THREE LECTURES GIVEN IN TORUN ON: SELECTED TOPICS IN MODERN ERGODIC THEORY

JEAN-PAUL THOUVENOT
UNIVERSITÉ PARIS 6

1. Fundamentals of the Ergodic Theory

The first section contains basic notions and facts from ergodic theory. Throughout the notes we will denote by $(X, \mathcal{A}, m)$ a Lebesgue probability space and by T a bijective, bi-measurable, measure-preserving transformation of X (called an *automorphism* of the space $(X, \mathcal{A}, m)$); altogether, $(X, \mathcal{A}, m, T)$ is an (*abstract*) *dynamical system* and the basic object of study in ergodic theory.

Example (1.1).
We will start with some examples of dynamical systems:

- $\mathbb{S}^1$ with Lebesgue σ-algebra and normalised measure, and rotation as the transformation;
- multidimensional tori with the product σ-algebra, product measure and transformation given by an algebraic automorphism. (Lebesgue measure is preserved by unicity of the Haar measure).
- for dynamical systems, the Hamiltonian flow ($\frac{\partial H}{\partial p} = -\dot{p}, \frac{\partial H}{\partial q} = \dot{q}$) equipped with the Liouville measure $dp.dq$ which it leaves invariant (finite when the phase space is compact).
- consider a probability space and a bi-infinite sequence (X_n) of random variables such that, for all $n, k > 0$, the law of $(X_n, X_{n+1}, ..X_{n+k})$ is independent of k; (this is called a stationary stochastic process) then the shift transformation $\Theta(X_n)_{n\in\mathbb{Z}} := (X_{n+1})_{n\in\mathbb{Z}}$ also generates a dynamical system. Of particular importance are the examples where the (X_n) form an independent family and where the common law of the X_n is a finite probability vector π. The corresponding dynamical system is the Bernoulli shift $B(\pi)$.

We will present now some properties of dynamical systems.

Poincaré Recurrence Theorem (1.2).
Let $(X, \mathcal{A}, m, T)$ be a dynamical system. Then almost every point of any non-negligible set returns to this set along its future trajectory: if $m(A) > 0$, then for almost every $x \in A$ there exists $n > 0$, for which $T^n(x) \in A$.

Sketch of the proof. Consider $N \subset A$, the set of points that do not return to A. One can check that its iterated images $T^n(N)$ are pairwise disjoint. But they are of the same measure, which is therefore zero. This implies that almost every point in A returns infinitely often to A. □

To a transformation can be associated a unitary operator acting on L^2-functions (the *Koopman* operator) :

$$L^2(X) \ni f \mapsto U_T f := f \circ T \in L^2(X),$$

which is well-defined and unitary due to T being measure-preserving.

One considers a special sub-σ-algebra as well: all the measurable T-invariant sets:

$$\mathcal{I} := \{A \in \mathcal{A} \colon T^{-1}A = A \text{ at least a.e. (i.e. } m(T^{-1}A \triangle A) = 0)\}^1$$

In particular, when this σ-algebra is trivial (i.e. $\mathcal{I} = \{\emptyset, X\} = \{A \in \mathcal{A} \colon m(A) = 0 \text{ or } 1\}$ – remember the a.e. convention!), the transformation is said to be ergodic. Observe also that the $\mathcal{I}$-measurable functions are precisely the T-invariant functions, and that the set of $L^2(X)$-invariant functions is exactly the eigenspace of the Koopman operator U_T corresponding to the eigenvalue 1. Ergodicity then means that this space is reduced to the constant functions. More generally, in the ergodic case the set of eigenvalues λ for which there exists f such that $U_T f = \lambda f$ satisfy $|\lambda| = 1$ and they form a group (as a consequence of ergodicity). $\mathcal{E}$, the smallest algebra which makes these functions measurable is called the Kronecker algebra.

2. Estimating time averages

Results on convergence. In what follows we will state theorems on convergence and estimation of the averages of functions along trajectories. Those theorems are landmarks of the beginning of the ergodic theory. We start with the theorem of Von Neumann of which we shall give three proofs.

Definition (2.1).
Given a dynamical system (S, $\mathcal{A}$, m, T) and a measurable function $f \colon X \to \mathbb{R}$, we call

$$M_N f := \frac{1}{N} \sum_{i=0}^{N-1} f \circ T^i$$

(for any $N \in \mathbb{N}$) the *time average* of f.

Von Neumann Theorem (2.2).
If, additionally, $f \in L^2(X)$, then $M_n f \xrightarrow{n\to\infty} E(f|\mathcal{I})$ in the L^2-norm.

[1]Throughout this summary the identities and inequalities between sets or functions are usually meant to be satisfied at least almost everywhere

Sketch of the first proof of the von Neumann Theorem. Consider the following two subspaces of $L^2(X)$: all $\mathcal{I}$-measurable functions $L^2(\mathcal{I})$ and the closure of the *coboundaries* $H := \overline{\{f - f \circ T \colon f \in L^2(X)\}}^{L^2}$. They sum up to the whole space: $L^2(X) \ominus L^2(\mathcal{I}) = H$, i.e. $H^\perp = \{f - f \circ T \colon f \in L^2(X)\}^\perp = L^2(\mathcal{I})$. Indeed if $g \in H^\perp$,

$\int fg = \int f \circ Tg$ implies $\int f(g - g \circ T) = 0$ for all f and therefore $g \in L^2(\mathcal{I})$. Due to the preceding, given $g \in L^2$ and $\varepsilon > 0$, there exists f and $h \in L^2$ such that

$$L^2(X) \ni g = \bar{g} + (f - f \circ T) + h \quad (*)$$

where $\|h\| < \epsilon$. (where $\bar{g}$ is the orthogonal projection of g on $L^2(\mathcal{I})$, which is the same as $E(g|\mathcal{I})$ the conditional expectation of g with respect to $\mathcal{I}$). From (*), one sees easily that the limit of the Birkhoff averages $M_N g$ is away from $\bar{g}$ by less than ε for any ε, which is the desired conclusion. □

second proof of the Von Neumann theorem. Given $f \in L^2$, we construct a measure σ_f on S^1 such that for all $n \in Z$,

$$\int f \circ T^n f dm = \int \exp 2i\pi n\theta d\sigma_f$$

σ_f, the spectral measure of f is easily obtained as the weak limit of the measures

$$\sigma_n = 1/n(\int \mid \sum_{k=0}^{k=n-1 f \circ} T^k(x) f(x) \exp 2i\pi k\theta \mid^2 dm(x))d\theta$$

Restricted to the L^2 closure of the linear span of the $T^n f, n \in Z$ the Koopman operator acts therefore exactly as the multiplication by $\exp 2i\pi\theta$ on L^2 of S^1 equipped with σ_f. The Von Neumann theorem is now a consequence of Lebesgue theorem and of the convergence towards the characteristic function of 0 of the averages $1/n \sum_{k=0}^{k=n-1} \exp 2i\pi k\theta$. (Note that this proof can be used to prove the L^2 convergence of averages like $1/n \sum_{k=0}^{k=n-1} T^{n^2} f$). □

Another application of the preceding proof is that if one considers $T \times T$ the cartesian square of T (this means that $T \times T$ acts on $(X, \mathcal{A}) \times (X, \mathcal{A})$ equipped with the product measure $m \times m$) and two functions f and g in $L^2(X)$, each one measurable with respect to one factor, $\sigma_{f \times g} = \sigma_f * \sigma g$. This implies as a consequence of a theorem of Wiener that the algebra of invariant sets of $T \times T$ is $\mathcal{E} \times \mathcal{E}$ measurable ($\mathcal{E}$ the Kronecker algebra).

Joinings, isomorphisms, and factors. We will introduce now joinings which are very important notions of the modern ergodic theory and describe, using them, isomorphisms and factors. As a preparation for future use, we shall put them at work to give a third proof of the von Neumann Theorem in the ergodic case.

Definition (2.3).
Consider two dynamical systems $(X, \mathcal{A}, m, T), (Y, \mathcal{B}, \mu, S)$. A *joining* of them is a $T \times S$-invariant measure λ on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ such that $\lambda(A \times Y) = m(A)$ and $\lambda(X \times B) = \mu(B)$ for relevant A and B. (This same definition obviously extends to the product of finitely many transformations). The σ-algebra $\mathcal{V} := \mathcal{A} \times Y$ is called the *vertical algebra*, and $\mathcal{H} := X \times \mathcal{B}$ – the *horizontal* one.

If the joining λ of $(X, \mathcal{A}, m, T)$ and $(Y, \mathcal{B}, \mu, S)$. satisfies that

$$\mathcal{V} = \mathcal{H} \quad (mod\lambda)$$

(i.e. for every $A \in \mathcal{V}$ there exists $B \in \mathcal{H}$ such that $\lambda(A \Delta B = 0$ and the symmetrical holds), then T and S are isomorphic.

If

$$\mathcal{H} \subset \mathcal{V} \quad (mod\lambda)$$

then $(Y, \mathcal{B}, \mu, S)$ is a factor of $(X, \mathcal{A}, m, T)$.

The link with the classical definition is that there exists a mapping $\Phi \colon (X, \mathcal{A}, m) \to (Y, \mathcal{B}, \mu)$, which carries m to μ, $(\Phi^*(m) := m \circ \Phi^{-1} = \mu)$ such that the following diagram is commutative:

$$\begin{array}{ccc} X & \xrightarrow{\Phi} & Y \\ {\scriptstyle T}\downarrow & & \downarrow{\scriptstyle S} \\ X & \xrightarrow[\Phi]{} & Y \end{array}$$

The isomorphism corresponds to the case where Φ is invertible. The corresponding joinings are given by $\lambda(A \times B) = m(A \cap \Phi^{-1}(B))$. Conversely, the existence of Φ is a consequence of the fact that $(X, \mathcal{A}, m)$ is assumed to be a Lebesgue space.

We also see that we can identify factors to invariant subalgebras (an algebra $\mathcal{A}$ is said to be invariant if, for every $A \in A$, $T(A) \in \mathcal{A}$ and $T^{-1}(A) \in \mathcal{A}$).

We refer to joinings λ where $(X, \mathcal{A}, m, T) = (Y, \mathcal{B}, \mu, S)$ as self-joinings of $(X, \mathcal{A}, m, T)$. In this case the self-joinings contain the important *diagonal joining* Δ defined by $\Delta(A \times B) = m(A \cap B)$.

The set of joinings between two dynamical systems is always nonempty – it contains $m \otimes \mu$. It is also compact with respect to the following metric: one takes a countable dense subsets of $\mathcal{A}$ and $\mathcal{B}$ (with the symmetric difference metric: $\mathrm{dist}(A_1, A_2) := m(A_1 \triangle A_2)$), say (A_m) and

(B_n),[2] and puts for two joinings

$$d(\lambda_1, \lambda_2) := \sum_{m,n=1}^{\infty} \frac{1}{2^{m+n}} |\lambda_1(A_m \times B_n) - \lambda_2(A_m \times B_n)|.$$

When $m \otimes \mu$ is the only joining, the dynamical systems are said to be *disjoint*. In particular, one can check that T and the identity map $\mathrm{id}\colon X \to X$ are disjoint whenever T is ergodic. This is because if we consider λ a joining between Id and T, the conditional expectation with respect to the vertical algebra (corresponding to T) takes any L^2 function measurable with respect to the horizontal algebra to a constant (as it has to be T-invariant). We have been using the fact that the Koopman operator and the projection on the L^2 space of a factor algebra (the conditional expectation operator) commute. We are going to use several times this observation.

We will give now the announced third proof:

Von Neumann Theorem in the ergodic case – a sketch of the proof. We assume that T is ergodic; we thus want to show the convergence $\frac{1}{N}\sum_{i=0}^{N-1} f \circ T^i \xrightarrow{L^2} \int_X f \,\mathrm{d}m$. Consider the *diagonal joining* of T, and its averages $\Delta_N := \frac{1}{N}\sum_{i=0}^{N-1} (\mathrm{id} \times T^i)^* \Delta$. These are also self-joinings, and, as the set of joinings is compact, have a cluster point. One may check that any cluster point is $\mathrm{id}\times T$-invariant, so it is a joining of $(X, \mathcal{A}, m, \mathrm{id})$ and T, but then, since id and T are disjoint, it equals $m \otimes m$. Therefore

$$\int_X (g \otimes f) \,\mathrm{d}\Delta_N \xrightarrow{N\to\infty} \int_X (g \otimes f) \,\mathrm{d}(m \otimes m)$$

for any $g \in L^2(X)$,

which yields the required convergence in weak sense:

$$\int_X g \cdot (\frac{1}{N} \sum_{i=0}^{N-1} f \circ T^i) \,\mathrm{d}m \to \int_X g \cdot (\int_X f \,\mathrm{d}m) \,\mathrm{d}m.$$

The L^2-convergence to 0 of $M_N(f)$ if $\int_X f \,\mathrm{d}m = 0$ is an easy consequence of the weak convergence (summing $\sum_{0\le i,j\le N} \int f T^{|i-j|} f$ on vertical and horizontal lines starting from the diagonal, one obtains the conclusion, omitting a Noth-West square of fixed size, and using that if a_n converges to 0, $\frac{1}{N}\sum_{i=0}^{i=N} i a_i$ also converges to 0). The L^2-convergence for any f follows then easily. □

Birkhoff Theorem (2.4).
If $f \in L^1(X)$, then $M_n f \xrightarrow{n\to\infty} E(f|\mathcal{I})$ both in the L^1-norm and almost everywhere.

[2]Of course, the spaces $\mathcal{A}$ and $\mathcal{B}$ are regarded here as quotient spaces under the equivalence relation of equality almost everywhere. Dense sequences exist in Lebesgue spaces .

We shall give two proofs of the Birkhoff ergodic theorem: one brings it very close to analysis and to the Hardy-littlewood theorem; the other is pure measure theory.

Maximal Lemma. The following important lemma is at the basis of the first proof. It is proved in steps.

Vitali Covering Lemma (2.5).
Let E be a subset of integers covered by finitely many finite intervals of integers:

$$E \subset \mathbb{Z},\ E = \sum_{i \in J} I_i,\ \#J < \infty.$$

Then one can choose, from that family, pairwise disjoint intervals that altogether sum up to at least $\frac{1}{4}\#E$:

$$\exists J_1 \subset J\colon\ (\#\sum_{i \in J_1} I_i \geq \frac{1}{4}\#E,\ \forall i, i' \in J_1 \text{ and } i \neq i'\colon\ I_i \cap I_{i'} = \emptyset).$$

Idea of the proof. Take as I_{i_1} the longest interval (or one of maximal length), as I_{i_2} – the longest not intersecting I_{i_1}, as I_{i_3} – the longest not intersecting I_{i_1} and I_{i_2}, and repeat until possible. Let $J_1 = (i_1, i_2, ...)$ Now the conclusion is satisfied, because any other interval intersects one in J_1 that is longer than it, and thus E is contained in $\sum_{j \in J_1} I'_{i_j}$ where I'_{ij} is the interval with same center as I_{ij} and length $4\#I_{ij}$. □

The l^1 version of the maximal lemma (2.6).
Consider a sequence of reals $(x_n)_{n=0}^{\infty}$ in l^1
For a fixed N put

$$\bar{x}_n := \sup_{1 \leq K \leq N} \frac{1}{K} \sum_{k=0}^{K-1} |x_{n+k}|$$

for every $n \geq 0$ and consider, given $\lambda > 0$, the set $E = \{n \geq 0 \mid \bar{x}_n > \lambda\}$. Then:

$$\#E \leq \frac{4}{\lambda} \sum_{n=0}^{\infty} |x_n|.$$

Idea of the proof. E is finite since $x_n \in l^1$. Denote by I_n, for $n \in E$, an interval $[n, n+K]$ $(K < N)$, such that $\frac{1}{K}\sum_{k=0}^{K-1} |x_{n+k}| > \lambda$. One can estimate its size: $\#I_n \leq \sum_{i \in I_n} |x_i|/\lambda$, and thus the size of E, using the pairwise disjoint subfamily of $\{I_n\}_{n \in E}$ chosen by the Vitali Lemma, since the sum of the $|\ x_n\ |$ on these intervals is smaller than $\|x_n\|_1$. □

The proof of the next lemma is an illustration of the "transference principle" which allows to extend inequalities from orbits to the whole space.

Maximal Lemma (2.7).
Consider a dynamical system $(X, \mathcal{A}, m, T)$. Take a function $f \in L^1(X)$ and put

$$f^* := \sup_{\bar{M}>0} \frac{1}{\bar{M}} \sum_{i=0}^{\bar{M}-1} |f \circ T^i| = \sup_{\bar{M}>0} M_{\bar{M}}|f|$$

Then, for all $\lambda > 0$, :

$$m(\{x \in X : f^*(x) > \lambda\}) \leq \frac{4}{\lambda}\|f\|_1.$$

Sketch of the proof. Put, for $\bar{N} \in \mathbb{N}$,

$$f_{\bar{N}} := \sup_{1\leq K\leq \bar{N}} \frac{1}{K} \sum_{k=0}^{K-1} |f \circ T^k| = \sup_{1\leq K\leq \bar{N}} M_K|f|.$$

The set $\{x \in X : f^*(x) > \lambda\}$ is the limit of the increasing sequence $F_{\bar{N}} := \{x \in X \colon f_{\bar{N}}(x) > \lambda\}$,

We therefore just need to prove the inequality for fixed (arbitrary) $\bar{N}$. Having fixed x and $N \gg \bar{N}$, use the l_1 version of the lemma for the sequence

$$x_n := \begin{cases} f(T^n(x)) & \text{for } n = 0, \ldots, N, \\ 0 & \text{for } n > N. \end{cases}$$

Observe that then $f_{\bar{N}} = \bar{x}_n$, for all $n < N - \bar{N}$ so the result gives in fact an inequality for $\#\{n \leq N - \bar{N} \colon f_{\bar{N}}(T^n x) > \lambda\}$, which is equal to $\sum_{n=0}^{N-\bar{N}} \mathbb{1}_{F_{\bar{N}}}(T^n x)$. Therefore,

$$\sum_{n=0}^{N-\bar{N}} \mathbb{1}_{F_{\bar{N}}}(T^n x) \leq \frac{4}{\lambda} \sum_{n=0}^{N} |f(T^n x)|.$$

By integrating this over $x \in X$ (remember that T is m-invariant) and letting N go to infinity, so that $\frac{N-\bar{N}}{N} \to 1$, one arrives at the required conclusion. □

The meaning of the maximal lemma is that the set of L^1 functions for which the pointwise convergence holds is closed in L^1. The general protocol for this type of proof of pointwise ergodic theorems is, after showing a maximal lemma, to exhibit a dense class of functions for which the convergence occurs. From what we have seen in the first proof of the Von Neumann theorem, a good dense class for the Birkhoff ergodic theorem is the set of functions $f(x) - f \circ T(x) + \phi(x)$ (where $f, \phi \in L^2$ and $\phi \circ T(x) = \phi(x)$ a.e.). Note that in some circumstances, the maximal lemma is easy to obtain, while the difficulty is to exhibit a dense class.

First proof of the Birkhoff ergodic theorem

let g and ε be given. Consider f such that the averages $M_N(f)$ converge almost everywhere and for which $\|f-g\| = h$ satisfies $\|h\| < \epsilon^2$. From the maximal lemma, $m\{x \in X\colon sup\frac{1}{N}\sum_1^N |f\circ T^n - g\circ T^n| \geq \epsilon\} \leq \epsilon$ and hence $\limsup M_N(g) - \liminf M_N(g) < \epsilon$ on a set of measure $> 1-\epsilon$. Since the functions f for which $M_N(f)$ converges almost everywhere are dense, ε can be taken as small as we want, the convergence almost everywhere is proved; the limit is easily identified as $E(g|\mathcal{I})$ and the L^1 convergence follows, from instance, of the uniform integrability of the $M_N(g)$.

Second proof of the Birkhoff ergodic theorem.

Proof. We keep the notations of the previous paragraph. Let $f > 0$ in L^1. We shall first show that

$$\limsup M_N(f) \leq E(f|\mathcal{I}) \text{ a.e. } (*)$$

If $f_K = f \vee (-K)$, (*) is satisfied for f_K, and by the monotone convergence theorem, (when $K \to +\infty$) , (*) is satisfied for all $f \in L^1$. But $\limsup M_N(f)$ is T invariant, hence $\mathcal{I}$ measurable, and (*) applied to $-f$ implies that $M_N(f) \to E(f|\mathcal{I})$ a.e. As before L^1 convergence follows from the uniform integrability of the $M_N(f)$.

We now prove (*) for positive functions f. If (*) is not true there exist $\varepsilon > 0$ and an $\mathcal{I}$ measurable set A with measure $> \epsilon$ such that, for every $x \in A$, $\limsup M_N(f) > E(f|\mathcal{I}) + \epsilon$.

We choose $\bar{N}$ such that for a set $A' \subset A$ with $m(A') > m(A)(1-\delta)$, ($\delta > 0$ to be precised later), for all $x \in A'$, $\sup_{1\leq N\leq \bar{N}} M_N f(x) > E(f|\mathcal{I}) + \epsilon/2$. A direct consequence of the Tchebicheff inequality applied to the function $\frac{1}{L}\sum_{i=0}^{i=L}(1_A - 1_{A'})\circ T^i$ (whose integral is smaller that $\delta m(A)$) implies that except for a set $E^c \subset A$ of size $\delta^{1/2}m(A')$, a fraction $> 1-\delta^{1/2}$ of the orbit of length L of points $x \in E$ can be covered by disjoint intervals (of length $< \bar{N}$) such that, on any of these intervals $[T^m x, T^{m+N}x]$ the average $M_N f\circ T^m(x) > (1+\epsilon/2)E(f|\mathcal{I})$. We then get, as $f > 0$ that $M_L(f) > E(f|\mathcal{I})(1+\epsilon/2)(1-\delta^{1/2})$ on E. Integrating on A, and choosing δ small enough, we get a contradiction (as $\int_A M_L(f) = \int_A f$) □

Open question on "mixed" Birkhoff averages. Consider two commuting measure preserving transformations S, T on $(X, \mathcal{A}, m)$ and two functions $f, g \in L^\infty(X)$. Does the sequence $\frac{1}{N}\sum_{i=0}^{N-1} f\circ S^i(x)\cdot g\circ T^i(x)$ converge almost everywhere? (the L^2- convergence of the preceding averages is known).

The question has a positive answer for two dynamical systems which are powers of the same transformation (this includes negative powers).

3. Entropy

In the present section we study the notions of the information and the entropy and their properties. To begin with, we will restrict to "static" situation, without any dynamics.

We will consider finite measurable partitions of X (and denote them $P = \{P_1, \dots, P_k\}$, Q, R) and sub-σ-algebras of $\mathcal{A}$ (these will be $\mathcal{B}$, $\mathcal{B}_n$).

Definition (3.1).
The *information function* with respect to $\mathcal{B}$ is defined as:

$$I(P|\mathcal{B}) := -\sum_{i=1}^{k} \mathbb{1}_{P_i} \log E(\mathbb{1}_{P_i}|\mathcal{B}); \quad I(P|\mathcal{B}) \geq 0$$

and the *entropy* with respect to $\mathcal{B}$:

$$H(P|\mathcal{B}) := \int_X I(P|\mathcal{B})\,\mathrm{d}m = -\int_X \sum_{i=1}^{k} E(\mathbb{1}_{P_i}|\mathcal{B}) \log E(\mathbb{1}_{P_i}|\mathcal{B})\,\mathrm{d}m.$$

Important is the case when $\mathcal{B}$ is the trivial algebra ν whose elements have measure 0 or 1, then $H(P|\nu)$ is noted $H(P)$ and:

$$H(P) = -\sum_{i=1}^{i=k} m(P_i) \log m(P_i)$$

A good intuitive view point is to consider that $H(P|\mathcal{B})$ measures how much information we receive from an experiment whose events are P measurable (and satisfy the probability law given to them by the measure m) once all the events which are $\mathcal{B}$ measurable are known.

We will give some simple properties of those functions. Recall that $P \vee Q$ stands for common refinement of the partitions: $\{p \cap q \colon p \in P,\ q \in Q\}$. Moreover, by $I(\cdot|P)$, $H(\cdot|P)$ we mean the functions with respect to the σ-algebra generated by P.

Properties of the information and the entropy (3.2).

(1) $I(P) := I(P, \{\emptyset, X\}) = -\sum_{i=1}^{k} \mathbb{1}_{P_i} \log m(P_i)$;
(2) $H(P) := H(P, \{\emptyset, X\}) = -\sum_{i=1}^{k} m(P_i) \log m(P_i)$;
(3) $I(P \vee Q) = I(P) + I(Q|P)$;
(4) $H(P \vee Q) = H(P) + H(Q|P)$;
(5) when P is already $\mathcal{B}$-measurable, then $I(P|\mathcal{B}) = H(P|\mathcal{B}) = 0$;
(6) Knowing more, the experiment modeled after P gives less information: namely, if $\mathcal{B}_1 \supset \mathcal{B}_2$, then $H(P|\mathcal{B}_1) \leq H(P|\mathcal{B}_2)$;
(7) $H(P|\mathcal{B}) = H(P)$ is equivalent to the fact that P and $\mathcal{B}$ are independent. (Knowing $\mathcal{B}$ brings no information concerning the experiment associated to P).
(8) more generally: if $\mathcal{B}_n \nearrow \mathcal{B}$, then $H(P|\mathcal{B}_n) \nearrow H(P|\mathcal{B})$; the same for decreasing sequences;
(9) $I(P \vee Q|R) = I(P|R) + I(Q|P \vee R) = I(P \vee Q \vee R) - I(R)$;

(10) $H(P\vee Q|R)=H(P|R)+H(Q|P\vee R)=H(P\vee Q\vee R)-H(R)$.
(11) $H(P\vee Q|\mathcal{B})=H(P|\mathcal{B})+H(Q|P\vee\mathcal{B})$
(12) entropy is also monotonous with respect to partitions: for $P\subset Q$, $H(P|\mathcal{B})\leq H(Q|\mathcal{B})$;
(13) if P has k elements, $H(P)\leq\log k$

Proof. (3), (6), and (8) are the only ones which need proof (all other statements being then trivially deduced).

For (3), $I(P\vee Q))$ equals $\log m(P_i\cap Q_j)$ on $P_i\cap Q_j$.

$E(\mathbb{1}_{Q_j}|P)$ equals $\frac{m(P_i\cap Q_j)}{m(P_i)}$ on P_i. Therefore $I(Q|P)$ equals $-\log\frac{m(P_i\cap Q_j}{m(Q_j)}$ on $P_i\cap Q_j$. Everything compensates to yield (3)

(6) comes from the fact that $H(P|\mathcal{B})=\sum_{i=1}^{k}\int\Phi(E(\mathbb{1}_{P_i})|\mathcal{B}))$ where $\Phi(x)=-x\log x$. As $\Phi(x)$ is concave, if $\mathcal{B}_1\supset\mathcal{B}_2$, Jensen's inequality implies that $E(\Phi(E(\mathbb{1}_{P_i})|\mathcal{B}_1)|\mathcal{B}_2)>\Phi(E(\mathbb{1}_{P_i})|\mathcal{B}_2)$, for all i, whence the result.

Since $\Phi(x)$ is bounded for $x\in[0,1]$ and since $E(\mathbb{1}_{P_i}|\mathcal{B}_n)\to E(\mathbb{1}_{P_i}|\mathcal{B})$, in L^2, a subsequence n_k yields pointwise convergence (and then norm convergence) of $\Phi(E(\mathbb{1}_{P_i}|\mathcal{B}_{n_k}))$ and (8) follows from the monotonicity (6), while (7) is deduced from the equality in Jensen's inequality. □

We will move to the dynamic setting now. We are given a dynamical system $(X,\mathcal{A},m,T)$. One considers, given a finite partition P:

$$\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}T^iP\right)$$

From (4) and the measure preservance of T:

$$\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}T^iP\right)=\frac{1}{n}\sum_{i=-1}^{i=-n+1}H\left(P\,\middle|\,\bigvee_{k=-1}^{k=i}T^kP\right)$$

and then, from (7)

$$\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}T^iP\right)\xrightarrow{n\to\infty}H\left(P\,\middle|\,\bigvee_{i=-1}^{-\infty}T^iP\right);$$

this limit is denoted by $H(P,T)$. It satisfies the so-called *Pinsker formula*:

$$H(P\vee Q,T)=H(P,T)+H\left(Q\,\middle|\,\bigvee_{i=-\infty}^{\infty}T^iP\vee\bigvee_{i=-1}^{\infty}T^iQ\right).$$

easily deduced from (9), developing

$$\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}T^i(P\vee Q)\right)$$

and

$$\frac{1}{n}H\left(\bigvee_{i=0}^{i=n-1}T^iQ\;\middle|\;\bigvee_{i=0}^{i=n-1}T^iP\right)$$

As a direct consequence, if $\bigvee_{i=-\infty}^{\infty}T^iP=\mathcal{A}$, ($P$ is then called a generator), for every finite partition Q the inequality $H(Q,T)\leq H(P,T)$ holds. This motivates the definition of entropy for the transformation itself, originating back to Kolmogorov:

$$H(T):=\sup\{H(P,T)\colon P\text{ – a finite partition of }X\}.$$

This least upper bound is attained at any partition generating $\mathcal{A}$. From this definition it is clear that entropy is an isomorphism invariant. Another consequence is that, if P_n is an increasing sequence of finite partitions which exhaust the algebra $\mathcal{A}$, i.e. $P_n\uparrow\mathcal{A}$, then $h(P_n,T)\uparrow h(T)$.

Properties of the entropy of transformations (3.3).

(1) $H(T^n)=|n|h(T)$
(2) if S is a factor of T, $H(S)\leq H(T)$
(3) $H(T\times S)=H(T)+H(S)$

Example (3.4).
Bernoulli shifts $B(\frac{1}{2},\frac{1}{2})$, $B(\frac{1}{3},\frac{1}{3},\frac{1}{3})$ are of entropy $\log 2$ and $\log 3$, respectively. These two transformations are therefore not isomorphic.

There is a converse to this, the Ornstein isomorphism theorem:

Ornstein Isomorphism Theorem (3.5).
Two Bernoulli shifts with the same entropy are isomorphic.

The Pinsker formula implies that, given a dynamical system

$$(X,\mathcal{A},m,T)$$

if P and Q are two finite partitions such that $H(P,T)=H(Q,T)=0$, then $H(P\vee Q,T)=0$. There is therefore a biggest algebra $\pi(T)$ which has the property that for every set $A\in\pi(T)$, $H(P_A,T)=0$, P_A being the two sets partition whose elements are A and A^c. This algebra is called the Pinsker algebra. When in a transformation the Pinsker algebra is reduced to the trivial algebra ν it is called a K-automorphism. Then for every partition P, $H(P,T)>0$. This is equivalent to the following: For every finite partition P the algebras $T^{-n}P^-\downarrow\nu$ when $n\to+\infty$. ($P^-=\bigvee_{i=0}^{i=-\infty}T^iP$). The reason for this is that if R is a finite partition in $\pi(P)=\lim\downarrow T^{-n}P^-$, H(R,T)=0 (a direct consequence of the Pinsker formula applied to $P\vee R$). Furthermore if P is a generator, then $\pi(P)=\pi(T)$. Therefore Bernoulli processes are K-automorphisms, as a consequence of the $0,1$ law. Ornstein has proved:

theorem (3.6).
There exists a K-automorphism which is not isomorphic to a Bernoulli shift.

We show the following very important lemma:
K-automorphisms are disjoint from systems with 0 entropy.

Proof. Assume $(X, \mathcal{A}, m, T)$ is a K automorphism while $(Y, \mathcal{B}, \mu, S)$ has 0 entropy. Let λ be a joining of these two systems. From now on all our computations are going to be made on the product space equippped with λ. Because T is a K automorphism, $H(P, T^n) \to H(P)$ $(n \to +\infty)$.

Because of Pinskers formula together with the fact that $H(S^n) = 0$ for all n, taking P and Q partitions $\mathcal{A}$ and $\mathcal{B}$ measurable respectively: $H(P \vee Q, (T \times S)^n) \to H(P)$ But this can also be evaluated (by Pinsker's formula once more) as $< H(P|Q)$ This implies that $H(P|Q) = H(P)$ (for the joining measure λ) whence independence (by (7)). □

4. Ergodic theory and combinatorial number theory

Ergodic theory and topological dynamics (the dynamical study of homeomorphisms of compact metric spaces) have been linked to important developments in combinatorial number theory. We shall here give the dynamical proofs to the Van der Waerden theorem and to the Roth theorem. We will start with Furstenberg's result on regular recurrence of compact continuous dynamical systems.

Theorem (Furstenberg; 4.1).
Let (X, ϱ) be a compact metric space and T an homeomorphism. Then T has points of regular recurence

$$\forall d \in \mathbb{N}_{>0}\ \forall \varepsilon > 0\ \exists x \in X\ \exists n \in \mathbb{N}_{>0}:$$
$$\varrho(x, T^n x) < \varepsilon,\ \varrho(x, T^{2n} x) < \varepsilon, \ldots,\ \varrho(x, T^{dn} x) < \varepsilon \quad (A)$$

Proof. An application of the Zorn lemma gives the existence of a compact $Y \subset X$ which is T invariant $(T(Y) = Y)$ and does not contain any proper non empty T-invariant subset. We restrict ourself to the action of T on Y. This implies that, for every $y \in Y, \overline{\{T^m y\}_{m \in \mathbb{N}}} = Y$, and that for every y, for every ϵ, there exists $n > 0$ such that $\varrho(y, T^n y) < \varepsilon$. (This is the topological equivalent to the Poincaré recurrence theorem). We work by induction and assume that the theorem has been proved for all integers $\leq d$. We first notice that for all ε, the set of $y \in Y$ such that (A) holds is dense. Let $B(z, \varepsilon) = V$ be the ball inside which we want to find x satisfying (A). As $\cup_{i=0}^{i=\infty} T^{-i} V = Y$, we can extract finitely many $T^{-n_i} V$ which cover Y. Let δ be the common uniform modulus of ε continuity of the T^{n_i}. Pick a point w satisfying (A), with ε changed into δ. If $w \in T^{-n_i} V$, $T^{n_i} w$ will satisfy (A). We now construct by induction

a sequence x_n for which, given $p < n$, there exist an integer k such that $\varrho(T^k x_n, x_p) < \varepsilon, \varrho(T^{2k} x_n, x_p) < \varepsilon, ..., \varrho(T^{dk} x_n, x_p) < \varepsilon$. By density, τ close to x_n there is $\bar{x}_n$ satisfying (A). Let m be the "common recurrence time" at $\bar{x}_n$, and take $x_{n+1} = T^{-m}\bar{x}_n$. Consider again $p \leq n$. Clearly, if τ was choosen sufficiently small, we shall get, $\varrho(T^{k+m} x_{n+1}, x_p) < \varepsilon, \varrho(T^{2(m+k)} x_{n+1}, x_p) < \varepsilon, ..., \varrho(T^{d(m+k)} x_n + 1, x_p) < \varepsilon$. By compactness there exist $n_1 < n_2$ such that $\varrho(x_{n_1}, x_{n_2}) < \varepsilon$, which is (A) for $d+1$ ($x = x_{n_2}$).

□

This purely dynamical result turns out to be equivalent to a number-theoretical one, the van der Waerden Theorem:

Van der Waerden Theorem (4.2).
If $\mathbb{Z}$ is divided into finite number of parts ("out of finitely many colours, every integer is painted with one colour"), one of them contains arbitrarily long arithmetic progressions ("there are arbitrarily long monochromatic progressions ").

We will now prove their equivalence (both of the following are sketches of the proofs).

Van der Waerden's implies Furstenberg's. With some distance ε and length d fixed, one can cover X with finitely many $\varepsilon/2$-balls, which may be turned into partition of X into finite number of subsets of diameter at most ε. This yields a finite partition of any orbit $\{T^n y\}_{n\in\mathbb{Z}}$, and by the van der Waerden Theorem one can find a and n such that $T^a y$, $T^{a+n} y, \ldots, T^{a+dn} y$ all lie in one the subsets. The conclusion is satisfied with $x := T^a y$. □

Furstenberg's implies Van der Waerden's. Fix a partition $\mathbb{Z} = C_1 \cup \cdots \cup C_K$ and length d. Observe that it suffices to find suitable subsets for d's separately – due to the number of the subsets being finite, one can then choose one of them fitting *every* d.

We will use the shift on alphabet $\{1, \ldots, K\}$.[3] Take $\varkappa = (\varkappa_n)$ which codifies the chosen partition (i.e. $\varkappa_n = $ this k that $n \in C_k$) and restrict to the compact subspace $X := \overline{\{T^m \varkappa\}_{m\in\mathbb{Z}}}$. To prove the conclusion in this setting it suffices to find a and n such that $\varkappa_a = \varkappa_{a+n} \cdots = \varkappa_{a+dn}$. By the Furstenberg Theorem (for d as above and $\varepsilon = 1$) one can find $x \in X$ and $n \in N_{>0}$ such that $x_0 = (T^n x)_0 = \cdots = (T^{dn} x)_0$, and,

[3]We remind the definition and basic facts about shift. It is defined on the set $\{1, \ldots, K\}^{\mathbb{Z}}$. The metric counts equal terms and is given by

$$\varrho((x_n), (y_n)) = \sum_n \frac{1}{2^{|n|}} \delta_{x_n, y_n},$$

where $\delta_{x,y} := 1$ if $x = y$ and $\delta_{x,y} := 0$ otherwise, and yields a compact space. Roughly speaking, two sequences are ϱ-close precisely if they coincide on $[-N, N]$ for N large enough; in particular, if $\varrho(x, y) < 1$, then $x_0 = y_0$. The shift is the homeomorphism $T(x_n)_n := (x_{n+1})_n$.

consequently, $x_0 = x_n = \cdots = x_{dn}$. Now x may be approximated by some $T^a\varkappa$ close enough to get $(T^a\varkappa)_0 = x_0, \ldots, (T^a\varkappa)_1 = x_1, \ldots, (T^a\varkappa)_{dn} = x_{dn}$, which yields the conclusion with thereby found a and n. □

Even though the van der Waerden theorem gives the existence of arbitrarily long arithmetic progressions which are monochromatic, it does not give any indication of which colour they are going to be. This is answered by the theorem of Szemeredi. To state it, we have to make precise he "size" of a set of integers.

Definition (4.3).
For $E \subset \mathbb{N}$ its *upper density* is defined as

$$d^*(E) := \limsup_{N\to\infty} \frac{1}{N} \sup_M \#(E \cap [M, M+N]).$$

Szemeredi Theorem (4.4).
Every subset of $\mathbb{N}$ *of positive upper density contains arbitrarily long arithmetic progressions.*

(If $\mathbb{N}$ is partitioned into finitely many sets, at least one of them has positive upper density). We are going to prove the less general

Roth Theorem (4.5).
Every subset of $\mathbb{N}$ *of positive upper density contains a three-term arithmetic progression.*

This statement can be proved using another ergodic result of Furstenberg:

Theorem (Furstenberg; 4.6).
Given an (abstract) dynamical system $(X, \mathcal{A}, m, T)$ *and a non-negligible set* $m(A) > 0$*, one can find* $n > 0$ *for which* $m\left(A \cap T^n(A) \cap T^{2n}(A)\right) > 0$.

In fact, one can even prove that

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} m\left(A \cap T^n(A) \cap T^{2n}(A)\right) > 0.$$

Proof. We may assume that T is ergodic, as the conclusion then implies the general case by an application of Fatou's theorem. We consider the diagonal joining Δ of three copies X_1, X_2, X_3 of X (here X is short for $(X\mathcal{A}, m, T)$) and we define the joinings $\Delta_N := \frac{1}{N}\sum_{i=0}^{i=N-1}(id \times T \times T^{2i})^*\Delta$. The theorem will be proved if we show that, letting Δ_∞ be a limit value of Δ_N, $\Delta_\infty(A \times A \times A) > 0$. Let $\mathcal{E}$ be the Kronecker algebra of T. We first notice that (1) Δ_∞ is $Id \times T \times T^2$ invariant and that (2) Δ_∞ restricted to $X_2 \times X_3$ is the product measure (a consequence of the ergodicity of T). The algebra $\mathcal{I}$ of invariant sets of $T \times T^2$ with the product measure is $\subset \mathcal{E} \times \mathcal{E}$. Let f, g, h be mesurable with respect to

X_1, X_2, X_3 respectively. We use the notation $\bar{f} = E(f|\mathcal{I})$. The disjointness argument already mentioned implies that, relative to $\mathcal{I}$, X_1 is independent of $X_2 \times X_3$. Therefore $\int fghd\Delta_\infty = \int E(f|\mathcal{I})E(gh|\mathcal{I})d\Delta_\infty$. As $\mathcal{I} \subset \mathcal{E} \times \mathcal{E}$, $E(gh|\mathcal{I}) = E(\bar{f}\bar{g}|\mathcal{I})$. We see then that $\int fghd\Delta_\infty = E(f|\mathcal{I})\bar{g}\bar{h})$, and, from the way Δ_N is constructed, we deduce that $\int fghd\Delta_\infty = \int \bar{f}\bar{g}\bar{h}d\Delta_\infty$. We consider the case $f = g = h = \mathbb{1}_A$. $E(\mathbb{1}_A|\mathcal{E})$ can be approximated arbitrarily well by linear combinations of finitely many eigenfunctions of U_T, and T there acts as a translation on a Torus for which there exists a sequence of integers n_i such that restricted to this Torus T^{n_i} is as close as we want to the identity, Hence $\int fghdm_{n_i}$ is very close to $\int fghd\Delta$ which is > 0 for n_i sufficiently large when $f = g = h = \mathbb{1}_A$. (Here $m_{n_i} = (Id \times T^{n_i} \times T^{2n_i}) * \Delta$). Whence $\int \mathbb{1}_A\mathbb{1}_A\mathbb{1}_A d\Delta_\infty > 0$ for every limit joining Δ_∞, which is the result.

□

We now prove that the preceding theorem which is of ergodic type implies the Roth theorem,

Theorem (transference theorem; 4.7).
The theorem of Furstenberg implies the theorem of Roth

Proof. Let E be a subset of N such that $d^*(E) = \alpha > 0$. There exists a sequence N_i such that $\alpha = lim_{N_i} \frac{1}{N_i} \sup_M \#[E \cap [M, M + N_i]$. By considering an ultafilter $\mathcal{U}$ finer than the filter $[N_i, +\infty]$, we define, for any $G \subset N$,

$$L(G) = \lim_{\mathcal{U}} \frac{1}{N} \sup_M \#(G \cap [M, M + N])$$

.

Clearly $L(E) = \alpha$ and L defines a finitely additive measure, invariant by the translation S ($Sx = x + 1$). It is also clear that L is not σ-additive. However, restricted to the S invariant sub σ-algebra of sets generated the partitions $S^n P_E$, $n \in Z$, where P_E is the two sets partition (E, E^c), L is σ-additive as a consequence of the Kolomogorov extension theorem. The Furstenberg theorem implies the existence of an integer n for which $L(E \cap S^n(E) \cap S^{2n}(E) > 0$, which is the desired conclusion.

□

Note that the Roth theorem is equivalent to a finite statement: Given $\varepsilon > 0$, there exists N such that any set E of integers in the interval $[1, N]$ with cardinality $> \varepsilon N$ contains a three term arithmetic progression. The preceding proof is (in essence) ineffective and cannot give any indication upon what N is, once ε is given. However effective proofs of the Roth (and of the Szemeredi) theorems exist,which do not make any appeal to ergodic theory.

There remain many questions

Erdős conjecture on arithmetic progressions (4.8).
If $A \subset \mathbb{N}$ is large enough to make $\sum_{a \in A} \frac{1}{a} = \infty$, then it contains arbitrarily long arithmetic progressions.

The Green-Tao Theorem solves this conjecture for the primes.