesic determined by z. The hyperbolic geodesic with positive end this point of

tangency and negative end identical with the negative end of the hyperbolic geodesic determined by y yields a unique element $[y, z] \in T^1\mathbb{H}^2$ with point of application its intersection with H_z, and such that $[y, z] = h_t(z)$ for some unique $t \in \mathbb{R}$.

In this way we get a function $k_{xy}(s)$ determined by

$$h_{k_{xy}(s)}(x) = [h_s(y), x].$$

We shall examine the properties of k_{xy}, if y is close to x. First of all we see that $k_{xy}(s) = v_{xy}(s) + \lambda_{xy}$ and $v_{xy}(s)$ is strictly increasing. Conjugating with a suitable orientation preserving hyperbolic isometry, we may assume that the ends of the hyperbolic geodesic detremined by y are ξ, the positive, and 1, the negative. Let a be the negative end of the hyperbolic geodesic determined by $h_s(y)$.

The hyperbolic isometry

$$T(z) = \frac{(a + 1 - \xi)(z - 1) + 1 - \xi}{z - \xi}$$

maps ξ to ∞ and fixes 1 and a. It also maps the horocycle H_y at ξ with euclidean radius $r > 0$ to the horocycle

$$\mathrm{Im} z = \frac{(1 - \xi)(a - \xi)}{2r}$$

and the two hyperbolic geodesics determined by y and $h_s(y)$ to vertical lines at 1 and a, respectively. It follows that

$$s = 2r \cdot \frac{a - 1}{(1 - \xi)(a - \xi)}.$$

Similarly,

$$v_{xy}(s) = 2t \cdot \frac{a-1}{(1-\eta)(a-\eta)}$$

where η is the positive end of the hyperbolic geodesic detrmined by x and $t > 0$ is the euclidean radius of the horocycle at η through the point of application of x. Solving with respect to a in the first and substituting in the second we find

$$v_{xy}(s) = \frac{-2t(1-\xi)^2 s}{(\xi-\eta)(1-\xi)(1-\eta)s - 2r(1-\eta)^2}, \qquad s \neq \frac{2r(1-\eta)}{(\xi-\eta)(1-\xi)}.$$

Of course t, r, ξ and η depend only on x, y and not on s. It is now evident from this expression of $v_{xy}(s)$ that $\lim_{y\to x} |k'_{xy}(s) - 1| = 0$ and $\lim_{y\to x} |k_{xy}(s) - s| = 0$, uniformly for every $-1 \leq s \leq 1$.

Since $k_{xy}(s) = k_{\gamma(x)\gamma(y)}(s)$ for every $\gamma \in \Gamma$ and x, $y \in T^1\mathbb{H}^2$, it follows that if x, $y \in T^1M$ are sufficiently close to each other then $k_{xy}(s)$ is well defined and has the above properties. We are now ready to proceed with the proof of the main result of this section.

6.4.4. Theorem. *The horocycle flow of a compact hyperbolic surface is uniquely ergodic.*

Proof. Using the notations as above, let $f \in C(T^1M)$, $x \in T^1M$ and $\epsilon > 0$. If y is sufficiently close to x, then

$$|f(g_{-t}(h_s(y))) - f(g_{-t}(h_{k_{xy}(s)}(x)))| < \epsilon$$

for every $|s| \leq 1$ and $t \geq 0$. Moreover, we may assume that $|k_{xy}(s) - s| < \epsilon$ and $|k'_{xy}(s) - 1| < \epsilon$ for all $|s| \leq 1$. Note that

$$|R_n f(y) - \int_0^1 f(g_{-t_n}(h_{k_{xy}(s)}(x)))ds| < \epsilon$$

and

$$|\int_0^1 f(g_{-t_n}(h_{k_{xy}(s)}(x)))ds - \int_0^1 k'_{xy}(s) f(g_{-t_n}(h_{k_{xy}(s)}(x)))ds| \leq \epsilon \|f\|.$$

From the change of variables formula we have

$$\int_0^1 k'_{xy}(s) f(g_{-t_n}(h_{k_{xy}(s)}(x)))ds = \int_{k_{xy}(0)}^{k_{xy}(1)} f(g_{-t_n}(h_s(x)))ds$$

and

$$|\int_{k_{xy}(0)}^{k_{xy}(1)} f(g_{-t_n}(h_s(x)))ds - R_n f(x)| \leq \|f\|(|k_{xy}(0)| + |k_{xy}(1) - 1|) < 2\epsilon\|f\|.$$

It follows that

$$|R_n f(y) - R_n f(x)| < (1 + 3\|f\|)\epsilon. \quad \square$$