## III. Recurrent, Ergodic and Minimal Dynamical Systems

"Ergodic theory is the study of transformations from the point of view of recurrence properties" (Walters [1975], p. 1). Sometimes, you meet such properties in daily life: If you walk in a park just after it has snowed, you will have to step into your own footprints after a finite number of steps. The more difficult problem of the reappearance of certain celestial phenomena led Poincaré to the first important result of ergodic theory at the end of the last century.

**III.1 Definition:**
Let $(X, \Sigma, \mu; \varphi)$ be an MDS and take $A \in \Sigma$. A point $x \in A$ is called recurrent to $A$ if there exists $n \in \mathbb{N}$ such that $\varphi^n(x) \in A$.

**III.2 Theorem (Poincaré, 1890):**
Let $(X, \Sigma, \mu; \varphi)$ be an MDS and take $A \in \Sigma$. Almost every point of $A$ is (infinitely often) recurrent to $A$.

*Proof.* For $A \in \Sigma$, $\varphi^{-n}A$ is the set of all points that will be in $A$ at time $n$ (i.e. $\varphi^n(x) \in A$). Therefore, $A_{\text{rec}} := A \cap (\varphi^{-1}A \cup \varphi^{-2}A \cup \dots)$ is the set of all points of $A$ which are recurrent to $A$.

If $B := A \cup \varphi^{-1}A \cup \varphi^{-2}A \cup \dots$ we obtain $\varphi^{-1}B \subseteq B$ and $A \backslash A_{\text{rec}} = B \backslash \varphi^{-1}B$. Since $\varphi$ is measure-preserving and $\mu$ finite, we conclude

$$\mu(A \backslash A_{\text{rec}}) = \mu(B) - \mu(\varphi^{-1}B) = 0,$$

and thus the non-recurrent points of $A$ form a null set. For the statement in brackets, we notice that $(X, \Sigma, \mu; \varphi^k)$ is an MDS for every $k \in \mathbb{N}$. The above results implies

$$\mu(A_k) = 0 \quad \text{for} \quad A_k := \big\{x \in A : (\varphi^k)^n(x) \notin A \text{ for } n \in \mathbb{N}\big\}.$$

Hence, $A_\infty := \bigcup_{k=1}^{\infty} A_k$ is a null set, and the points of $A \backslash A_\infty$ are infinitely often recurrent to $A$. ∎

We explained in the physicist's answer in Lecture I that the dynamics can be described by the MDS $(X, \Sigma, \mu; \varphi)$ on the state space

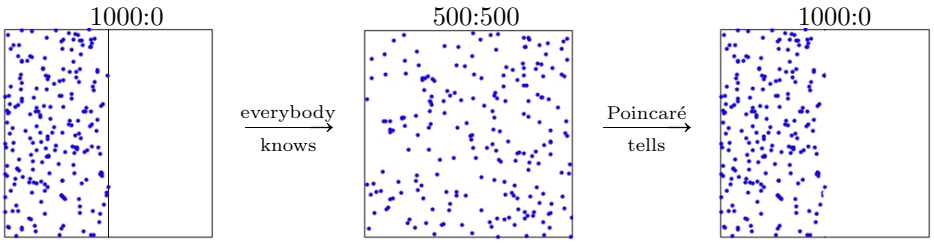
$$X := \{\text{coordinates of the possible locations and impulses of the}$$
$$\text{1000 molecules in the box}\}$$

As the set $A$ to which recurrence is expected we choose

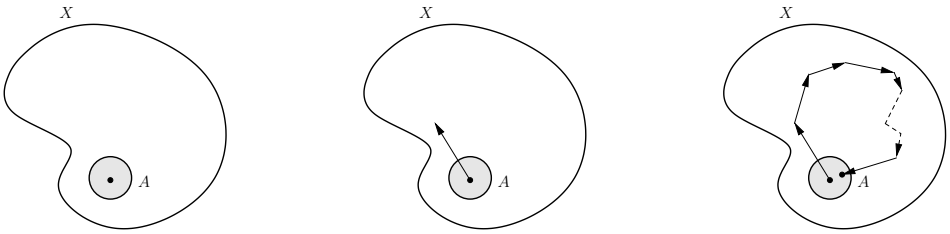
$$A := \{\text{all 1000 molecules are located on the left hand side}\}.$$

Since $\mu(A) > 0$, we obtain from Poincaré's recurrence theorem a surprising conclusion contradicting somehow our daily life experience.

**gas container**




**state space**




"Ergodic theory is the study of transformations from the point of view of *mixing properties*" (Walters [1975] p. 1), where "mixing" can even be understood literally (see Lecture IX). In a sense, ergodicity and minimality are the weakest possible "mixing properties" of dynamical systems. Another, purely mathematical motivation for the concepts to be introduced below is the aim of defining (and then classifying) the "indecomposable" objects, e.g. simple groups, factor von Neumann algebras, irreducible polynomials, prime numbers, etc.. From these points of view the following basic properties (III.3) and (III.6) appear quite naturally.

**III.3 Definition:**
An MDS $(X, \Sigma, \mu; \varphi)$ is called *ergodic* if there are no non-trivial $\varphi$-invariant sets $A \in \Sigma$, i.e. $\varphi(A) = A$ implies $\mu(A) = 0$ or $\mu(A) = 1$.

It is obvious that an MDS which is not ergodic is "reducible" in the sense that it can be decomposed into the "sum" of two MDSs. Therefore the name "irreducible" instead of "ergodic" would be more intuitive and more systematic. Still, the use of the word "ergodic" may be justified by the fact that ergodicity in the above sense implies the validity of the classical "ergodic hypothesis": time mean equal space mean (see III.D.6), and therefore gave rise to "ergodic theory" as a mathematical theory. Our first proposition contains a very useful criterion for ergodicity and shows for the first time the announced duality between properties of the transformation $\varphi : X \to X$ and the induced operator $T_\varphi : L^p(\mu) \to L^p(\mu)$.

**III.4 Proposition:**
For an MDS $(X, \Sigma, \mu; \varphi)$ the following statements are equivalent:

(a) $(X, \Sigma, \mu; \varphi)$ is ergodic.
(b) The fixed space $F := \{f \in L^p(X, \Sigma, \mu) : T_\varphi f = f\}$ of $T_\varphi$ is one-dimensional, or: 1 is a simple eigenvalue of $T_\varphi \in L^p(\mu)$ for $1 \leqslant p \leqslant \infty$.

*Proof.* We observe, first, that the constant functions are always contained in $F$, hence 1 is an eigenvalue of $T_\varphi$. Moreover, we shall see that the proof does not depend on the choice of $p$.

(b) $\Rightarrow$ (a): If $A \in \Sigma$ is $\varphi$-invariant, then $\mathbf{1}_A \in F$ and $\dim F \geqslant 2$.

(a) $\Rightarrow$ (b): For any $f \in F$ and any $c \in \mathbb{R}$ the set

$$[f > c] := \{x \in X : f(x) > c\}$$

is $\varphi$ invariant, and hence trivial. Let $c_0 := \sup\{c \in \mathbb{R} : \mu[f > c] = 1\}$. Then for $c < c_0$ we have $\mu[f \leqslant c] = 0$, and therefore $\mu[f < c_0] = 0$. For $c > c_0$ we have $\mu[f > c] \neq 1$, hence $\mu[f > c] = 0$, and therefore $\mu[f > c_0] = 0$, too. This implies $f = c_0$ a.e.. ∎

**III.5 Examples:**

(i) The *rotation* $(\Gamma, \mathcal{B}, m; \varphi_a)$ is ergodic, iff $a \in \Gamma$ is not a root of unity: If $a^n = 1$ for some $n \in \mathbb{N}$, then $\mathbf{1}$ and $f : z \to z^n$ are in $\Gamma$, and so $\varphi_a$ is not ergodic. On the other hand, if $a^n \neq 1$ for all $n \in \mathbb{N}$, assume $T_{\varphi_a} f = f$ for some $f \in L^2(m)$. Since the functions $f_n$, $n \in \mathbb{Z}$, with $f(z) = z^n$ form an orthonormal basis in $L^2(\mu)$ we obtain

$$f = \sum_{n=-\infty}^{\infty} b_n f_n \quad \text{and} \quad T_{\varphi_a} f = \sum_{n=-\infty}^{\infty} b_n T_{\varphi_a} f_n = \sum_{n=-\infty}^{\infty} b_n a^n f_n.$$

The comparison of the coefficients yields $b_n(a^n - 1) = 0$ for all $n \in \mathbb{Z}$, hence $b_n = 0$ for all $n \in \mathbb{N}$, i.e. $f$ is constant.

(ii) The Bernoulli shift $B(p_0, \ldots, p_{k-1})$ is ergodic: Let $A \in \hat{\Sigma}$ be $\tau$-invariant with $0 < \hat{\mu}(A)$ and let $\varepsilon > 0$. By definition of the product $\sigma$-algebra, there exists $B \in \hat{\Sigma}$ depending only on a finite number of coordinates such that $\hat{\mu}(A \triangle B) < \varepsilon$, and therefore $|\hat{\mu}(A) - \hat{\mu}(B)| < \varepsilon$. Choose $n \in \mathbb{N}$ large enough such that $C := \tau^n B$ depends on different coordinates than $B$. Since $\mu$ is the product measure, we obtain $\hat{\mu}(B \cap C) = \hat{\mu}(B) \cdot \hat{\mu}(C) = \hat{\mu}(B)^2$, and $\tau(A) = A$ gives $\hat{\mu}(A \triangle B) = \hat{\mu}(\tau^n(A \triangle B)) = \hat{\mu}(A \triangle C)$. We have $A \triangle (B \cap C) \subseteq (A \triangle B) \cup (A \triangle C)$ and therefore $\hat{\mu}(A \triangle (B \cap C)) < 2\varepsilon$. This implies

$$\begin{aligned} |\hat{\mu}(A) - \hat{\mu}(A)^2| &\leqslant |\hat{\mu}(A) - \hat{\mu}(B \cap C)| + |\hat{\mu}(B \cap C) - \hat{\mu}(A)^2| \\ &\leqslant \hat{\mu}(A \triangle (B \cap C)) + |\hat{\mu}(B)^2 - \hat{\mu}(A)^2| \\ &= \hat{\mu}(A \triangle (B \cap C)) + |\hat{\mu}(B) - \hat{\mu}(A)| \cdot |\hat{\mu}(B) + \hat{\mu}(A)| \\ &\leqslant 4\varepsilon, \quad \text{which proves } \hat{\mu}(A) = \hat{\mu}(A)^2 = 1. \end{aligned}$$

In the last third of this lecture we introduce the concept of "irreducible" TDSs. Formally, this will be done in complete analogy to III.3, but due to the fact that in general the complement of a closed $\varphi$-invariant set is not closed, the result will be quite different.

**III.6 Definition:**
A TDS $(X; \varphi)$ is called *minimal*, if there are no non-trivial $\varphi$-invariant closed sets $A \subseteq X$, i.e. $\varphi(A) = A$, $A$ closed, implies $A = \varnothing$ or $A = X$.

Again, "irreducible" seems to be the more adequate term (see III.D.11) but "minimal" is the term used by the topological dynamics specialists. It is motivated by property (ii) in the following proposition.

**III.7 Proposition:**

(i) If $(X;\varphi)$ is minimal, then the fixed space $F := \{f \in C(X) : T_\varphi f = f\}$ is one-dimensional.

(ii) If $(X;\varphi)$ is a TDS, then there exists a non-empty $\varphi$-invariant, closed subset $Y$ of $X$ such that $(Y;\varphi)$ is minimal.

*Proof.* We observe that the orbit $\{\varphi^n(x) : n \in \mathbb{Z}\}$ of any point $x \in X$ and also its closure are $\varphi$-invariant sets. Therefore, $(X;\varphi)$ is minimal iff the orbit of every point $x \in X$ is dense in $X$.

(i) For $f \in F$ we obtain $f(x) = f(\varphi^n(x))$ for all $x \in X$ and $n \in \mathbb{Z}$. If $(X;\varphi)$ is minimal, the continuity of $f$ implies $f =$ constant.

(ii) The proof of this assertion is a nice, but standard application of Zorn's lemma and the finite intersection property of compact spaces. ∎

**III.8 Examples:**

(i) Take $X = [0,1]$ and $\varphi(x) = x^2$. Then $(X;\varphi)$ is not minimal (since $\varphi(0) = 0$) but $\dim F = 1$

(ii) A property analogous to (III.7.ii) is not valid for MDSs: in $([0,1],\mathcal{B},m;\mathrm{id})$ there exists no "minimal" invariant subset with positive measure.

(iii) The rotation $(\Gamma;\varphi_a)$ is minimal iff $\in \Gamma$ is not a root of unity: If $a^{n_0} = 1$ for some $n_0 \in \mathbb{N}$, then $\{z \in \Gamma : z^{n_0} = 1\}$ is closed and $\varphi$-invariant. For the other implication, we show that the orbit of every point in $\Gamma$ is dense. To do this we need only prove that $\{1, a, a^2, \dots\}$ is dense in $\Gamma$. Choose $\varepsilon > 0$. Since by assumption $a^{n_1} \neq a^{n_2}$ for $n_1 \neq n_2$, there exist $l < k \in \mathbb{N}$ such that $0 < |a^l - a^k| < \varepsilon$. $0 < |a^l - a^k| = |1 - a^{k-l}| = |a^{(k-l)n} - a^{(k-l)(n+1)}| < \varepsilon$ for all $n \in \mathbb{N}$. Since the set of "segments" $\{(a^{(k-l)n}, a^{(k-l)(n+1)}) : n \in \mathbb{N}\}$ covers $\Gamma$, we proved that there is at least one power of $a$ in every $\varepsilon$-segment of $\Gamma$.

(iv) The shift $\tau$ on $\{0,1,\dots,k-1\}$ is not minimal, since $\tau(x) = x$ for $x = (\dots,0,0,0,\dots)$.

We state once more that ergodicity and minimality are the most fundamental properties of our measure-theoretical or topological dynamical systems. On the other hand they gave us the first opportunity to demonstrate how dynamical properties of a map $\varphi : X \to X$ are reflected by (spectral) properties of the induced linear operator $T_\varphi$ (see III.4 and III.7.i). In particular, it can be expected that the set $P\sigma(T_\varphi)$ of all eigenvalues of $T_\varphi$ has great significance in ergodic theory (see Lectures VIII and IX). Here we show only the effect of ergodicity or minimality on the structure of the point spectrum $P\sigma(T_\varphi)$.

**III.9 Proposition:**
Let $(X;\varphi)$ be a minimal TDS (resp. $(X,\Sigma,\mu;\varphi)$ an ergodic MDS). Then the point spectrum $P\sigma(T_\varphi)$ of the induced operator $T_\varphi$ on $C(X)$ (resp. $L^p(X,\Sigma,\mu)$) is a subgroup of $\Gamma$, and each eigenvalue is simple.

*Proof.* Since $T_\varphi$ is a bijective isometry the spectrum of $T_\varphi$ is contained in $\Gamma$. Let $T_\varphi f = \lambda f$, $\|f\| = 1 = |\lambda|$. Since $T_\varphi$ is a lattice homomorphism we conclude

$$T_\varphi|f| = |T_\varphi f| = |\lambda f| = |\lambda| \cdot |f| = |f|,$$

and hence $|f| = \mathbf{1}$ by (III.7.i), resp. (III.4), i.e. every normalized eigenfunction is unimodular and the product of two such eigenfunctions is non-zero. Since $T_\varphi$ is also an algebra homomorphism (on $L^\infty(X)$, resp. $C(X)$) we conclude from $T_\varphi f = \lambda_1 f \neq 0$ and $T_\varphi g = \lambda_2 g \neq 0$ that

$$T_\varphi(f \cdot g^{-1}) = T_\varphi f \cdot T_\varphi g^{-1} = \lambda_1 \cdot \lambda_2^{-1}(f \cdot g^{-1}) \neq 0$$

which shows that $P\sigma(T_\varphi)$ is a subgroup of $\Gamma$. If $\lambda_1 = \lambda_2$, it follows $T_\varphi(f \cdot g^{-1}) = f \cdot g^{-1}$ and, again by the one-dimensionality of the fixed space, $f \cdot g^{-1} = c \cdot \mathbf{1}$ or $f = c \cdot g$, i.e. each eigenvalue is simple. ∎

## III.D Discussion

**III.D.1. The "original" Poincaré theorem:**
Henri Poincaré ([1890], p. 69) formulated what later on was called the recurrence theorem:

> "Théorème I. Supposons que le point $P$ reste à distance finie, et que le volume $\int \mathrm{d}x_1\, \mathrm{d}x_2\, \mathrm{d}x_3$ soit un invariant intégral; si l'on considère une région $r_0$ quelconque, quelqe petite que soit cette région, il y aura des trajectoires qui la traverseront une infinité de fois."

In the corollary to this theorem he mentioned some kind of probability distribution for the trajectories:

> "Corollaire. Il résulte de ce qui précède qu'il existe une infinité de trajectoires qui traversent une infinité de fois la région $r_0$; mais il peut en exister d'autres qui ne traversent cette région qu'un nombre fini de fois. Je me propose maintenant d'expliquer pourquoi oes dernières trajectoires peuvent être regardées oomme exceptionnelles."

**III.D.2. Recurrence and the second law of thermodynamics:**
As we explained in Lecture I the time evolution of physical "states" is adequately described in the language of MDS and therefore "states" are "recurrent". This (and the picture following (III.2)) seems to be in contradiction with the second law of thermodynamics which says that entropy can only increase, if it changes at all, and thus we can never come back to a state of entropy $h$, once we have reached a state of entropy higher than $h$. One explanation lies in the fact that the second law is an empirical law concerning a quantity, called entropy, that can only be determined through measurements that require *time averaging* (in the range from milliseconds to seconds). In mathematical models of "micro"-dynamics, which were the starting point of ergodic theory, such time averages should be roughly *constant* (and equal to the space mean by the ergodic hypothesis). Therefore entropy should be constant for dynamical systems (like the constant defined in Lecture XII, although at least to us it is unclear whether the two numbers, the Kolmogoroff-Sinai entropy and the physical entropy can be identified or compared in such a model). In this case there is no contradiction to Poincaré's theorem, because entropy does not really depend

on the ("micro"-)state $x$.

The second law of thermodynamics applies to *changes* in the underlying physical "micro"-dynamics, i.e. in the dynamical system or in the mapping $\varphi$. Such changes can occur for example if boundary conditions are changed by the experimenter or engineer; they are described on a much coarser time scale, and as a matter of fact, they can only lead in a certain direction, namely toward higher entropy.

Another way of turning this argument is the following: The thermodynamical (equilibrium) entropy is a quantity that is based on thermodynamical measurements, which always measure time averages in the range from milliseconds to seconds. In particular, such an unusual momentary state as in the picture following (III.2) cannot be measured thermodynamically, in fact the ergodic hypothesis states that we shall usually measure a time average which is close to the "space mean". Therefore a thermodynamical measurement of the number of atoms (i.e. the "pressure") in the left chamber will almost always give a result close to 500. In some branches of thermodynamics ("non-equilibrium" thermodynamics), however, a variable $e(x)$ is associated with micro states $x \in X$, which is also interpreted as the "entropy" of $x$, but is not constant on $X$. In this case Poincaré's theorem shows that the second law for this variable $e$ cannot be strictly true, but still it is argued that a big decrease of $e$ is very improbable. For example, we can try to capture the *momentary* state of the gas in the box, by quickly inserting a separating wall into the box at some arbitrary moment (chosen at random). Then the thermodynamical calculations of the invariant measure on the state space tell us, that we have a chance of $2^{-1000}$ of catching the gas in a position with all 1000 atoms in the left half of the box (low "entropy"), and a chance of 27.2% of having 495 to 505 atoms in the left half of the box (high "entropy").

**III.D.3. Counterexamples:**
The recurrence theorem (III.2) is not valid without the assumption of *finite* measure spaces or *measure-preserving* transformations:

(i) Take $X = \mathbb{R}$ and the Lebesgue measure $m$. Then the shift

$$\tau : x \mapsto x + 1$$

on $X$ is bi-measure-preserving, but no point of $A := [0, 1)$ is recurrent to $A$.

(ii) The transformation

$$\varphi : x \to x^2$$

on $X = [0, 1]$ is bi-measurable, but not measure-preserving for the Lebesgue measure $m$. Clearly, no point of $A := [\frac{1}{2}, \frac{2}{3}]$ is recurrent to $A$.

**III.D.4. Recurrence in random literature:**
A usual typewriter has about 90 keys. If these keys are typed at random, what is the probability to type for example this book? Let us say, this book has $N$ letters including blanks. Then the probability of typing it with $N$ random letters is $p = 90^{-N}$. The Bernoulli shift $B(\frac{1}{90}, \ldots, \frac{1}{90})$ is an MDS $(\widehat{X}, \widehat{\Sigma}, \widehat{\mu}; \tau)$ whose state space consists of sequences $(x_k)_{k\in\mathbb{Z}}$ which can be regarded as the result of infinite random typing. What is the probability, that such a sequence contains this book,

i.e. the sequence $R_1, \dots, R_N$ of letters? From

$$\begin{aligned}
&\widehat{\mu}[\text{there exists } k \in \mathbb{Z} \text{ such that } x_{k+1} = R_1, \dots, x_{k+N} = R_N] \\
&\quad = 1 - \widehat{\mu}[\text{for every } k \in \mathbb{Z} \text{ there exists } i \in \{1, \dots, N\} \text{ such that } x_{k+i} \neq R_i] \\
&\quad \geqslant 1 - \prod_{k=1}^{n} \widehat{\mu}[\text{there exists } i \in \{1, \dots, N\} \text{ such that } x_{k+i} \neq R_i] \\
&\quad = 1 - (1-p)^n \quad \text{for every } n \in \mathbb{N}
\end{aligned}$$

we conclude that this probability is 1. Now consider $A := [x_1 = R_1, \dots, x_N = R_N]$ having $\widehat{\mu}(A) = 0$. We have just shown that for almost every $x \in \widehat{X}$ there is a number $k$ such that $\tau^k(x) \in A$ for the shift $\tau$. Poincaré's theorem implies that there are even infinitely many such numbers, i.e. almost every sequence contains this book infinitely often!

By Kac's theorem (Kac [1947], Petersen [1983]) and the ergodicity of $B(\frac{1}{90}, \dots \frac{1}{90})$ the average distance between two occurrences of this book in random text is $\frac{1}{p} = 90^N$ digits. The fact that this number is very large, may help to understand the strange phenomenon depicted in (III.2)

**III.D.5. Invariant sets:**
The transformations $\varphi : X \to X$ which we are considering in these lectures are bijective. Therefore it is natural to call a subset $A \subseteq X$ $\varphi$-*invariant* if $\varphi(A) \subseteq A$ and $\varphi^{-1}(A) \subseteq A$, i.e. $\varphi(A) = A$. With this definition, a closed $\varphi$-invariant set $A \subseteq X$ in a TDS $(X; \varphi)$ always leads to the restricted TDS $(A; \varphi|_A)$, while $([0,1]; \varphi)$, $\varphi(x) := x^2$ and $A = [0, \frac{1}{2}]$ gives an example such that $\varphi(A) \subseteq A$ but $\varphi|_A$ is not a homeomorphism of $A$.

For MDSs $(X, \Sigma, \mu; \varphi)$ the situation is even simpler: $\varphi(A) \subseteq A$ implies $A \subseteq \varphi^{-1}(A)$ and $\mu(A) = \mu(\varphi^{-1}(A))$ since $\varphi$ is measure-preserving. Therefore $A = \varphi^{-1}(A)$ and $\varphi(A) = A$ $\mu$-a.e..

In agreement with the definition above we define the orbit of a point $x \in X$ as $\{\varphi^k(x) : k \in \mathbb{Z}\}$. If $(X; \varphi)$ is a TDS, the smallest closed invariant set containing a point $x \in X$ is clearly the "closed orbit" $\overline{\{\varphi^k(x) : k \in \mathbb{Z}\}}$. However, the closed orbit is, in general, not a minimal set: For example consider the one point compactification of $\mathbb{Z}$

$$X := \mathbb{Z} \cup \{\infty\}$$

$$\text{and the shift} \quad \tau : \begin{cases} x \mapsto x+1 & \text{if } x \in \mathbb{Z} \\ \infty \mapsto \infty \end{cases}.$$

Then $\overline{\{\tau^k(0) : k \in \mathbb{Z}\}} = X$ is not minimal since $\tau(\infty) = \infty$.
In many cases, however, the closed orbit is minimal as can be seen in the following.

**Lemma:** Let $(X; \varphi)$ be a TDS, where $X$ is a metric space (with metric $d$) and assume that $X = \overline{\{\varphi^s(a) : s \in \mathbb{Z}\}}$ for some $a \in X$. If for every $\varepsilon > 0$ there exists $k \in \mathbb{N}$ with

$$d(a, \varphi^{ks} a) < \varepsilon \quad \text{for all } s \in \mathbb{Z},$$

then $(X; \varphi)$ is minimal.

*Proof.* It suffices to show that $a \in \overline{\{\varphi^s(x) : s \in \mathbb{Z}\}}$ for every $x \in X$. Let be $x \in X$, $\varepsilon > 0$, and choose $k \in \mathbb{N}$ such that

(i) $d(a, \varphi^{ks}a) < \varepsilon$ for all $s \in \mathbb{Z}$.
Since the family of mappings $\{\varphi^0, \varphi^1, \ldots, \varphi^k\}$ is equicontinuous at $x$ there is $\delta > 0$ such that

(ii) $d(\varphi^t x, \varphi^t y) < \varepsilon$ if $t \in \{0, \ldots, k\}$ and $d(x, y) < \delta$. The orbit of $a$ is dense in $X$. Therefore, we find $r \in \mathbb{Z}$ with

(iii) $d(x, \varphi^r a) < \delta$ and by (i) a suitable $t \in \{0, \ldots, k\}$ with

(iv) $d(\varphi^{t+r}a, a) < \varepsilon$.

Combining (ii), (iii) and (iv) we conclude that

$$d(\varphi^t x, a) \leqslant d(\varphi^t x, \varphi^t(\varphi^r a)) + d(\varphi^{t+r}a, a) \leqslant 2\varepsilon.$$

∎

**Remark:** Minimality in metric spaces is equivalently characterized by a property weaker than that given above (see Jacobs [1960], 5.1.3.).

**III.D.6. Ergodicity implies "time mean equal space mean":**
The physicists wanted to replace the time mean

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

of an "observable" $\varphi$ in the "state" $x$ by the space mean

$$\int_X f \, \mathrm{d}\mu \qquad \text{(see Lecture I)},$$

i.e. the above limit has to be equal the constant function $(\int_X f \, \mathrm{d}\mu) \cdot \mathbf{1}$. Obviously the time mean is a $\varphi$-invariant function, and we conclude by (III.4) that "time mean equal space mean" holds for every observable $f$ (at least: $f \in L^p(\mu)$) if and only if (!) the dynamical system is ergodic. In this way the original problem of ergodic theory seems to be solved, but there still remains the task for the mathematician to prove the existence of the above limit (see Lecture IV and V). Even more important (and more difficult) is the problem of finding physical systems and their mathematical models, which are ergodic. The statement of Birkhoff-Koopmann [1932] "the outstanding unsolved problem in ergodic theory is the question of the truth or falsity of metrical transitivity (= ergodicity) for general Hamiltonian systems" is still valid, even if important contributions have been made for the so-called "billiard gas" by Sinai [1963] and Gallavotti-Ornstein [1974] (see Gallavotti [1975]).

**III.D.7. Decomposition into ergodic components:**
As indicated it is a mathematical principle to decompose an object into "irreducible" components and then to investigate these components. For an MDS this is possible (with "ergodic" for "irreducible"). In fact, such a decomposition is based on the geometrical principle of expressing a point of a (compact) convex set as a convex sum of extreme points (see books on "Choquet theory", e.g. Phelps [1966] or Alfsen [1971]), but the technical difficulties, due to the existence of null sets, are considerable, and become apparent in the following example:

Consider the MDS $(X, \mathcal{B}, m; \varphi_a)$ where $X := \{z \in \mathbb{C} : |z| \leqslant 1\}$, $\mathcal{B}$ the Borel algebra, $m$ the Lebesgue measure $m(X) = 1$ and $\varphi_a$ the rotation

$$\varphi_a(z) = a \cdot z$$

for some $a \in \mathbb{C}$ with $|a| = 1$, $a^n \neq 1$ for all $n \in \mathbb{N}$. Its ergodic "components" are the circles $X_r := \{z \in \mathbb{C} : |z| = r\}$ for $0 \leqslant r \leqslant 1$ and $(X, \mathcal{B}, m; \varphi_a)$ is "determined" by these ergodic components. For more information we refer to von Neumann [1932] or Rohlin [1966].

**III.D.8. One-dimensionality of the fixed space:**
Ergodicity is characterized by the one-dimensionality of the fixed space (in the appropriate function space) while minimality is not (III.4 and III.8.1). The fixed space of the induced operator $T_\varphi$ in $C(X)$ is already one-dimensional if there is at least one point $x \in X$ having dense orbit $\{\varphi^n(x) : n \in \mathbb{Z}\}$ in $X$ (see III.7, Proof). This property of a TDS, called "topological transitivity" or "topological ergodicity", is another topological analogue of ergodicity as becomes evident from the following characterizations (see Walters [1975] p. 22 and p. 117):

1. For an MDS $(X, \Sigma, \mu; \varphi)$the following are equivalent:
   a. $\varphi$ is ergodic.
   b. For all $A, B \in \Sigma$, $\mu(A) \neq 0 \neq \mu(B)$, there is $k \in \mathbb{Z}$ such that $|\mu(\varphi^k A \cap B) > 0$.
2. For a TDS $(X; \varphi)$, $X$ metric, the following assertions are equivalent:
   a. $\varphi$ is topologically ergodic.
   b. For all $A, B$ open, $A \neq \varnothing \neq B$ there is $k \in \mathbb{Z}$ such that $\varphi^k A \cap B \neq \varnothing$

But even topological transitivity, although weaker than minimality, is not characterized by the fact that the fixed space is one-dimensional in $C(X)$, see (III.8).i. The reason is that $T_\varphi$ in $C(X)$ lacks a certain convergence property which is automatically satisfied in $L^p(X, \Sigma, \mu)$ (see VI.7 and IV.8; for more information see IX.D.7.

**III.D.9. Ergodic and minimal rotations on the $n$-torus:**
The rotation

$$\varphi_a : z \mapsto a \cdot z$$

on the $n$-dimensional torus $\Gamma^n$ with $a = (a_1, \ldots, a_n) \in \Gamma^n$ is is ergodic (minimal) if and only if $\{a_1, \ldots, a_n\}$ are linearly independent in the $\mathbb{Z}$-module $\Gamma$

*Proof.* (i) In the measure-theoretical case use the $n$-dimensional Fourier expansion and argue as in (III.5.i).
(ii) In the topological case we argue as in (III.8.iii) observing that for an $a = (a_1, \ldots, a_n) \in \Gamma^n$ the set $\{a^k : k \in \mathbb{Z}\}$ is dense in $\Gamma^n$ iff $\{a, \ldots, a_n\}$ is linearly independent in the $\mathbb{Z}$-module $\Gamma$ (see D.8).

■

**III.D.10. Ergodic vs. minimal:**
Let $(X; \varphi)$ be a TDS and $\mu$ a $\varphi$-invariant probability measure on $X$ (see also App. S). Then $(X, \mathcal{B}, \mu; \varphi)$ is an MDS for the Borel algebra $\mathcal{B}$. In this situation, is it possible that if is ergodic but not minimal, or vice versa? The positive answer to the first part or our question is given by the Bernoulli shift, see (III.5.ii) and (III.8.iv). The construction of a dynamical system which is minimal but not ergodic is much more difficult and needs results of Lecture IV. We come back to this problem in IV.D.9.

**III.D.11. Irreducible operators on Banach lattices:**
Let $T$ be a positive operator on some Banach lattice $E$. It is called irreducible if it leaves no non-trivial closed lattice ideal invariant. If $E = C(X)$, resp. $E = L^1(X, \sigma, \mu)$, every closed lattice ideal is of the form

$$I_A := \{f \in E : f(A) \subseteq \{0\}\}$$

where $A \subseteq X$ is closed, resp. measurable, (Schaefer [1974], p. 157). Therefore, it is not difficult to see that an induced operator $T_\varphi$ on $C(X)$, resp. $L^p(X, \Sigma, \mu)$ is irreducible if and only if $(X; \varphi)$ is minimal, resp. if $(X, \Sigma, \mu; \varphi)$ ergodic. In contrast to minimal TDSs the ergodicity of an MDS $(X, \Sigma, \mu; \varphi)$ is characterized by the one-dimensionality of the $T_\varphi$-fixed space in $L^p(X, \Sigma, \mu)$, $1 \leqslant p < \infty$, (see III.4). The reason for this is the fact that the induced operators are mean ergodic on $L^p(\mu)$ but not on $C(X)$ (see Lecture IV). More generally, the following holds (see Schaefer [1974], III.8.5).

**Proposition:** Let $T$ be a positive operator on a Banach lattice $E$ and assume that $T$ is mean ergodic with non-trivial fixed space $F$. The following are equivalent:

(a) $T$ is irreducible.
(b) $F = \langle u \rangle$ and $F' = \langle \mu \rangle$ for some quasi-interior point $u \in E_+$ and a strictly positive linear form $\mu \in E'_+$.

If $E$ is finite-dimensional, we obtain the classical concept of irreducible (= indecomposable) matrices (see IV.D.7 and Schaefer [1974], I.6).

**Example:** The matrix

$$\begin{pmatrix} p_0 & \cdots\cdots & p_{k-1} \\ \vdots & & \vdots \\ p_0 & \cdots\cdots & p_{k-1} \end{pmatrix}$$

of (II.6), Exercise is irreducible whereas the Bernoulli shift $B(p_0, \ldots, p_{k-1})$ is ergodic (see (III.5.ii)). This gives the impression that irreducibility is preserved under dilation (see App. U) at least in this example. In fact, this turns out to be true (App. U), and in particular in (IV.D.8) we shall show that any Markov shift is ergodic iff the corresponding matrix is irreducible. Frobenius discovered in 1912 that the point spectrum of irreducible positive matrices has nice symmetries. The same is true for operators $T_\varphi$, as shown in (III.9).

This result has been considerably generalized to irreducible positive operators on arbitrary Banach lattices. We refer to Schaefer [1974], V.5.2 for a complete treatment and quote the following theorem.

**Theorem (Lotz, 1968):** Let $T$ be a positive irreducible contraction on some Banach lattice $E$. Then $P\sigma(R) \cap \Gamma$ is a subgroup of $\Gamma$ or empty, and every eigenvalue in $\Gamma$ is simple.

References: Lotz [1968], Schaefer [1967/68], Schaefer [1974].

**III.D.12. The origin of the word "Ergodic Theory":**
In the last decades of the $19^{\text{th}}$ century mathematicians and physicists endeavoured to explain thermodynamical phenomena by mechanical models and tried to prove the laws of thermodynamics be mechanical principles or, at least, to discover close analogies between the two. The Hungarian M.C. Szily [1872] wrote:

> "The history of the development of modern physics speaks decidedly in favour of the view that only those theories which are based on mechanical principles are capable of affording a satisfactory explanation of the phenomena."

Those efforts were undertaken particularly in connection with the second law of thermodynamics; Szily [1876] even claimed to have deduced it from the first, whereas a few years earlier he had declared:

> "What in thermodynamics we call the second proposition, is in dynamics no other than Hamilton's principle, the identical principle which has already found manifold applications in several branches of mathematical physics."

(see Szily [1872]; see also the subsequent discussion in Clausius [1872] and Szily [1873].)

In developing the Mechanical Theory of Heat three fundamentally different hypotheses were made; besides the hypothesis of the stationary or quasi-periodic motions (of R. Clausius and Szily) and the hypothesis of monocyclic systems (of H. von Helmholtz, cf. Bryan-Larmor [1892]), the latest investigations at that time concerned considerations which were based on a very large number of molecules in a gas and which established the later Kinetic Theory of Gases. This was the statistical hypothesis of L. Boltzmann, J.B. Maxwell, P.G. Tait and W. Thomson, and its fundamental theorem was the equipartition theorem of Maxwell and Boltzmann: When a system of molecules has attained a stationary state the time-average of the kinetic energy is equally distributed over the different degrees of freedom of the system. Based on this theorem there are some proofs of the second law of thermodynamics (Burbury [1876], Boltzmann [1887]), but which was the exact hypothesis for the equipartition theorem itself? In Maxwell [1879] we find the answer:

> "The only assumption which is necessary for the direct proof (of the equipartition theorem) is that the system, if left to itself in its actual state of motion, will, sooner or later, pass through every phase which is consistent with the equation of energy."

Boltzmann [1871], too, made use of a similar hypothesis:

> "Von den zuletzt entwickelten Gleichungen können wir unter einer Hypothese, deren Anwendbarkeit auf warme Körper mir nicht unwahrscheinlich scheint, direkt zum Wärmegleichgewicht mehratomiger Gasmoleküle je noch allgemeiner zum Wärmegleichgewicht eines beliebigen mit einer Gasmasse in Berührung stehenden Körpers gelangen. Die große Unregelmäßigkeit der Wärmebewegung und die Mannigfaltigkeit der Kräfte, welche von außen auf die Körper wirken, macht es wahrscheinlich, daß die Atoms derselben vermöge der Bewegung, die wir Wärme nennen, alle möglichen mit der Gleichung der lebendigen Kraft vereinbare Positionen und Geschwindigkeiten durchlaufen, daß wir also die zuletzt entwickelten Gleichungen auf die Koordinaten und die Geschwindigkeitskomponenten der Atome warmer Körper anwenden können."

Sixteen years later, Boltzmann mentioned in [1887]

> "... (Ih habe für derartige Inbegriffe von Systemen den Namen *Ergoden* vorgeschlagen.)..."

This may have induced P. and T. Ehrenfest to create the notion of "Ergodic Theory" by writing in "Begriffliche Grundlagen der statistischen Auffassung" [1911]:

> "... haben Boltzmann und Maxwell eine Klasse von mechanischen Systemen durch die folgende Forderung definiert:
> Die einzelne ungestörte Bewegung des Systems führt bei unbegrenzter Fortsetzung schließlich *durch jeden Phasenpunkt hindurch*, der mit der mitgegebenen Totalenergie verträglich ist. – Ein mechanisches System, das diese Forderung erfüllt, nennt Boltzmann ein ergodisches System."

The notion "ergodic" was explained by them in a footnote:

> " ἔργον = Energie, ὁδός= Weg : Die G-Bahn geht durch alle Punkte der Energiefläche. Diese Bezeichnung gebraucht Boltzmann das erste Mai in der Arbeit [15] (1886) " (here Boltzmann [1887])

But this etymological explanation seems to be incorrect as we will see later. The hypothesis quoted above, i.e. that the gas models are ergodic systems, they called the "Ergodic Hypothesis". In the sequel they doubted the existence of ergodic systems, i.e. that their definition does not contradict itself. Actually, only few years later A. Rosenthal and M. Plancherel proved independently the impossibility of systems that are ergodic in this sense (cf. Brush [1971]). Thus, "Ergodic Theory" as a theory of ergodic systems hardly survived its definition. Nevertheless, from the explication of the "Ergodic Hypothesis" and its final negation, "Ergodic Theory" arose as a new domain of mathematical research (cf. Brush [1971], Birkhoff-Koopmann [1932].

But, P. and T. Ehrenfest were mistaken when they thought that Boltzmann used the notion "Ergodic" and "Ergodic Systems" in Boltzmann [1887] for the first time. In 1884 he had already defined the notion "Ergode" as a special type of "Monode". In his article (Boltzmann [1885]) first of all he wrote:

> "Ich möchte mir erlauben, Systeme, deren Bewegung in diesem Sinne stationär ist, als monodische Oder kürzer als Monoden zu bezeichnen. (Mit dem Namen stationär wurde von Herrn Clausius jede Bewegung bezeichnet, wobei Koordinaten und Geschwindigkeiten immer zwischen endlichen Grenzen eingeschlossen bleiben). Sie sollen dadurch charakterisiert sein, daß die in jedem Ptmkte derselben herrschende Bewegung unverändert fortdauert, also nicht Funktion der Zeit ist, solange die äußeren Kräfte unverändert bleiben, und daß auch in keinem Punkte und keiner Flche derselben Masse oder lebendige Kraft oder sonst ein Agens ein- oder austritt."

In a modern language a "Monode" is a system only moving in a finite region of phase space described by a dynamic system of differential equations; a simple example is a mathematical pendulum. From Boltzmann's definition we can understand the name: μόνος means "unique", "Monode" probably comes from μονώδης which is composed of μόνο–ώδης where the suffix –ώδης means "–like".

Having specified some different kinds of "Monoden" as "Orthoden" and "Holoden", Boltzmann turned towards collections (ensembles) of systems which were all of the same nature, totally independent of each other and each, of them fulfilling a number of equations $\varphi_1 = a_1, \ldots, \varphi_k = a_k$. Of special interest to him were those collections of systems fulfilling only one equation $\varphi = a$ concerning the energy of all systems in the collection.

> "... so wollen wir den Inbegriff aller $N$ Systeme als eine Monode bezeichnen, welche durch die Gleichungen $\varphi_1 = a_1, \ldots$ beschränkt ist ... Monoden, welche nur durch die Gleichung der lebendigen Kraft beschränkt sind, will ich als *Ergoden*, solche, welche außer dieser Gleichung auch noch durch andere beschränkt sind, als *Supergoden* bezeichnen.... Für Ergoden existiert also nur ein $\varphi$, welches gleich der für alle Systeme gleichen und während der Bewegung jedes Systems konstanten Energie eines Systems $\chi + \psi = \frac{(\phi+L)}{N}$ ist".

(Boltzmann [1885]; $\chi$, $\phi$ mean the potential energy, $\psi$, $L$ the kinetic energy of one system, of the collection of $N$ systems, respectively.) The last sentence of that quotation helps us to understand the name "Ergode" in the right way: The word ἔργον = "work, energy" is used, but in a sense different from that presumed by the Ehrenfests who also did not mention Boltzmann's article [1885] in their bibliography [1911].

Boltzmann also had knowledge of "Monoden" fulfilling the "Ergodic Hypothesis" of the Ehrenfests. In the fourth paragraph of Boltzmann [1885] we read in a footnote:

> "Jedesmal, wenn jedes einzelne System der Monode im Verlaufe der Zeit alle an den verschiedenen Systemen gleichzeitig nebeneinander vorkommenden Zustände durchläuft, kann an Stelle der Monode ein einziges System gesetzt werden.... Für eine solche Monode wurde schon früher die Bezeichnung "isodisch" vorgeschlagen"

In summary an "Ergode" is a special kind of "Monode", namely one which is determined by " ἔργον" = "energy" or "work", and the word "Monode" stems from μόνος = "one" or "unique" and the suffix –ώδης = "–like" or "–full". Therefore a "Monode" is literally "one-like" i.e. atomary or indecomposable, which is just the modern meaning of ergodic. Taken literally, however, the word "Ergode" means "energy-like" or "work-full", which brings us back to our first etymological answer in Lecture I:

" difficult "!

References: Boltzmann [1885], [1887], Brush [1971], Ehrenfest [1911]

P.S. The above section originated from a source study by M. Mathieu. The Ehrenfests' explanation of the word "ergodic" is still advocated by A. LoBello:

> The etymology of the word ergodic, in: Conference on modern Analysis and Probability, New Haven 1982, Contempt.Math. 26, Amer. Math. Soc. Providence R.I., 1984, p.249.