Foreword. Ergodic Theory is a branch of the theory of Dynamical Systems, namely the branch in which the reference to a measure plays an essential role. Ergodic theory was started as a mathematical theory at the end of twenties/beginning of thirties of last century by G.D. Birkhoff and J. von Neumann, who provided the good mathematical frame and the first nontrivial results. The roots of the theory, however, and some leading ideas, go back to the work of L. Boltzmann and J.W. Gibbs, in the second half of 19th century, namely to their program to provide a microscopic dynamical fundation of thermodynamics, through statistical mechanics. Ergodic Theory had a strong development in the sixties and seventies of last century, after the introduction of the notion of Entropy by A.N. Kolmogorov and Ya.G. Sinai (end of fifties), the introduction by Sinai of the first really nontrivial examples of ergodic dynamical systems (Sinai billiards, 1962), and the connection with the theory of Lyapunov Characteristic Exponents (Oseledec 1968, Pesin 1974).

This short note is more a guide to the literature than a real introduction to Ergodic Theory. It includes the mathematical frame of the theory (sect. 1), the very central notions of “ergodicity”, which gives the name to the theory, and of “mixing” (sect. 2), an introduction to the Kolmogorov-Sinai entropy, with the main ideas and a few hints on selected nontrivial results (sect. 3), and finally, an introduction to Lyapunov Characteristic Exponents in Ergodic Theory (sect. 4). Appendix A reports a few comments on the physical roots of Ergodic Theory in Boltzmann and Gibbs, while the other appendices are devoted to technical parts.

Books on Ergodic Theory are abundant. Very classical readings include:


A suggestes very general book on dynamical systems is:

1 Introduction

1.1 Dynamical systems

Loosely speaking, a dynamical system is composed by two entities: a space $M$ — a set with some structure, traditionally called phase space; a point of $M$ is assumed to define completely the state of a system — and a dynamics on $M$. The typical entry level is that $M$ is a (separable, complete) metric space, usually assumed to be compact; $M$ has then a topology, and Borel sets form a $\sigma$–algebra. The so–called smooth theory of dynamical systems, in which the ergodic properties of a system are related to the expansion/contraction of tangent dynamics, requires more, namely that $M$ is a smooth manifold of some class $C^r$. Concerning the dynamics, reference is made to an independent variable $t$, usually called time, which belongs either to $\mathbb{R}$ (continuous dynamical systems) or to $\mathbb{Z}$ (discrete dynamical systems); for any $t$, a homeomorphism (non smooth case) or a diffeomorphism of class $C^r$ ($\Phi^t : M \to M$ is given, in such a way that

$$\Phi^0 = \text{Id}, \quad (\Phi^t)^{-1} = \Phi^{-t}, \quad \Phi^s \circ \Phi^t = \Phi^{s+t}. $$

This is indeed the typical situation one meets in physical systems, for example in Lagrangian or Hamiltonian mechanics, where one is given a vector field $X$ on some manifold $M$, and $\Phi^t(x)$ is the solution at time $t$, with initial datum $x \in M$, of the differential equation

$$\dot{x} = X(x).$$

In the discrete case, the group property implies that $\Phi$ is formed by the iterates of a map $\Phi^1$ or its inverse $\Phi^{-1}$; $\Phi^1$ is usually shortly denoted $\Phi$. A generalization is given by non invertible dynamical systems, where $t$ is nonnegative, $t \in \mathbb{R}^+$ or $\mathbb{N}$, and correspondingly $\Phi$ is a semigroup.

**Definition 1** Let $M$ be a compact metric space and

$$\Phi = \{\Phi^t, \; t \in \mathbb{R} \text{ or } \mathbb{Z}\}$$

be a one–parameter group ($t \in \mathbb{R}$ or $\mathbb{Z}$) or semigrup ($t \in \mathbb{R}^+$ or $\mathbb{N}$) of homeomorphisms ($t \in \mathbb{R}$ or $\mathbb{Z}$) or continuous maps ($t \in \mathbb{R}^+$ or $\mathbb{N}$) $M \to M$. The pair $(M, \Phi)$ is called dynamical system: continuous if $t \in \mathbb{R}$ or $\mathbb{R}^+$, discrete if $t \in \mathbb{Z}$ or $\mathbb{N}$, invertible if $t \in \mathbb{R}$ or $\mathbb{Z}$.

If in addition $M$ is a manifold of class $\mathcal{C}^r$ and each $\Phi^t$ is a diffeomorphism or a differentiable map $M \to M$ of class $\mathcal{C}^r$, the pair $(M, \Phi)$ is called smooth dynamical system of class $\mathcal{C}^r$.

Occasionally, with some attention, piecewise regular maps are also considered, for example in the study of iterations of maps on the unit interval, or in connection with the so–called billiards.

In Ergodic Theory, a further structure on $M$, namely an invariant measure defined on the Borel sets of $M$, plays a crucial role.

**Definition 2** Let $(M, \Phi)$ be a dynamical system, and let $\mu$ be a measure defined on the $\sigma$–algebra of the Borel sets of $M$, normalized in such a way that $\mu(M) = 1$; $\mu$ is said to be an invariant measure for $(M, \Phi)$, if for any $t$ and any measurable $A \subset M$

$$\mu(\Phi^{-t}(A)) = \mu(A),$$

(1)
where
\[ \Phi^{-t}(A) = \{ x \in M : \Phi^t(x) \in M \} . \]

In the non-invertible case, (1) is different from \( \mu(\Phi^t(A)) = \mu(A) \).

**Proposition 1** (Krylov–Bogoliubov theorem). Any continuous map on a compact metric space admits at least one invariant measure.

For the proof, not much interesting for us, see for example [1], sect. 1.8; the idea is to start with any measure, and construct from it an invariant one by a suitable averaging in time.

**Definition 3** Let \((M, \Phi)\) be a dynamical system and \(\mu\) an invariant measure for it; the triplet \((M, \mu, \Phi)\) is called dynamical system endowed with a measure; \((M, \mu, \Phi)\) is said to be smooth, or also classical, if \((M, \Phi)\) is smooth and moreover in each chart of \(M\) it is \(d\mu = \rho dV\), where \(dV\) is the Lebesgue measure and \(\rho\) is a smooth density.

It is common to call \((M, \mu, \Phi)\) simply a dynamical system, omitting the heavy specification “endowed with a measure”. Dynamical systems \((M, \varphi)\) not endowed with a measure are often called topological dynamical systems.

- As we shall see, making reference to a measure is crucial to introduce sensible definitions and formulate interesting propositions. In particular, it will be possible to consider properties which are satisfied almost everywhere, i.e. with the possible exception of a zero-measure subset of \(M\). “Almost everywhere” will be abbreviated in “a.e.”.

Some advanced results in Ergodic Theory require smoothness. The cornerstones of the theory, however, like the Birkhoff ergodic theorem or the notions of ergodicity and mixing, do not, and in fact are meaningful even without making reference to a metrics on \(M\), only assuming that \(M\) is a measure space i.e. an abstract set of points endowed with a \(\sigma\)-algebra of measurable sets and a measure \(\mu\) on it.

**Definition 4** The triplet \((M, \mu, \Phi)\), where \(M\) is a measure space endowed with normalized measure \(\mu\), \(\Phi : M \to M\) is as above, and (1) holds, is called abstract dynamical system.

Some elementary examples will be useful to take familiarity with the notion of dynamical system in Ergodic Theory.

**1.2 Examples**

**Example 1** (Free motion on \(T^2\)). Let \(M = T^2 = \mathbb{R}^2/\mathbb{Z}^2\), \(\mu = \text{Lebesgue measure}\); for any \(x \in T^2\) and \(t \in \mathbb{R}\) let
\[ \Phi^t(x) = x + vt \pmod{1} , \]
where \(v\) is some given velocity, i.e. vector of \(\mathbb{R}^2\). This is the system associated to the differential equation \(\dot{x} = v\) on \(T^2\); see figure 1a. If the ratio \(\alpha = v_1/v_2\) is rational, \(\alpha = p/q \in \mathbb{Q}\), then all motions are periodic with period \(T = p/v_1 = q/v_2\). If \(\alpha\) is irrational, all trajectories are open; as we shall see in a moment (proposition 2) they are dense on the torus.
Example 2 (Discrete translation on $S^1$). Let $M = S^1$, $\mu$ = Lebesgue measure, and let $\Phi$ be the map defined by

$$\Phi(x) = x + \alpha \pmod{1}.$$  

(2)

The connection with the previous example is immediate: if $N$ is a section of $T^2$ given by $x_2 = \text{const}$, then, see figure 1b, the free motion on $T^2$ with velocity $v$ induces a discrete translation of the form (2) on $N$, with $\alpha = v_1/v_2$, and the Lebesgue measure on $N$ is preserved.

**Proposition 2** If $\alpha$ is irrational, then all trajectories are dense on $S^1$; correspondingly, trajectories of example 1 are dense on $T^2$.

**Proof.** We show that for any $\varepsilon > 0$ and any $x \in S^1$, the trajectory $\Phi^t(x)$, $t \in \mathbb{Z}$, enters all intervals of length $\varepsilon$ (“almost everywhere” here is not necessary). Let $T > \varepsilon^{-1}$; among the $T$ points $x, \Phi(x), \ldots, \Phi^{T-1}(x)$, which for irrational $\alpha$ are all different, two at least, say $\Phi^{t'}(x)$ and $\Phi^{t''}(x)$, $t' > t''$, have distance smaller than $\varepsilon$. But the translation $\Phi$ is rigid, i.e. preserves the distance; going back $t''$ iterations, it follows $\text{dist}(\Phi^{t'}(x), x) < \varepsilon$, with $t = t' - t''$. This means the sequence $x, \Phi^{t}(x), \ldots, \Phi^{kt}(x), \ldots$ proceeds by steps smaller than $\varepsilon$, and thus enters any interval of length $\varepsilon$.

Example 3 (Free motion on $T^n$). Example 1 naturally generalizes to higher dimension. Let $M = T^n = \mathbb{R}^n/\mathbb{Z}^n$, $\mu$ = Lebesgue measure and

$$\Phi^t(x) = x + vt \pmod{1}, \quad v \in \mathbb{R}^n.$$

---

1 A more accurate notation for the abstract case, frequently found in books, is $(M, \mathcal{S}, \mu, \Phi)$, where $\mathcal{S}$ is the $\sigma$–algebra on which $\mu$ is defined.

2 Properly speaking, it is the Haar measure on the torus. We shall not make such a distinction.
The vector $v$ is said to be **resonant** with the integer vector $k \in \mathbb{Z}^n$, $k \neq 0$, if

$$k \cdot v = 0,$$

(3)

where the dot denotes the usual scalar product in $\mathbb{R}^n$; $v$ is said to be **nonresonant** if (3) is satisfied only by $k = 0$. The set $L(v)$ of integer vectors $k$ that satisfy (3) with a given $v$, is a subgroup of $\mathbb{Z}^n$, called **resonant lattice** or **resonant modulus** of $v$; its dimension is called the **multiplicity** of the resonance.

**Proposition 3** (Jacobi, 1835). If $v$ is nonresonant, each trajectory is dense in $\mathbb{T}^n$.

We shall prove the proposition later, as an exercise on the notion of ergodicity.

**Proposition 4** If the dimension of the resonant lattice of $v$ is $r$, $1 \leq r \leq n - 1$, then each trajectory is confined to an invariant manifold $N$ of dimension $n - r$ which is diffeomorphic to the torus $\mathbb{T}^{n-r}$.

**Proof.** We provide only a hint. One should preliminarily show that for any lattice $\mathcal{L}$ of dimension $r$, there exists a matrix $J$ with integer entries and $\det J = 1$ (so that $J^{-1}$ also has integer entries, and $J : \mathbb{Z}^n \to \mathbb{Z}^n$ is invertible), such that for any $k \in \mathcal{L}$ it is $Jk = (k'_1, \ldots, k'_r, 0, \ldots, 0)$. This implies that if $\mathcal{L}$ is the resonant lattice of $v$, then the change of variable $x = J^T y$ on $\mathbb{T}^n$ ($J^T$ denoting the transposed of $J$) turns the equation of motion $\dot{x} = v$ into $\dot{y} = u$, with

$$u = J^Tv = (0, \ldots, 0, u_{r+1}, \ldots, u_n).$$

So, the trajectory issuing from $y$ is confined to a torus $\mathbb{T}^{n-r}$ and the original trajectory is confined to $N = J^T \mathbb{T}^{n-r}$.

- Since $\mathcal{L}$ is assumed to be the resonant lattice of $v$, no further resonances do exist among $u_{r+1}, \ldots, u_n$. According to Jacobi theorem, $\mathbb{T}^{n-r}$ is densely filled by trajectories; in other words, $\mathbb{T}^n$ is naturally decomposed (foliated) in invariant tori of dimension $n - r$, one for each value of $y_1, \ldots, y_r$ (which are left constant by the dynamics), and this is the finest possible decomposition compatible with the dynamics.

**Example 4** (Hamiltonian systems). As is known (Liouville theorem), the Hamiltonian flow $\Phi^t$, $t \in \mathbb{R}$, preserves the Euclidian volume in canonical coordinates, $dV = dp_1 \cdots dp_n dq_1 \cdots dq_n$. So, if $M$ is the layer between two compact constant energy surfaces $\Sigma_E$ and $\Sigma_{E'}$ (think, to fix the ideas, they are $(n-1)$-dimensional spheres) and $\mu$ is the volume in $M$ normalized to one, $(M, \mu, \Phi)$ is a smooth dynamical system. However, this is not a deep view, since energy is conserved and so the flow will never mix different constant energy surfaces. It is not difficult, however, to see that the conservation of the volume in the phase space induces an invariant measure on each single
constant energy surface $\Sigma_E$. Such a measure, called Liouville measure and usually denoted $\mu_L$, can be defined as follows (fig. 2): let $A$ be any disc on $\Sigma_E$; consider the thick disc with base $A$, comprised between $\Sigma_E$ and $\Sigma_{E+\varepsilon}$, with any choice of the (transversal) lateral walls; let $V(A, \varepsilon)$ be its volume. Then define

$$\mu_L(A) = C \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} V(A, \varepsilon)$$

(4)

(the choice of the lateral walls, if transversal, is clearly irrelevant). The measure $\mu_L(A)$ is obviously invariant (both $\varepsilon$ and $V(A, \varepsilon)$ are, even before the limit in (4), and transversal lateral walls remain transversal). Once the measure is given on all discs on $\Sigma_E$, it naturally extends to all measurable sets of $\Sigma_E$. It is worthwhile to observe that $\mu_L$ is continuous with respect to the Euclidean area $d\Sigma$ on $\Sigma_E$, namely it is

$$d\mu_L = C \frac{d\Sigma}{\|\nabla H\|},$$

where $\|\cdot\|$ denotes Euclidian norm; this is an immediate consequence of the fact that the height $h(x)$ of the thick disc at $x \in A$ is $\varepsilon/\|\nabla H(x)\| + O(\varepsilon^2)$.

**Example 5** (Algebraic automorphism of $\mathbb{T}^2$, better known as Arnol’d cat). Let $M = \mathbb{T}^2$, $\mu =$ Lebesgue measure, and

$$\Phi \left( \begin{array}{c} x \\ y \end{array} \right) = A \left( \begin{array}{c} x \\ y \end{array} \right) \mod 1, \quad A = \left( \begin{array}{cc} 1 & 1 \\ 1 & 2 \end{array} \right);$$

see figure 3. Matrix $A$ can be replaced by any $2 \times 2$ matrix with integer entries, $\det A = 1$ (so that $A^{-1}$ also has integer entries), $\Tr A > 2$. The fact that both $A$ and $A^{-1}$ have integer entries, implies that $\Phi$ is a diffeomorphism $\mathbb{T}^2 \to \mathbb{T}^2$; $\det A = 1$ implies that $\mu$ is preserved. So, $(\mathbb{T}^2, \mu, \Phi)$ is a.

Figure 3: The Arnol’d cat: the action of $A$, the folding on $\mathbb{T}^2$. 
discrete invertible smooth dynamical system. The fact that \( \text{Tr } A > 2 \) implies that the eigenvalues are real and different from 1 and the eigenvectors have irrational slope. In the example, the eigenvalues are \( \lambda_1 = \lambda \) and \( \lambda_2 = \lambda^{-1} \), with

\[
\lambda = \frac{1}{2}(3 + \sqrt{5}) > 1 ;
\]

the eigenvectors are

\[
u_1 = (1, \lambda - 1) , \quad u_2 = (1, \lambda^{-1} - 1) .
\]

The map \( \Phi \) contracts in the direction of \( u_2 \) and expands (more precisely: \( \Phi^{-1} \) contracts) in the direction of \( u_1 \), at exponential rate.

The eigenspaces \( E_1, E_2 \), reported on the torus by mod 1, thanks to the irrational slope fill it densely (Proposition 2). Now, let \( B \) be a no matter how small ball; for simplicity imagine it centered in the origin, which is a fixed point, although this is not relevant. From the very definition of the map \( \Phi \), it follows that

\[
\Phi^t \left( \begin{array}{c} x \\ y \end{array} \right) = A^t \left( \begin{array}{c} x \\ y \end{array} \right) \mod 1
\]

(the reduction to the torus can be made at the end), and so

\[
\Phi^t(B) = A^t B \mod 1 .
\]

But \( A^t B \), for large positive \( t \), is a thin strip along \( E_1 \) of length \( \lambda^t d \) and width \( \lambda^{-1} d \), \( d \) being the diameter of \( B \); the reduction to the torus via the operation mod 1, scatters \( \Phi^t(B) \) everywhere in \( \mathbb{T}^2 \). This is of course vague and heuristic, but is a good view. Later on we shall formalize this vague idea into a precise notion, called mixing. This is a toy–model, but is the prototype of a class of systems, called Anosov’s systems, where contraction–dilatation of sets at exponential rate plays a determinant role. Figure 4 shows the evolution of 20,000 points randomly chosen inside a small box adjacent to the origin, up to 8 iterations.

**Exercise 1** Show that in the Arnol’d cat the set of periodic points is dense in \( \mathbb{T}^2 \). [Hint: all points with rational coordinates are periodic]

**Example 6** (Rubber-band map). This is a very simple discrete non–invertible system, with \( M = S^1, \mu = \text{Lebesgue} \), \( \Phi(x) = 2x \mod 1 \) (the common folding of a rubber band). Distances are stretched, without any contracting direction; measure is preserved because \( \Phi^{-1}(A) \) is the disjoint union of two subsets, whose measure is one half of the measure of \( A \). See figure 5.

**Example 7** (Bernoulli shifts). This is a dynamical system (thus, a deterministic system) which nevertheless is a model for a process as random as the throwing of a dice. Let \( I \) be a finite alphabet with \( n \) symbols, that we denote here \( 0, \ldots, n-1 \). The phase space, that we shall denote \( \Sigma \), is \( I^\mathbb{Z} \), i.e. the points of \( \Sigma \) are the sequences

\[
\sigma = (\ldots, \sigma_{-2}, \sigma_{-1}, \sigma_0, \sigma_1, \sigma_2, \ldots) , \quad \sigma_k \in I
\]

(think each of them be an a priori possible exit of an infinite sequence of throwings of a dice with \( n \) faces). A distance on \( \Sigma \) can be defined by following the idea that \( \sigma, \sigma' \) are close, if they coincide for a large interval around \( k = 0 \); for example

\[
\text{dist} (\sigma, \sigma') = \sum_{k \in \mathbb{Z}} 2^{-|k|} \delta(\sigma_k, \sigma'_k) , \quad \delta(l, l') = \begin{cases} 0 & \text{if } l = l' \\ 1 & \text{if } l \neq l' \end{cases}
\]
Figure 4: The evolution of 20,000 points initially confined in a small square near the origin, in the Arnol’d cat, for $t = 0, 1, 2, 4, 6, 8$. 
The $\sigma$–algebra of measurable sets is the one generated by the “elementary cylinders”

$$C_k^l := \{ x \in \Sigma : x_k = l \}, \quad k \in \mathbb{Z}, \quad l \in I;$$

the measure is assigned on it as follows: first, each symbol in $I$ is assigned a weight $p_l > 0$, with $\sum_{l \in I} p_l = 1$ (the a priori probability of the faces of the dice). Then one assigns

$$\mu(C_k^l) = p_l$$  \hspace{1cm} (5)

(the time $k$ of the individual throwing is irrelevant) and extends the measure to all “cylinders”

$$C_{k_1, \ldots, k_m}^{l_1, \ldots, l_m} = \bigcap_{j=1}^{m} C_{k_j}^{l_j}$$

of “base” $k_1, \ldots, k_m$ and “specification” $l_1, \ldots, l_m$, by posing

$$\mu(C_{k_1, \ldots, k_m}^{l_1, \ldots, l_m}) = p_{l_1} \cdots p_{l_m}$$

(different throwings are independent). It can be seen that such a prescription is sufficient to define $\mu$ on all measurable sets. Concerning the dynamics, we define $\Phi : \Sigma \rightarrow \Sigma$ by saying that $\sigma' = \Phi(\sigma)$ is $\sigma$ itself, left shifted by one position:

$$\sigma'_k = \sigma_{k+1}, \quad k \in \mathbb{Z}.$$  

Quite clearly, $\Phi$ is a homeomorphism and preserves $\mu$. The nonsmooth discrete dynamical system $(\Sigma, \mu, \Phi)$ constructed in this way is called Bernoulli shift, and is usually denoted $B_{p_0, \ldots, p_{n-1}}$. Bernoulli shifts play a quite relevant role in Ergodic Theory.

Exercise 2 Show that in any Bernoulli shift the periodic points are dense in $\Sigma$.  

Example 8 (Baker’s map). This is piecewise smooth system. Let $M$ be the square $[0, 1) \times [0, 1)$, $\mu =$ Lebesgue measure, and

$$\Phi(x, y) = \begin{cases} (2x, \frac{1}{2}y) & \text{for } x < \frac{1}{2} \\ (2x - 1, \frac{1}{2}y + \frac{1}{2}) & \text{for } x \geq \frac{1}{2} \end{cases} ;$$

see figure 6. $\Phi$ contracts vertically and expands horizontally; figure 7 shows what happens to 20,000 points originally filling a disc in the center of the square. $\Phi^{-1}$ is the same as $\Phi$, with $x$ and $y$ exchanged.

1.3 Isomorphism

Any interesting notion in Ergodic Theory must be invariant by isomorphism.

Definition 5 The dynamical systems $(M, \mu, \Phi)$ and $(N, \nu, \Psi)$ are said to be isomorphic, if there exist a map $h : M \to N$, defined and invertible almost everywhere, such that (i) for any measurable $A \subset M$, $h(A) \subset N$ is measurable and $\nu(h(A)) = \mu(A)$, and conversely; (ii) the two dynamics commute:

$$h \circ \Phi^t = \Psi^t \circ h . \quad (6)$$

An example is the following:

Proposition 5 The Baker’s map and the Bernoulli shift $B_{\frac{1}{2}, \frac{1}{2}}$ are isomorphic.

Proof. Let $M$ be the phase space (the square) of the Baker’s map; let $z = (x, y) \in M$, and write $x$ and $y$ in binary digits:

$$x = 0.a_0a_1a_2 \ldots , \quad y = 0.b_0b_1b_2 \ldots ;$$

the isomorphism is $\sigma = h(x, y) \in \Sigma$ with

$$\sigma = (\ldots, \sigma_{-3}, \sigma_{-2}, \sigma_{-1}, \sigma_0, \sigma_1, \sigma_2, \ldots) = (\ldots, b_2, b_1, b_0, a_0, a_1, a_2 \ldots) .$$

It is not difficult to check that the measures behave appropriately (look at the generators) and (6) holds. Note that $h$ is not one to one on strings that end, in either direction, by a sequence of symbols “1”. \qed
Figure 7: The evolution of 20,000 points initially filling a disc in the center of the square, by applying the Baker’s map: $t = 0, 1, 4, 6, 8, 10$. 
• An equivalent definition of the isomorphism is that $\sigma = h(z)$ is the *story* of $z$, observed through the natural partition of $M$

$$M = M_0 \cup M_1, \quad M_0 = \{ z = (x, y) \in M : x < \frac{1}{2} \};$$

this means $\sigma_k = l$ iff $\Phi^k(z) \in M_l$. Indeed, the above introduced map $h$ is such that $z \in M_l$ iff $\sigma = h(z) \in \mathcal{C}^l_0$; it follows

$$\Phi^k(z) \in M_l \iff h(\Phi^k(z)) \in \mathcal{C}^l_0 \iff \Psi^k(h(z)) \in \mathcal{C}^l_0 \iff h(z) \in \mathcal{C}^l_k \iff \sigma_k = l.$$ 

Quite in general, for any dynamical system $(M, \mu, \Phi)$, if a partition $M = M_0 \cup \cdots \cup M_{l-1}$ is introduced, then a map $h : M \to \Sigma = (0, \ldots, l-1)^{\mathbb{Z}}$ gets defined, such that $h(z)$ is the story of $z$ through the partition, and the shift $\Psi$ in $\Sigma$ obviously satisfies (6). But in general it is far from trivial to understand which is the image $h(M) \in \Sigma$ (which stories are effectively realized), if $h$ is invertible in $h(M)$, and which is the measure $\nu$ in $\Sigma$ corresponding to $\mu$ in $M$. The dynamics in a space of (sequences of) symbols, conjugate to the dynamics one is interested to study, is called *symbolic dynamics*. If one is able to find a good partition of $M$, such that $h$ is invertible on its image and $\nu$ is easy (for example: a Markov measure, with well defined transition probabilities), then $(M, \mu, \Phi)$ is completely understood. Not easy.

### 1.4 General results

Ergodic Theory deals, basically, with the statistical behavior of typical trajectories; this was indeed the very motivation of Boltzmann and Gibbs, who started it at the end of 19th century (Appendix A). The most elementary notion, in this perspective, is averaging. Given any function $f : M \to \mathbb{R}$, one wishes to know something about its *time average* $\bar{f} : M \to \mathbb{R}$, defined in the natural way as the limit (if it exists)

$$\bar{f}(x) = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(\Phi^t(x))dt \quad \text{or} \quad \bar{f}(x) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} f(\Phi^t(x))$$

respectively in the continuous and in the discrete case. For invertible systems, the backward time average is also defined in the obvious way, namely

$$\bar{f}_{(-)}(x) = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(\Phi^{-t}(x))dt \quad \text{or} \quad \bar{f}_{(-)}(x) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} f(\Phi^{-t}(x)).$$

For example, for any measurable set $A \subset M$, one might be interested in the fraction of time spent by a given trajectory in $A$, actually

$$\tau_A(x) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \chi_A(\Phi^t(x))dt \quad \text{or} \quad \tau_A(x) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \chi_A(\Phi^t(x)),$$ (7)

where $\chi_A$ denotes the characteristic function of $A$: $\chi_A(x) = 1$ for $x \in A$, otherwise $\chi_A(x) = 0$. The quantity $\tau_A(x)$ is naturally interpreted as the probability to find the system in $A$, for the motion with initial datum $x$, over a long time interval.

Another average, in a sense trivial for it does not involve the dynamics, is the *phase average* $\langle f \rangle$ of $f$, namely the number defined by

$$\langle f \rangle = \int_M f \, d\mu.$$
The phase average exists, by definition, for summable functions, i.e. \( f \in L_1(M, \mu) \). What about the time average? As is not trivial, the time average does exist for any dynamical system, for all summable functions and almost all points \( x \in M \). This is guaranteed by a fundamental theorem due to Birkhoff, usually called Birkhoff (or Birkhoff-Kinchin) ergodic theorem. The theorem does not make reference to distance, so it holds for abstract systems as well; to be definite, in the statement we refer to a discrete system, but the transposition to a continuous system is immediate.

**Proposition 6** Let \((M, \mu, \Phi)\) be any discrete dynamical system, and \( f \) be any function in \( L_1(M, \mu) \). For almost any \( x \in M \) the limit

\[
\bar{f}(x) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} f(\Phi^t(x))
\]

exists and moreover

\[
\bar{f}(\Phi^t(x)) = \bar{f}(x), \quad \langle \bar{f} \rangle = \langle f \rangle .
\]

(8)

If the system is invertible, \( \bar{f}(x) \) also exists almost everywhere, and almost everywhere coincides with \( \bar{f}(x) \).

The proof, not difficult but rather long and not much instructive, is deferred to the Appendix.

- One should stress the “almost everywhere” appearing in the statement. For example, for the Bernoulli shift \( B_{\{1,2\}} \) and for the characteristic function of the cylinder \( C_{(0,1)}^1 \), one easily finds initial data for which the forward average does not exist: for example,

\[
\sigma = (\ldots, \sigma_{-2}, \sigma_{-1}, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, (16 \times 0), (32 \times 1), \ldots)
\]

with any \( \sigma_k \) for \( k < 0 \). Quite clearly, the average truncated at a finite \( T \) oscillates, for no matter how large \( T \), between \( 1/3 \) and \( 2/3 \). It is worthwhile mentioning that by letting the above special string start in any position \( k > 0 \), rather than in \( k = 0 \), with arbitrary entries \( \sigma_j \) for \( j \leq k - 1 \), one obtains a dense set in \( \Sigma \). So, in this example (and in general) \( \bar{f} \) is not defined in any open set. This shows how crucial is making reference to a measure, and decide to disregard what happens to any set of points of zero measure. By similar tricks, one easily constructs a dense set of points \( x \) for which \( \bar{f}(x) \) and \( \bar{f}(x) \) are both defined, but do not coincide.

We conclude the Section with a second general result, due to Poincaré, showing a nontrivial subtle aspect of dynamics called “recurrence”.

**Definition 6** Let \((M, \mu, \Phi)\) be any dynamical system, and consider a measurable subset \( A \) of \( M \); a point \( x \in A \) is said to be recurrent in \( A \), if for any \( T > 0 \) there exists \( t \geq T \) such that \( \Phi^t(x) \in A \).

The set of non-recurrent points of \( A \) (points of \( A \) which after a certain time leave \( A \) forever) is then

\[
N_A = \{ x \in A : \exists T > 0 : \Phi^t(x) \notin A \forall t \geq T \} .
\]

(9)

For a pendulum, all points are recurrent in any open neighborhood (much more than recurrent: periodic), but those on the separatrices.

The following theorem holds:
Proposition 7 (Poincaré recurrence theorem) For any measurable $A \subset M$, $N_A$ is measurable and has zero measure.

Correspondingly, almost all points of $A$ are recurrent in $A$.

- An immediate consequence, when $M$ is a metric space, is that for any $\varepsilon > 0$ and almost any $x \in M$, it is $\text{dist} (\Phi^t(x), x) < \varepsilon$ for some arbitrarily large $t$ (consider a finite covering of $M$ with balls of diameter less than $\varepsilon$, and apply the recurrence theorem to each of them).

Proof. We can restrict ourselves to the case of discrete systems: the continuous case indeed reports to the discrete one by considering the time-one map (the set $N_A$ defined in (9), if one restricts $t$ to integers, possibly grows but does not shrink). For any $T \in \mathbb{N}$, denote

$$N_{A,T} = \{ x \in A : \Phi^t(x) \notin A \forall t \geq T \}$$

(“$T$-non recurrent” points of $A$); it is obviously

$$N_A = \bigcup_{T \in \mathbb{N}} N_{A,T} ,$$

so it is enough to show that for any measurable $A$ and any $T \in \mathbb{N}$ the set $N_{A,T}$ is measurable and has zero measure. The measurability of $N_{A,T}$ follows from its very definition, which can be rewritten

$$N_{A,T} = A \cap \left[ \bigcap_{t \geq T} \Phi^{-t}(M \setminus A) \right] .$$

Consider now the sets

$$N_{A,T}, \Phi^{-T}(N_{A,T}), \Phi^{-2T}(N_{A,T}), \ldots$$

They are all disjoint: indeed, should it be

$$x \in \Phi^{-kT}(N_{A,T}) \cap \Phi^{-lT}(N_{A,T}) \neq \emptyset ,$$

with for example $k < l$, then

$$\Phi^{kT}(x) \in N_{A,T} \cap \Phi^{-(l-k)T}(N_{A,T}) ,$$

against the definition of $N_{A,T}$. Due to the conservation of measure, all such sets have the same measure. The overall measure of the space being finite, it is necessarily $\mu(N_{A,T}) = 0$.

2 Ergodicity and mixing

2.1 Ergodicity

Ergodicity is the first nontrivial property one meets in ergodic theory, which discriminates among systems and starts a classification. It can be introduced in different equivalent ways, which stress different aspects of this deep notion.

E1. For any measurable $A \subset M$, the time spent in $A$ by a generic trajectory coincides with the measure of $A$:

$$\tau_A(x) = \mu(A) \quad \text{a.e. in } M ,$$

(10)

$\tau_A(x)$ being defined as in (7).
E2. For any $f \in L^1(M, \mu)$, it is 
\[
\bar{f}(x) = \langle f \rangle \quad \text{a.e. in } M.
\] (11)

E3. No constants of motion exist, but the trivial ones: namely if $f \in L^1(M, \mu)$ and 
\[
f(\Phi^t(x)) = f(x) \quad \forall t, \quad \text{a.e. in } M,
\] 
then $f$ is constant a.e. in $M$.

E4. The system is **metrically indecomposable**, namely for any measurable $A \subset M$
\[
\Phi^{-t}(A) = A \implies \mu(A) = 0 \text{ or } 1
\] 
(any decomposition $M = A \cup (M \setminus A)$, if measurable and invariant, is metrically trivial).

Properties E1–E4 are immediately seen to be invariant by isomorphism.

- E1 suitably formalizes the Boltzmann ergodic hypothesis, see Appendix A. E2 is proposed as definition of ergodicity in many textbooks of statistical mechanics. E3 formalizes the Gibbs idea, see again Appendix A, that the equilibrium distribution $\rho^* \equiv \text{unique}$ (if $f$ is a nontrivial constant of motion, then $\rho^* = cf$, $c$ providing normalization, is a Gibbs equilibrium distribution). Finally, E4 is the definition used by Birkhoff: very “elementary”, and very useful in proofs.

**Proposition 8** Properties E1, . . . , E4 are equivalent.

**Proof.** Many cross implications are easily proved. An easy path is the following:

(a) E2 $\Rightarrow$ E1  
(b) E1 $\Rightarrow$ E4  
(c) E4 $\Rightarrow$ E2  
(d) E4 $\Leftrightarrow$ E3 .

(a) is trivial, as (10) is a particular case of (11). To prove (b), assume by absurd that E4 is violated, i.e. there exists $A$ nontrivial and invariant; then $\tau_A(x) = 1 \neq \mu(A)$ for $x \in A$, against E1. Concerning (c), assume by absurd that E2 is violated; this means that for some function $f$ it is, for example, $\bar{f}(x) > \langle f \rangle$ in a set of positive measure. Denote
\[
A = \{x \in M : \bar{f}(x) > \langle f \rangle\}.
\]
By Birkhoff ergodic theorem $A$ is invariant (use the former of (8)) and nontrivial (use the latter: $\bar{f}(x)$ cannot be a.e. larger than $\langle f \rangle$), against E4. Finally (d) follows, in one direction, for, if E4 is violated by a set $A$, then $\chi_A$ is a nontrivial constant of motion and E3 is then violated; in the other direction for, if $f$ is a nontrivial constant of motion, then for some $c \in \mathbb{R}$ the invariant set
\[
A = \{x \in M : f(x) \leq c\}
\] is nontrivial, against E4. □
Definition 7 The dynamical system \((M, \mu, \Phi)\) is said to be ergodic, if any of E1–E4, and thus all, are satisfied.

- In E2 and E3, reference is made to all functions \(f \in L_1(M, \mu)\). Quite clearly, it is sufficient that properties are satisfied by all \(f \in L_2(M, \mu)\), to ensure they are satisfied by all functions in \(L_1(M, \mu)\); the restriction to a much smaller set of functions, like the set of characteristic functions of measurable sets, is in fact sufficient.

Proposition 9 Any Hamiltonian system with one degree of freedom, restricted to a compact connected line of constant energy which does not contain singular points, is ergodic.

Proof. Quite trivially, the trajectory runs on the line of constant energy, passing through all points. According to E4, the system is ergodic.

- Consider a set of \(n \geq 2\) independent harmonic oscillators, with Hamiltonian

\[
H(p, q) = \sum_{i=1}^{n} \frac{1}{2}(p_i^2 + \omega_i^2 q_i^2);
\]

the system is obviously not ergodic (the \(n\) energies are separately conserved). Understanding how to perturb the system, in such a way that it becomes ergodic, is among the most important still largely open problems of statistical mechanics (see the “Fermi-Pasta-Ulam problem”).

Proposition 10 The discrete translation on \(S^1\) (example 2) is ergodic iff \(\alpha\) is irrational.

Proof. If \(\alpha\) is rational, trajectories are periodic and the system is obviously non ergodic (or also: if \(\alpha = p/q\), then \(f(x) = \cos 2\pi qx\) is a non trivial constant of motion). Assume now \(\alpha\) is irrational; consider any \(f \in L_2(S^1, \mu)\) and let

\[
f(x) = \sum_{k \in \mathbb{Z}} \hat{f}_k e^{2\pi i kx}.
\]

Quite clearly,

\[
f(\Phi(x)) = \sum_{k \in \mathbb{Z}} (\hat{f}_k e^{2\pi i k\alpha}) e^{2\pi i kx},
\]

and so \(f(\Phi(x)) = f(x)\) a.e. if, for any \(k \in \mathbb{Z}\), \(\hat{f}_k(e^{2\pi i k\alpha} - 1) = 0\). This implies \(\hat{f}_k = 0 \forall k \neq 0\), i.e. \(f\) is a.e. constant on \(S^1\).

Exercise 3 Prove that if \(f\) is continuous, or is the characteristic function of an interval, then equalities (10) and (11) hold everywhere and not only a.e. (this it true in fact for all Riemann integrable functions, see [2]).

Proposition 11 The free motion on \(\mathbb{T}^n\) (examples 1, 3) is ergodic iff \(v\) is nonresonant. In particular (Jacobi’s theorem) if \(v\) is nonresonant, trajectories are dense on \(\mathbb{T}^n\).

Proof. The proof exploits the same idea used in Proposition 10; details are left as an exercise.
Exercise 4 Find the condition on $\alpha \in \mathbb{R}^n$ such that $(\mathbb{T}^n, \mu, \Phi)$, where $\mu$ is the Lebesgue measure and $\Phi : \mathbb{T}^n \to \mathbb{T}^n$ is the discrete translation $\Phi(x) = x + \alpha \ (\text{mod } 1)$, is ergodic. $[k \cdot \alpha \notin \mathbb{Z}, \forall k \neq 0.]$

Exercise 5 Table 1 reports the first 45 powers of 2; none of the numbers begins with the digit 7 or 9. Study the asymptotic frequencies $p_l, 1 \leq l \leq 9$, of the powers of 2 which begin with the digit $l$: do they exist? How are they ordered? Why do the digits 4 and 8 appear so regularly, every ten powers?

2.2 Mixing

The next relevant notion in ergodic theory is mixing. The underlying idea is that if one looks at sets of points, rather than at individual points, the dynamics is essentially irreversible; equivalently, any macroscopic initial information on the initial datum, is lost for large time. Formally, we can introduce two equivalent notions:

M1. For any pair of measurable sets $A, B \subset M$, it is

$$\lim_{t \to \infty} \mu(\Phi^{-t}(A) \cap B) = \mu(A)\mu(B). \quad (12)$$

M2. For any pair of functions $f, g \in L_2(M, \mu)$, it is

$$\lim_{t \to \infty} \langle (f \circ \Phi^t)g \rangle = \langle f \rangle \langle g \rangle, \quad (13)$$

that is

$$\lim_{t \to \infty} \int_M (f \circ \Phi^t)g \, d\mu = \int_M f \, d\mu \int_M g \, d\mu. \quad (14)$$
Properties M1, M2 are immediately seen to be invariant by isomorphirm.

- According to M1, the set $\Phi^{-t}(A)$, though preserving its measure, gets asymptotically diluted uniformly in $M$, more or less as when we put a drop of ink in a glass of water and stir. According to M2, correlations between any two functions (or physical observables) are lost with time. In general, the function $\mathbb{C} : \mathbb{R} \to \mathbb{R}$ defined by

$$\mathbb{C}(t) = \langle (f \circ \Phi^t) g \rangle - \langle f \rangle \langle g \rangle,$$

is called the correlation function of $f$ and $g$ (autocorrelation function of $f$, if $g = f$). If $\mathbb{C}(t) \neq 0$, then the knowledge of $g$ at time zero conditions the expectation of $f$ at time $t$.

- M2 formalizes Gibbs’ idea of convergence to equilibrium (take $\rho_0 = c g$, $c$ providing normalization, and introduce the change of variables $x = \Phi^{-t}(x')$ at the left hand side of (14)).

**Proposition 12** Properties M1 and M2 are equivalent.

**Proof.** M1 is a particular case of M2, when $f = \chi_A$ and $g = \chi_B$: indeed, it is $\chi_A \circ \Phi^t = \chi_{\Phi^{-t}(A)}$, $\langle (\chi_A \circ \Phi^t) \chi_B \rangle = \mu(\Phi^{-t}(A) \cap B)$, and so M2 turns into M1. To show that M1 implies M2, consider preliminarily the case in which $f$ and $g$ are simple functions, i.e. finite sums of characteristic functions of suitable measurable sets:

$$f = \sum_i f_i \chi_{A_i}, \quad g = \sum_j g_j \chi_{B_j};$$

for such functions, using $\chi_{A_i} \circ \Phi^t = \chi_{\Phi^{-t}(A_i)}$, it follows

$$\langle (f \circ \Phi^t) g \rangle = \sum_{ij} f_i g_j \langle \chi_{\Phi^{-t}(A_i)} \chi_{B_j} \rangle = \sum_{ij} f_i g_j \mu(\Phi^{-t}(A_i) \cap B_j)
\rightarrow \sum_{ij} f_i g_j \mu(A_i) \mu(B_j) = \sum_{ij} f_i g_j \langle \chi_{A_i} \rangle \langle \chi_{B_j} \rangle = \langle f \rangle \langle g \rangle.$$

For generic functions $f, g \in L_2(M, \mu)$, one exploits the fact that for arbitrary $\varepsilon$ they can be written $f = f_0 + f_1, g = g_0 + g_1$, with $f_0, g_0$ simple and $\|f_1\|, \|g_1\| < \varepsilon$. One then easily gets

$$|\langle (f \circ \Phi^t) g \rangle - \langle (f_0 \circ \Phi^t) g_0 \rangle| < C \varepsilon, \quad |\langle f \rangle \langle g \rangle - \langle f_0 \rangle \langle g_0 \rangle| < C \varepsilon$$

with some $C > 0$, and the conclusion follows.

**Definition 8** The dynamical system $(M, \mu, \Phi)$ is said to be mixing, if any of M1, M2, and thus both, are satisfied.

**Proposition 13** Mixing is stronger than ergodicity.

**Proof.** Assume $(M, \mu, \Phi)$ is mixing, and let $A$ be measurable and invariant. Using M1 with $B = A$, it follows

$$\mu(A) = \mu(\Phi^{-t}(A) \cap A) \to \mu(A)^2.$$

So $\mu(A) = 0$ or 1, and according to E4, the system is ergodic. An example of an ergodic non mixing system is the free motion on $\mathbb{T}^n$ ($\Phi^{-t}$ is a rigid motion, mixing cannot occur).

**Proposition 14** The Arnold’s cat (example 5) is mixing.
Proof. The functions \( u_k(x) = e^{2\pi ik \cdot x}, k \in \mathbb{Z}^2 \), are an orthogonal basis for \( L_2(M, \mu) \):

\[
\langle u_k u_l \rangle = 0 \quad \text{for} \quad k \neq l.
\]

From the definition of \( \Phi \), it follows

\[
u_k \circ \Phi = u_{\tilde{A}k}, \quad u_k \circ \Phi^t = u_{\tilde{A}^t k},
\]

where \( \tilde{A} \) denotes the transposed of \( A \) (we are not assuming here \( A \) is symmetric). Now, for \( k \neq 0 \), the trajectory of \( k \) under the action of \( \tilde{A}^t \), i.e. \( \{\tilde{A}^t k, t \in \mathbb{Z}\} \), is open, and with any norm in \( \mathbb{Z}^2 \) it is

\[
|\tilde{A}^t k| \to \infty \quad \text{for} \quad t \to \infty, \quad \forall k \in \mathbb{Z}^2, \quad k \neq 0;
\]

this follows because \( \tilde{A}^t k = k \) has no solutions, so \( \tilde{A}^t k \) cannot stay finite.\(^3\)

Mixing easily follows, making reference to M2. Consider preliminarily the case in which \( f, g \) are Fourier polynomials, say

\[
f(x) = \sum_{k \in K} \hat{f}_k e^{2\pi ik \cdot x}, \quad g(x) = \sum_{k \in K} \hat{g}_k e^{2\pi ik \cdot x},
\]

with some finite \( K \subset \mathbb{Z}^2 \). For such functions it is

\[
\langle (f \circ \Phi^t) g \rangle = \sum_{k \in \tilde{A}^t K, l \in K} \hat{f}_k \hat{g}_l \langle u_k u_l \rangle,
\]

but for large enough \( t \), \( \tilde{A}^t K \) and \( K \) intersect only in the origin, and so \( \langle (f \circ \Phi^t) g \rangle = \langle f \rangle \langle g \rangle \).

The results extends to generic \( f, g \in L_2(M, \mu) \), by approximating them, in \( L_2 \) norm, with Fourier polynomials.

- The idea of the proof is the existence of an orthogonal basis \( U = \{u_k, k \in \mathbb{Z}\} \) in \( L_2(M, \mu) \), such that (i) \( U \) is invariant under the dynamics induced by \( \Phi \) (i.e., composition \( u_k \circ \Phi \)), (ii) all functions but \( u_0 \) have an open non–recurrent behavior. The same idea will be used in the next proposition.

**Proposition 15** Bernoulli shifts are mixing.

**Proof.** We shall profit of notion M1 of mixing. Consider any cylinder \( C^{l_1, \ldots, l_K}_{-K, \ldots, K} \), with base between \( -K \) and \( K \) and any specification; it is clearly

\[
\Phi^{-t}(C^{l_1, \ldots, l_K}_{-K, \ldots, K}) = C^{l_1, \ldots, l_K}_{-K+t, \ldots, K+t},
\]

i.e. cylinders go into cylinders, the base going to infinity for \( t \to \infty \). (Cylinders here play the role of the basis functions \( u_k \) in the previous proof.) Let now

\[
A = \bigcup_{(l_1, \ldots, l_K) \in L} C^{l_1, \ldots, l_K}_{-K, \ldots, K}, \quad B = \bigcup_{(l_1', \ldots, l_K') \in L'} C^{l_1', \ldots, l_K'}_{-K, \ldots, K},
\]

\(^3\)In fact, the norm \( |\tilde{A}^t k| \) grows exponentially for large \( t \); profiting of this, one can understand more, namely that correlations in this example decay exponentially.
L and \( L' \) being any sets of indices in \( I^{2K+1} \). Quite clearly, it is

\[
\mu(\Phi^{-t}(A) \cap B) = \mu(A) \mu(B) \quad \text{for} \quad t > 2k + 1;
\]

indeed the base of cylinders forming \( A \), after translation, get disjoint from the base of cylinders forming \( B \), so the measure of the intersection factorizes. (Sets \( A \) and \( B \) play the same role as Fourier polynomials in the previous proof.)

The extension to any pair of measurable sets is straightforward: since the cylinders generate the \( \sigma \)-algebra of measurable sets, then for any \( \varepsilon > 0 \) there exist \( A_0, B_0 \) of the form above, with sufficiently large \( K(\varepsilon) \), such that

\[
\mu(A - A_0) < \varepsilon, \quad \mu(B - B_0) < \varepsilon;
\]

the symbol “−” denoting symmetric difference. One gets immediately

\[
|\mu(\Phi^{-t}(A) \cap B) - \mu(\Phi^{-t}(A_0) \cap B_0)| < (\text{const}) \varepsilon
\]

and so for \( t > 2K(\varepsilon) + 1 \) it is

\[
\mu(\Phi^{-t}(A) \cap B) - \mu(A) \mu(B) < (\text{const}) \varepsilon;
\]

M1 is thus satisfied.

\[\square\]

**Exercise 6** Prove that the rubber–band map (exercise 6) is mixing.

**Exercise 7** Show that \((M, \mu, \Phi^2)\) is mixing, if and only if \((M, \mu, \Phi)\) is mixing; show that if \((M, \mu, \Phi^2)\) is ergodic, then \((M, \mu, \Phi)\) is ergodic, but not conversely.

**Exercise 8** Show that in the Bernoulli shift \(B_{p,1-p}\), for almost all strings the frequency of the symbol “0” is \(p\).

**Exercise 9** For Arnol’d cat, find a continuous function \(f\) such that \(\overline{f}(x) \neq \langle f \rangle\) in a dense set. [Hint: take any function with nonvanishing average, which however vanishes in the origin, for example \(f(x_1,x_2) = \sin^2 x_1\); look at \(\overline{f}(x)\) for \(x\) in the contracting eigenspace \(E_2\) of the origin, reported to the torus via mod 1.]

Besides elementary examples as the above ones, not so many systems are known to be ergodic, or mixing. Among them, there are some billiards. Billiards, as dynamical systems, have been first studied by Birkhoff, in his celebrated book *Dynamical Systems*. Quickly: a billiard is a region \(M\) of the plain, or of the torus, with piecewise smooth border, on which a point freely moves, bouncing on the border according to the laws of elastic reflection. The measure \(d\mu = Cdx dy d\theta\) is easily seen to be preserved. The ergodic properties entirely depend on the shape of the border, and the question is highly nontrivial. The first billiard for which ergodicity and mixing have been proved (Sinai, 1962) is the Sinai billiard: a torus \(T^2\) with a circular obstacle inside. An equivalent problem is that of two discs on a torus, bouncing elastically. The proof has been extended to \(n \leq 5\) discs; a proof of ergodicity for generic \(n\) has been longly searched, but obstacles have been found (a model of \(n\) bouncing discs, or spheres in three dimensions, is a beautiful model of a gas in classical mechanics).
2.3 Comparing measures

Let $M$ be a compact metric space and $\Phi$ a dynamics on it; the Krylov–Bogoliobov theorem (Section 1.1) ensures the existence of at least one invariant measure. In special cases the invariant measure is unique (for example, the rotation $\dot{x} = 1$ on $S^1$), but in general it is not; an example is the set of the Bernoulli shifts $B_{p,1-p}$, $0 < p < 1$. Whenever there are more than one invariant measures, they are necessarily infinite: quite clearly, if $\mu'$ and $\mu''$ are invariant measures, the convex combination

$$\mu = c\mu' + (1-c)\mu'', \quad 0 \leq c \leq 1,$$

is also an invariant measure. This also shows that the set of invariant measures is convex.

The ergodic properties of the dynamical system $(M, \mu, \Phi)$ obviously depend on the choice of $\mu$. A good nontrivial question, for given $M$ and $\Phi$, is characterizing the set of the invariant measures, and among them the set of the ergodic measures, i.e. the measures such that $(M, \mu, \Phi)$ is ergodic. It is fairly possible that $(M, \mu, \Phi)$ and $(M, \nu, \Phi)$, with $\mu \neq \nu$, are both ergodic; this is precisely the case of $B_{p,1-p}$ and $B_{q,1-q}$ if $q \neq p$.

**Proposition 16** For given $M$ and $\Phi$,

i) If $\mu$ is not ergodic, there are infinitely many invariant measures, which are absolutely continuous with respect to $\mu$ (correspondingly, if there is only one invariant measure, it is necessarily ergodic).

ii) If $\mu$ is ergodic and $\nu \neq \mu$ is invariant, then $\nu$ is singular with respect to $\mu$ (there exists measurable $A$, such that $\mu(A) = 0$, $\nu(A) \neq 0$).

iii) If $\nu$ is also ergodic, there exist disjoint subsets $M_\mu$, $M_\nu$ such that

$$\mu(M_\mu) = 1, \quad \nu(M_\nu) = 1, \quad \mu(M_\nu) = 0, \quad \nu(M_\mu) = 0,$$

(each measure is supported by a subset which has zero measure for the other).

iv) If $\mu$ and $\nu$ are invariant, the strictly convex combination

$$c\mu + (1-c)\nu, \quad 0 < c < 1,$$

is not ergodic.

**Proof.** Point (i) is immediate: if $A$ is invariant and $\mu(A) \neq 0, 1$, then the restrictions of $\mu$ to $A$ and $M \setminus A$

$$\mu'(B) = \frac{\mu(B \cap A)}{\mu(A)}, \quad \mu''(B) = \frac{\mu(B \cap (M \setminus A))}{\mu(M \setminus A)},$$

are invariant, and so is any convex linear combination of them.

Let us now prove (iii). Since $\mu \neq \nu$, certainly there exists $A$ such that $\mu(A) \neq \nu(A)$. According to notion E1 of ergodicity, there exist $M_\mu$, $M_\nu$, of measure one respectively for $\mu$ and $\nu$, such that

$$\tau_A(x) = \mu(A) \quad \text{for} \quad x \in M_\mu, \quad \tau_A(x) = \nu(A) \quad \text{for} \quad x \in M_\nu.$$

The two sets are disjoint, and consequently $\nu(M_\mu) = \mu(M_\nu) = 0.$
To prove (ii), we can now assume $\nu$ is not ergodic, otherwise (ii) is already achieved. But then, according to E4, there exists $A$ invariant such that both $\nu(A)$ and $\nu(M \setminus A)$ are positive. On the other hand, since $\mu$ is assumed to be ergodic, either $A$ or $M \setminus A$ have zero $\mu$–measure.

The last point (iv) is immediate: let $\lambda_c = c\mu + (1 - c)\nu$; if both $\mu$ and $\nu$ are ergodic, then $M_{\mu}$ as constructed above is invariant and $\lambda_c(M_{\mu}) = c \neq 0, 1$. If instead for example $\mu$ is not ergodic, let $A$ be invariant with $0 < \mu(A) < 1$; no matter which $\nu(A)$ is, $\lambda_c(A) \neq 0, 1$.

• The situation (iii), which might appear paradoxical, occurs for the Bernoulli shifts. Take for simplicity shifts with only two symbols, and denote by $\mu_p$ the measure of $B_{p,1-p}$. Let $\Sigma_p$ be the subset of $\Sigma$, including strings $\sigma$ for which the frequency of the symbol “0” is defined and is $p$. For any $p$ (exercise (8)) it is $\mu_p(\Sigma_p) = 1$.

Now for $p \neq q$ the sets $\Sigma_p$, $\Sigma_q$ are by definition disjoint, and so $\mu_p(\Sigma_q) = 0 \forall q \neq p$.

We see that each $\mu_p$ is supported by a set $\Sigma_p$, which has zero measure for any other Bernoulli measure $\mu_q$ with $q \neq p$.

• According to (iv), ergodic measures cannot be expressed as convex combinations of other invariant measures: they stay at the border of the set of invariant measures. Instead, any non–ergodic measure is internal, namely can be expressed as convex combination of ergodic measures. In general infinitely many of them are necessary (an infinite sum and/or an integral). This is not trivial; see, in textboox, the **Ergodic decomposition**.

### 3 The Kolmogorov–Sinai entropy

In this section we shall introduce one of the most fundamental notions of ergodic theory, namely entropy. Entropy is a real number associated to any dynamical system $(M, \mu, \varphi)$, which turns out to be invariant by isomorphism. It has connections with information theory, and leads, in particular for smooth dynamical systems, to quite interesting developments, including an important connection (Pesin’s theorem) between entropy and dynamical quantities like the Lyapunov characteristic exponents.

#### 3.1 The notion of Entropy

**A. Entropy of a partition.** Let $\alpha = \{A_0, \ldots, A_{n-1}\}$ be a measurable partition of $M$ in $n$ atoms $A_0, \ldots, A_{n-1}$; in ergodic theory partition means “up to zero measure sets”, i.e. different atoms may intersect in a zero measure set and the union of all atoms may differ from $M$ by a zero measure set. Two partitions $\alpha = \{A_0, \ldots, A_{n-1}\}$ and $\alpha' = \{A'_0, \ldots, A'_{n-1}\}$ are considered to be the same partition if, after suitable ordering, $A_i$ and $A'_i$, $i = 1, \ldots, n$, are identical up to zero measure sets; zero measure sets can always be removed. A partition can be usefully thought of as an experiment, with a finite set of possible exits represented by the symbols $0, \ldots, n - 1$.  

22
In information theory, it is natural to assign to $\alpha$ the number

$$\eta(\alpha) = -\sum_{i=0}^{n-1} p_i \log p_i , \quad p_i = \mu(A_i) , \quad 0 \log 0 = 0 ,$$

where $\log$ denotes\(^4\) logarithm in base 2; $\eta$ is called the entropy of the partition $\alpha$. We shall use the notation

$$\eta = \sum_{i=0}^{n-1} \rho(p_i) , \quad \rho(p) = -p \log p ;$$

the graph of $\rho$ is shown in figure 8. The idea underlying the definition is that of the “a priori uncertainty” of the exit of the measurement, if the individual probabilities of the exits are $p_0, \ldots, p_{n-1}$; such an uncertainty then represents the expected information produced by the measurement. So, $\eta$ is zero if the exit is certain (one of the $p_i$’s is one, the other vanish), and is maximal if nothing is known, i.e. all exits have a priori the same probability; for $n = 2$, $\eta$ increases, left to right, for

$$(p_0, p_1) = (0, 1) , \quad (0.01, 0.99) , \quad \left(\frac{1}{3}, \frac{2}{3}\right) , \quad \left(\frac{1}{2}, \frac{1}{2}\right) .$$

To state a first proposition, which will help us to understand the meaning of $\eta$, we need a couple of definitions.

**Definition 9** Two partitions

$$\alpha = \{A_0, \ldots, A_{n-1}\} , \quad \beta = \{B_0, \ldots, B_{m-1}\}$$

are said to be independent, if for any pair of atoms $A_i$, $B_j$ it is $\mu(A_i \cap B_j) = \mu(A_i) \mu(B_j)$.

Examples of independent partitions are any partition of a rectangle in vertical and horizontal strips, or for Bernoulli shifts, the partitions $\alpha_k = \{C^0_k, \ldots, C^{n-1}_k\}$ for different $k$.

**Definition 10** The partition $\beta$ is said to be finer than $\alpha$, or to be a refinement of $\alpha$, denoted $\beta \succeq \alpha$, if (up to zero measure sets) each of the atoms of $\beta$ is entirely contained in one of the atoms of $\alpha$: for any $j$ there exists $i$ such that $\mu(B_j \cap A_i) = \mu(B_j)$ (and consequently $\mu(B_j \cap A_{i'}) = 0$ for $i' \neq i$).

\(^4\)The base of the logarithm plays no role at all, as it simply provides the unit of measure of the entropy. The logarithm in base 2 allows to use the language of “bits”, and is natural within information theory. Nowadays most books in ergodic theory use the natural logarithm.
**Definition 11** The composition \( \alpha \lor \beta \) of two partitions

\[
\alpha = \{A_0, \ldots, A_{n-1}\}, \quad \beta = \{B_0, \ldots, B_{m-1}\}
\]

is the partition

\[
\{A_i \cap B_j \neq \emptyset, \ 0 \leq i < n, \ 0 \leq j < m\}.
\]

**Proposition 17** The function \( \eta \) has the following properties:

i) \( \eta = 0 \) iff one of the \( p_i \) is one, and consequently the remaining ones are zero.

ii) For given \( n \), \( \eta \) is maximal if \( p_0 = \ldots = p_{n-1} = 1/n \), and \( \eta_{\text{max}} = \log n \).

iii) If \( \alpha \) and \( \beta \) are independent, then

\[
\eta(\alpha \lor \beta) = \eta(\alpha) + \eta(\beta).
\]

iv) If \( \beta \succeq \alpha \), then \( \eta(\beta) \geq \eta(\alpha) \).

In the proof we shall use an elementary inequality valid for concave functions, known as Jensen’s inequality:

**Lemma 18** (Jensen’s inequality). If \( f : \mathbb{R} \to \mathbb{R} \) is concave, then

\[
f(\sum_{i=1}^{n} c_i x_i) \geq \sum_{i=1}^{n} c_i f(x_i) \quad \text{for} \quad \sum_{i=1}^{n} c_i = 1.
\]

**Proof of the Lemma.** For \( n = 2 \) it is just the definition of concave function; for larger \( n \) it is an easy induction.

**Proof of the Proposition.** Point i) only depends on the fact that \( \rho \) vanishes in 0,1 and is positive in between. Point ii) follows from the concavity of \( \rho \), using Jensen’s inequality with \( c_i = 1/n, \ x_i = p_i \) (also recalling \( \sum_i p_i = 1 \)). Point iii) uses the explicit form of \( \rho \), with the logarithm inside: if \( p_0, \ldots, p_{n-1} \) and \( q_0, \ldots, q_{m-1} \) are respectively the measures of the atoms of \( \alpha \) and \( \beta \), for independent partitions it is

\[
\eta(\alpha \lor \beta) = -\sum_{i,j} p_i q_j (\log p_i + \log q_j) = \eta(\alpha) \sum_j q_j + \eta(\beta) \sum_i p_i = \eta(\alpha) + \eta(\beta).
\]

Finally, for point iv), it is enough to show that if an atom \( A \) of \( \alpha \) of measure \( p \) is the union of two atoms \( B \) and \( B' \) of \( \beta \), of measure respectively \( cp \) and \( (1-c)p \), then \( \rho(cp) + \rho((1-c)p) \geq \rho(p) \). This follows from concavity and \( \rho(0) = 0 \): indeed,

\[
\rho(cp) = \rho((1-c)0 + cp) \geq (1-c)\rho(0) + c\rho(p) = c\rho(p);
\]

similarly \( \rho((1-c)p) = (1-c)\rho(p) \), and the conclusion follows.

\[\Box\]

- It can be proved\(^5\) that properties i) — iv) completely characterize \( \eta \), up to a multiplicative constant (corresponding to the arbitrary choice of the base of the logarithm).

---

• If there are only two possible exits of a measurement, with equal a priori probability, then \( \eta = 1 \): this is the elementary information corresponding to one binary digit, or bit. Four, eight, . . . equally probable exits, give two, three . . . bits of information. If there are three exits \( A, B, C \) with a priori probability \( \frac{1}{2}, \frac{1}{4}, \frac{1}{4} \), then with probability \( \frac{1}{2} \) one gains one bit, with probability \( \frac{1}{4} \) one gains two bits, and again with probability \( \frac{1}{4} \) one gains two bits. Correspondingly it is \( \eta = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 = \frac{3}{2} \) bits (the uniform probability would give \( \eta = \log 3 \simeq 1.58 \)).

B. Entropy of \((M, \Phi)\) relative to a partition \(\alpha\). Let \((M, \mu, \Phi)\) be a discrete dynamical system and \(\alpha = \{A_0, \ldots, A_{n-1}\}\) be any measurable partition of \(M\). Let

\[
\beta_t = \alpha \vee \Phi^{-1}(\alpha) \vee \cdots \vee \Phi^{-t+1}(\alpha),
\]

where of course

\[
\Phi^{-1}(\alpha) = \{\Phi^{-1}(A_0), \ldots, \Phi^{-1}(A_{n-1})\}.
\]

**Definition 12** We shall call entropy of \((M, \Phi)\) relative to the partition \(\alpha\) the limit, which will be proved to exist,

\[
h(\Phi, \alpha) = \lim_{t \to \infty} \frac{1}{t} \eta(\beta_t). \tag{15}
\]

• Each atom of \(\beta\) can be interpreted as one of the possible exits of a sequence of \(t\) measurements of \(\alpha\), at times 0, 1, . . . , \(t-1\); indeed, if \(x\) belongs to the atom

\[
\bigcap_{s=0}^{t-1} \Phi^{-s}(A_{l_s})
\]

of \(\beta\), for some \(l_0, l+1, \ldots, l_{t-1}\), then

\[
x \in A_{l_0}, \quad \Phi(x) \in A_{l_1}, \quad \ldots, \quad \Phi^{t-1}(x) \in A_{l_{t-1}},
\]

and \(l_0, l+1, \ldots, l_{t-1}\) is the sequence of results. The entropy \(h(\Phi, \alpha)\) appears then as the expected information per measurement, in an infinite sequence of measurements.

• A positive value of \(h(\Phi, \alpha)\) indicates that \(\beta_t\) fragments, in a sense, with a rate “on the average, exponential” (remember the \(\log\) inside \(\eta\)). Indeed, let \(B(x, t)\) be the atom of \(\beta_t\) that contains \(x\); one easily checks that

\[
h(\Phi, \alpha) = \lim_{t \to \infty} \frac{1}{t} \int_M \log \mu(B(x, t)) \, d\mu. \tag{16}
\]

In this (weak) sense, it appears that the measure of atoms contracts, on the average, exponentially (this does not mean that all of them effectively do). A stronger result, in which we shall not enter, is provided by the Shannon–McMillan–Brian theorem.

**Proposition 19** The limit (15) exists.

To prove the proposition, we need a definition and a couple of lemmas.
**Definition 13** The quantity 
\[ \mu(A|B) = \frac{\mu(A \cap B)}{\mu(B)} \]
is called measure of \( A \) conditioned to \( B \).

Observe that for atoms \( A \) and \( B \) of two independent partitions, it is \( \mu(A|B) = \mu(A) \).

**Lemma 20** For any pair of measurable partitions 
\[ \alpha = \{A_0, \ldots, A_{n-1}\}, \quad \beta = \{B_0, \ldots, B_{m-1}\}, \]
it is 
\[ \eta(\alpha \vee \beta) \leq \eta(\alpha) + \eta(\beta), \]
the equality holding iff the partitions are independent.

**Proof.** From the definition of \( \eta \), one easily writes
\[
\eta(\alpha \vee \beta) = -\sum_{ij} \mu(A_i \cap B_j) \log (\mu(B_j) \mu(A_i|B_j)) \\
= -\sum_{ij} \mu(A_i \cap B_j) \log \mu(B_j) - \sum_{ij} \mu(A_i \cap B_j) \log \mu(A_i|B_j) \\
= \eta(\beta) - \sum_{ij} \mu(B_j) \mu(A_i|B_j) \log \mu(A_i|B_j)
\]
(use is made of \( \sum_i \mu(A_i \cap B_j) = \mu(B_j) \)). The second term on the r.h.s. has the form 
\[ \sum_i \left[ \sum_j c_j \rho(p_{ij}) \right], \]
with 
\[ c_j = \mu(B_j), \quad p_{ij} = \mu(A_i|B_j), \quad \rho(p) = -p \log p; \]
using Jensen’s inequality for each index \( i \), and then summing over \( i \), one immediately sees that such a term does not exceed \( \eta(\alpha) \), i.e. (17) is satisfied. On the other hand, due to the strict concavity of \( \rho \), Jensen’s inequality is strict, unless all \( p_{ij} \), for any \( i \), do not depend on \( j \); this means \( \mu(A_i|B_j) \) is independent of \( j \), and this in turn implies that partitions are independent.

**Lemma 21** If the sequence \( h_1, h_2, \ldots \) is bounded from below and satisfies the sub–additivity condition 
\[ h_{t+s} \leq \frac{t}{t+s} h_t + \frac{s}{t+s} h_s, \]
the limit \( \lim_{t \to \infty} h_t \) exists and is equal to \( h = \inf \{ h_t, t \in \mathbb{N} \} \).

The lemma extends to sub–additive functions the well known property of monotone non increasing functions.\(^6\)

**Proof.** We need to show that for large \( t \), the difference \( h_t - h \) is arbitrarily small. From the definition of \( h \), one has that for any \( \varepsilon > 0 \) there exists \( s = s(\varepsilon) \) such that 
\[ h_s < h + \varepsilon; \]
\(^6\)Monotone non increasing functions are immediately seen to be sub–additive; an example of sub–additive non monotone function is \( h_t = 0 \) for even \( t \), \( h_t = t^{-1} \) for odd \( t \).
from sub-additivity it follows immediately $h_{js} \leq h_s$, and so
\[ h_{js} < h + \varepsilon \quad \forall j > 0. \]

Let now $t$ stay between $js$ and $(j+1)s$, say $t = js + k$ with $0 < k < s$. From sub-additivity one has
\[ h_t \leq \frac{js}{js+k} h_{js} + \frac{k}{js+k} h_k < h + \varepsilon + \frac{1}{j} h_k; \]
on the other hand, once more for additivity, it is $h_k < h_1$: as a consequence,
\[ h_t < h + 2\varepsilon \]
for large enough $j \geq h_1/\varepsilon$, and so for large enough $t \geq T(\varepsilon) = h_1 s(\varepsilon)/\varepsilon$.

The proof of the proposition now gets trivial.

**Proof of Proposition 19.** From the definition of $\beta_t$ and from Lemma 20, it follows
\[ \eta(\beta_{t+s}) \leq \eta(\beta_t) + \eta(\Phi^{-1}(\beta_s)) = \eta(\beta_t) + \eta(\beta_s). \]
The sequence $h_t = \frac{1}{t} \eta(\beta_t)$ then satisfies the assumptions of Lemma 21, and so the limit (15) does exist.

**Exercise 10** Show that for the Bernoulli shift $B_{p_0,\ldots,p_{n-1}}$, denoting by $\alpha$ the partition in elementary cylinders $\alpha = \{ C_0^l, l \in I \}$, it is
\[ h(\Phi, \alpha) = \eta(\alpha) = -\sum_{i=0}^{n-1} p_i \log p_i. \]

**Exercise 11** Let $(M, \mu, \Phi)$ be any discrete dynamical system and $\alpha$ be any measurable partition of $M$; let $\beta_s = \alpha \lor \Phi^{-1}(\alpha) \lor \cdots \lor \Phi^{-s+1}(\alpha)$, for some $s > 1$. Show that $h(\Phi, \beta_s) = h(\Phi, \alpha)$. (In particular, for a Bernoulli shift $B_{p_0,\ldots,p_{n-1}}$, denoting $\beta_s = \{ C_0^l_0,\ldots,C_{s-1}^l \in I \}$, it is $h(\Phi, \beta_s) = h(\Phi, \alpha)$.) In the interpretation of measurements: by measuring $\beta_s$ several times, we repeat some already done measures, and do not achieve new information.

**C. The entropy of a system.** We are now ready to define the entropy of a system.

**Definition 14** Let $(M, \mu, \Phi)$ be any discrete dynamical system. The quantity
\[ h(\Phi) = \sup_{\alpha \text{ measurable}} h(\Phi, \alpha) \]
is called Kolmogorov–Sinai entropy of the system.

**Exercise 12** Show that $h$ is invariant by isomorphism.

**Exercise 13** Show that, for a discrete invertible system, $h(\Phi^{-1}) = h(\Phi)$.

**Exercise 14** Show that for any $s$ it is $h(\Phi^s) = |s|h(\Phi)$.  

27
Let us now come to continuous dynamical systems, that till now we disregarded. Consider any continuous dynamical system \((M, \mu, \Phi)\) and let \((M, \mu, \Phi^s)\) denote the discrete system corresponding to the time–s map of the continuous flow. It can be shown that for any \(s \in \mathbb{R}\), the entropy of \((M, \mu, \Phi^s)\) is \(h(\Phi^s) = |s|h(\Phi^1)\). This motivates the following definition:

**Definition 15** Let \((M, \mu, \Phi)\) be a continuous dynamical system. Its Kolmogorov–Sinai entropy is defined as the entropy \(h(\Phi^1)\) of the discretized system \((M, \mu, \Phi^1)\).

Definition 14 might appear obscure: taking the supremum among all measurable partitions — a quite huge set, far beyond the intuition — is apparently abstract and non constructive at all; moreover, since \(\eta(\alpha)\) and thus \(\eta(\beta_t)\) grow by taking finer and finer \(\alpha\), one could be afraid the supremum is infinite. Some relevant results, shortly sketched in the next paragraph, will help to make clear the situation. Here we only show, on the basis of an example, that taking a finer and finer \(\alpha\), does not obviously produce a growth of \(h(\Phi, \alpha)\). The example is that of a Bernoulli shift.

Let \(\alpha\) be the partition in elementary cylinders based on zero, 
\(\alpha = \{C^l_0, \ l \in I\}\), and denote 
\(\alpha_k = \{C_{-k, \ldots, k}^l, \ l_k, \ldots, l_k \in I\}\) 
\((\alpha_0 = \alpha)\). Quite clearly, \(\alpha_k\) gets finer for larger \(k\), namely \(\eta(\alpha_k) = (2k + 1)\eta(\alpha)\); however, 
\[
\begin{align*}
\alpha_k \lor \ldots \lor \Phi^{-t+1}(\alpha_k) &= \Phi^k(\alpha) \lor \ldots \lor \Phi^{-t-k+1}(\alpha) \\
\eta(\alpha_k \lor \ldots \lor \Phi^{-t+1}(\alpha_k)) &= (2k + t)\eta(\alpha)
\end{align*}
\]
and consequently \(h(\Phi, \alpha_k) = h(\Phi, \alpha) = \eta(\alpha)\).

### 3.2 Some results

**D. Some results on entropy.** The set \(\mathcal{P}\) of all measurable partitions is naturally endowed of a distance.

**Definition 16** Given any two measurable partitions \(\alpha = \{A_0, \ldots, A_{n-1}\}, \beta = \{B_0, \ldots, B_{m-1}\}\), the quantity 
\[
\eta(\alpha \mid \beta) = -\sum_j \mu(B_j) \sum_i \mu(A_i \mid B_j) \log \mu(A_i \mid B_j)
\]
is called entropy of \(\alpha\) relative to \(\beta\).

An equivalent definition is
\[
\eta(\alpha \mid \beta) = \eta(\alpha \lor \beta) - \eta(\beta)
\]
By means of relative entropy, we can introduce a distance in \(\mathcal{P}\), according to
\[
\text{dist} (\alpha, \beta) = \eta(\alpha \mid \beta) + \eta(\beta \mid \alpha)
\]

**Lemma 22**

i. The above defined quantity \(\text{dist} (\alpha, \beta)\) is a distance in \(\mathcal{P}\).
ii. For any dynamical system \((M, \mu, \Phi)\), \(h(\Phi, \alpha)\) is continuous in \(\alpha\), more precisely it is

\[
|h(\Phi, \alpha) - h(\Phi, \beta)| \leq \text{dist}(\alpha, \beta) .
\]

For the proof (annoying, but not difficult) see the references.

The above lemma is important: it allows to approximate partitions with “easier” ones, and the search of the supremum, in the definition 15 of entropy, can be restricted to any subset \(\mathcal{P}_0\) dense in \(\mathcal{P}\).

**Definition 17** Let \(M\) be a compact manifold. A partition of \(M\) is said to be smooth, if its atoms are polyhedra having as border a piecewise smooth submanifold of \(M\) of codimension 1.

**Lemma 23** The subset \(\mathcal{P}_0\) of smooth partitions of \(M\), is dense in the space \(\mathcal{P}\) of measurable partitions of \(M\).

For the (easy) proof, see the references.

**Proposition 24** (Kouchnirenko’s theorem). The entropy of smooth dynamical systems is finite.

**Proof.** We provide only a sketch, assuming \(\Phi\) is a diffeomorphism. Thanks to Lemma 23, it is enough to take into consideration smooth partitions of \(M\). One proceeds as follows:

i. It is always possible to introduce in \(M\) a metric adapted to the (smooth) measure \(\mu\), namely such that any polyhedron of volume \(V\) has measure \(V\). Let \(S(A)\) be the surface of atom \(A\) in such a metrics, and \(S(\alpha)\) the sum of the surfaces of all atoms of \(\alpha\). It is possible to prove that, denoting as usual \(\beta_t = \alpha \lor \cdots \lor \Phi^{-t+1}(\alpha)\), the surface of \(\beta_t\) grows at most exponentially with \(t\), namely

\[
S(\beta_t) \leq \text{const}\lambda^t
\]

where \(\lambda\) is the operator norm of the derivative \(D\Phi\) (the maximal dilatation coefficient of lengths, produced by \(\Phi\)). In the essence: take a thin disk, with base on the surface of an atom \(A\); since volumes are preserved by \(\Phi\), the base of the disc grows, by applying \(\Phi^{-1}\), only if its thickness decreases. But the maximal contraction produced by \(\Phi^{-1}\) is the maximal dilatation produced by \(\Phi\).

ii. For any \(\beta\), a bound on \(S(\beta)\) turns into a bound on \(\eta(\beta)\): fragmenting \(M\) in small atoms, so as to increase \(\eta(\beta)\), has a minimal cost in surface. The relation one finds is

\[
\eta(\beta) \leq \text{const} + m\text{Log} S(\beta) , \quad m = \dim M .
\]

This is true because in any Riemannian manifold an isoperimetric inequality holds, of the form

\[
\text{Vol}(A) \leq C S(A)^{\frac{m}{m-1}} ,
\]

with suitable \(C\) (depending on the minimal curvature on \(M\)).

Putting together, one easily finds \(\eta(\beta_t) \leq \text{const} + tm\text{Log} \lambda\), and thus, for any smooth \(\alpha\),

\[
h(\Phi, \alpha) \leq m\text{Log} \lambda . \quad (18)
\]

As a consequence, \(h(\Phi) \leq m\text{Log} \lambda\).  

\[\square\]
• Equation (18), concluding the proof, establishes a connection between entropy and dilatation of distances, here found as a bound from above to entropy. For smooth systems, there exists an exact formula, called Piezin’s formula, connecting entropy to the so-called Lyapunov characteristic exponents; see the next section. For the Arnol’d cat, (18) provides the inequality $h \leq 2 \log \lambda$, where $\lambda = \frac{1}{2}(3 + \sqrt{5})$ is the maximal eigenvalue of matrix $A$; the exact value is $h = \log \lambda$.

Inequality (18) implies that any rigid translation (examples 1 and 2 in Section 1) has zero entropy.

We close the section by just mentioning one more result; the proof can be found in [1] and [2].

**Definition 18** A partition $\alpha = (A_0, \ldots, A_{n-1})$ is said to be generating for a dynamical system $(M, \mu, \Phi)$, if the atoms of the partition, together with all their iterated $\Phi^{-t}(A_i)$ for any $t$ and $i$, generate the $\sigma$–algebra of the measurable sets of $M$.

This is the case, for example, of the partition $\alpha = \{C_l^0, \ l \in I\}$ for a Bernoulli shift.

**Proposition 25** If $\alpha$ is a generating partition, then $h(\Phi) = h(\Phi, \alpha)$.

So, in presence of a generating partition, the supremum in definition 14 is in fact a maximum. For a Bernoulli shift it is $h(\Phi) = -\sum_i p_i \log p_i$. This means that, for example, $B_{\frac{1}{2}, \frac{1}{2}}$ and $B_{\frac{1}{2}, \frac{1}{2}}$ cannot be isomorphic. Instead, $B_{\frac{1}{2}, \frac{1}{2}}$ and $B_{\frac{1}{2}, \frac{1}{2}}$ have the same entropy, so it is not excluded they are isomorphic. In fact, they are. (Ornstein, 1970: Bernoulli shifts are isomorphic iff they have the same entropy.)

4 **The Lyapunov Characteristic Exponents**

(To be written)
APPENDICES

A The roots of Ergodic Theory in Boltzmann and Gibbs

As already remarked in the Foreword, the roots of Ergodic Theory go back to the work of the founding fathers of Statistical Mechanics, in particular Boltzmann and Gibbs, who understood that the key idea to connect thermodynamics and microscopic dynamics, in systems with many degrees of freedom, is a statistical treatment, in which the volume of the phase space acquires the meaning of probability.

A.1 Macroscopic vs. microscopic description

Thermodynamics is a macroscopic experimental science, in which every quantity entering the game is conceptually defined with reference to an ideal experimental procedure. Take for simplicity a gas in a box of volume $V$; the thermodynamical state, at equilibrium, is completely characterized by specifying the density $\varrho = n/V$, $n$ being the amount of gas in moles, the pressure $p$ and the temperature $T$; such variables however are not independent but related by an “equation of state”, namely a relation (depending on the substance at hand) of the form $F(\varrho, p, T) = 0$. For ideal gases it is, as is well known,

$$pV = nRT,$$

i.e.

$$p/\varrho = RT,$$

$R$ being the so-called universal constant of gases. For real gases the equation of state is more complicated, but does exist; more complex systems (mixtures, coexistence of different phases...) require more variables to define the state, but conceptually the situation is not different. Non equilibrium situations can also be considered, in which $\varrho$, $p$ and $T$ are not uniform in the box and an equation of state $F(\varrho(x), p(x), T(x)) = 0$ is satisfied locally in any point $x$ of the box.

Besides such variables, thermodynamics introduces further important quantities depending on the state of the system, among them the internal energy $U$ and the entropy $S$, and describes what happens to such quantities during a thermodinamical transformation, namely a change of the thermodynamical state, the system possibly interacting with nearby systems (making work, absorbing heat). The statements include the fact that in any isolated system $U$ is conserved (1\textsuperscript{st} principle), while $S$ can only grow, and does grow in any irreversible (or spontaneous, or natural) process (2\textsuperscript{nd} principle). An easy example to have in mind is a gas initially confined in a corner of a box and let free to expand: it will expand, and so raise its entropy, while the opposite process is not natural and cannot take place. More generally, if $\varrho$, $p$ and $T$ are not initially uniform in an isolated system, they become, and correspondingly entropy increases.

On the other hand, one knows or believes\footnote{This was still a little conjectural throughout the 19\textsuperscript{th} century, before the determination of Avogadro’s number and so of the mass of atoms and molecules. Avogadro’s number is approximately $6 \times 10^{23}$; this is indeed the number of molecules in 2 grams of hydrogen or 18 grams of water.} that thermodynamic systems have a microscopic internal structure, namely are composed of a huge number of very small subsystems, which individually obey the laws of mechanics. The highly relevant and deep question then arises, whether it is
possible to deduce the macroscopic behavior of a system from the microscopic mechanical laws governing its elementary constituents. A quite ambitious goal: reducing thermodynamics to mechanics. Boltzmann was certainly the person who more strongly felt, and even suffered, the necessity of such a reduction.

Apparently the purpose is hopeless, while instead some paradoxes suggest that thermodynamics and mechanics are incompatible.

- **The Loschmidt paradox** (or reversibility paradox): mechanical processes are reversible, namely for each process the reversed one is possible as well, while instead, as remarked above, thermodynamic processes in general are irreversible and entropy definitely grows. How is it then possible the reduction?

- **The Zermelo paradox** (or recurrence paradox): conservative mechanical systems, like Hamiltonian systems, are recurrent, that is for most initial data, after a convenient time — possibly huge but finite — the system comes back near the initial state (see above the Poincaré recurrence theorem). Nothing similar, however, is expected in thermodynamics. How does recurrence disappear, passing from microscopic to macroscopic?

The question is indeed subtle, and a crucial role is played by the distinction, to be stressed, between a microscopic mechanical state (a point in a suitable phase space) and a macroscopic or thermodynamical state, defined with reference to experimental procedures, to be better understood in a theoretical frame. The ideas of Boltzmann and Gibbs can be regarded, in a sense, as attempts to give a meaning to the notion of macroscopic state, such that the paradoxes in principle solve and the reduction is at least conceptually possible.

It is of course impossible to enter here the not easy work and thought of Boltzmann and Gibbs, and we shall (drastically) limit ourselves to a few comments, in which the ideas of these authors are oversimplified aalso a little reinterpreted.

### A.2 Boltzmann’s equiprobability of microscopic states

Let us follow Boltzmann and consider a gas in a box composed of $N$ identical weakly interacting molecules, each having $l$ degrees of freedom, so that the complete system has $n = lN$ degrees of freedom; denote by $\gamma$ and $\Gamma$, respectively, the phase space of the single molecule and of the complete system, $\Gamma = \gamma^N$ (for a gas of point masses in a box $B$, it is $n = 3$; $\gamma = B \times \mathbb{R}^3$). Let

$$ (p^{(i)}, q^{(i)}) = (p_1^{(i)}, \ldots, p_l^{(i)}, q_1^{(i)}, \ldots, q_l^{(i)}) \in \gamma $$

denote the canonical coordinates of the $i$-th molecule. The state of the whole system is then described by $N$ (ordered) points in $\gamma$ or equivalently by a single point

$$ x = (p, q) = (p_1, \ldots, p_n, q_1, \ldots, q_n) \in \Gamma ; $$

the evolution of the system appears then equivalently as a single movement in $\Gamma$, or an evolving cloud of $N$ points in $\gamma$. The evolution is supposed to be governed by a Hamiltonian of the form

$$ H(p, q) = \sum_{i=1}^{N} h(p^{(i)}, q^{(i)}) + V(q) , $$

---

8The traditional notation for the phase of a single molecule is $\mu$; we are using $\gamma$ to avoid the conflict with the measure.
\( h \) being the individual Hamiltonian of a molecule, while \( V \) is some interaction potential; \( V \) is supposed to be small and not much relevant in the energy balance (almost free molecules).

Boltzmann’s thought then develops more or less as follows:

i. A macroscopic state of the system is identified with a distribution of the cloud of the \( N \) points in \( \gamma \). In a rough but deep analysis, Boltzmann imagines \( \gamma \) is partitioned into very small cells of identical volume \( \omega \) — so small that moving molecules inside their cell is not appreciable physically — and identifies a macroscopic state with the “occupation numbers” \( N_1, N_2, \ldots \) of the cells, \( \sum_j N_j = N \). Each cell has an energy \( \varepsilon_i \) (that of any chosen point in it) and so, denoting by \( E \) the total energy, it is

\[
E \simeq \sum_j N_j \varepsilon_j .
\]

The macroscopic state does not change by moving molecules inside a cell, or exchanging molecules among different cells. In place of the numbers \( N_j \) one could use the densities\(^9\)

\[
f_j = \frac{N_j}{N \omega} , \quad \sum_j f_j \omega = 1 .
\]

ii. A macroscopic state (given occupation numbers) occupies a volume \( W(N_1, N_2, \ldots) \) in \( \Gamma \), actually in a layer \( \Gamma_{E \pm \Delta E} \) around the constant energy surface \( \Sigma_E \), with \( \Delta E \) related to the size of the cells in \( \gamma \). One immediately gets

\[
W(N_1, N_2, \ldots) = \frac{N!}{N_1! N_2! \ldots} \omega^N , \quad \text{(A.1)}
\]

where \( \omega^N \) accounts for the displacements of the \( N \) molecules inside their cell, while the combinatorial coefficient counts the exchanges among different cells.

The volume \( W \) of a state (up to a normalization) is given the strong meaning of the a priori probability that the state is realized. This is often referred to as the “a priori equiprobability of the microscopic states (points) in \( \Gamma^\prime \)”: it is not important where points are, only their overall volume is relevant.

A standard computations then shows that

ii. The maximum \( W^* \) of \( W \), for fixed \( N \) and \( E \), is found for\(^10\)

\[
N_j^* = C' N e^{-\beta \varepsilon_j} , \quad C'^{-1} = \sum_j e^{-\beta \varepsilon_j} ,
\]

where \( \beta \) is a Lagrange multiplier determined by the energy per molecule \( E/N \); equivalently for \( f_j = f_j^* \), with

\[
f_j^* = \frac{C}{\omega} e^{-\beta \varepsilon_j} . \quad \text{(A.2)}
\]

---

\(^9\)The underlying idea, difficult unfortunately to formalize in well defined limit procedure, is to make a finer and finer partition of \( \gamma \) in smaller and smaller cells, letting correspondingly \( N \) go to infinity. Boltzmann in fact did not care too much of the continuum limit. Also the microscopic dynamics in \( \Gamma \) is replaced, if useful, by a discretized dynamics: \( \Gamma \) itself is discretized (decomposed into cells), and a finite time step is introduced, more or less as we do today in computer simulations. The fundamental intuition is that the details of the dynamics should be irrelevant, provided a few essential features, like the conservation of the volume in phase space and the conservation of energy, are saved.

\(^10\)Stirling approximation for the factorials, the \( N_j \)'s being treated as real numbers; computing the constrained maximum of \( \log W \) is straightforward.
Such a state is commonly called the Maxwell-Boltzmann state; the exponential is called the Boltzmann exponential factor.

Moreover: small changes $\delta N_j$ of the occupation numbers drastically reduce $W$, namely\(^{11}\)

$$ W(N_1^* + \delta N_1, N_2^* + \delta N_2, \ldots) \simeq W^* \prod_j e^{-\frac{1}{2} \frac{(\delta N_j)^2}{N_j^*}}; $$

for example, for minor changes of the occupation numbers, say

$$ \delta N_j = \sqrt{N_j^*} \ll N_j, $$

the volume $W$ reduces by a huge factor, actually the exponential of the number of the occupied cells; for $\delta N_j$ proportional to $N$ (assigned fluctuations $\delta f_j$), $W$ reduces as much as the exponential of $N$.

Finally, and this is the most important point for us, the interpretation of the volume as probability is supported by a fundamental dynamical assumption, known as the Boltzmann ergodic hypothesis:\(^{12}\)

iii. A typical trajectory, observed over a long time interval, wanders erratically in $\Gamma_{E\pm \Delta E}$, spending asymptotically in any subset of volume $W$ a time proportional to $W$. In this sense, the chance a macroscopic state is dynamically realized, in a long time while, is proportional to $W$, in agreement with the idea of equiprobability of microscopic states.

The overall picture is summarized in figure 9, representing symbolically the layer $\Gamma_{E\pm \Delta E}$. The big set, of overwhelming volume, is the Maxwell-Boltzmann state, together with the states similar to it within $\delta N_j = \sigma_j$. Small sets correspond instead to states well distinguishable from it. Practically,

\(^{11}\)For this, it is enough to compute the second derivatives of $\log W$ at $N_j^*$ (the first derivatives obviously vanish).

no matter how the initial state is chosen, the trajectory is expected to reach soon the Maxwell-Boltzmann state, and spend there the overwhelming majority of time, up to very short stays, unlike but possible, again in small sets. The transient, i.e. the time needed to enter the Maxwell-Boltzmann state, represents, in this view, the approach to equilibrium, starting from an exceptional state far from it. Properly speaking, however, the equilibrium is not the Maxwell-Boltzmann state, but the collection of all macroscopic states, each having probability to be realized proportional to its volume. Equilibrium is the equiprobability of microscopic states in $\Gamma$, or equivalently the distribution (A.2) in $\gamma$. Equilibrium includes fluctuations.

• Among the fluctuations, there is the one that reports a system close to the initial datum in $\Gamma$: extremely unlike, and so not expected in any physically conceivable time interval, but dynamically not excluded; this solves, in principle, the recurrence paradox. Concerning the reversibility paradox, the answer is similar: reversed trajectories are included in the equilibrium state, but like any other trajectory, they spend the overwhelming majority of time in the Maxwell-Boltzmann state, and their transit in the extremely small set corresponding to the chosen non–uniform initial conditions, is too unlike to be observed.

Since Boltzmann, the equiprobability of microscopic states is the very basis of statistical mechanics. Boltzmann himself could deduce from it beautiful results. He was indeed able to provide a mechanical interpretation of the fundamental thermodynamical variables $p$, $T$, $U$, $S$, thus constructing a model of thermodynamics, in which the second principle of thermodynamics, in the form $dU + pdV \leq TdS$, is satisfied. Let us recall that $T$ is connected to the multiplier $\beta$ via

$$\beta = \frac{1}{k_B T},$$

where $k_B > 0$ is the Boltzmann constant. In elementary models, $\frac{1}{2}k_B T$ is also the average kinetic energy per degree of freedom. Concerning $S$, it turns out to be defined, microscopically, by $S = k_B \log W$; so, up to an inessential additive constant, it is

$$S = -k_B N \omega \sum_j f_j \log f_j.$$

A.3 Gibbs’ statistical ensembles

Let us shortly describe an alternative view, commonly referred to as Gibbs’ view. Gibbs’ view is somehow more abstract, and probability plays a more primitive role in it. The space $\gamma$ of single molecules does not play a role, and attention is addressed only to $\Gamma$.

The basic idea is that a macroscopic state is any given probability distribution $\rho$ in $\Gamma$. While Boltzmann focuses the attention on a single trajectory, which assigns the probability to a set $W$ of states through the fraction of time spent in it in a long time interval, Gibbs instead imagines to deal, at any time, with a family, or ensemble of evolving points, independent mental replicas of the same system, distributed in $\Gamma$ with some density $\rho_t$ evolving in time. One should think that in each experiment the way the system is prepared does not correspond to a single mechanical initial datum, rather to a spot of initial data, more precisely to a convenient initial distribution $\rho_0$ in $\Gamma$ (the initially prepared macroscopic state). Each replica of the system then evolves independently.

---

13 It is $k_B = 1.380649 \times 10^{23}$ Joules/Kelvin degree: an exact value fixed conventionally, nowadays used in the “International System of units” as a primitive value; the Kelvin degree, not anymore primitive, remains consequently defined.
according to Hamilton equations, and correspondingly $\rho_t$ evolves in time, as in a fluid of non-interacting particles. From the conservation of the volume of the phase space one immediately deduces the evolution law
denotes energy, and $\rho(\Phi^{-t}(x))$ is just the state of $x$ at time $t$, namely $\rho_t(x)$. Thus,$\rho_t(x)$ evolves in time, as in a fluid of non-interacting particles. From the conservation of the volume of the phase space one immediately deduces the evolution law
\begin{equation}
\rho_t(x) = \rho_0(\Phi^{-t}(x)) , \quad x \in \Gamma
\end{equation}
($\rho_t$ is constant along trajectories).

It is now natural to search for equilibrium states, namely states such that $\rho_t(x)$ at any $x$ is independent of $t$. An easy example is
\[\rho^\ast(x) = \begin{cases}
  c & \text{in } \Gamma_{E\pm\Delta E} \\
  0 & \text{elsewhere}
\end{cases},
\]
with $c$ such as to ensure normalization; so, Boltzmann’s equiprobability of microscopic states is, in Gibbs’ view, an equilibrium state. It is obviously not unique: for the conservation of energy, any $\rho(x) = F(H(x))$, with any $F : \mathbb{R} \to \mathbb{R}$ up to the normalization, is an equilibrium state.

Let us then focus the attention on a single constant energy surface $\Sigma_E$, passing from volume densities to surface densities without changing the notation; a macroscopic state is then a surface probability distribution $\rho_t$, evolving in time, such that the probability of any subset $A \in \Sigma_E$ at time $t$ is
\[\int_A \rho_t(x) \, d\mu_L , \]
$\mu_L$ being the Liouville measure introduced in the previous section; $\mu_L$ being preserved, $\rho_t$ evolves in time according to (A.3). Quite clearly,
\[\rho^\ast(x) = 1 \quad \forall x \in \Sigma_E \]
is an equilibrium state (this is indeed Boltzmann’s equilibrium, rewritten in a more transparent mathematical notation). Two nontrivial question now are well posed:

i. whether $\rho^\ast$ is unique, or there is a multiplicity of equilibria;

ii. whether, in addition, the system, prepared in a non-equilibrium state $\rho_0$, does reach asymptotically equilibrium:
\[\lim_{t \to \infty} \rho_t = \rho^\ast . \]

Due to (A.3), the limit cannot be pointwise but should be understood in the weak sense
\[\int_{\Sigma_E} f(x) \rho_t(x) \, d\mu_L \to \int_{\Sigma_E} f(x) \, d\mu_L . \]

In the next section we shall see how the ideas of Boltzmann and Gibbs have been formalized in Ergodic Theory, through the basic notions of ergodicity and mixing.

---

\[\text{The system is in } W \text{ at time } t \text{ iff it is in } \Phi^{-t}(W) \text{ at } t = 0, \quad \int_W \rho_t(x) \, dV = \int_{\Phi^{-t}(W)} \rho_0(x) \, dV . \]

Introduce now at the r.h.s. the change of variable $x = \Phi^{-t}(x')$, which reports the integration volume to $W$; for the conservation of volume in Hamiltonian dynamics the Jacobian is 1, and so, dropping the prime, $\int_W \rho_t(x) \, dV = \int_W \rho_0(\Phi^{-t}(x)) \, dV$. The conclusion follows from the arbitrariness of $W$.\]
B Proof of the Birkhoff ergodic theorem

The proof of Birkhoff’s ergodic theorem is based on a technical Lemma, known as maximal ergodic theorem:

**Lemma 26 (Maximal ergodic Theorem)** For \( f \in L_1(M, \mu) \), let

\[
F_t(x) = f(x) + \cdots + f(\Phi^{t-1}(x)), \quad A = \{x \in M : \sup_{t \geq 1} F_t \geq 0\}.
\]

Then \( A \) is measurable and

\[
\int_A f \, d\mu \geq 0.
\]

**Proof.** Let

\[
\mathcal{F}_t(x) = \max_{1 \leq s \leq t} F_s(x), \quad A_t = \{x \in M : \mathcal{F}_t \geq 0\},
\]

so that

\[
A_t \subset A_{t+1}, \quad A = \bigcup_{0 \leq t < \infty} A_t.
\]

It is then

\[
\int_A f \, d\mu = \lim_{t \to \infty} \int_{A_t} f \, d\mu
\]

and to prove the lemma it is enough to show that for any \( t \geq 0 \) it is

\[
\int_{A_t} f \, d\mu \geq 0.
\]

To this purpose, observe that

\[
\mathcal{F}_t(x) = \max \{f(x), \ldots, f(x) + \cdots + f(\Phi^{t-1}(x))\} = f(x) + \max(0, \mathcal{F}_{t-1}(\Phi(x))),(\text{i.e.})
\]

\[
f(x) = \mathcal{F}_t(x) - \mathcal{F}^+_{t-1}(\Phi(x)),
\]

having denoted \( \mathcal{F}^+_{t}(x) = \max(0, \mathcal{F}_{t}(x)) \). It follows

\[
f(x) \geq \mathcal{F}_t(x) - \mathcal{F}^+_{t}(\Phi(x))
\]

and consequently

\[
\int_{A_t} f \, d\mu \geq \int_{A_t} \mathcal{F}_t \, d\mu - \int_{A_t} \mathcal{F}^+_{t} \circ \Phi \, d\mu
\]

\[
\geq \int_M \mathcal{F}^+_{t} \, d\mu - \int_M \mathcal{F}^+_{t} \circ \Phi \, d\mu = 0
\]

(for the latter inequality: \( \mathcal{F}^+_{t} = \mathcal{F}_t \) in \( A_t \), \( \mathcal{F}^+_{t} = 0 \) in the complement).

We can now prove Birkhoff’s theorem.

**Proof.** (a) **Existence of the limit a.e.** For \( a, b \in \mathbb{R} \), \( a < b \), let

\[
E_{a,b} = \left\{ x \in M : \liminf_{t \to \infty} \frac{1}{t} F_t(x) < a < b < \limsup_{t \to \infty} \frac{1}{t} F_t(x) \right\};
\]

\[
a - 7
\]
$E_{a,b}$ is measurable (lim inf and lim sup of sequences of measurable functions are measurable) and invariant. The set for which the time average of $f$ does not exist is then

$$E = \bigcup_{a,b} E_{a,b},$$

but since a denumerable union is sufficient (sets are conveniently nested), it is enough to show that any of the $E_{a,b}$ has zero measure. To this purpose, let us show that, thanks to the lemma, the “reversed” inequality

$$b \mu(E_{a,b}) \leq \int_{E_{a,b}} f \, d\mu \leq a \mu(E_{a,b})$$

holds, which implies $\mu(E_{a,b}) = 0$. To prove, for example, the left inequality, apply the lemma to

$$g(x) = \begin{cases} f(x) - b & \text{per } x \in E_{a,b} \\ -1 & \text{per } x \notin E_{a,b} \end{cases};$$

it is easy to see that the set $A$ appearing in the statement of the lemma, namely the set such that $G_t(x) = \sum_{s=0}^{t-1} g(\Phi^s(x))$ is non negative for at least one $t > 0$, is precisely $E_{a,b}$; indeed, if $x \in E_{a,b}$, then (definition of $E_{a,b}$) there exists $t > 0$ such that $F_t(x)/t > b$, and so $G(x) > 0$; conversely, if $x \notin E_{a,b}$, then $\Phi^t(x) \notin E_{a,b}$ for any $t > 0$ ($E_{a,b}$ is invariant), and correspondingly $G_t(x) < 0$. The lemma than says that

$$\int_{E_{a,b}} g \, d\mu = \int_{E_{a,b}} f \, d\mu - b \mu(E_{a,b}) \geq 0,$$

and the left inequality is satisfied. Similarly, using

$$g(x) = \begin{cases} a - f(x) & \text{per } x \in E_{a,b} \\ -1 & \text{per } x \notin E_{a,b} \end{cases},$$

the right inequality can be proved.

(b) Proof that $\bar{f}(\Phi^t(x)) = \bar{f}(x)$. This is a trivial consequence of the definition of $\bar{f}(x)$, whenever the limit exists.

(c) Proof that $\langle \bar{f} \rangle = \langle f \rangle$. First of all, let us observe that $\bar{f} \in L^1(M, \mu)$, as follows from

$$\int_M \left| \frac{1}{t} \sum_{s=0}^{t-1} f(\Phi^s(x)) \right| \, d\mu(x) \leq \frac{1}{t} \sum_{s=0}^{t-1} \int_M |f(\Phi^s(x))| \, d\mu(x) = \int_M |f| \, d\mu.$$ 

Let now

$$C_{a,b} = (\bar{f})^{-1}[a,b) = \{ x \in M : a \leq \bar{f}(x) < b \};$$

it is then

$$a \mu(C_{a,b}) \leq \int_{C_{a,b}} \bar{f} \, d\mu \leq b \mu(C_{a,b}),$$

while from the lemma, proceeding as above with a convenient $g$, one deduces

$$a \mu(C_{a,b}) \leq \int_{C_{a,b}} f \, d\mu \leq b \mu(C_{a,b});$$

a – 8
it follows
\[ \left| \int_{C_{a,b}} \tilde{f} \, d\mu - \int_{C_{a,b}} f \, d\mu \right| \leq (b - a) \mu(C_{a,b}). \]
From this inequality, thanks to the arbitrariness of \( a \) and \( b \), it is not difficult to conclude that
\[ \int_M \tilde{f} \, d\mu - \int_M f \, d\mu = 0. \]
Indeed, given \( \varepsilon > 0 \), divide \( \mathbb{R} \) in intervals \([k\varepsilon, (k+1)\varepsilon), k \in \mathbb{Z}\); it is clearly \( \bigcup_k C_{k\varepsilon, (k+1)\varepsilon} = M \), and so
\[ \left| \int_M \tilde{f} \, d\mu - \int_M f \, d\mu \right| \leq \sum_{k \in \mathbb{Z}} \left| \int_{C_{k\varepsilon, (k+1)\varepsilon}} \tilde{f} \, d\mu - \int_{C_{k\varepsilon, (k+1)\varepsilon}} f \, d\mu \right| \leq \varepsilon \sum_{k \in \mathbb{Z}} \mu(C_{k\varepsilon, (k+1)\varepsilon}) = \varepsilon, \]
and this is enough.

(d) \textit{Proof that in the invertible case} \( \tilde{f}_- \) \textit{exists and coincides a.e. with} \( \tilde{f} \). The existence is obvious (just replace \( \Phi \) with \( \Phi^{-1} \)). Suppose now it is, for example, \( \tilde{f}_- > \tilde{f} \) in a set of positive measure. Denoting
\[ A := \{ x \in M : \tilde{f} - \tilde{f}_- > 0 \}, \]
it is
\[ \int_A (\tilde{f} - \tilde{f}_-) \, d\mu > 0. \]
But since (point b) both \( \tilde{f} \) and \( \tilde{f}_- \) are constants of motion, \( A \) is invariant: so, denoting \( g(x) = \chi_A(x)f(x) \), it is \( \tilde{g}(x) = \tilde{f}(x) \) for \( x \in A \), \( \tilde{g}(x) = 0 \) elsewere, and similarly for the backwards average. It follows
\[ \langle \tilde{g} \rangle - \langle \tilde{g}_- \rangle = \int_A (\tilde{f} - \tilde{f}_-) \, d\mu > 0. \]
But this is a contradiction because, according to point (c), \( \langle \tilde{g} \rangle = \langle \tilde{g}_- \rangle = \langle g \rangle \). This concludes point (d) and the proof of the theorem.

\( \square \)